ComSIS
Computer Science and Information Systems
Volume 22, Number 4, September 2025

Contents

ISSN: 2406-1018 (Online)

Computer Science and Information Systems

Vol 22, No 4, September 2025

ComSIS

Computer Science and Information Systems

Published by ComSIS Consortium

Volume 22, Number 4
September 2025

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Editorial Office:
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Informatics
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
Phone: +381 21 458 888; **Fax:** +381 21 6350 458
www.comsis.org; Email: comsis@uns.ac.rs

Volume 22, Number 4, 2025
Novi Sad

Computer Science and Information Systems

ISSN: 2406-1018 (Online)

The ComSIS journal is sponsored by:

Ministry of Education, Science and Technological Development of the Republic of Serbia
http://www.mpn.gov.rs/

Computer Science and Information Systems

AIMS AND SCOPE

Computer Science and Information Systems (ComSIS) is an international refereed journal, published in Serbia. The objective of ComSIS is to communicate important research and development results in the areas of computer science, software engineering, and information systems.

We publish original papers of lasting value covering both theoretical foundations of computer science and commercial, industrial, or educational aspects that provide new insights into design and implementation of software and information systems. In addition to wide-scope regular issues, ComSIS also includes special issues covering specific topics in all areas of computer science and information systems.

ComSIS publishes invited and regular papers in English. Papers that pass a strict reviewing procedure are accepted for publishing. ComSIS is published semiannually.

Indexing Information

ComSIS is covered or selected for coverage in the following:

- Science Citation Index (also known as SciSearch) and Journal Citation Reports / Science Edition by Thomson Reuters, with 2024 two-year impact factor 1.8,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Summon (Serials Solutions),
- EBSCO bibliographic databases,
- IET bibliographic database Inspec,
- FIZ Karlsruhe bibliographic database io-port,
- Index of Information Systems Journals (Deakin University, Australia),
- Directory of Open Access Journals (DOAJ),
- Google Scholar,
- Journal Bibliometric Report of the Center for Evaluation in Education and Science (CEON/CEES) in cooperation with the National Library of Serbia, for the Serbian Ministry of Education and Science,
- Serbian Citation Index (SCIndeks),
- doiSerbia.

Information for Contributors

The Editors will be pleased to receive contributions from all parts of the world. An electronic version (LaTeX), or three hard-copies of the manuscript written in English, intended for publication and prepared as described in "Manuscript Requirements" (which may be downloaded from http://www.comsis.org), along with a cover letter containing the corresponding author's details should be sent to official journal e-mail.

Criteria for Acceptance

Criteria for acceptance will be appropriateness to the field of Journal, as described in the Aims and Scope, taking into account the merit of the content and presentation. The number of pages of submitted articles is limited to 20 (using the appropriate LaTeX template).

Manuscripts will be refereed in the manner customary with scientific journals before being accepted for publication.

Copyright and Use Agreement

Computer Science and Information Systems

Volume 22, Number 4, September 2025

CONTENTS

Special section: Emergences in Computing and Information Technologies: Towards a Sustainable Wellbeing Environment

Computer Science and Information Systems 22(4): i–ii https://doi.org/10.2298/CSIS250400iI

Editorial

Mirjana Ivanović, Miloš Radovanović, and Vladimir Kurbalija

University of Novi Sad, Faculty of Sciences
Novi Sad, Serbia
{mira,radacha,kurba}@dmi.uns.ac.rs

Welcome to Volume 22, Issue 4 of the Computer Science and Information Systems journal, which encompasses 10 regular articles and one special section, "Emergences in Computing and Information Technologies: Towards a Sustainable Wellbeing Environment," which features 9 articles. As is customary, we acknowledge the efforts and enthusiasm of our authors, reviewers, and guest editors, without whom the current issue and the publication of the journal itself would not be possible.

The first regular article, "FAP: A Time Series Analysis and Mining Framework for Scientific and Practical Applications" by Zoltán Gellér et al. presents the main capabilities of the Framework for Analysis and Prediction (FAP), a free and open source Java library designed for processing and mining time series data that has been successfully applied both in research and education since its initial presentation. In this moment, the FAP library contains implementations of all main concepts needed for time-series mining: a significant number of distance measures, various variants of the NN classifier, multiple representations of time series, the main techniques for evaluating classifier performance, as well as classes for training classifiers and tuning parameters of distance measures.

In the second regular article, "VSAF: Verifiable and Secure Aggregation Scheme for Federated Learning in Edge Computing," Shiwen Zhang et al. focus two issues in federated learning (FL): (1) privacy protection of the parameters uploaded by clients, and (2) verification of the correctness of the aggregated result from a cloud server. To this end, the article proposes VSAF, a verifiable and secure aggregation scheme for federated learning in edge computing by designing a single masking protocol which combines the Bloom filter and Shamir's secret sharing, and introducing a lightweight verification algorithm for aggregated gradients based on a linear homomorphic hash function.

"Augmented Reality Mobile Application as a Support in Presentation of Orthodox Iconography," by Dušan Tatić et al. presents a mobile application based on augmented reality technology that facilitates and speeds up access to iconographic content stored on the Virtual Encyclopedia of Bulgarian Iconography (BIDL) platform. The main goal, based on image recognition by a specially designed augmented reality module, is providing instantaneous and on-site information about the concrete icon observed by visitors, while avoiding classical search over a large database (requiring keywords such as geographical location, name of the church, etc.) since the icons are immediately recognized.

Lang Wu and Yi Dong, in "Federated Learning with Committee Mechanism for Class Imbalance," introduce FedCCSM, a federated learning framework designed to address class imbalance and malicious client behavior. Firstly, to accelerate model optimization, a client selection mechanism is introduced based on specific criteria. Secondly, the adoption of a committee mechanism involves selecting a client committee to screen the model before aggregation, enhancing system security. And finally, by simulating mechanisms for unbalanced clients, the algorithm's practical application effectiveness is strengthened.

"Data-Driven Traffic Management: Enhancing Road Safety through Integrated Digital Twin Technology," by Miloš Durković et al. proposes a data-driven approach to enhancing traffic safety through the integration of digital twins, in-vehicle monitoring system, and machine learning. The main goal of the approach is to contribute to solving problems related to driver behavior, inadequate road signage infrastructure, and delayed maintenance, by developing a digital twin model that leverages real-time data for predictive analysis, coaching, and maintenance.

Shunxiang Zhang et al., in their article "Fire Detection Models Based on Attention Mechanisms and Multiscale Features," propose the attention mechanisms and multiscale features (AMMF) model for fire detection, which integrates an attention mechanism and multi-scale feature fusion to improve accuracy and real-time performance. The model incorporates a dynamic sparse attention mechanism in the backbone network to enhance feature capture and restructures the neck network using CepBlock and MPFusion modules for better feature fusion.

The article "Defining the Attractiveness Concept for Cyber Incidents Forecasting," authored by Javier García-Ochoa et al., presents a methodology that defines the attractiveness concept to address challenges in analysing the proneness of an entity to be attacked by an adversary evaluating the relevance of different target features or behaviours. Attractiveness is the possession of features or the exhibition of behaviours in entities that raise interest for potential adversaries. Thus, the more significant the attractiveness value is, the greater the proneness to being attacked is to be considered.

In "Hyperparameter Optimisation in Differential Evolution Using Summed Local Difference Strings, A Rugged but Easily Calculated Landscape for Combinatorial Search Problems," Husanbir Singh Pannu and Douglas B. Kell analyse the effectiveness of differential evolution hyperparameters in large-scale search problems, i.e., those with very many variables or vector elements, using a novel objective function that is easily calculated from the vector/string itself. A neural network is trained by systematically varying three hyper-parameters, viz population (NP), mutation factor (F) and crossover rate (CR).

"Digital Transformation in Public Accounting and Finance Management: A Clusters Literature Review," by Ambrósio Teixeira et al., investigates the literary corpus on the role and potential of digital transformation in public accounting and finance management, encompassing 890 relevant research papers, out of which 24 publications, divided into two clusters, were selected for an in-depth analysis. The findings demonstrate that technologies have significantly transformed accounting and public finance by automating processes to reduce errors and save time, increasing transparency and accountability, preventing fraud with analytical tools, improving budget planning and monitoring, and integrating systems for a comprehensive financial view.

Finally, Xin Su et al., in "HRSP: A High-Risk Social Personnel Risk Assessment Model Based on Graph Attention Label Propagation Algorithm," first analyze and construct a knowledge graph of high-risk individuals based on their backgrounds, trajectories, and related information. Subsequently, they propose a high-risk personnel risk assessment model based on a graph attention-label propagation algorithm. The model employs a multi-label feature selection method, a basic classifier based on a graph attention network for the label propagation algorithm, and an adversarial data augmentation algorithm to enhance the gradient-based adversary during training.

Computer Science and Information Systems 22(4):iii–v https://doi.org/10.2298/CSIS250400iiiC

Guest Editorial: Emergences in Computing and Information Technologies: Towards a Sustainable Wellbeing Environment

Jia-Wei Chang[1], Hwa-Young Jeong[2], Nigel Lin[3], Mirjana Ivanovic[4]

[1] Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, Taichung City, Taiwan
jwchang@nutc.edu.tw

[2] Humanitas College, Kyung Hee University, Republic of Korea
hyjeong@khu.ac.kr

[3] Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA, USA
nigel@mail.topwise.com

[4] University of Novi Sad, Faculty of Sciences, Novi Sad, Serbia
mira@dmi.uns.ac.rs

With rapid advances in computing and information technologies, society faces an inflection point: the integration of these capabilities must demonstrably advance human well-being while safeguarding environmental sustainability. As digital infrastructures reshape the contours of interaction and decision-making, their design, development, and deployment must adopt a socio-technical lens that centers people and the environment. This special issue aims to chart a responsible innovation agenda—one that aligns technical progress with sustainability-by-design principles. Our goal is to catalyze a shared effort toward inclusive, equitable, and resilient digital ecosystems that are consistent with global sustainability goals. Through open calls, this special issue received 41 submissions; approximately one third passed editorial pre-screening and proceeded to double-blind peer review under rigorous editorial standards, after which nine papers were accepted.

Against this backdrop, the nine accepted contributions articulate complementary, operational directions: (1) advances in perception and computational imaging that enable robust downstream inference and decision support; (2) humane, fatigue-aware immersive systems for extended reality (XR) that integrate human factors and ergonomics; (3) data-driven methods for health and the life sciences, spanning multimodal analytics and clinical decision support; (4) real-time, embodied interaction for remote participation via edge computing, real-time systems, and telepresence; (5) explainable and accountable modeling (XAI) to increase transparency and trust in digital services; (6) user-centered approaches to equitable data discovery and governance, including FAIR principles and participatory design; (7) privacy-preserving collaborative learning at scale (e.g., federated and split learning with differential privacy and secure aggregation); (8) analytics for sustainable mobility within intelligent transportation

systems to inform resource-efficient planning; and (9) inclusive multilingual language technologies that reduce cross-cultural miscommunication, with attention to low-resource settings. Collectively, these threads illustrate credible pathways by which technical innovation can translate into improved quality of life, greater social inclusion, and more sustainable resource use. The following summaries introduce these contributions in turn.

The first paper, titled "A study on Multi-scale Attention dense U-Net for image denoising method," by Mingshou An, Xuhang Zhao, Dae-Seong Kang, and HyeYoun Lim, proposes an enhanced denoising architecture that layers multi-scale attention atop a dense U-Net backbone. By improving perceptual fidelity and structural consistency, the method supports more reliable downstream analysis in sensing-intensive pipelines; such quality gains can, in principle, reduce reprocessing and error propagation in large-scale imaging workflows.

The second paper, "The Intersection of Digital Wellbeing and Collection Exhibition: A Study on the Impact of AR Interactive Display Models on Visitor Experience," by Min-Feng Lee, Guey-Shya Chen, Hui-Chien Chen, and Jian-Zhi Chen, examines augmented-reality interaction models for cultural exhibitions. Through a mixed-method study with a sizable visitor cohort, the work reports enhanced engagement, comprehension, and satisfaction, alongside indications of reduced digital fatigue—evidence toward humane, fatigue-aware immersive systems in XR settings.

The third paper, "Application of the Inception-ResNet-V2 algorithm to the analysis of embryo microscope images for the prediction model of assisted reproduction," by Yu-Yu Yen, Shao-Ping Weng, Li-Jen Su, Jui-Hung Kao, and Woei-Chyn Chu, applies a state-of-the-art deep vision backbone to embryo microscopy for outcome prediction in assisted reproduction. The approach leverages high-capacity feature learning to improve discriminative performance in a clinically consequential setting, pointing to decision support that may enhance effectiveness and reduce unnecessary interventions.

The fourth paper, "A Study of Real-Time Operations by Converting Human Skeleton Coordinates to Digital Avatars," by Fei-lung Lin, Jui-Hung Kao, Yu-Yu Yen, Kuan-Wen Liao, and Pu Huang, investigates a real-time pipeline that transforms skeletal coordinate data into responsive digital avatars. The system advances fidelity and responsiveness for motion-capture-based interaction, opening practical opportunities in rehabilitation, remote collaboration, and education through accessible, low-latency embodiment.

The fifth paper, "Implementing Persona in the Business Sector by A Universal Explainable AI Framework Based on Byte-Pair Encoding," by Zhenyao Liu, Yu-Lun Liu, Wei-Chang Yeh, and Chia-Ling Huang, introduces an explainable persona-modeling framework grounded in byte-pair encoding. By clarifying feature attributions and decision rationales, the study supports transparent and auditable deployment in data-driven business settings; we highlight its contribution primarily as an advance in interpretability and accountability.

The sixth paper, "Formative Interviews for a User-Centered Design Study on Developing an Effective Gateway for Health Research Data Search – Towards a Sustainable Wellbeing Environment," by Hsiu An Lee, Tung Lin, Hsin-I Chen, Wei-Chen Liu, Yen-Ju Shen, Wen-Chang Tseng, Chien-Yeh Hsu, and Yi-Hsin Yang, reports formative interviews that surface requirements for an effective, user-centered gateway to health research data. The findings map pain points and design principles for findability,

usability, and transparency—laying groundwork for equitable data access and reproducible research.

The seventh paper, “Elastic-Trust Hybrid Federated Learning,” by Yi-Cheng Chen, Lin Hui, and Yung-Lin Chu, presents a federated learning scheme that hybridizes training modes with an elastic-trust mechanism to address client heterogeneity and variable data quality. The framework calibrates contribution and enhances aggregation robustness while safeguarding privacy, aiming to preserve utility under real-world non-IID conditions.

The eighth paper, “A Comparative Study of Key Factors in Travel Time Prediction for Sustainable Mobility and Well-Being,” by Chuang-Chieh Lin, Min-Chu Ho, and Chih-Chieh Hung, systematically evaluates modeling choices—such as preprocessing, temporal windows, and exogenous signals (e.g., weather)—that materially affect accuracy and stability. The analysis offers practitioners actionable guidance for building efficient, reliable travel-time predictors and discusses implications for congestion mitigation and commute reliability.

The ninth paper, “Cultural Pragmatics and Causal Connectives: A Contrastive Study of Korean and English Using the AI-Hub Parallel Corpus,” by Sujeong Choi and Sinhye Nam, provides a contrastive analysis of causal connectives through a cultural-pragmatics lens. The results illuminate systematic cross-lingual differences with implications for natural-language understanding and generation, informing more context-sensitive multilingual technologies.

Acknowledgments. The guest editors extend their sincere appreciation to all authors who submitted interesting and challenging papers; their creativity and willingness to engage with rigorous feedback have substantially enriched the scope and depth of this special issue. We are equally indebted to the reviewers for their insightful evaluations—their constructive critiques materially improved the clarity and rigor of the accepted manuscripts. We also gratefully acknowledge the leadership of the Editor-in-Chief, Prof. Mirjana Ivanovic, whose guidance and high editorial standards shaped the vision and execution of this issue, as well as the journal’s editorial assistants for their professional support. Finally, we are grateful to our readers for their continued engagement; we hope that the contributions gathered here will stimulate further research and practice at the intersection of computing, human well-being, and sustainability.

Computer Science and Information Systems 22(4):1379–1403 https://doi.org/10.2298/CSIS240910029G

FAP: A Time Series Analysis and Mining Framework for Scientific and Practical Applications

Zoltán Gellér[1], Vladimir Kurbalija[2], and Mirjana Ivanović[2]

[1] University of Novi Sad, Faculty of Philosophy, Department of Media Studies
Dr Zorana Đinđića 2, 21000 Novi Sad, Serbia
zoltang@ff.uns.ac.rs
[2] University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics
Trg D. Obradovića 4, 21000 Novi Sad, Serbia
{kurba, mira}@dmi.uns.ac.rs

Abstract. Given the exponential growth of data in modern society, data analysis tools have become increasingly pivotal in a wide range of fields, such as business, advertising, economy, medicine, biology, meteorology, astronomy, agriculture, and others. As the time component often plays an essential role in data analysis, the application and research of different methods for examining temporal data is among the current interests of both practitioners and researchers. This paper presents the main capabilities of the Framework for Analysis and Prediction (FAP), a free and open source Java library designed for processing and mining time series data that has been successfully applied both in research and education since its initial presentation.

Keywords: time series, data mining, open source, software library, Java.

1. Introduction

Past two decades influenced significant changes in processing huge amounts of data. The need to process such ever-growing amounts of data from different sources all over the world, importance of developing and applying different approaches of data mining and machine learning has gained more and more attention. They are unavoidable instruments of applications in computer science, ICT, and education [3, 57].

An emerging sub-field of data mining is temporal data mining that is focused on knowledge discovery from huge amounts of temporal data [51]. Time series are the most common form of temporal data which are composed of real values usually sampled at regular time intervals [28]. The chronologically represented arrays of numbers are collected in different domains and they are used to express the change of the observed phenomena over time, like financial sector, economics, engineering, meteorology, medicine [58, 31] as well as in other areas of natural and social sciences [9].

Statistical analysis of time series [37] is mainly focused on identifying patterns, trend analysis, seasonality and forecasting [17]. On the other hand, data mining of time series is focused on tasks like prediction, classification, clustering, indexing, anomaly detection, data representation, distance measures and others [20, 63]. Laxman and Sastry [45] considered and presented significant differences between statistical analysis and temporal data mining: data-mining approaches effectively analyze much larger volumes of data.

More important is that their field of interest exceeds the scope and limitations of statistical time-series analysis.

The numerous possibility of applying time series for storage, analysis, and visualization of big data collections influenced a significant growth of interest in researching in significant aspects and tasks of time-series data mining. The methods presented in [63] have always claimed a particular superiority over previously achieved results.

Contemporary research in domain of time series data mining inspired authors to produce free and open sourced support and services that could assist and facilitate researching new and comparing existing techniques in this domain. Usefulness of such freely available high-quality services could help in more productive and quality research but also educational processes.

All mentioned highly motivated us to significantly improve our previously developed framework for time series processing and analysis [24]. The extended version of FAP (Framework for Analysis and Prediction) implements significant number of essential algorithms in the field of time-series data mining: time-series representations, distance/similarity measures, preprocessing, classification, classifier evaluation techniques with a focus on efficiency, multi-threading, and resumability of time-consuming tasks.

In this paper, we will give a comprehensive analysis of the newly developed functionalities and capabilities of FAP that are essential for researchers and educators to facilitate their research and applications of time series data mining. All the features provided by FAP were implemented from scratch; it does not utilize any other underlying libraries.

The rest of the paper is organized as follows. The second Section is devoted to an extensive review of related work. Central Section 3 deals with various concepts of time series analysis and data mining and their realization within FAP. Concluding remarks are given in last Section.

2. Related Work

The interest in time series has increased dramatically in the past decade. The main reason is the exponential growth of data available for various machine learning and decision support system. A significant part of this data is available in form of time series. Consequently, a large number of frameworks and systems which can help in time-series analysis was developed and improved recently. This section will give an overview of these systems, and also will elaborate the context and motivation for developing FAP.

The investigation of time series is usually based on two important methodologies: statistical analysis and data mining. Time series statistical analysis is an approach used to analyze data points collected or recorded at specific time intervals using well established methods from statistics and econometrics. This type of analysis helps identify patterns, trends, and other characteristics within the data over time. On the other hand, data mining is a newer discipline focused on analyzing complex, massive datasets to extract useful knowledge. Time-series data mining involves analyzing time-dependent data for tasks like forecasting and anomaly detection. Both approaches (statistical analysis and data mining) have numerous systems which can offer assistance in corresponding time-series analysis. Additionally, several high-level programming languages offer powerful packages and libraries which can considerably help in various time-series tasks.

In the market, there is a considerable number of systems that enable time-series analysis relying on statistical and econometric concepts. Probably the most widely used is SAS. SAS[3] (Statistical Analysis System) is a complete software suite widely used for statistical analysis and data management [38]. It offers several tools for time series modeling and forecasting. Key components for time series analysis include:

1. SAS/ETS (Econometric Time Series): A specialized module designed for time series analysis and forecasting, supporting models such as ARIMA, exponential smoothing, state space models, and multivariate time series analysis. It includes functions for model estimation, diagnostics, forecasting, and simulation.
2. SAS/STAT: Provides a wide range of statistical procedures, including time series decomposition, autocorrelation analysis, spectral analysis, and structural time series models. It also incorporates ARIMA and GARCH model fitting, unit root tests, and handling missing values.
3. SAS Forecast Studio: A graphical interface that simplifies building and evaluating time series forecasting models. It enables visual exploration, model selection, parameter specification, and forecast accuracy assessment, integrating with SAS/ETS and SAS/STAT for seamless model estimation and forecasting.

SPSS[4] is also very widely used and influential statistical software which provides tools for statistical analysis, data management, and documentation. It offers following key features for time series analysis: data management (import, merge, clean, recode, and handle missing values in time series data), descriptive statistics (calculate measures like mean, median, standard deviation, and percentiles to summarize time series data), time series visualization, autocorrelation analysis, and forecasting (methods like ARIMA and exponential smoothing, with accuracy assessment options).

GRETL[5] (GNU Regression, Econometrics and Time-series Library) is a platform-independent, open source software package for econometric analysis [6]. It offers a very intuitive interface, parallelization, and an integrated powerful scripting language. In time-series analysis several concept are provided; ARIMA, GARCH-type models, VARs and VECMs (including structural VARs), unit-root and cointegration tests, Kalman filter, etc.

Stata[6] is a widely used statistical software developed by StataCorp for data manipulation, visualization, statistics, and automated reporting. It is widely used by researchers in various fields, such as: economics, epidemiology, biomedicine, and sociology. Stata also offers comprehensive tools for time series analysis: time series data management, descriptive statistics, graphical analysis, and time series modeling.

On the other side of the spectrum of available time-series software there is general data mining software with support for time-series. Since the number of these systems is huge, we will limit our analysis only on non-commercial, research oriented software.

Weka[7] (Waikato Environment for Knowledge Analysis) [64] is an open-source software suite designed for machine learning and data mining tasks. It provides a collection of algorithms and tools for data preprocessing, classification, regression, clustering, association rules, and visualization. Weka is particularly popular in educational and research

[3] http://www.sas.com
[4] https://www.ibm.com/spss
[5] http://gretl.sourceforge.net/
[6] https://www.stata.com/
[7] https://ml.cms.waikato.ac.nz/weka

communities due to its ease of use and comprehensive documentation. It has good capabilities for time series analysis and forecasting through a dedicated environment that can be accessed via its graphical user interface. This environment allows users to develop, evaluate, and visualize forecasting models. Additionally, there are packages like `TS-Classification` that facilitate time series classification tasks in Weka.

RapidMiner[8] is a Java-based data mining and machine learning platform offering features like data loading, transformation, preprocessing, visualization, predictive analytics, statistical modeling, evaluation, and deployment. It provides a graphical user interface to design and execute analytical workflows called "Processes," which consist of multiple "Operators" each performing a specific task. The output of one operator serves as input for the next. RapidMiner can also be accessed via an API or command line, and its functionality can be extended using R and Python scripts for custom operations. It provides strong capabilities for time series analysis through its integrated time series extension.

ELKI[9] (Environment for DeveLoping KDD-Applications Supported by Index Structures) is an open-source data mining software written in Java. It is primarily designed for research in algorithms, with a strong focus on unsupervised methods such as cluster analysis and outlier detection. ELKI facilitates time series analysis through its ability to evaluate various distance measures and algorithms specifically designed for time series data. Several time-series concepts are implemented in ELKI: distance measures, various time-series algorithms and visualization tools.

KNIME[10] (KoNstanz Information MinEr) [8] is a free, open-source platform for data analytics, reporting, and integration. It uses a modular, "Building Blocks of Analytics" concept for machine learning and data mining, allowing users to blend data sources and perform tasks like preprocessing, modeling, and visualization through a graphical interface, minimizing the need for programming. KNIME provides comprehensive tools for time series analysis through its various components and extensions.

The third large group of time-series software is the group of programming languages with powerful libraries for time-series analysis. Here we will give an overview of modern and actively used languages with this property.

Python is a general-purpose programming language [69] which can be used for time series analysis due to its extensive ecosystem of libraries and tools. These libraries offer a variety of tools and models that can help in analysis and forecasting time series. Libraries with functionalities for time-series forecasting, anomaly detection, and feature extraction include: `Tsfresh`, `Darts`, `Kats`, `GreyKite` and `AutoTS`.

R[11] is a comprehensive language that is well-suited for a wide variety of statistical analyses. It also offers a variety of packages designed for handling time series data. Key packages include: `forecast` (methods like exponential smoothing, ARIMA, and state space models for time series forecasting), `tseries` (tools for unit root tests, seasonality tests, time series decomposition, detrending, and differencing), and `zoo` (support for irregularly spaced time series, with efficient data structures for manipulation, subsetting, merging, handling missing values, and aggregating data over irregular intervals).

[8] https://altair.com/altair-rapidminer
[9] https://elki-project.github.io/
[10] https://www.knime.org/
[11] https://www.r-project.org

MATLAB is a high-level programming and numeric computing platform developed by MathWorks [25]. It is widely used by engineers and scientists for a variety of applications, including data analysis, algorithm development, and modeling. Key features for time-series manipulation include: time series objects (`Timeseries` and `timetable` objects for efficient manipulation, indexing, and visualization of time series data), signal processing toolbox (functions for filtering, spectral analysis, Fourier and wavelet analysis, and time-frequency analysis on time series data), econometrics toolbox (tools for econometric time series analysis, including model estimation, forecasting, unit root tests, panel data analysis, and multivariate time series models), and financial toolbox (focuses on financial time series analysis, offering tools for analyzing market data, portfolio optimization, and risk measurement).

Julia [59] is a high-level, high-performance programming language designed for technical computing, particularly in areas like data science, machine learning, and numerical analysis. It also has strong capabilities for time series analysis. Two popular packages are: (1) `TimeSeries.jl` - A comprehensive package for handling time series data with efficient structures like TimeArray and TS. It provides tools for data manipulation, visualization, resampling, merging, differencing, and rolling window calculations, models such as AR, MA, and ARIMA; (2) `Econometrics.jl` - Focused on econometric modeling, it offers functions for time series models like ARDL, VAR, GARCH, and structural time series. The package includes tools for model estimation, hypothesis testing, diagnostic checking, and forecasting in econometrics.

Clearly, three types of software packages for time-series analysis and mining can be distinguished:

1. Statistical and econometric software systems that provide methods and tools for time-series data.
2. General-purpose data mining and machine learning systems that have extensions for time-series tasks.
3. General-purpose programming languages with libraries for time-series analysis.

Evidently, all of these packages, frameworks and systems have some disadvantages. Some of them are not free or open sourced, many of them are not primarily made for time series and the systems from the third group can't be used by non-programmers.

The system FAP, presented in this paper, tries to overcome all of these disadvantages. It is free and open-sourced. It is designed to work with time series and encompasses all main concepts for time-series analysis (pre-processing tasks, distance measures, time-series representations); and for time-series mining (indexing, classification, prediction). Finally, it can be used by experts from various fields since it can be used without any programming experience. In the past 15 years we have developed and constantly upgraded FAP system making it up-to-date with modern findings in time-series mining field. Furthermore, we successfully applied FAP in various domains both for research [42] and educational [41] purposes.

3. Essential functionalities of FAP for high quality time-series data mining

The `core` sub-packages of the FAP library (Fig. 1) define basic interfaces and classes for implementing various time series analysis and data mining concepts such as data

points, time series, datasets, representations (`data`), distance measure (`distance`), classifiers (`classifier`), classifier performance evaluators (`evaluator`), classifier trainers and distance measure tuners (`trainer`), predictors (`predictor`), as well as for loading data points from strings (`input`), and basic classes for checked and unchecked exceptions thrown within the library (`exception`).

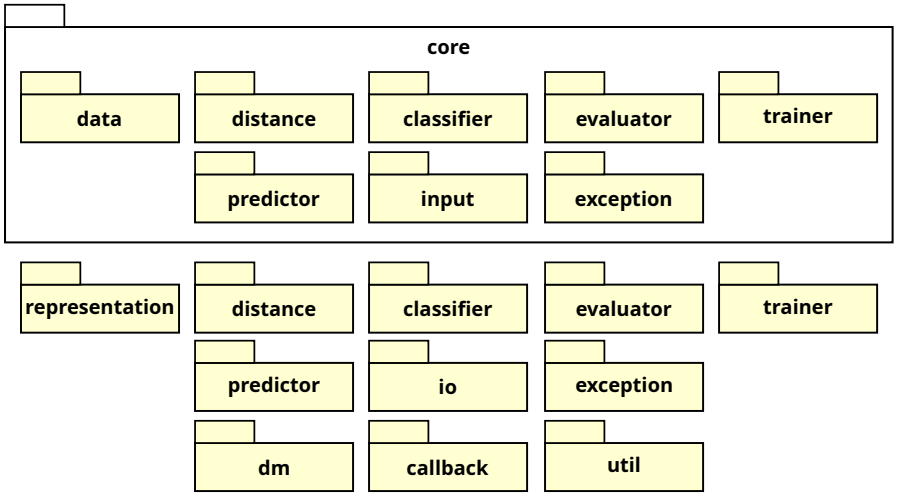


Fig. 1. Sub-packages of the FAP library

Specific implementations of the various time series processing tasks are provided in the appropriate first-level sub-packages. For example, the `fap.core.distance` package defines the `Distance` interface along with the auxiliary abstract class `AbstractDistance`, and classes that represent various distance measures by implementing that interface or extending the auxiliary class are placed in the `fap.distance` sub-package.

Since all base interfaces extend the `Serializable` interface and all classes offer parameterless constructors as well as public getter and setter methods for their properties, the FAP library also supports the JavaBeans standard.

In the rest of this section, we will provide an overview of the capabilities of the FAP library. The review will not cover the predictors (since they are in early stage of development) nor the `fap.core.input`, `fap.io`, and `fap.dm` sub-packages (since they deal with technical issues that are not necessary to understand the main functionalities of the framework). Furthermore, for brevity, the utility classes of the `fap.util` sub-package, which contain mathematical, statistical, and accessory methods intended to facilitate working with threads, strings, files, datasets, and time series (including preprocessing algorithms such as shifting, scaling, z-normalization, mean normalization, min-max normalization, maximum absolute normalization, and decimal scaling), will also be omitted.

To provide a more comprehensive view, the UML diagrams show only a representative subset of the constructors and methods of the classes, and only the types of their parameters.

3.1. Time Series and Representations

Time series are implemented in the form of a list of two-dimensional data points (Fig. 2) where the y coordinate describes the observed phenomenon at the timestamp specified by the x coordinate. A general assumption of all FAP classes that perform tasks related to time series processing (for example, calculating the distance between them) is that the data points are chronologically ordered (i.e., the x coordinate of the i-th element of the time series is less than the x coordinate of the $(i + 1)$-th element).

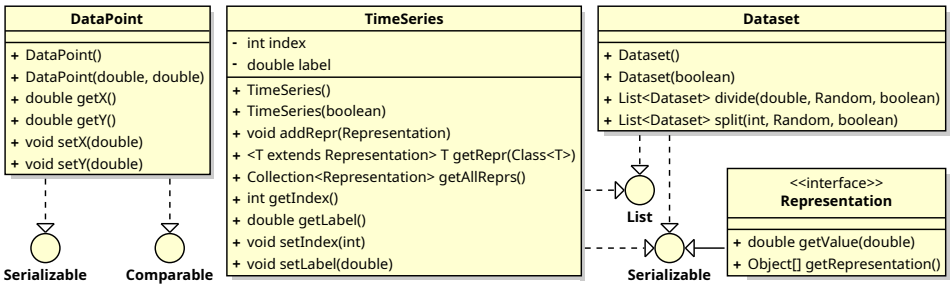


Fig. 2. Basic classes and interfaces

Each time series can be assigned a class label, an index, and a collection of representations. The index represents the unique identifier of the time series within the dataset to which it belongs (providing unique indices is the user's responsibility). It is used by distance measures to store calculated distances in memory (an optional feature that, when performing certain tasks, allows avoiding multiple calculation of distances between the same pairs of time series and thereby speeding up execution) and by kNN classifiers in combination with distance and neighbor matrices. A more detailed insight into how the index is used is given in the corresponding subsections.

To create an (indexed) time series, it is sufficient to specify the class label, index, and chronological list of its values. The x coordinates will be automatically initialized with values from 0 to n-1, where n is the number of elements in the list. For example, the following line of code creates a new time series with label 1.0, index 0, and values 2.0, 3.0, and 4.0 at timestamps 0.0, 1.0, and 2.0:

```
TimeSeries ts = new TimeSeries(1.0, 0, 2.0, 3.0, 4.0);
```

By default, time series are stored in `ArrayLists`. If we want to use `LinkedLists` instead, we just need to add `true` as the first parameter of the constructor:

```
TimeSeries ts = new TimeSeries(true, 1.0, 0, 2.0, 3.0, 4.0);
```

Datasets represent lists of time series, where, analogously to time series, `ArrayLists` are used for storage by default. By specifying the `boolean` value `true` as the first parameter of the constructor, the data structure for storing the time series will be `LinkedList`.

The `Dataset` class offers several methods for partitioning datasets into two (`divide`) or two or more subsets of approximately the same size (`split`), with the ability to control shuffling and stratification. Thus, by applying the `divide(20.0)` method, a list containing

two stratified subsets of the given dataset is obtained: the first subset will contain 20% and the second 80% of its time series. Similarly, a list of 10 stratified subsets of approximately the same size is obtained utilizing the `split(10)` method. The obtained subsets will be stored in the same type of list as the original dataset.

According to the definition given by Esling and Agon [20], a representation of a time series A of length n is a model $\bar{A}$ of length m (where $m \ll n$) that closely approximates A. FAP's interface requires classes implementing time series representations to be able to return the value of the time series at a given timestamp according to their model, and also the representation itself in the form of an array (Fig. 2).

The `fap.representation` sub-package offers implementation of several time-series representations: based on discrete Haar wavelet transform [36, 23, 1, 62, 12], discrete Fourier transform [4, 21, 54, 55, 66], spline [40], Piecewise Linear Approximation (PLA) [34], Piecewise Aggregate Appoximation (PAA) [32, 67, 35, 47], Indexable Piecewise Linear Approximation (IPLA) [15], Piecewise Aggregate Appoximation (PAA) [32, 67, 35, 47], Adaptive Piecewise Constant Approximation (APCA) [33], and Symbolic Aggregate Approximation (SAX) [47, 46].

3.2. Distance Measures

Distance measures should implement the `Distance` interface, declaring a single method whose task is to return the distance between two time series passed as its parameters (Fig. 3).

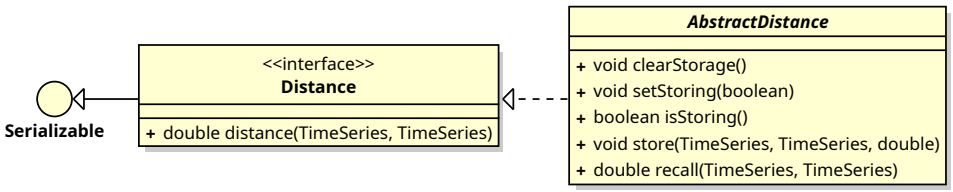


Fig. 3. Distance measures

All the distance measures in the `fap.distance` sub-package inherit the abstract class `AbstractDistance` offering the ability to store calculated distances between pairs of time series for reuse, which can speed up the execution of some tasks. The basic condition for using this mechanism is that the time series are assigned unique indices (see the previous subsection), namely, distances are stored in a (thread-safe) hash map whose keys are the indices of the time series. In addition, all of the distance measures implement the `Copyable` auxiliary interface discussed in subsection 3.6.

Listing 1 shows the implementation of the Manhattan distance as an example of utilizing this mechanism. Before calculating the distance between two time series, it should be checked whether the distance has already been calculated and saved in memory. This is achieved by relying on the `recall` method which returns `NaN` if the requested distance is not yet stored in the hash map. Memorizing new distances is achieved by calling the `store` method, and the utilization of the storage mechanism is controlled via the `setStoring`

Listing 1. Implementation of the Manhattan distance relying on the mechanism for storing calculated distances

```
double distance = recall(series1, series2);
if (!Double.isNaN(distance))
        return distance;

int len = IncomparableTimeSeriesException.checkLength(series1, series2);

distance = 0;

for (int i = 0; i < len; i++) {

        double y1 = series1.getY(i);
        double y2 = series2.getY(i);

        distance += Math.abs(y1 - y2);

}

store(series1, series2, distance);

return distance;
```

method. Additionally, any distance measure that uses this feature must clear the contents of the underlying hash map by calling the `clearStorage` method if the value of any of its parameters that affect the distance between time series changes.

Currently, the following distance measures based on linear matching of time series data points are implemented [18, 11, 2]: Euclidean, Manhattan, Chebyshev, Minkowski, Canberra, Kulczynski, Lorentzian, Soergel, Sørensen (Bray-Curtis), and Wave-Hedges. Within them, 0/0 is treated as 0, and the zero denominator is replaced with the value provided by the `getZeroDenominator` method of the `MathUtils` auxiliary class of the `fap.util` sub-package, as recommended in [11].

The list of implemented elastic distance measures includes Dynamic Time Warping (DTW) [7], Longest Common Subsequence (LCS) [61], Edit distance with Real Penalty (ERP) [13], Edit Distance on Real sequence (EDR) [14], and Time Warp Edit Distance (TWED) [49]. Their elasticity can be adjusted by applying the Sakoe-Chiba [56] or Itakura [30] global constraints.

3.3. Classifiers

In order for a class to be used to classify time series, it must implement the `Classifier` interface (Fig. 4), which declares two methods. The `initialize` method serves for (optional) initialization of the classifier and is not intended for its training. Classification is realized through the `classify` method, which should return the predicted class label.

Distance-based classifiers should implement the `DistanceBasedClassifier` interface that extends the base interface with getter/setter methods to access and update the distance measure to rely on. The abstract convenience class `AbstractDistanceBasedClassifier` stores the distance measure in the `distance` field with protected access level.

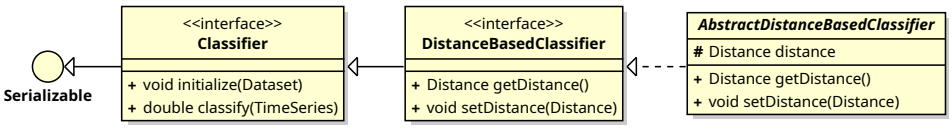


Fig. 4. Classifiers and distance-based classifiers

Presently, the `fap.classifier.NN` sub-package offers the implementation of the nearest neighbor (1NN) rule [16], the majority-voting kNN classifier [22, 50] and several of its weighted variants relying on the inverse of the distances [19], the inverse of the squared distances [50, 44, 60], Dudani's weighting scheme [19], the dual distance-weighted function [26], the uniform and dual-uniform weighting techniques [27], Zavrel's weighting scheme [68], Macleod's weighting function [48], the neighbours' ranks [19], and the Fibonacci weighting function [52]. In the case of weighted kNN variants based on the inverse and inverse of the squared distances, to avoid division by zero, a small value is added to the denominator. By default, it is initialized with the value returned by the `getZeroDenominator` method of the utility class `MathUtils` of the `fap.util` sub-package.

By implementing the `Multithreaded` and `Copyable` auxiliary interfaces, all NN classifiers enable multi-threaded execution of the classification process and making copies of themselves (see subsection 3.6 for details).

For additional acceleration of classification, NN classifiers also support the use of pre-generated distance and neighbor matrices which can be set and accessed through the `setDistances`, `setNeighbours`, `getDistances`, and `getNeighbours` methods.

A distance matrix is a (diagonal) matrix that in the intersection of the i-th row and the j-th column contains the distance between the time series with indices i and j of the given dataset. As an example, part of the distance matrix generated by applying the DTW distance measure to the *SyntheticControl* dataset from the UCR Time Series Classification Archive [5] is given in Table 1. Sub-package `fap.dm` contains classes for (multi-threaded) generation of distance matrices.

The intersection of the i-th row and the j-th column of the neighbor matrix contains the index of the j-th nearest neighbor of the time series with index i in the given dataset. Table 2 shows part of the neighbor matrix generated by applying the DTW distance measure to the *SyntheticControl* dataset.

3.4. Evaluators

Classifier evaluators should implement the `Evaluator` interface depicted in Fig. 5. The classifier performance evaluation algorithm should be implemented within the `evaluate` method, which has three parameters: the trainer (see the next subsection), the classifier, and the whole dataset. The evaluator should split the dataset into test and training subsets, train the classifier using the training set and the trainer, and evaluate the performance of the trained classifier on the test set. As the result, the method should return the classification error rate. This same value should be returned by the `getErrorRate` method, and the result of the call to the `getMisclassified` method should be the number of misclassified time series.

Table 1. The first ten rows and columns of the distance matrix of the *SyntheticControl* dataset generated by applying the DTW distance measure

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0									
2	24.42	0								
3	25.33	22.43	0							
4	22.90	26.85	25.74	0						
5	31.99	23.16	30.41	30.67	0					
6	35.06	22.19	33.38	31.65	23.53	0				
7	34.63	25.50	31.03	28.95	25.81	24.12	0			
8	26.52	32.72	30.17	24.70	34.85	35.59	31.54	0		
9	32.69	26.65	25.22	33.74	41.46	41.56	31.07	30.77	0	
10	21.22	26.66	31.62	23.26	32.28	32.32	28.81	26.14	27.32	0

Table 2. Ten nearest neighbors of the first ten time series of the SyntheticControl dataset obtained by applying the DTW distance measures

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	322	305	320	10	12	14	26	49	321	4
2	342	322	338	17	348	313	324	49	12	31
3	310	323	38	347	333	2	340	315	349	316
4	12	310	305	22	28	343	21	1	10	322
5	309	345	12	311	47	48	50	330	319	2
6	44	2	36	5	30	7	346	339	311	50
7	41	324	330	348	44	12	48	311	309	47
8	18	12	28	16	27	305	328	350	4	341
9	303	334	38	329	307	337	315	350	306	37
10	305	317	1	49	20	28	17	4	320	350

The abstract convenience class `AbstractEvaluator` stores the classification error in the `errorRate` and the number of missclassified time series in the `misclassified` field with protected access level.

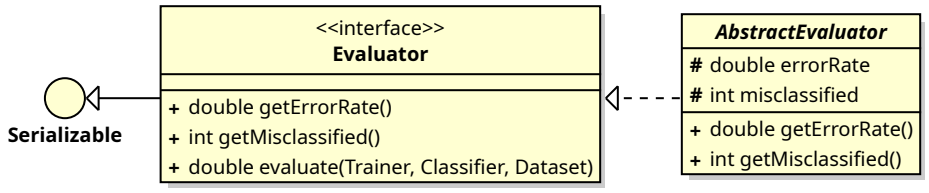


Fig. 5. Classifier evaluators

The `fap.evaluator` sub-package provides the following classes that implement the most common evaluation techniques [1, 60, 29]:

- `HoldoutEvaluator` - applies the Holdout method: a given percentage of the dataset constitutes the training set, and the rest is used as the test set.
- `CrossValidationEvaluator` - performs the cross-validation algorithm: the dataset is divided into k approximately equal subsets of which the union of $k-1$ subsets is used for training and one for testing. The procedure is repeated until each of the k subsets has been used as a test set (exactly) once. The classification error is calculated as the average of the errors obtained over the k test subsets.
- `LeaveOneOutEvaluator` - executes the leave-one-out procedure: the classifier is trained on a training set that contains all the time series of the original dataset except for one that is reserved for testing. The procedure is repeated until each time series of the initial dataset is excluded from the training process (and used for testing) exactly once.

The parameters of the `HoldoutEvaluator` and `CrossValidationEvaluator` classes allow the choice between stratified and non-stratified partitioning. In addition, by specifying an array of random seed values, a repeated variant of these two methods can be applied, where before each application, random shuffling of the initial dataset is performed based on the corresponding seed value. The final error rate is obtained by averaging over all repeated evaluations.

All three classes implement the `Callbackable`, `Resumable`, `Multithreaded`, and also the `Copyable` auxiliary interfaces (see subsection 3.6). Evaluator multi-threading takes precedence over trainer and classifier multi-threading, which means that in the case of multi-threaded evaluation, the number of threads of the trainer (when the evaluator performs the training multi-threaded) and the classifier will be set to 1 (if they implement the `Multithreaded` interface).

In the case of the `LeaveOneOutEvaluator` class multi-threaded execution requires that both the trainer and the classifier implement the `Copyable` interface (otherwise it will revert to single-threaded execution). This is necessary because the evaluation of the classifier in each iteration is reduced to the classification of a single time series, while all other

time series of the dataset are used for training the classifier, i.e. multi-threaded evaluation requires parallelization of training: each thread must have its own trainer and classifier.

The leave-one-out and cross-validation methods also offer the possibility of sequential evaluation of the classifier over individual splits of the dataset into training and test subsets. In each iteration, the classifier is tested multi-threaded on the test set after training on the training set. This approach does not require the implementation of the `Copyable` interface by either the trainer or the classifier.

In Figures 6 and 7, which illustrate the difference between full and partial parallelization on the example of m-fold cross-validation, `C` denotes the classificator, `C(i)` the i-th copy of `C` trained on the training set `trainset(i)` constructed in the i-th iteration, and `testset(i, j)` the j-th time series of the test set corresponding to the i-th iteration. If both the trainer and the classifier implement the `Copyable` interface (or if no trainer is specified and the classifier implements it), the holdout and cross-validation evaluators apply full parallelization by default.

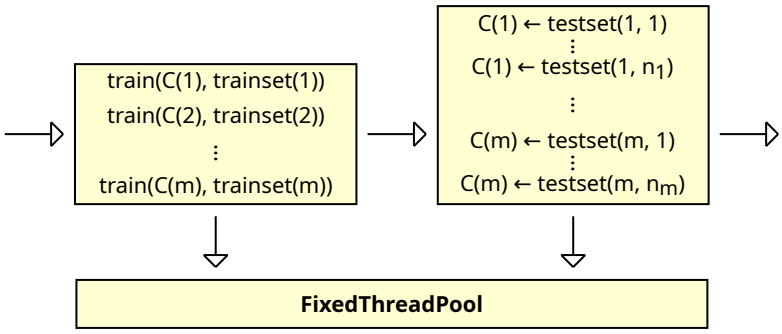


Fig. 6. Full parallelization of m-fold cross-validation

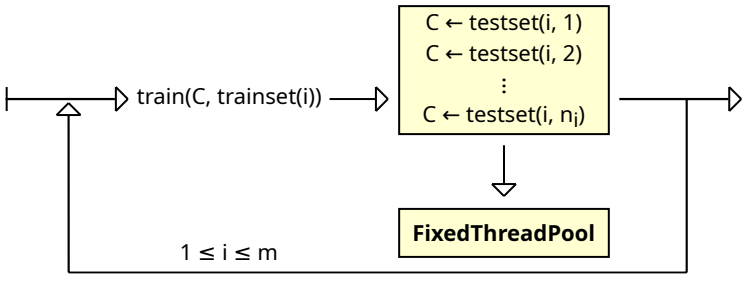


Fig. 7. Partial parallelization of m-fold cross-validation

Listing 2 demonstrates the evaluation of the 1NN classifier paired with the DTW distance measure over the *FiftyWords* dataset using 3 times repeated 10-fold cross-validation. The first parameter of the constructor of the evaluator determines the number of folds (10),

Listing 2. Evaluating the 1NN classifier paired with the DTW distance measure on the FiftyWords dataset with 3 times repeated 10-fold stratified cross-validation

```
Dataset dataset = DatasetUtils.loadDataset("FiftyWords");
Distance distance = new DTWDistance();
Classifier classifier = new NNClassifier(distance);
Trainer trainer = null;
Evaluator evaluator =
          new CrossValidationEvaluator(10, new long[] {1, 2, 3}, 0);

double error = evaluator.evaluate(trainer, classifier, dataset);
ThreadUtils.shutdown(evaluator);
```

the second parameter is an array of seed values (1, 2, 3) that should be used to initialize the random number generator utilized for shuffling the dataset before each run, and the last parameter defines the number of threads (0 means that it should use as many threads as processors are available to the Java Virtual Machine). Due to optimization, the underlying executor service is not automatically shut down after the evaluation is completed. Currently, shutdown should be initiated by the user.

The `getResults()` method of the `HoldoutEvaluator` and `CrossValidationEvaluator` classes returns an array of `FoldResult` objects (Fig. 8) that describe the results of individual iterations. In the case of the Holdout evaluator, iterations represent runs, and in the case of cross-validation, they correspond to folds. The `FoldResult` class defines only public fields that store the training and test sets, the number of misclassified time series of the test set along with the error rate, as well as the expected error and the list of the optimal parameter values found by the trainer (provided that the trainer supports retrieving them by implementing the `ParameterTrainer` interface presented in the next subsection).

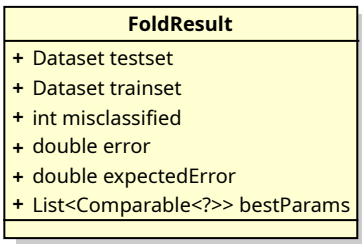


Fig. 8. A class for storing the results of individual iterations of the holdout and cross-validation evaluators

3.5. Trainers

Training a classifier on a given training set is the task of the `train` method declared by the `Train` interface shown in Fig. 9. Its result should be the expected classification error,

which should also be returned by the `getExpectedError` method. In the case of distance-based classifiers, through the `affectsDistance` method, the trainer should report whether it changes the parameters of the distance measures (this information is used by the evaluators described in the previous subsection to optimize the evaluation process when storing calculated distances between pairs of time series is enabled). Its default return value is `false`.

For convenience, the `AbstractTrainer` class stores the expected error and information about whether the training affects the distance measure in fields with protected access level.

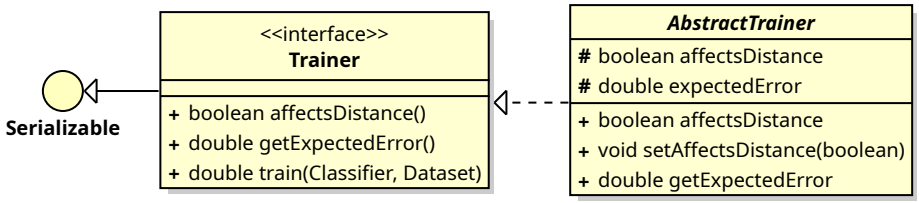


Fig. 9. Classifier trainers and distance measure tuners

The `ParameterTrainer` interface (Fig. 10) of the `fap.trainer` sub-package extends the `Trainer` interface by declaring methods for trainers that tune the value of a single parameter of a classifier or a distance measure (the type `T` of the parameter must implement the `Comparable` interface). Such trainers should provide getter and setter methods for the list of possible parameter values, the evaluator that evaluates their impact on classifier performance, and the sub-trainer that tunes some other parameter of the classifier or distance measure. In this way, specifying sub-trainers opens up the possibility of chaining a series of trainers.

After completing the training, the `getBestValue` method should return the optimal value of the parameter (the one that generated the smallest classification error). The return value of the `getParameters` method should be a list of optimal values of all the parameters tuned by the chained trainers: the first element of the list is the optimal value of the parameter tuned by the given trainer (and returned by the `getBestValue` method), the second element is the optimal value of the parameter tuned by the sub-trainer, and so on. Furthermore, when the `setParameters` method is called, the parameter trainer should set the parameter value to the first value of the specified list, and pass the rest of the list via the same method to the sub-trainer.

The `AbstractParameterTrainer` abstract class (Fig. 10) provides basic fields and methods for parameter trainers (it implements the `ParameterTrainer` interface and extends the `AbstractTrainer` class), including both sequential and parallel finding the optimal value. Since such a general implementation has no knowledge of which parameter it is tuning, nor whether it is a classifier parameter or a distance measure parameter, it is necessary to provide an auxiliary object that will assign the current value with the corresponding parameter. Such an object should implement the `Modifier` interface (Fig. 11), which defines two methods: `set` for assigning a given value to the parameter, and `affectsDistance`, which should report whether the parameter belongs to a distance measure (`true`) or a clas-

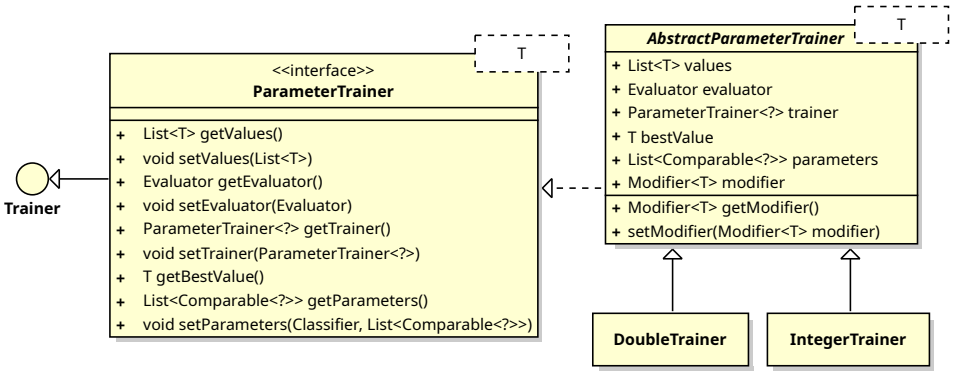


Fig. 10. Types for tuning the values of individual parameters of classifier or distance measures

sifier (`false`). The `DistanceModifier` and `ClassifierModifier` sub-interfaces contain only the corresponding (default) implementation of the `affectsDistance` method.

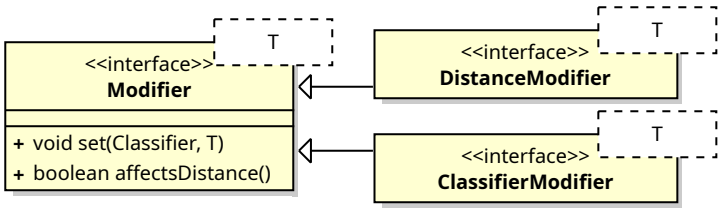


Fig. 11. Interfaces for parameter-modifier classes

An example of using trainers and modifiers to evaluate the weighted kNN classifier based on the Dudani's weighting function [19] and paired with the Sakoe-Chiba [56] constrained DTW [7] distance measure using nested cross-validation [65] is given in Listing 3. The optimal combination of the number of nearest neighbours and the width of the warping window is determined by applying 9-fold cross-validation within each iteration of the 10-fold cross-validation algorithm used to evaluate the classifier performance. The number of nearest neighbors is chosen from the interval between 1 to 10, and the (relative) width of the warping window from the interval between 0% to 25% of the time series length, both values are increased in unit steps.

Based on the result shown in Listing 4, it can be seen that the (average) classification error (rounded to 3 decimal places) was 0.050, i.e. approximately 5% of the time series of the *MoteStrain* dataset were misclassified (the dataset was preprocessed using Paparizzo's script [53]). In the first iteration of the evaluation process, the smallest error (0.046) over the training subset was obtained by a combination of one (1) nearest neighbor and a warping-window width that was 14% of the length of the time series. The actual classification error was 0.023 (calculated over the test subset by applying the classifier and distance measure trained over the training subset).

Listing 3. Evaluation of the weighted kNN classifier using nested cross-validation (utilizing Dudani's weighting function in combination with the Sakoe-Chiba constrained DTW distance measure)

```
Dataset dataset = DatasetUtils.loadDataset("MoteStrain");
Distance distance = new SakoeChibaDTWDistance(true);
Classifier classifier = new DudaniKNNClassifier(distance);

Evaluator subEvaluator = new CrossValidationEvaluator(9);
DoubleTrainer subTrainer =
             new DoubleTrainer(Modifiers.ELASTICITY, 0d, 25d);
subTrainer.setEvaluator(subEvaluator);

IntegerTrainer trainer = new IntegerTrainer(Modifiers.KNN, 1, 10);
trainer.setTrainer(subTrainer);

Evaluator evaluator = new CrossValidationEvaluator(10, 0);
double error = evaluator.evaluate(trainer, classifier, dataset);

ThreadUtils.shutdown(evaluator);

System.out.format("%.3f\n", error);

System.out.println("error, expected, parameters");
for (FoldResult fr : ((CrossValidationEvaluator) evaluator).getResults())
  System.out.format("%.3f, %.3f, " + fr.bestParams + "\n",
                    fr.error, fr.expectedError);
```

The `fap.trainer.Modifiers` class defines a modifier for each parameter of each classifier of the sub-package `fap.classifier.NN` and each distance measure of the sub-package `fap.distance` that relies on parameters. Listing 3 uses two of these modifiers: `KNN` to set the number of nearest neighbors and `ELASTICITY` to set the relative width of the warping window. Their source codes are given in Listing 5.

3.6. Auxiliary Interfaces

The `fap.util` and `fap.callback` sub-package auxiliary interfaces and classes briefly described in this subsection are intended to support mechanisms for monitoring, terminating, and resuming long-running processes, as well as their parallelization.

By implementing the `Resumable` interface (Fig. 12), classes that perform long-running tasks indicate that their execution can be interrupted and resumed from near the breakpoint. Via the `isDone` method, they should report whether the task has already been completed, and the `isInProgress` method should report whether it is still in progress (if the result of both methods is `false`, it means that the execution has not yet started). The function of the reset method is to reset the internal state of the object for reuse (for example, if the same trainer is to be used to train another classifier after finishing training the previous one).

Classes supporting multi-threaded execution must implement the `Multithreaded` interface (Fig. 12), which declares getter/setter methods to set and read the number of threads,

Listing 4. The output of the code shown in Listing 3

```
0.050
error, expected, parameters
0.023, 0.046, [1, 14.0]
0.055, 0.042, [6, 23.0]
0.039, 0.043, [4, 17.0]
0.071, 0.041, [6, 17.0]
0.055, 0.045, [1, 25.0]
0.039, 0.045, [4, 21.0]
0.063, 0.049, [6, 6.0]
0.063, 0.033, [4, 16.0]
0.024, 0.045, [4, 16.0]
0.063, 0.039, [6, 21.0]
```

Listing 5. Implementation of the modifiers of the number of nearest neighbors of the kNN classifier, and the width of the warping window of constrained elastic distance measures

```
public static final ClassifierModifier<Integer> KNN =
                    new ClassifierModifier<>() {

  @Override
  public void set(Classifier classifier, Integer value) {
    ((KNNClassifier) classifier).setK(value);
  }

};

public static final DistanceModifier<Double> ELASTICITY =
                    new DistanceModifier<>() {

  @Override
  public void set(Classifier classifier, Double value) {
    Distance distance =
             ((DistanceBasedClassifier) classifier).getDistance();
    ((ConstrainedDistance) distance).setR(value);
  }

};
```

and for stopping them (threads might not be stopped automatically after completing a task in order to optimize resource usage in case of reusing the same object for executing multiple tasks).

Parallel execution of some tasks requires that each thread be provided with a copy of the objects involved in the process of the task realization. For example, for each parallel partitioning of the dataset into testing and training subsets within repeated holdout evaluation, it is necessary to provide a copy of the trainer and classifier, and if the trainer changes the parameters of the distance measure used by the classifier, then also a copy of the distance measure. The ability of a class to make copies of objects of its type is indicated by implementing the `Copyable` interface (Fig. 12). Whether it is necessary to make a deep copy of the object is indicated by the value of the `boolean` parameter of the `makeACopy` method. For example, if a trainer changes the parameters of a distance measure, different copies of a classifier cannot share the same distance measure, and when copying a classifier, a copy of the associated distance measure must also be made. The parameterless form of this method is a shortcut for deep copying. The result of calling these methods should be a copy of the corresponding object.

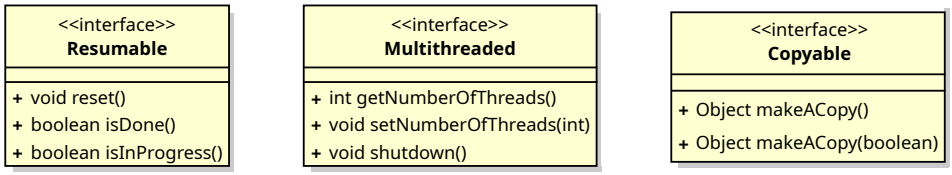


Fig. 12. Interfaces for resumable, multi-threaded and copyable tasks

Supporting monitoring the progress of task execution is indicated by implementing the `Callbackable` interface (Fig. 13) and it should be realized by regularly calling the `callback` method of the provided `Callback` object. The `Callback` object should report the desired number of callbacks via the `getDesiredCallbackNumber` method, and through `setPossibleCallbackNumber`, the `Callbackable` object can indicate the maximum number of callbacks it can perform. For example, if a `Callback` object requests 100 callbacks, but the task in question consists of only 60 steps, the number of possible callbacks should be reported as 60. Initialization and reading of the current number of callbacks should be enabled via the `setCallbackCount` and `getCallbackCount` methods. The purpose of the `getProgress` methods is to map the number of callbacks from the range of the possible number to the range of the desired number of callbacks.

The `AbstractCallback` abstract class provides basic data structures and a basic implementation of the methods of the `Callback` interface. The `fap.callback` sub-package besides it (and the `Callback` and `Callbackable` interfaces) also contains two concrete implementations: the `SystemOutCallback` class prints a specified character on the standard output with each callback, and the `ProgressBarCallback` class displays the progress through a given progress bar.

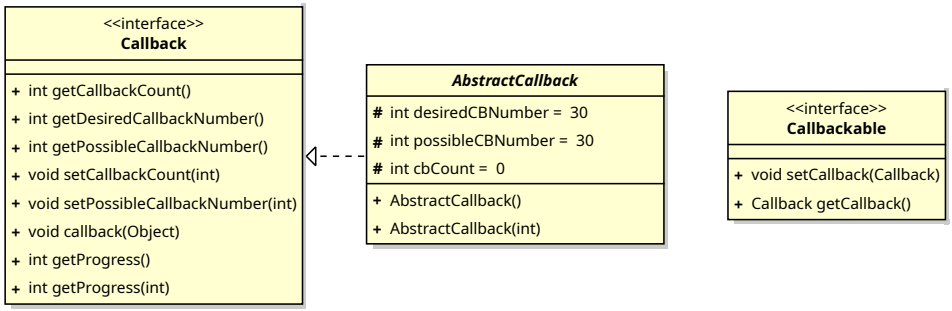


Fig. 13. Interfaces for callbackable tasks

4. Conclusions

In this paper, we presented the basic interfaces and classes around which our FAP library was built and demonstrated the ease of its utilization through examples of applying repeated and nested stratified cross-validation [65] to evaluate the performance of the 1NN and the variant of the weighted kNN classifier based on Dudani's scheme [19], paired with the unconstrained and the Sakoe-Chiba [56] constrained Dynamic Time Warping [7] (dis)similarity measure. In addition, we gave an insight into some more advanced capabilities of the framework, such as storing calculated distances between time series in memory to avoid multiple calculations, multi-threaded classification, evaluation and training of classifiers, and tuning of distance measure parameters for more efficient use of modern, multi-core processors, using pre-generated distance and neighbour matrices to speed up NN classifiers, as well as mechanisms for monitoring, interrupting and continuing interrupted long-term processes (considering environments where the availability of computers to run long-term experiments is not continuous - for example, university computer centers and classrooms).

Motivated by the need to develop a new representation of time series based on cubic splines [40, 43], the library was gradually expanded with new capabilities that enabled its application in both research [42, 39, 24, 10] and education [41]. Currently, the FAP library contains implementations of a number of distance measures based on linear matching of time series data points, the basic elastic measures whose elasticity can be constrained by applying either the Sakoe-Chiba band or the Itakura parallelogram [30], various variants of the NN classifier, multiple representations of time series, the main techniques for evaluating classifier performance (holdout, leave-one-out, cross-validation) with the possibility of multiple repetitions and nested evaluation, as well as classes for training classifiers and tuning parameters of distance measures.

In the future, we plan not only to expand the already existing sub-packages of the FAP library with additional capabilities, but also to implement solutions related to other areas of time series analysis and mining (such as, for example, clustering, anomaly detection, and prediction). Furthermore, believing that it may also be useful to other researchers and practitioners, the FAP library is open source and freely available via GitHub (https://github.com/zgeller/FAP.git).

Acknowledgments. The authors would like to thank Eamonn Keogh for collecting and making available the UCR time-series datasets, as well as everyone who contributed data to the collection. The authors would also like to thank the contribution of Brankica Bratić, Miklós Kálózi, and Aleksa Todorivić to the development of certain parts of the library. Vladimir Kurbalija and Mirjana Ivanović gratefully acknowledge the financial support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grants No. 451-03-66/2024-03/200125 & 451-03-65/2024-03/200125) and the Science Fund of the Republic of Serbia, project #7462, Graphs in Space and Time: Graph Embeddings for Machine Learning in Complex Dynamical Systems – TIGRA.

References

1. Abonyi, J.: Adatbányászat a hatékonyság eszköze. ComputerBooks, Budapest, 1st edn. (2006)
2. Abu Alfeilat, H.A., Hassanat, A.B., Lasassmeh, O., Tarawneh, A.S., Alhasanat, M.B., Eyal Salman, H.S., Prasath, V.S.: Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. Big Data 7(4), 221–248 (dec 2019), https://www.liebertpub.com/doi/10.1089/big.2018.0175
3. Aggarwal, C.C.: Data Mining: The Textbook. Springer Publishing Company, Incorporated (2015)
4. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: David B. Lomet (ed.) Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms (FODO '93), Lecture Notes in Computer Science, vol. 730, pp. 69–84. Springer Berlin Heidelberg (1993), http://link.springer.com/10.1007/3-540-57301-1_5
5. Anh Dau, H., Keogh, E., Kamgar, K., Michael Yeh, C.C., Zhu, Y., Gharghabi, S., Ann Ratanamahatana, C., Chen, Y., Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., Hexagon-ML: The UCR Time Series Classification Archive (2019), https://www.cs.ucr.edu/∼eamonn/time_series_data_2018/
6. Baiocchi, G., Distaso, W.: Gretl: Econometric software for the gnu generation. Journal of Applied Econometrics 18(1), 105–110 (2003), http://www.jstor.org/stable/30035190
7. Berndt, D., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Usama M. Fayyad, R.U. (ed.) Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop. vol. 10, pp. 359–370. AAAI Press, Seattle, Washington (1994), http://dblp.uni-trier.de/rec/bib/conf/kdd/BerndtC94
8. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: Knime: The konstanz information miner. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications. pp. 319–326. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
9. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. John Wiley & Sons, Inc., Hoboken, New Jersey, 5th editio edn. (2015)
10. Bratić, B.: Approximation algorithms for k-NN graph construction. Phd thesis, University of Novi Sad, Serbia (2021), https://nardus.mpn.gov.rs/handle/123456789/18059
11. Cha, S.H.: Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International Journal of Mathematical Models and Methods in Applied Sciences 1(4), 300–307 (2007), http://www.gly.fsu.edu/∼parker/geostats/Cha.pdf
12. Chaovalit, P., Gangopadhyay, A., Karabatis, G., Chen, Z.: Discrete wavelet transform-based time series analysis and mining. ACM Computing Surveys 43(2), 1–37 (jan 2011), https://dl.acm.org/doi/10.1145/1883612.1883613
13. Chen, L., Ng, R.: On The Marriage of Lp-norms and Edit Distance. In: Nascimento, M.A., Özsu, M.T., Kossmann, D., Miller, R.J., Blakeley, J.A., Schiefer, K.B.

(eds.) Proceedings 2004 VLDB Conference, vol. 04, pp. 792–803. Elsevier (2004), https://linkinghub.elsevier.com/retrieve/pii/B978012088469850070X
14. Chen, L., Özsu, M.T., Oria, V.: Robust and fast similarity search for moving object trajectories. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data - SIGMOD '05. pp. 491–502. SIGMOD '05, ACM Press, New York, New York, USA (2005), http://doi.acm.org/10.1145/1066157.1066213
15. Chen, Q., Chen, L., Lian, X., Liu, Y., Yu, J.X.: Indexable PLA for Efficient Similarity Search. In: Proceedings of the 33rd International Conference on Very Large Data Bases. pp. 435–446. VLDB '07, VLDB Endowment (2007), http://dl.acm.org/citation.cfm?id=1325851.1325903
16. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (jan 1967), http://ieeexplore.ieee.org/document/1053964/
17. Das, G., Gunopulos, D.: Time Series Similarity and Indexing. In: Ye, N. (ed.) The Handbook of Data Mining, chap. 11, pp. 279–304. Human factors and ergonomics, Lawrence Erlbaum Associates, Mahwah, N.J. (2003)
18. Deza, M.M., Deza, E.: Encyclopedia of Distances. SpringerLink : Bücher, Springer Berlin Heidelberg, Berlin, Heidelberg (2016), http://link.springer.com/10.1007/978-3-662-52844-0
19. Dudani, S.A.: The Distance-Weighted k-Nearest-Neighbor Rule. IEEE Transactions on Systems, Man, and Cybernetics SMC-6(4), 325–327 (apr 1976), http://ieeexplore.ieee.org/document/5408784/
20. Esling, P., Agon, C.: Time-series data mining. ACM Computing Surveys 45(1), 12:1–12:34 (nov 2012), http://doi.acm.org/10.1145/2379776.2379788
21. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast subsequence matching in time-series databases. ACM SIGMOD Record 23(2), 419–429 (jun 1994), http://dl.acm.org/citation.cfm?id=191843.191925
22. Fix, E., Hodges, J.L.: Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. International Statistical Review / Revue Internationale de Statistique 57(3), 238–247 (dec 1989), http://www.jstor.org/stable/1403797?origin=crossref https://www.jstor.org/stable/1403797?origin=crossref
23. Fu, A.W.c., Leung, O.T.W., Keogh, E., Lin, J.: Finding Time Series Discords Based on Haar Transform. pp. 31–41 (2006), http://link.springer.com/10.1007/11811305_3
24. Geler, Z.: Role of Similarity Measures in Time Series Analysis. Phd thesis, University of Novi Sad, Serbia (2015), https://nardus.mpn.gov.rs/handle/123456789/1703
25. Gilat, A.: MATLAB: An Introduction with Applications. John Wiley & Sons, Inc. (2017)
26. Gou, J., Du, L., Zhang, Y., Xiong, T.: A New distance-weighted k-nearest neighbor classifier. Journal of Information & Computational Science 9(6), 1429–1436 (2012)
27. Gou, J., Xiong, T., Kuang, Y.: A Novel Weighted Voting for K-Nearest Neighbor Rule. Journal of Computers 6(5), 833–840 (may 2011), http://ojs.academypublisher.com/index.php/jcp/article/view/4056
28. Grossmann, W., Rinderle-Ma, S.: Data Mining for Temporal Data. In: Fundamentals of Business Intelligence SE - 6, pp. 207–244. Data-Centric Systems and Applications, Springer Berlin Heidelberg (2015), http://dx.doi.org/10.1007/978-3-662-46531-8_6
29. Han, J., Pei, J., Tong, H.: Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, 4 edn. (2022)
30. Itakura, F.: Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 23(1), 67–72 (feb 1975), http://ieeexplore.ieee.org/document/1162641/
31. Ivanovic, M., Autexier, S., Kokkonidis, M., Rust, J.: Quality medical data management within an open AI architecture – cancer patients case. Connection Science 35(1) (dec 2023), https://www.tandfonline.com/doi/full/10.1080/09540091.2023.2194581
32. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Knowledge and Information Systems 3(3), 263–286 (aug 2001), http://link.springer.com/10.1007/PL00011669

33. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of the 2001 ACM SIGMOD international conference on Management of data. pp. 151–162. SIGMOD '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/375663.375680
34. Keogh, E.J., Chu, S., Hart, D., Pazzani, M.: Segmenting Time Series: A Survey and Novel Approach. In: Last, M., Kandel, A., Bunke, H. (eds.) Data Mining In Time Series Databases, Series in Machine Perception and Artificial Intelligence, vol. 57, chap. 1, pp. 1–22. World Scientific Publishing Company (2004)
35. Keogh, E.J., Pazzani, M.J.: Scaling up dynamic time warping for datamining applications. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 285–289. ACM, New York, NY, USA (aug 2000), https://dl.acm.org/doi/10.1145/347090.347153
36. Kin-Pong Chan, Ada Wai-Chee Fu: Efficient time series matching by wavelets. In: Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337). pp. 126–133. IEEE (1999), http://ieeexplore.ieee.org/document/754915/
37. Kirchgässner, G., Wolters, J., Hassler, U.: Introduction to Modern Time Series Analysis. Springer Texts in Business and Economics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2 edn. (2013), http://link.springer.com/10.1007/978-3-642-33436-8
38. Konasani, V.R., Kadre, S.: Practical Business Analytics Using SAS: A Hands-on Guide. Apress Berkeley, CA (2015), https://link.springer.com/book/10.1007/978-1-4842-0043-8
39. Kurbalija, V.: Time series analysis and prediction using case based reasoning technology. Phd thesis, University of Novi Sad, Novi Sad (oct 2009), http://www.doiserbia.nb.rs/phd/university.aspx?theseid=NS20091005KURBALIJA
40. Kurbalija, V., Ivanović, M., Budimac, Z.: Case-based curve behaviour prediction. Software: Practice and Experience 39(1), 81–103 (jan 2009), http://doi.wiley.com/10.1002/spe.891
41. Kurbalija, V., Ivanović, M., Geler, Z., Radovanović, M.: Two Faces of the Framework for Analysis and Prediction, Part 1 - Education. Information Technology And Control 47(2), 249–261 (jun 2018), http://itc.ktu.lt/index.php/ITC/article/view/18746
42. Kurbalija, V., Ivanović, M., Geler, Z., Radovanović, M.: Two Faces of the Framework for Analysis and Prediction, Part 2 - Research. Information Technology And Control 47(3), 489–502 (sep 2018), http://itc.ktu.lt/index.php/ITC/article/view/18747
43. Kurbalija, V., Radovanović, M., Geler, Z., Ivanović, M.: A Framework for Time-Series Analysis. In: Dicheva, D., Dochev, D. (eds.) Artificial Intelligence: Methodology, Systems, and Applications SE - 5. Lecture Notes in Computer Science, vol. 6304, pp. 42–51. Springer Berlin Heidelberg (2010), http://link.springer.com/10.1007/978-3-642-15431-7_5
44. Larose, D.T., Larose, C.D.: Discovering Knowledge in Data. Wiley, 2 edn. (jun 2014), https://onlinelibrary.wiley.com/doi/book/10.1002/9781118874059
45. Laxman, S., Sastry, P.S.: A survey of temporal data mining. Sadhana 31(2), 173–198 (apr 2006), http://dx.doi.org/10.1007/BF02719780 http://link.springer.com/10.1007/BF02719780
46. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. pp. 2–11. ACM, New York, NY, USA (jun 2003), https://dl.acm.org/doi/10.1145/882082.882086
47. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery 15(2), 107–144 (2007), http://dx.doi.org/10.1007/s10618-007-0064-z
48. Macleod, J., Luk, A., Titterington, D.: A Re-Examination of the Distance-Weighted k-Nearest Neighbor Classification Rule. IEEE Transactions on Systems, Man, and Cybernetics 17(4), 689–696 (jul 1987), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4075685
49. Marteau, P.F.: Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), 306–318 (feb 2009), https://ieeexplore.ieee.org/document/4479483

50. Mitchell, T.M.: Machine Learning. McGraw-Hill, Inc., New York, NY, USA (1997)
51. Mitsa, T.: Temporal Data Mining. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, Taylor & Francis (2010), http://books.google.rs/books?id=4P_7ydvW7cAC
52. Pao, T.L., Chen, Y.T., Yeh, J.H., Cheng, Y.M., Lin, Y.Y.: A Comparative Study of Different Weighting Schemes on KNN-Based Emotion Recognition in Mandarin Speech. In: Huang, D.S., Heutte, L., Loog, M. (eds.) Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, Lecture Notes in Computer Science, vol. 4681, pp. 997–1005. Springer Berlin Heidelberg, Berlin, Heidelberg (2007), http://dx.doi.org/10.1007/978-3-540-74171-8_101 http://link.springer.com/10.1007/978-3-540-74171-8_101
53. Paparrizos, J.: 2018 UCR Time-Series Archive: Backward Compatibility, Missing Values, and Varying Lengths (2019), https://github.com/johnpaparrizos/UCRArchiveFixes
54. Rafiei, D.: On similarity-based queries for time series data. In: Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337). pp. 410–417. IEEE (1999), http://ieeexplore.ieee.org/document/754957/
55. Rafiei, D., Mendelzon, A.: Similarity-based queries for time series data. ACM SIGMOD Record 26(2), 13–25 (jun 1997), https://dl.acm.org/doi/10.1145/253262.253264
56. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26(1), 43–49 (feb 1978), http://ieeexplore.ieee.org/document/1163055/
57. Savaglio, C., Ganzha, M., Paprzycki, M., Bădică, C., Ivanović, M., Fortino, G.: Agent-based Internet of Things: State-of-the-art and research challenges. Future Generation Computer Systems 102, 1038–1053 (jan 2020), https://linkinghub.elsevier.com/retrieve/pii/S0167739X19312282
58. Savic, M., Kurbalija, V., Ilic, M., Ivanovic, M., Jakovetic, D., Valachis, A., Autexier, S., Rust, J., Kosmidis, T.: The application of machine learning techniques in prediction of quality of life features for cancer patients. Computer Science and Information Systems 20(1), 381–404 (2023), https://doiserbia.nb.rs/Article.aspx?ID=1820-02142200061S
59. Sherrington, M.: Mastering Julia - Second Edition: Enhance your analytical and programming skills for data modeling and processing with Julia. Packt Publishing, USA (2024)
60. Tan, P.N., Steinbach, M., Kumar, V., Karpatne, A.: Introduction to Data Mining. Pearson Education, 2 edn. (2019)
61. Vlachos, M., Kollios, G., Gunopulos, D.: Discovering similar multidimensional trajectories. In: Proceedings 18th International Conference on Data Engineering. pp. 673–684. IEEE Comput. Soc (2002), http://ieeexplore.ieee.org/document/994784/
62. Vlachos, M., Lin, J., Keogh, E.J., Gunopulos, D.: A Wavelet-Based Anytime Algorithm for K-Means Clustering of Time Series. In: Workshop on Clustering High Dimensionality Data and Its Applications, at the 3rd SIAM International Conference on Data Mining. San Francisco, CA, USA (2003), https://api.semanticscholar.org/CorpusID:18338443
63. Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E.: Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery 26(2), 275–309 (mar 2013), http://link.springer.com/10.1007/s10618-012-0250-5
64. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edn. (2011)
65. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 4 edn. (2017)
66. Yi, B.K., Jagadish, H.V., Faloutsos, C.: Efficient retrieval of similar time sequences under time warping. In: Data Engineering, 1998. Proceedings., 14th International Conference on. pp. 201–208 (1998)

67. Yi, B.K., Faloutsos, C.: Fast Time Sequence Indexing for Arbitrary Lp Norms. In: Proceedings of the 26th International Conference on Very Large Data Bases. pp. 385–394. VLDB '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000)
68. Zavrel, J.: An Empirical Re-Examination of Weighted Voting for k-NN. In: Proceedings of the 7th Belgian-Dutch Conference on Machine Learning. pp. 139–148 (1997)
69. Zelle, J.: Python Programming: An Introduction to Computer Science 2nd Edition. Franklin, Beedle & Associates Inc., USA (2010)

Zoltán Gellér is an Associate Professor at the Department of Media Studies, Faculty of Philosophy, University of Novi Sad, Serbia. He authored or co-authored 2 textbooks, 1 international monograph, and over 30 publications in data mining, machine learning, computer literacy, and related fields. He was a member of Program Committees of several international conferences and a reviewer in several international journals.

Vladimir Kurbalija holds the position of Full Professor from 2021 at the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Serbia, where he received his B.Sc., M.Sc. and Ph.D. degrees. He was/is a member of several international projects supported by DAAD, TEMPUS, Horizon and bilateral and national programs. Vladimir (co)authored over 60 papers in Case-Based Reasoning, Time-Series Analysis, Medical decision-support systems, and related fields. He was a member of Program Committees of numerous international conferences, and a reviewer in more than 30 international journals.

Mirjana Ivanović is a Full Professor at the Faculty of Sciences, University of Novi Sad, Serbia, since 2002, and a corresponding member of the Serbian Academy of Sciences and Arts since 2024. She is a member of the Board of Directors of the Institute for Artificial Intelligence Research and Development of Serbia. Mirjana has authored or co-authored 17 textbooks, 30 edited proceedings, 4 monographs, and more than 540 research articles on multi-agent systems, e-learning and web-based learning, applications of intelligent techniques (CBR, data and web mining), software engineering education, most of which are published in international journals and proceedings of high-quality international conferences. She has served as a member of program committees for more than 500 international conferences and has chaired numerous international conferences as general chair and program committee chair. Additionally, she has been an invited speaker at numerous international conferences and a visiting lecturer in Australia, Thailand, and China. As a leader and researcher, she has participated in highly regarded international projects.

Received: September 10, 2024; Accepted: January 08, 2025.

Computer Science and Information Systems 22(4):1405–1432 https://doi.org/10.2298/CSIS250221057Z

VSAF: Verifiable and Secure Aggregation Scheme for Federated Learning in Edge Computing

Shiwen Zhang[1,2], Feixiang Ren[1,2], Wei Liang[1,2], Kuanching Li[1,2,⋆], and Al-Sakib Khan Pathan[3]

[1] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
[2] Sanya Research Institute, Hunan University of Science and Technology, Sanya 572024, China
shiwenzhang@hnu.edu.cn
rfx.point@mail.hnust.edu.cn
wliang@hnust.edu.cn
aliric@hnust.edu.cn
[3] Department of Computer Science and Engineering, United International University, Dhaka 1212, Bangladesh
spathan@ieee.org

Abstract. Federated Learning (FL) has gained attention for its promising privacy protection. In FL, clients train local gradients on their data without sharing raw data to update the global model. However, security issues persist. Attackers can infer original data from local gradients, compromising privacy, while a malicious cloud server may tamper with uploaded parameters, leading to incorrect aggregation. Considering this, we focus on the above issues in FL: (1) privacy protection of the parameters uploaded by clients and (2) verification of the correctness of the aggregated result from a cloud server. In response to these issues, this article proposes VSAF, a verifiable and secure aggregation scheme for federated learning in edge computing. Using a linear homomorphic hash function, we design a lightweight verification algorithm for aggregated gradients. To protect gradient privacy, we combine the Bloom filter and Shamir's secret sharing to design a single masking protocol. Detailed analyses and experiments demonstrate the security and efficiency of the proposed scheme.

Keywords: Federated Learning, Privacy-preserving, Correctness Verification, Edge Computing.

1. Introduction

As the Internet of Things (IoT), along with mobile devices, becomes more widespread, more and more computing tasks can be processed on the edge devices [6,33,37]. In recent years, edge devices have become increasingly intelligent and more powerful, enabling the use of edge computing [26]. This allows for transferring computing tasks and stored data from central servers to devices at the network's edge [34, 35]. As a result, both computing efficiency and data privacy protection are improved. Therefore, how to effectively use

⋆ Corresponding author

these large numbers of IoT devices and the data they generate has become a hot research topic in academia and industry [26, 36]. Many institutions and enterprises are conducting machine learning on edge nodes [5, 7, 11]. For example, the Google team has used its users' smartphones for training to predict the next word on the virtual keyboard and perform a music recognition search [11]. However, a critical and sensitive issue is that users would not like Google to access their private data for these services.

To address such an issue, Federated Learning (FL) has received widespread attention since it has performed well in privacy protection and ensured data security. FL is a technology that can achieve distributed machine learning whilst protecting data privacy [25, 27, 31]. In edge computing scenarios like vehicular networking, healthcare, and finance, FL has been widely used to achieve cross-device, cross-platform, and cross-institutional machine learning cooperation. Nonetheless, FL still has two issues that need to be addressed.

The first is how to prevent privacy leakage [12, 18, 24, 41]. Attackers or servers can infer information about the dataset used for training by the client from the gradients uploaded by the client, resulting in privacy leakage of the client. Some works depict this attack. For instance, Melis *et al.* [18] proved that the uploaded gradients may expose the privacy of clients' local data. Zhu *et al.* [41] trained on an image dataset and proved that the client's gradients would leak information about the images in its private dataset. Again, the second issue effectively verifies the aggregated result's correctness [29, 38]. It is possible for a malicious central server to modify the aggregated result of gradients and return incorrect results, leading to a failure of convergence of the training model. In addition to this, a lazy server may deliberately omit some users' gradients to save computational overhead and only aggregate the gradients of some other users, resulting in an inaccurate global model and affecting the model's convergence and efficiency [16, 29]. Hence, verifying the correctness of aggregated results while simultaneously protecting users' data privacy in an edge computing environment is challenging.

To solve the privacy protection problem, some researchers have contributed their mechanisms [1, 2, 4, 20] on privacy-protected FL. Phong *et al.* [20] used additive homomorphic encryption to protect model parameters and achieve secure aggregation. However, homomorphic encryption generates higher communication and computational overheads. M. Abadi *et al.* [1] designed a deep learning framework that integrates differential privacy and a gradient descent algorithm to protect users' data privacy. Nevertheless, differential privacy can cause an accuracy loss problem. Another more direct method is to blind the gradient directly. Keith Bonawitz *et al.* [4] introduced a double masking scheme, based on the (t, n) threshold secret sharing, to protect users' gradients. However, this scheme has a considerable restriction on the threshold t; that is, if the threshold t is less than or equal to $\lfloor \frac{n}{2} \rfloor$, the cloud server may divide all users into two sets: A and B on average, deceive users in set A (or B) to obtain the secret shares of users in set B (or A); thereby infringing on the privacy of users in set B (or A). Of course, if the value of t is too small, the cloud server can divide more sets and then resist such a privacy attack. In addition, the double masking scheme masks the user's local gradient twice to protect user privacy, and two masks will generate more computational overhead and communication overhead to a certain extent.

To verify the aggregated result's correctness in FL, existing schemes [16, 29] use homomorphic hash functions HH to verify the aggregated result. Each user uses HH to

generate proof for local gradients and uploads the hash values to the cloud server. The users receive the aggregated result of proofs from the server and verify the correctness of the gradients' aggregated result by determining if the aggregated result of proofs is consistent with the proof of the gradients' aggregated result. However, researches [16,29] allows the cloud server to collude with some users, and then the cloud server can obtain the homomorphic hash function owned by the users. Thus, after maliciously modifying the gradients' aggregated result, the server can change the aggregated result of hash values to prevent users from detecting this malicious behavior [9].

To avoid this problem, researchers [8,13,40] delegate the authority to aggregate proofs to each user, as each user sends gradient's proof to the cloud server, and the cloud server broadcasts them to the other users. Next, each user verifies the other users' proofs. After verification passes, all proofs are aggregated and used to verify the correctness of the aggregated result. These approaches increase the users' computational and communication overhead and cannot resist lazy servers' deletion attacks (Deletion attack refers to the sluggish behavior of lazy servers to save computing resources and communication overhead by only summarizing or broadcasting some users' data. From the user's perspective, it is as if the server has 'deleted' some users' data). Therefore, these approaches may cause inaccurate aggregated results and affect model convergence efficiency.

To address the above issues, we propose a verifiable and secure aggregation scheme for federated learning in edge computing called VSAF. To protect the privacy of user gradients, we design a single masking protocol based on the (t, n) threshold secret sharing mechanism and the Bloom technique. This protocol supports the dropout of some users while also protecting their privacy. To verify the correctness of the aggregated result while also discovering lazy servers, we developed a lightweight verification algorithm. This algorithm combines a homomorphic hash function with dual servers, reducing the computation and communication overheads of the user for verification. The users send the proofs for verification to the Trusted Authority (TA) and local gradients to the aggregation server, preventing the lazy and tampering behavior of the server by leveraging the mutual distrust between these two servers. To optimize verification efficiency, we plan to outsource the verification operation to the TA to reduce the user's computation overhead. In addition, since we use the homomorphic hash function for verification, the overhead for verification is independent of the gradient dimension, which can reduce the user's computation and communication loads.

The key contributions of this work are as follows:

(1) Design a single masking scheme to protect the privacy of user gradient and also tolerate the dropout of some users. Compared with the double masking scheme, we lift the restriction on the threshold t and simultaneously reduce some communication and computation overheads.
(2) Put forward a lightweight verification algorithm that leverages linear homomorphic hash function to realize the verification for the correctness of aggregated results. This method gives the aggregation authority of hash values to a trusted authority, which reduces the computation overhead of user verification, and it can also detect the lazy aggregation server in time.
(3) Achieve that the communication overhead for verification is independent of the gradient dimension, the dropout rate, and the number of users; thereby diminishing the communication overhead for verification.

(4) Implementation and evaluation VSAF. The comprehensive theoretical analysis and experimental results of the proposed scheme demonstrate its security and efficiency.

The remainder of this article is organized as follows: we first review the related work in Section 2 and introduce the preliminaries in Section 3. Then, we depict the problem statement in Section 4, followed by the description of our scheme VSAF in Section 5. After that, we analyze the security of VSAF in Section 6, and evaluate the performance in Section 7, and finally, concluding remarks and future directions are given in Section 8.

2. Related Work

2.1. Privacy Protection Schemes in Federated Learning

To address the privacy leakage problem caused by intermediate parameters in federated learning, many privacy protection schemes [1,4,15,20,23] have been proposed. To prevent attackers from recovering the training set from intermediate parameters through numerical methods, Phong *et al.* [20] use additive homomorphic encryption to protect model parameters and achieve secure aggregation. However, all participants employed the same key for the encryption and decryption of the model parameters. If any participant leaks the key pair to the attackers, the privacy of all participants will be at risk of being revealed. In addition, homomorphic encryption has a high computational overhead. Li *et al.* [15] use homomorphic encryption to encrypt the training data and directly train on the ciphertext, so thus, the data privacy is protected, but the computational overhead is still significant.

Bonawitz *et al.* [4] propose a double masking scheme, implemented based on secure multi-party computation and pseudo-random generator. This scheme can achieve privacy protection for the parameters uploaded by participants while also achieving robustness for users who are dropping out. However, it incurs high communication and computational overheads. Shokri *et al.* [23] propose a joint deep learning framework that prevents the server from directly accessing the training dataset and uses differential privacy to perturb some of the gradients. Abadi *et al.* [1] introduce a deep learning scheme that integrates differential privacy with a gradient descent algorithm, adding appropriate Laplace noise during gradient descent so that the local gradient satisfies differential privacy. However, the differential privacy can lead to the model's accuracy loss. Although these works [1,23] achieve privacy protection with smaller computational and communication overheads, it is necessary to balance privacy and accuracy.

2.2. Verifiable Aggregation Schemes in Federated Learning

Several researches on verifiable federated learning have been done in recent years, such as [10, 16, 19, 30, 32, 39, 40], and many of them are verifiable federated learning schemes focused on verifying the correctness of aggregated results [10, 16, 19, 30, 40]. Specifically, these schemes detect malicious or lazy dishonest behavior of aggregation servers by verifying the correctness of aggregated results. Other schemes [32, 39] focus on various aspects of verification, which are mainly related to detecting server failure issues and verifying the integrity of the gradients.

In [10,19,30], the server returns the aggregated result and its proof to the users, so they can utilize the proofs to verify the correctness of aggregated results and justify whether

the server is trusted or malicious. Zhou *et al.* [40] utilized homomorphic hashing combined with signature techniques to verify the aggregated result, in which all clients must take part in the verification process to verify the correctness of the parameters of other clients. However, with the increase in the number of participants, the time cost of the verification process increases. Li Lin *et al.* [16] proposed a discrete logarithm-based verification scheme that verifies the correctness of the aggregated result and discovers inert cloud servers simultaneously. However, in the above schemes, each user must validate other users' data before verifying the accuracy of the aggregated result. Furthermore, the increase in the number of users will lead to an increase in verification operations. Therefore, the larger the number of users, the higher the verification costs. At the same time, the communication overhead of the verification operation is also increased linearly with the dimension of the gradient.

To detect the server failure issues and verify the integrity of the gradients, Zhao *et al.* [39] and Zhang *et al.* [32] have proposed different schemes. Zhao *et al.* [39] introduced the PVD-FL framework that employs a cryptographic-based matrix multiplication (EVCM) algorithm for the encryption and verification of parameters. However, it can only verify the incorrect aggregated result caused by problems such as insufficient computing power and device failure of honest users and does not support detection and verification of malicious behavior by the dishonest cloud server. Zhang *et al.* [32] designed a verifiable federated learning scheme based on an online/offline signature method that realizes the integrity verification of gradients during the transmission process. However, this scheme cannot verify the incorrect aggregated result or detect the malicious behavior of the cloud server.

Unlike previous works, we propose a verifiable and secure aggregation scheme for federated learning in edge computing (VSAF). The proposed scheme can protect users' privacy, detect the aggregation server's tampering and lazy behavior, and effectively reduce verification communication and computational overheads.

3. Preliminaries

3.1. Federated Learning

The general federated learning framework is shown in Fig. 1, existing one cloud server and N clients. Each client is a user, denoted as $u_{i:1\leq i\leq N}$, who trains a local gradient $x_{i:1\leq i\leq N}$ based on the local dataset $D_{i:1\leq i\leq N}$ and sends x_i to server.

During this training process, the user u_i's gradient is typically computed using the Stochastic Gradient Descent (SGD) algorithm. Specifically, u_i first uses the global model W and the local dataset D_i to compute the gradient, x_i

$$x_i = \nabla L(W, D_i) \tag{1}$$

Here, x_i represents the direction of the steepest descent. The loss function is denoted by $L(\bullet)$, and its derivative is represented by $\nabla L(\bullet)$. The inputs to this function are the global model W and the dataset D_i.

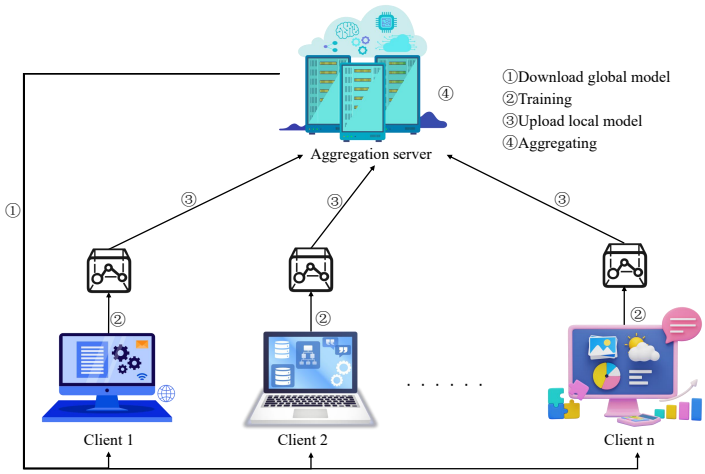


Fig. 1. The general federated learning framework

After receiving enough gradients, the cloud server acts as an aggregation server to aggregate the gradients and send users the aggregated result, which is computed as follows:

$$z = \sum_{i=1}^{N} x_i \tag{2}$$

in which z is the aggregated result.

The users update the global model W based on z and use W to carry out the next round of training. The global model update is computed as follows:

$$W = W - \eta \frac{z}{N} \tag{3}$$

in which, η is the learning rate.

Finally, repeat the above steps until the model converges or reaches the desired training accuracy.

3.2. Secret Sharing

This scheme employs Shamir's (t, N)-threshold secret sharing protocol [22]. The secret is divided into N shares, where N represents the number of users. A pre-set threshold, t, is established. The secret can only be reconstructed by gathering at least t shares. Precisely, the (t, N) threshold secret sharing protocol consists of the following steps:

(1) $\{(u_i, s_i)\}_{u_i \in U} \leftarrow S.share(s, t, U)$: This sharing algorithm divides the secret s into N shares. The inputs include the secret s, the threshold t (satisfying $t \leq \|U\|$), and the user set U ($\|U\| = N$ represents the number of users in the user set). The output is the share s_i for each user u_i.

(2) $s \leftarrow S.recon(\{(u_i, s_i)\}_{u_i \in U'}, t)$: This is a reconstruction algorithm. The inputs include the secret share s_i from users in the set U' and the threshold t, where $u_i \in U' \subseteq U$ and $t \leq u'$. The output of the algorithm is the secret s.

3.3. Homomorphic Hash

We use homomorphic hash functions to construct our verification scheme to achieve verification for the correctness of the aggregated result from the server and to defend against forgery attacks and deletion attacks from the server. Here, forgery attacks refer to when the Cloud Server (CS) forges aggregated results and sends them to the users.

If a file is divided into several file blocks, the homomorphic hash function can independently compute the hash value for each block. By aggregating these hash values, we can derive the hash value for the entire file. This scheme's homomorphic hash algorithm comprises three main components:

(1) $h_i \leftarrow HH.hash(x)$: the Hash algorithm, input, and output are a d-dimensional vector x and a hash of x. In this scheme, the d-dimensional vector x refers to the user u_i local gradient, and h_i is the gradient's hash value of the user u_i. The specific calculation process of the hash value h_i can be expressed as follows:

$$h_i = \prod_{j=1}^{d} g_j^{x_j} \ mod \, p \tag{4}$$

Here, we randomly select d distinct elements from the cyclic group G of prime order q, where g_j is the jth element. x_j denotes the jth dimensional element of vector x and p is a large prime number.

(2) $HH\_proof \leftarrow HH.agregate(h_i)$: This is the aggregation algorithm for hash values. The input is the hash value of each user, and the output is the aggregated value of all the user hashes, called HH_proof in this scheme, which is used to achieve verification for the accuracy of aggregated results from the server.

(3) $HH\_va \leftarrow HH.verify(HH\_proof, z)$: This is a verification algorithm designed to check the correctness of the aggregated result from the cloud server. The inputs are the aggregated result HH_Proof of all the user hashes and the aggregated result z of all the user local gradients. The output is HH_va, representing the evaluation of the aggregated result, and its value is either 0 or 1: the result is correct, and verification passes with value 1, and 0 if otherwise. The above process is specifically formulated as follows:

$$HH\_proof \stackrel{?}{=} HH.hash(z) \tag{5}$$

3.4. Bloom Filter

This work employs the Bloom filter [3] to efficiently determine whether a query element belongs to a given set S of n elements. Using k independent hash functions $BFH_1, BFH_2, \ldots, BFH_k$, each element in S is mapped to k positions in an m-bit vector BF, initially set to 0. Adding an element involves setting the k mapped positions to 1. To check membership of an element w, $BFH_i(w)$ $(1 \leq i \leq k)$ is used to verify if all k positions are 1. If any position is 0, $w \notin S$; otherwise, w is assumed to be in S. While Bloom filters optimize query efficiency and memory usage, they are prone to false positives, where $w \notin S$ but is falsely identified as a member. The false positive rate is $(1 - e^{\frac{nk}{m}})^k$, minimized to 2^{-k} when $k = (\ln 2) \cdot \frac{m}{n}$.

In this scheme, we use the Bloom filter to defend the server against deceiving attacks by verifying that the user requested by the server is dropped. The process unfolds as follows:

(1) $BF_i \leftarrow BFH(u_i)$: This is a hash algorithm, where the input is the identity number u_i of the online user and is mapped onto the vector BF_i through k independent hash functions $BFH_{i:1\leq i\leq k}$. BF_i represents the online status of the user u_i and is the proof that the user is online.
(2) $BF \leftarrow BF.aggregate(BF_i)$: This aggregation algorithm will map all online users to a vector BF. The input is the mapping vector for each user, and the output vector represents the online status of all users. The calculation process of vector BF is:

$$BF = \bigcup_{u_i \in \mu} BF_i \tag{6}$$

where μ denotes the set of online users.
(3) $BF\_va \leftarrow BF.verify(BF, S)$: This is a verification algorithm. The input is the set S of some users and the vector BF that reflects the user's online status. This algorithm is used to verify whether the users in set S belong to online users. The output value is 0 or 1. If $BF\_va = 0$, it indicates that users in set S are not online. If $BF\_va = 1$, it indicates that the users in set S are online users.

3.5. Key Agreement

This scheme uses the Diffie-Hellman (DH) key agreement protocol to create a secure channel between any two users, and this channel is used to negotiate the generation of shared random numbers. Specifically, we acknowledge that a group G has a prime order q, and g is the generator of G. Subsequently, the DH protocol in this scheme is composed of these two algorithms:

(1) $(sk_i, g^{sk_i}) \leftarrow DH.gen(G, p, q)$: This algorithm is a key pair generation algorithm. The output key pair (sk_i, g^{sk_i}) is the public and private keys of the user u_i respectively.
(2) $sk_{i,j} \leftarrow DH.agree(sk_i, g^{sk_j})$: The algorithm is a shared key generation algorithm. The inputs are the private key sk_i and the public key g^{sk_j}, which belong to user u_i and user u_j, respectively. The output is the shared key $s_{i,j}$, enabling secure communication between users u_i and u_j.

4. Problem Statement

In this section, we initially present the system model, followed by an introduction to the threat model, our design goals, and an overview of our VSAF.

4.1. System Model

The proposed system contains three types of entities: N users provide data at the edge, a cloud server (CS), and a trusted authority (TA), as depicted in Fig.2 the system model of the proposed VSAF.

Users: The participants who join federated learning are typically computing nodes at the network edge, such as smartphones, computers, and other IoT devices. Each user has a local dataset and is trained based on the local dataset.

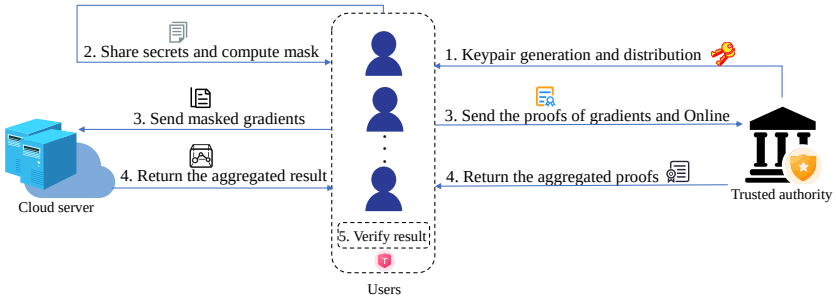


Fig. 2. System model of VSAF

Cloud Server (CS): In this system, the cloud server collects the local gradients that the users upload and aggregates the gradients. Then, the server sends the aggregated result to the users. Using this result, users update the global model and proceed to the next round of training.

Trusted Authority (TA): TA mainly generates and distributes initialization parameters in this system. Specifically, TA generates some initialization parameters for the user, such as the initial global model and the key pairs for encryption. In addition, it is responsible for generating the users' online proofs, which can help the users resist the deceiving attack of the server. TA also generates the aggregation verification proof, assisting the users to verify the correctness of the aggregated result from the server.

4.2. Threat Model

We assume the following threat model:

(1) **Malicious and lazy CS:** We consider that CS may tamper with users' data or omit some users' data to save computational load during the aggregation process. Therefore the CS in this scheme is not honest but a curious server.
(2) **Semi-honest Users:** In this system, the users perform the protocol honestly despite interested in other users' data, and attempting to get access to other users' data.
(3) **TA:** TA is trustworthy, and will strictly adhere to the protocol, and will strive to maintain the privacy of users.

In this scheme, we have the following assumptions for the threat model: the CS cannot collude with the TA but may collude with fewer than t users and tamper with or delete user-uploaded data during aggregation. The CS can also launch deception and deletion attacks during parameter aggregation and verification to extract users' local gradients or compromise the aggregated result.

Deceiving attacks: While resolving the dropout problem, since the CS is not trustworthy, it may deceive the online users by falsely claiming that certain users have dropped out to obtain their secret shares. Then, CS reconstructs the masks of these users, enabling it to compute their local gradients and thus violate their privacy.

Deletion attacks: In typical verification schemes, users send their local gradients and corresponding proofs to the CS, which then aggregates and distributes the results along

with the proofs to the users for verification. However, to reduce computational and communication overhead, the lazy CS may aggregate only the gradients from a subset of users and send the aggregated result along with their proofs. It appears the server has "deleted" the data from specific users, and users cannot detect this deletion through the verification algorithm.

4.3. Design Goals

Specifically, this work should meet the following design goals:

- **Privacy.** Our proposed VSAF aims to protect users' gradients' privacy, preventing adversaries from recovering users' sensitive data from the gradients and defending against deceiving attacks.
- **Verification.** VSAF should be able to detect deletion attacks to ensure the correctness of the aggregated results.
- **Efficiency.** VSAF aims to achieve efficient communication and computation for verification, saving on communication and computational overheads.

4.4. Overview of VSAF

The processes of VSAF are divided into four rounds: negotiation and distribution of keys, generation and distribution of shares, generation and uploading of ciphertexts and proofs, and aggregation and verification of parameters. We will explain these four rounds in detail in the next section.

(1) **Negotiation and Distribution of Keys (Round 1):** TA generates key pairs for each user and public parameters HH_{pp} and BF_{pp} for the homomorphic hash function and bloom filter algorithm. Then, TA distributes key pairs and public parameters to the corresponding users.
(2) **Generation and Distribution of Shares (Round 2):** All users share their one private key through Shamir's secret sharing technique. Then, the users encrypt the shares and distribute the ciphertext of the shares to the corresponding users with the help of the server.
(3) **Generation and Uploading of Ciphertexts and Proofs (Round 3):** According to the DH protocol, any two users negotiate a shared key and then use this key as a seed to generate a mask for encrypting the local gradient. Each user uses the bloom filter and homomorphic hash function to generate online proof and verification proof. Finally, the ciphertexts and proofs are sent to CS and TA.
(4) **Aggregation and Verification of Parameters (Round 4):** The users verify the CS's request based on the online proof and upload the secret shares of the dropped users. CS recovers the masks using the secret shares, then derives the unbiased aggregated result based on these masks, and finally sends this result to online users. The users verify the correctness of the aggregated result based on the proof of verification.

Implementation of VSAF

Round 1 (*Negotiation and Distribution of Keys*)

TA :

- Generate two key pairs $(N_i^{pk}, N_i^{sk}) \leftarrow DH.gen(G,p,q)$ and $(M_i^{pk}, M_i^{sk}) \leftarrow DH.gen(G,p,q)$ for each user u_i.
- Generate public parameter HH_{pp} of the linear homomorphic hash algorithm.
- Generate public parameter BF_{pp} of the bloom filter algorithm.
- Send (N_i^{pk}, N_i^{sk}), (M_i^{pk}, M_i^{sk}), HH_{pp} and BF_{pp} to the corresponding users.

$User\ u_i$:

- Receive $\{(N_i^{pk}, N_i^{sk}), (M_i^{pk}, M_i^{sk}), u_i^{id}, HH_{pp}, BF_{pp}\}$ from TA.
- Send two public keys (N_i^{pk}, M_i^{pk}) to CS, and use the CS to broadcast the public keys (N_i^{pk}, M_i^{pk}) to other users.

CS :

- Receive (N_i^{pk}, N_i^{sk}) from users. Set μ_1 as the set of users whose messages are received by CS, in which $\mu_1 \subseteq \mu$ and μ_1 is the set of all users. Moreover, ensure that $|\mu| \geq t$.
- Generate the id number u_i for user $u_i \in \mu_1$, so that CS can number the users who sent the messages, where all the u_i are different from each other and used to generate users' online proofs.
- Send $\{u_i, N_i^{pk}, M_i^{pk}\}_{u_i \in \mu_1}$ to each user $\in \mu_1$.

Round 2 (*Generation and Distribution of Shares*)

$User\ u_i$:

- Receive $\{u_i, N_i^{pk}, M_i^{pk}\}_{u_i \in \mu_1}$ from CS.
- Generate the shares of N_i^{sk} as $\{(u_i, s_{i,j})\}_{u_j \in \mu_1} \leftarrow S.share(N_i^{sk}, t, \mid \mu_1 \mid)$, where $s_{i,j}$ is the share of user u_i to user u_j.
- Compute $C_{i,j} \leftarrow DH.Enc(DH.agree(M_i^{sk}, M_j^{pk}), u_i \mid\mid u_j \mid\mid s_{i,j})$, where $C_{i,j}$ is the ciphertext of user u_i to user u_j.
- Send the ciphertext $\{C_{i,j}\}_{u_j \in \mu_1}$ to CS.

CS :

- Receive the ciphertext $\{C_{i,j}\}_{u_j \in \mu_1}$ from user $u_i \in \mu_2$, in which $\mu_2 \subseteq \mu_1$ and μ_2 is the set of users whose messages are received by CS. In addition, ensure that $\mid \mu_2 \mid \geq t$.
- Broadcast the ciphertext $\{C_{i,j}\}_{u_i \in \mu_1}$ to user $u_j \in \mu_2$.

Round 3 (*Generation and Uploading of Ciphertexts and Proofs*)

$User\ u_i$:

- Receive the ciphertext $\{C_{j,i}\}_{u_i \in \mu_2, u_j \in \mu_1}$ and ensure that $\mu_2 \subseteq \mu_1$ and $\mid \mu_2 \mid \geq t$.
- Negotiate the shared key $S_{i,j} \leftarrow DH.agree(N_i^{sk}, N_j^{pk})$ with user $u_j \in \mu_2$ according to the DH protocol, where $S_{i,j}$ is the shared key between user u_i and user u_j.
- Compute the masks $msk_{i,j} \leftarrow PRG(S_{i,j})$, where $PRG()$ is a pseudorandom generator.
- Get a local gradient x_i, after training on local dataset.
- Mask the gradient x_i as $\widehat{x}_i \leftarrow x_i + \sum_{u_j \in \mu_2 : i<j} msk_{i,j} - \sum_{u_j \in \mu_2 : i>j} msk_{i,j}$.
- Compute the verification proof of the local gradient $h_i \leftarrow HH(x_i)$.
- Compute the online proof $BF_i \leftarrow BFH(u_i^{id})$, where u_i^{id} is the identity number for user u_i.
- Send the encrypted gradient $\{\widehat{x}_i\}$, and proofs $\{h_i, BF_i\}$ to CS and TA, respectively.

CS :

- Receive $\{\widehat{x}_i\}$ from user $u_i \in \mu_3$, where μ_3 is the set of users who send messages to CS and TA. Besides, ensure that $\mu_3 \subseteq \mu_2$, $\mid \mu_3 \mid \geq t$.
- Send a list of $\mu_2 \backslash \mu_3$ to each user in μ_3.

TA :

- Receive $\{h_i, BF_i\}$ from user $u_i \in \mu_3$, where μ_3 is the set of users who send messages to CS and TA. Furthermore, ensure that $\mu_3 \subseteq \mu_2$, $\mid \mu_3 \mid \geq t$.
- Compute the verification proof of the aggregated result as $HH\_proof \leftarrow HH.aggregate(h_i)$.
- Compute the online proof of all the users in the set μ_3 as $BF \leftarrow BFH.aggregate(BF_i)$.
- Send the proofs $\{HH\_proof, BF\}$ to each user $u_i \in \mu_3$.

Round 4 (*Aggregation and Verification of Parameters*)

$User\ u_i$:

- Receive a list of $\mu_2 \backslash \mu_3$ form CS and the proofs $\{HH\_proof, BF\}$ from TA.
- Compute $BF\_va \leftarrow BF.verify(BL, \mu_2 \backslash \mu_3)$, where BF_va is used to determine whether the users in $\mu_2 \backslash \mu_3$ are online or dropout. If the verification passes, then continue. Otherwise, abort and start over.
- Decrypt $C_{i,j}$ from dropped users as $\{u_i \parallel u_j \parallel s_{i,j}\} \leftarrow DH.Dec(DH.agree(M_i^{sk}, M_j^{pk}), C_{i,j})$.
- Send $\{u_i \mid\mid u_j \mid\mid s_{i,j}\}_{u_j \in \mu_2 \backslash \mu_3}$ to CS.

CS :

- Receive $\{s_{i,j}\}_{u_j \in \mu_2 \backslash \mu_3}$ form users $u_i \in \mu_4$, where $\mu_4 \subseteq \mu_3$ and $\mid \mu_4 \mid \geq t$, otherwise, abort and start over.
- Reconstruct private keys $N_i^{sk} \leftarrow S.recon(\{u_i,\ s_{i,j}\}_{u_j \in \mu_4}, t)$ for users $u_i \in \mu_2 \backslash \mu_3$.
- Compute the shared key $S_{i,j} \leftarrow DH.agree(N_i^{sk}, N_j^{pk})$ and the masks $msk_{i,j} \leftarrow PRG(S_{i,j})$, where $u_i \in \mu_2 \backslash \mu_3$ and $u_j \in \mu_3$.
- Compute the aggregated result as

$$\sum_{u_i \in \mu_3} x_i \leftarrow \sum_{u_i \in \mu_3} \widehat{x}_i - \sum_{\substack{u_i \in \mu_3, \\ u_j \in \mu_2 \backslash \mu_3 : i<j}} msk_{i,j} + \sum_{\substack{u_i \in \mu_3, \\ u_j \in \mu_2 \backslash \mu_3 : i>j}} msk_{i,j}$$

- Broadcast the aggregated result $z = \sum_{u_i \in \mu_3} x_i$ to the user $u_i \in \mu_4$.

$User\ u_i$:

- Receive the aggregated result z form CS and verify the correctness of the aggregated result z as $HH\_va \leftarrow HH.verify(HH\_proof, z)$.
- Accept z and move to Round 1, if the $HH\_va = 1$. Otherwise, abort and start over.

Fig. 3. The detailed description of VSAF

5. The Proposed VSAF

The proposed scheme VSAF includes three types of entities: N users at the edge, a cloud server (CS), and a trusted authority (TA). In addition, it also includes four rounds: **Negotiation and Distribution of Keys (Round 1)**, **Generation and Distribution of Shares**

(Round 2), **Generation and Uploading of Ciphertexts and Proofs (Round 3)**, **Aggregation and Verification of Parameters (Round 4)**. The interaction details of each entity in different rounds are shown in Fig. 3.

5.1. Negotiation and Distribution of Keys (Round 1)

TA mainly generates key pairs (N_i^{pk}, N_i^{sk}) and (M_i^{pk}, M_i^{sk}) for users, as well as the necessary public parameters HH_{pp} and BF_{pp} for the homomorphic hash function and the bloom filter algorithm. Then, the TA assigns an identifier to each user. After that, it sends the key pairs and the public parameters to the corresponding users. Specifically, the key pairs generated by TA for user u_i are as follows:

$$(N_i^{pk}, N_i^{sk}) \leftarrow DH.gen(G, p, q) \tag{7}$$

Similarly, (M_i^{pk}, M_i^{sk}) is also generated in this way. N_i^{pk} and M_i^{pk} are public keys, and N_i^{sk} and M_i^{sk} are private keys.

After receiving messages from TA, all users will use the CS to broadcast the public keys N_i^{pk} and M_i^{pk} to other users. Set μ_1 as the users who send messages to the CS. After receiving the messages, the CS generates the identifier u_i to number the users who sent these messages and then sends $\{u_i, N_i^{pk}, M_i^{pk}\}_{u_i \in \mu_1}$ to each user. Note that all users have only two states: online or dropped.

5.2. Generation and Distribution of Shares (Round 2)

Let μ_2 be the set of users who perform secret sharing in Round 2. After completing the Round 1, the user $u_i (u_i \in \mu_2)$ secretly shares private key N_i^{sk} to other users. The purpose of secret sharing is to allow the CS to request the secret shares of the dropped users from the online users, thereby recovering the dropped users' private keys. Subsequently, the CS calculates the masks to correct its aggregated result. The secret share procedure is shown below:

$$\{(u_i, s_{i,j})\}_{u_j \in \mu_1} \leftarrow S.share(N_i^{sk}, t, \mid \mu_1 \mid) \tag{8}$$

where $S.share(\bullet)$ is the Shamir's (t, N) threshold secret sharing algorithm. In the input, N_i^{sk} is the private key of user u_i, $|\mu_1|$ is the number of shares into which N_i^{sk} needs to be divided, and *t* represents the minimum number of shares required to reconstruct N_i^{sk}. The output includes $|\mu_1|$ secret shares $s_{i,j}$ that user u_i sends to user u_j.

User u_i sends shares to the CS, which forwards them to the corresponding users. The message sent is shown below:

$$C_{i,j} \leftarrow DH.Enc(DH.agree(M_i^{sk}, M_j^{pk}), u_i \mid\mid u_j \mid\mid s_{i,j}) \tag{9}$$

where $DH.Enc(\bullet)$ is an encryption algorithm based on the DH protocol that encrypts the user identity numbers u_i, u_j, and the secret share $s_{i,j}$ using $DH.agree(M_i^{sk}, M_j^{pk})$ as the encryption key. $C_{i,j}$ denotes the ciphertext sent from user u_i to user u_j.

5.3. Generation and Uploading of Ciphertexts and Proofs (Round 3)

After Round 2, the users receive $C_{i,j}$ from CS and will decrypt them in Round 4.

After that, any two users negotiate the shared key $S_{i,j}$ between them according to the DH protocol. The negotiation procedure is as follows:

$$S_{i,j} \leftarrow DH.agree(N_i^{sk}, N_j^{pk}) \tag{10}$$

Here, $DH.agree(\bullet)$ is the shared key negotiation algorithm, which takes as input the private key N_i^{sk} of user u_i and the public key N_j^{pk} of user u_j. The output is the shared key $S_{i,j}$ between user u_i and user u_j. The shared key $S_{i,j}$ serves as the seed for a pseudo-random generator responsible for producing masks. The generation formula is as follows:

$$msk_{i,j} \leftarrow PRG(S_{i,j}) \tag{11}$$

where $PRG(\bullet)$ is a pseudo-random generator. It takes as input a shared key, which is negotiated between users u_i and u_j following the DH protocol. The output is the mask $msk_{i,j}$, which is used by user u_i and user u_j for masking the local gradient; moreover, the mask is equal in length with the gradient.

Every user performs training on their local dataset and obtains their local gradient, denoted as x_i.

After local training, user u_i encrypts the local gradient x_i as follows:

$$\widehat{x}_i \leftarrow x_i + \sum_{u_j \in \mu_2 : i<j} msk_{i,j} - \sum_{u_j \in \mu_2 : i>j} msk_{i,j} \tag{12}$$

In the abovementioned formula, $msk_{i,j}$ is the mask shared by users u_i and u_j, where both u_i and u_j belong to μ_2. u_i uses $msk_{i,j}$ to mask the local gradient x_i, and then obtains the encrypted gradient x_i by adding $msk_{i,j}$ to the local gradient x_i (where i<j), subtracting $msk_{i,j}$ (where i>j).

Next, the user u_i calculates both the local gradient's verification proof and the online proof:

$$h_i \leftarrow HH(x_i) \tag{13}$$

$$BF_i \leftarrow BFH(u_i) \tag{14}$$

Finally, u_i sends the encrypted gradient $\{\widehat{x}_i\}$ to the CS, the verification proof, and the online proof $\{h_i, BF_i\}$ to the TA.

Let μ_3 be the set of users who send messages to CS and TA. The CS sends a list of $\mu_2 \backslash \mu_3$ to each user in μ_3, requesting the offline users' shares for unmasking the aggregated result. Here, $\mu_2 \backslash \mu_3$ represents the users in set μ_3 but not in set μ_2.

TA aggregates the verification proofs h_i and online proofs BF_i of all users and broadcasts the two aggregated proofs to users. The aggregation mechanism is as follows:

$$HH\_proof \leftarrow HH.aggregate(h_i) \tag{15}$$

$$BF \leftarrow BFH.aggregate(BF_i) \tag{16}$$

where HH_proof denotes the verification proof of the aggregated result, which is used to verify the correctness of the gradient aggregated by CS. BF denotes the online proof of all users in the set μ_3, and any user can verify whether a particular user is online through BF.

5.4. Aggregation and Verification of Parameters (Round 4)

After receiving the list $\mu_2 \backslash \mu_3$ from the CS, to defend against the CS's deceiving attack, the users will verify whether the users in $\mu_2 \backslash \mu_3$ are dropped users. The verification procedure is as follows:

$$BF\_va \leftarrow BF.verify(BL, \mu_2 \backslash \mu_3) \tag{17}$$

The $BL.verify(\bullet)$ algorithm is a verification algorithm, used to verify whether the users in $\mu_2 \backslash \mu_3$ are dropped out. The inputs are the set of dropped users $\mu_2 \backslash \mu_3$ sent by CS and the users' online proof BL sent by TA. If the output is 1, it indicates that all users in the set $\mu_2 \backslash \mu_3$ are offline, and the verification is successful. If not, the verification fails, and users decline to send the shares of users in the set $\mu_2 \backslash \mu_3$ to the CS.

In Round 3, each online user receives the ciphertext $C_{i,j}$ of the secret shares from other users. Therefore, after the verification is passed, users will only decrypt the $C_{i,j}$ of those who have been dropped, specifically those who have shared their secret shares but have not uploaded the ciphertext of their gradients. The decryption algorithm is expressed as follows:

$$\{u_i \mid\mid u_j \mid\mid s_{i,j}\} \leftarrow DH.Dec(DH.agree(M_i^{sk}, M_j^{pk}), C_{i,j}) \tag{18}$$

where $DH.Dec(\bullet)$ is a decryption algorithm based on the DH protocol that decrypts the ciphertext $C_{i,j}$ using $DH.agree(M_i^{sk}, M_j^{pk})$ as the decryption key.

Next, users send the shares of the dropped users to CS. Let μ_4 represent the set of users who send information to the CS. CS receives shares from at least t users; otherwise, it stops. After receiving enough shares, CS reconstructs the private key N_i^{sk} of the dropped user u_i, where $u_i \in \mu_2 \backslash \mu_3$. The reconstruction algorithm is as follows:

$$N_i^{sk} \leftarrow S.recon(\{u_i,\ s_{i,j}\}_{u_j \in \mu_4}, t) \tag{19}$$

CS calculates the masks of the dropped users based on N_i^{sk}, and calculated as follows:

$$msk_{i,j} \leftarrow PRG(DH.agree(N_i^{sk}, N_j^{pk})) \tag{20}$$

For the subscripts i and j, where $u_i \in \mu_2 \backslash \mu_3$, and $u_j \in \mu_3$
Afterwards, CS uses $msk_{i,j}$ to correct the aggregated result:

$$\sum_{u_i \in \mu_3} x_i \leftarrow \sum_{u_i \in \mu_3} \widehat{x_i} \ - \sum_{\substack{u_i \in \mu_3, \\ u_j \in \mu_2 \backslash \mu_3 : i < j}} msk_{i,j} \ + \sum_{\substack{u_i \in \mu_3, \\ u_j \in \mu_2 \backslash \mu_3 : i > j}} msk_{i,j} \tag{21}$$

Lastly, CS sends the aggregated result $z = \sum_{u_i \in \mu_3} x_i$ to each user. After receiving z, the users will verify it. The verification process is as follows:

$$HH\_va \leftarrow HH.verify(HH\_proof, z) \tag{22}$$

where $HH.verify(\bullet)$ represents the aggregated result verification algorithm, and the inputs are the verification proof HH_proof and the aggregated result z. If the output is 1, it indicates the aggregated result is accurate, the verification is successful, and users accept the aggregated result z, moving on to the next training round. If not, the verification fails, leading to a halt in training and a restart.

6. Security Analysis

This section will analyze and prove the security of our scheme. Firstly, we will demonstrate that our scheme will protect user local gradients' privacy (Input Privacy). Secondly, we will conduct a security analysis on server forgery and deletion attacks. Finally, we will demonstrate that our verification scheme is correct.

6.1. Privacy Protection of the Gradients

Firstly, as can be inferred from the previous text, we employ a single masking scheme to ensure the privacy and security of the user's local gradient. Every user masks the local gradient as:

$$\widehat{x_i} = x_i + \sum_{u_j \in \mu_2, i<j} msk_{i,j} - \sum_{u_j \in \mu_2, i>j} msk_{i,j} \tag{23}$$

There is a lemma that, if we have some uniformly random numbers added to the inputs of users, the result will appear uniformly random.

In our threat model, the server is honest but curious. Moreover, it may collude with fewer than $t-1$ users to infer a specific user's input privacy. In addition, if the server is malicious, the subsequent sections on correction of the aggregated result and correction of verification will ensure that our scheme is secure. Next, before proving our scheme's input privacy, we must introduce some useful notation. We denote Cloud Server by the set S, the n users participating in federated learning by the set U, and introduce a security parameter k for the cryptographic primitives, using t to denote the threshold in Shamir's secret sharing. Because of the user dropout problem, we denote by U_i the set of users whose local parameters are received by the CS in round $i-1$. The number of users may change each round as users can drop out of training anytime. Hence, we have $U_3 \subseteq U_2 \subseteq U_1 \subseteq U$. We denote by $U_i \backslash U_{i+1}$ the users whose messages were received by the server in round $i-1$, but not received in round i.

Let $W \subseteq U \cup S$ be a set of corrupt parties. The combined perspective of all parties within W is characterized by a random variable $REAL_W^{U,t,k}(x_U, U_1, U_2, U_3)$, where k stands for a security parameter and t is the threshold in our protocol. This view includes the parties' input in W, randomness, and all communications received from parties outside of W. Additionally, the party will remain receiving messages until it drops out and stops receiving messages. Then, we will postulate two theorems to discuss the security of the input privacy in our protocol. In these theorems, one considers the collusion of active adversarial users, and the other is based on the collusion between Cloud Servers and users. Ultimately, they can show that any collusion between these parties cannot infringe on the privacy of others.

Theorem 1: (Safeguarding against collaborative assaults from multiple users) For all k, t, x_U and $U_3 \subseteq U_2 \subseteq U_1 \subseteq U$, a PPT simulator SIM exists with an output indistinguishable from $REAL_W^{U,t,k}$:

$$REAL_W^{U,t,k}(x_U, U_1, U_2, U_3) \equiv SIM_W^{U,t,k}(x_U, U_1, U_2, U_3) \tag{24}$$

Proof: As we only consider the collusion between multiple users excluding the Cloud Server, the joint view of the parties in set W is independent of the inputs from users not

in W. A perfect simulation can be achieved by having the simulator operate all honest users on false inputs while running the honest but curious users on their actual inputs. As the messages received by users from Cloud Server only include the set of the online users and final aggregation but do not (contain) the true value of $\widehat{x_n}$, the simulator can utilize random numbers to mask all honest users' inputs, instead of using true values. Hence, the parties in W will be unable to determine if the input from honest users is true or dummy. Ultimately, the simulated perspective of parties in set W is identical to the actual perspective $REAL_W^{U,t,k}$.

Theorem 2: (Guarding against collaborative assaults from users and the Cloud Server) For all $k, t, U, x_U, W \subseteq U \cup S, n_W = |W \backslash S|, n_W < t$, and $U_3 \subseteq U_2 \subseteq U_1 \subseteq U$, a PPT simulator SIM exists, generating an output that, computationally, cannot be differentiated from the output of $REAL_W^{U,t,k}$:

$$REAL_W^{U,t,k}(x_U, U_1, U_2, U_3) \equiv SIM_W^{U,t,k}(x_U, U_1, U_2, U_3) \tag{25}$$

Proof: We will employ a conventional hybrid argument to provide proof for the above theorem. The approach is gradually executing an array of secure alterations on the actual view, which eventually results in the output of the simulated view being computationally identical to the output of the real view.

Hyb1: In this hybrid, regarding the interaction among users in Round 1, we employ a random numeral to substitute the shared key among any interacting entities for the message's encryption/decryption. Specifically, assuming that we fix any two users $u_i, u_j \in U_2 \backslash W, u_i \neq u_j$, u_i and u_j are honest users, so then the simulator modifies the conduct of all upright participants by employing a uniformly random numeral $r_{i,j}$ as a replacement for the shared key $DH.agree(N_i^{sk}, N_j^{pk})$ between u_i and u_j. Subsequently, u_i and u_j will use a random number $r_{i,j}$ to encrypt and decrypt the messages based on the symmetric encryption system. Eventually, the DDH assumption will ensure this hybrid is computationally identical to the real protocol.

Hyb2: In this hybrid, the simulator will use a random number (this random number is the same length as the data that honest users need to encrypt) to replace the data that all honest users $u_i (u_i \in U_2 \backslash W)$ want to encrypt. In the subsequent Rounds, when the CS requires users to upload the offline users' shares to unmask the ciphertexts of the offline users, all honest users will upload the real shares (*i.e.*, the shares of random numbers used by this hybrid). By altering the data to be encrypted, we ensure, through the properties of symmetric authenticated encryption, that this hybrid is distinguishable from the actual protocol.

Hyb3: Here, the simulator will use shares of random numbers with appropriate length instead of N_i^{SK}'s shares from all honest users u_i who are in the set U_2 but not in W. Hence, the security of Shamir's secret sharing ensures the indistinguishability of this hybrid from the actual protocol.

Hyb4: In this hybrid, we first select any two users u_i and $u_{j'}$, who are from the set $U_2 \backslash W$ and $u_i \neq u_{j'}$. Then, for the shared key $S_{i,j'} = DH.agree(N_i^{sk}, N_{j'}^{pk})$ between u_i and $u_{j'}$, the simulator selects a random number $S_{i,j'}^{'}$ uniformly for replacement. Specifically, for the user u_i, instead of sending $\widehat{x}_i = x_i + \sum_{u_j \in U_2 : i<j} PRG(S_{i,j}) - \sum_{u_j \in U_2 : i>j} PRG(S_{i,j})$, SIM submits

$\widehat{w}_i = w_i + \sum_{u_j \in U_2 \setminus \{u_{j'}\}: i<j} PRG(S_{i,j}) - \sum_{u_j \in U_2 \setminus \{u_{j'}\}: i>j} PRG(S_{i,j}) + \triangle_{i,j'} PRG(S'_{i,j})$,

where $\triangle_{i,j'} = \begin{cases} 1, i < j' \\ -1, i > j' \end{cases}$

For $u_{j'}$, there exists $\widehat{x}_{j'} = x_{j'} + \sum_{u_i \in U_2} \triangle_{i,j'} PRG(S'_{i,j})$

Subsequently, the DDH assumption certifies this hybrid as indistinguishable from the authentic protocol.

Hyb5: Based on the previous hybrid, the simulator, in this hybrid, replaces the output of $PRG(s'_{i,j'})$ with a random number that is uniformly selected. The simulator merely replaces the output of PRG, thus the security of the pseudo-random generator makes this hybrid indistinguishable from the actual protocol.

As can be seen from the previous hybrids, the distribution of these hybrids cannot be differentiated from the actual protocol, so thus, this completes and finalizes the proof.

6.2. Correctness of the Aggregated Result

By following our scheme honestly, the server can ensure that an accurate aggregated result z is ultimately obtained by the users. If the users and Cloud Server (CS) follow our scheme honestly, users can ultimately get a correct aggregated result z.

In Round 4, the CS requests the shares of dropout users after receiving parameters uploaded by the online users. After the request from the CS, is verified by the users, the CS receives the secret shares from the users who have dropped out. The CS retrieves the dropout users' private key $N_i sk$ and calculates its mask $msk_{i,j}$. Finally, the CS calculates the aggregated result:

$$z = \sum_{u_i \in U_3} x_i = \sum_{u_i \in U_3} \widehat{x}_i - \sum_{u_i \in U_3, u_j \in U_2 \setminus U_3} \triangle_{i,j} msk_{i,j} \tag{26}$$

where $\triangle_{i,j} = \begin{cases} 1, i > j \\ -1, i < j \end{cases}$

Therefore, according to the above formula, CS can ultimately obtain the correct aggregated result for online users.

6.3. Correctness of Verification

Up to $t-1$ users are allowed to collude with the server in our scheme, which means that the CS can obtain the homomorphic hash function HH and users' verification scheme. Therefore, if we make the CS aggregate the hash values of the users' gradients or make the CS broadcast the hash values to each user to aggregate the hash values, a malicious CS may tamper with the aggregated result of the users' gradients and modify the aggregated result of the hash values at the same time, or a lazy CS may "delete" the local gradients and the hash values of some users in order to save computational resources, which may result in inaccurate aggregated result.

Under our assumptions above, neither of these two malicious behaviors of the CS will be detected by users, which will violate the original intention of verifiable federated learning. Therefore, our scheme delegates the operation of merging the hash values of the local

gradient updates from the users to the TA. In this scenario, the CS may tamper with the users' gradients or omit some users' gradients during aggregation to save computational overhead. Alternatively, the CS may add noise to affect the aggregated result's accuracy. However, since the TA and CS are not colluding, these actions will cause the hash value of the aggregated result to differ from the aggregated result of hash values. By comparing these two values, each user can detect these malicious behaviors by the server. Therefore, this scheme can resist malicious attacks such as forgery and deletion from CS.

7. Performance Analysis and Evaluation

7.1. Performance Analysis

In this subsection, we choose five MPC (Multi-Party Computation) based schemes, which are VerifyNet [29], VerSA [9], PFLM [13], VERIFL [8] and PVFL [40] as well as a HE (Homomorphic Encryption) based scheme (VPFL [32]) to analyze and compare the communication and computation overheads. In Table 1, n is used to represent the total number of users, d stands for the gradient dimension, and ϕ signifies the count of users who have dropped out.

Communication Overhead Analysis In Table 1, we analyze the outgoing communication of the user and the server in these schemes. [29] first proposed a secure and verifiable federated learning scheme, but it did not achieve the independence of the communication overhead for verification from the gradient dimension. Analysis of this scheme reveals that the communication overheads for each user and the server are $O(n + d)$ and $O(n^2 + nd + n + d)$, respectively. In [9], the main communication overhead lies in the user uploading the encrypted gradient and n secret shares and the server forwarding the data uploaded by the user. Thus, the communication overhead for each user is $O(n + d)$, while for the server, it is $O(n^2 + nd)$. [13] adopts a double masking protocol to protect privacy, and the user needs to receive $n - 1$ masked messages from the server to verify the correctness of the aggregated result. Thus, the communication overheads for each user and the server are $O(nd)$ and $O(n^2 + nd)$, respectively.

Again, in the work [8], all users need to receive the commitments and hash values of other online users from the server, which leads to some communication overhead, we can derive that the communication overhead for each user is $O(n + d)$, while for the server is $O(n^2 + nd)$. In [40], the communication overhead is mainly caused by the differential privacy mechanism, and the user needs to negotiate noise with other users. Therefore, its communication overhead is $O(nd)$ for each user and is $O(n^2 d)$ for the server. [32] proposes a distributed encryption of gradients algorithm, which can reduce the computational overhead of encryption but increases the communication overhead. Therefore, the communication overheads for each user and the server are $O(nd)$ and $O(n^2 d)$, respectively.

Computation Overhead Analysis. Since VPFL [32] realizes the integrity verification of gradients during the transmission process, but it cannot verify the incorrect aggregated result or detect the malicious behavior of the cloud server, we only analyze and compare the computational overhead of these schemes [8,9,13,29,40]. The comparison of the computation overheads between these schemes is shown in Table 1. The privacy-preserving

Table 1. computation and outgoing communication overhead

	Outgoing Communication		Computation Overhead	
Schemes	Each user	Server	Each user	Server
VerifyNet [29]	$O(n+d)$	$O(n^2+nd+n+d)$	$O(nd+n)$	$O(n^2+nd+d)$
VerSA [9]	$O(n+d)$	$O(n^2+nd)$	$O(nd+n)$	$O(n^2+nd)$
PFLM [13]	$O(nd)$	$O(n^2+nd)$	$O(n^2+nd+n)$	$O(n^2d)$
VERIFL [8]	$O(n+d)$	$O(n^2+nd)$	$O(nd+n+d)$	$O(n^2+nd)$
PVFL [40]	$O(nd)$	$O(n^2d)$	$O(nd+n+d)$	$O(nd)$
VPFL [32]	$O(nd)$	$O(n^2d)$	\	\
Our scheme	$O(n+d)$	$O(n^2)$	$O(nd)$	$O(n^2)$

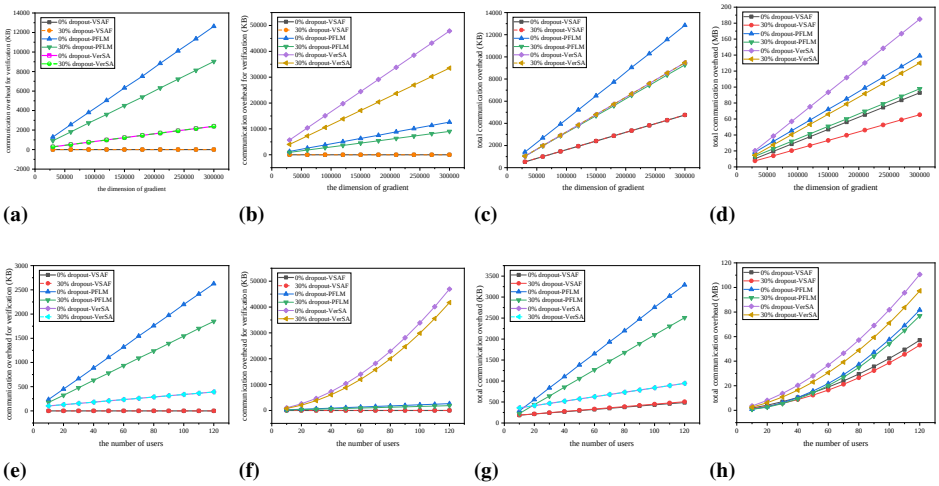


Fig. 4. The comparison of our scheme VSAF and PFLM [13] and VerSA [9] with respect to communication overhead for verification and total communication overhead. (a), (b), (c), and (d) compare the three schemes under 0% and 30% dropout rates, showing how the communication overhead varies with gradient dimension. Specifically, (a) illustrates the communication overhead for verification for each user per iteration, (b) illustrates the communication overhead for verification of CS per iteration, (c) illustrates the total communication overhead for each user per iteration, and (d) illustrates the total communication overhead of CS per iteration. (e), (f), (g), (h) compare the three schemes under 0% and 30% dropout rates, showing how the communication overhead varies with the number of users. Specifically, (e) illustrates the communication overhead for verification for each user per iteration, (f) illustrates the communication overhead for verification of CS per iteration, (g) illustrates the total communication overhead for each user per iteration, and (h) illustrates the total communication overhead of CS per iteration

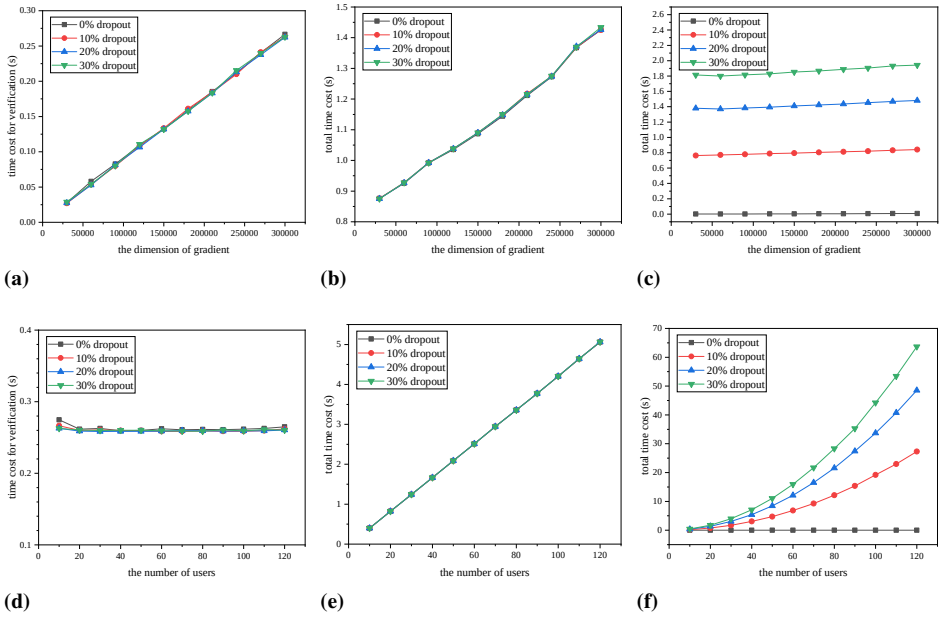


Fig. 5. Comparison of the time costs for each user and server under different dropout rates in our scheme. (a), (b), (c) show the time cost variations with gradient dimension for each user and CS in our scheme under 0%, 10%, 20%, and 30% dropout rates. Specifically, (a) illustrates the time cost for verification of each user per iteration, (b) illustrates the total time cost for each user per iteration, and (c) illustrates the total time cost for CS at each iteration. (d), (e), (f) show the time cost variations with the number of users for each user and CS in our scheme under 0%, 10%, 20%, and 30% dropout rates. Specifically, (d) illustrates the time cost for verification of each user per iteration, (e) illustrates the total time cost for each user per iteration, and (f) illustrates the total time cost for CS at each iteration

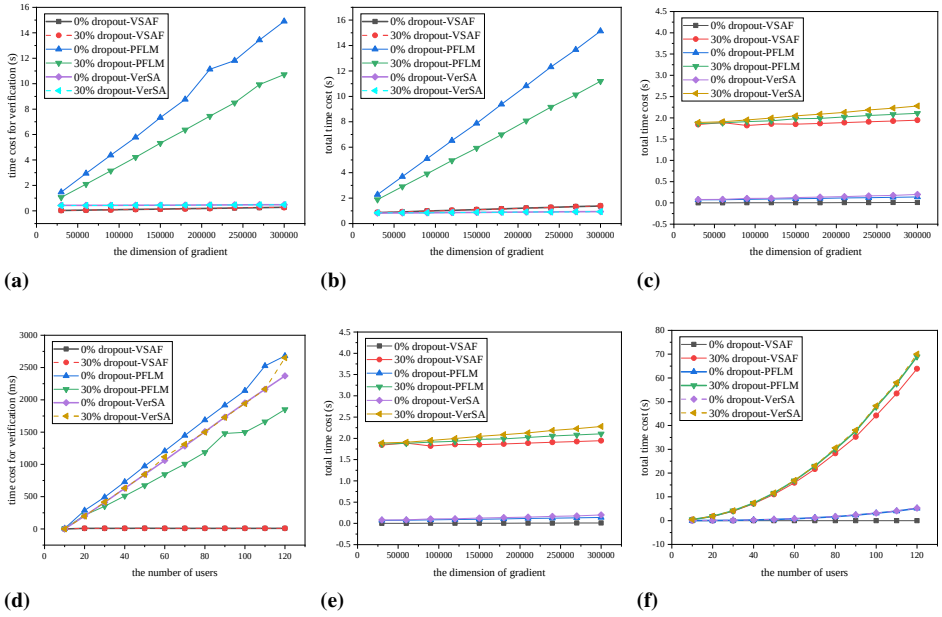


Fig. 6. The comparison of our scheme VSAF and PFLM [13] and VerSA [9] in terms of time cost for verification and total time cost. (a), (b), (c) compare the three schemes under 0% and 30% dropout rates, showing the time cost varies with gradient dimension. Specifically, (a) illustrates the time cost for verification of each user per iteration, (b) illustrates the total time cost for each user per iteration, and (c) illustrates the total time cost for CS at each iteration. (d), (e), (f) compare the three schemes under 0% and 30% dropout rates, showing the time cost varies with the number of users. Specifically, (d) illustrates the time cost for verification of each user per iteration, (e) illustrates the total time cost of each user per iteration, and (f) illustrates the total time cost of CS at each iteration

schemes in [8, 9, 13, 29] are mainly based on the double masking protocol, so the computation overhead of these schemes mainly comes from generating secret shares, masks, and ciphertexts.

[29] designs a verification scheme based on a homomorphic hash function and pseudo-random technology. The computation overhead for each user and server is $O(nd+n)$ and $O(n^2 + nd + d)$, respectively. In [9], the proposed scheme of the computation overhead for verification lies in generating secret shares, masks, and proofs. Therefore, the computation overhead is $O(nd+n)$ for each user and $O(n^2+nd)$ for the server. [13] proposes a verification scheme based on a variant of ElGamal encryption, the computation overhead amounts to $O(n^2 + nd + n)$ for each user and $O(n^2d)$ for the server. Using homomorphic hash technology and commitment scheme, [8] develops a verification scheme. The computation overhead of this scheme is $O(nd+n+d)$ for each user and $O(n^2+nd)$ for the server. [40] combines a variant of double masking protocol and differential privacy to design a privacy-preserving scheme, and uses a linear homomorphic hash to design a verification scheme. Thus, the computation overhead for each user is $O(nd + n + d)$, while for the server, it is $O(nd)$.

7.2. Experimental Settings

We implemented a prototype of our scheme, VSAF, by Python 3 and Charm-Crypto. Since our encryption scheme is based on MPC (Multi-Party Computation), the model accuracy is unaffected. Therefore, we set the gradient data in the same way as PFLM [13], randomly selecting data from a normal distribution $N(50, 20)$. We implemented the DH protocol based on the discrete logarithm problem. Then, using this DH protocol and Shamir's (t, N)-threshold secret sharing protocol, we implemented the relevant parts of our VSAF. In our scheme, we also implemented a linear homomorphic hash function for verification using the charm-crypto library. This function is irreversible and does not leak gradient information. Every experiment was conducted on a 64-bit Ubuntu OS 20.04.6 version, equipped with an Intel i3-10105 CPU and 4GB memory.

7.3. Experimental Results

Since VerifyNet is a classic secure and verifiable FL scheme, PFLM [13] and VerSA [9] are more representative than the other four masking-based schemes. Therefore, we compare the simulation experimental results of the proposed scheme with the PFLM and VerSA. We analyzed the performance of these three schemes by continuously adjusting the number of participating users N, the gradient dimension d, and the users' dropout rate.

Communication Overhead. In our scheme, the communication overhead for verification of each user mainly consists of the following two aspects: (1) In the Generation and Uploading of Ciphertexts and Proofs phase (Round 3), the user must upload proof of the gradient. (2) In the Aggregation and Verification of Parameters phase (Round 4), each user must receive a cryptographic proof from the TA validating the aggregated result. In Fig.4(a) (b) (c) (d), these subfigures illustrate the communication overhead varies with the gradient dimension and we set the number of users $N = 20$, with a threshold of

$t = \lfloor \frac{N}{2} \rfloor + 1$, *i.e.*, $t = 11$. In Fig.4(e) (f) (g) (h), these subfigures illustrate the variation in communication overhead with the number of users. and we set gradient dimension $d = 10000$ and threshold $t = \lfloor \frac{N}{2} \rfloor + 1$.

In Fig.4(a) (b), in both the PFLM [13] and VerSA [9] schemes, we see that, for both users and CS, the communication overhead for verification linearly varies with the dimension of the gradient. However, in the case of the VSAF scheme, the communication overhead for verification does not vary with the gradient dimension. This is due to our verification scheme utilizing a linear homomorphic hash function. It should be noted that the function we designed can compress d-dimensional data into one-dimensional data, thereby achieving independence of the communication overhead for verification from the gradient dimension.

As shown in Fig.4(e), in both the PFLM [13] and VerSA [9] schemes, both users and CS, the communication overhead for verification grows directly with the number of users. We can see that the communication overhead for verification of our VSAF almost remains constant regardless of the number of users. The reason is that, in PFLM and VerSA, each user needs to process the proofs of other users, which will result in the number of users affecting each user's communication. However, the users of our scheme only send their verification proofs and receive the aggregated verification proofs, so the change in the number of users has no impact on the per-user communication overhead for verification. Thus, our VSAF realizes that the verification's communication overhead for users does not depend on the number of users.

As we can see from Fig.4(a) (e), whether the dropout rate is 0% or 30%, the communication overhead of users for verification does not significantly change with the variation of the dropout rate. In VSAF, the users only send their verification proofs and receive the aggregated verification proofs, so the change in the dropout rate does not affect each user's communication overhead for verification. Thus, our proposed VSAF ensures the independence of users' communication overhead for verification from the dropout rate.

From Fig.4, we can see that the communication overhead of VSAF is more negligible than the PFLM [13] and VerSA [13]. Both PFLM [13] and VerSA [9] use a double mask scheme for privacy protection, while we use a single masking scheme, which reduces the users' communication overhead by $O(n)$. In addition, in PFLM [13], each user is required to receive other users' proofs, while in our scheme, the users do not need to receive the proofs of other users. Thus, the users of our scheme have an additional $O(n)$ reduction in communication overhead compared to PFLM [13].

Computation Overhead In Fig.5 and Fig.6, for sub-figures (a)(b)(c), we set the number of users N = 20, and the threshold $t = \lfloor \frac{N}{2} \rfloor$+1, that is, $t = 11$. For sub-figures (d)(e)(f), we set the gradient dimension d = 10000, and the threshold $t = \lfloor \frac{N}{2} \rfloor$+1.

As shown in Fig.5(d), each user's time cost for verification of these schemes does not significantly vary with a rise in the number of users. In our scheme, users only need to generate verification proofs of their own gradients through a homomorphic hash function, without requiring other users' data. Hence, the user's computation overhead for verification of our scheme should be independent of the number of users. Furthermore, as we can see from Fig.5(a) (d), the time cost for verification per user does not change with the variation of the dropout rate. Since the dropout rate reflects the changes in the number of users, and each user's computation overhead for verification is independent of the number

of users, the computation overhead for verification of each user is also independent of the dropout rate.

In Fig.5(b) (e) (c) (f), it is clear that the total time cost for users and CS increases with the growth in either the gradient dimension or the number of users. The reason is that, as the gradient dimension and the number of users increases, the data volume that users and CS need to process also increases, leading to an increase in total time cost. Furthermore, in these four sub-figures, the total time cost for users or CS increases as the dropout rate increases. This is due to the fact that our scheme uses the Shamir secret-sharing technique, and a rise in dropout users results in more secrets needing recovery, thereby increasing the total time cost. Additionally, as in sub-figures (b) (e), the changes in user time cost are not evident with the increases in the dropout rate, because we use a single masking scheme that is improved from the double mask scheme. In our scheme, online users only need to decrypt the shares' ciphertext of the dropout users and send it to CS; unlike the double mask scheme, where online users need to decrypt the ciphertext of all users.

From Fig.6, compared with the VerSA [9] and PFLM [13] schemes, it is clear that our VSAF is superior in terms of computation overhead, because our scheme is more lightweight than VerSA and PFLM concerning privacy protection and verification. In terms of privacy protection, VerSA [9] and PFLM [13] use a double mask scheme, while our VSAF uses a lightweight single masking scheme improved from the double mask scheme. Furthermore, in terms of verification, PFLM [13] uses Identity-Based Aggregate Signature technology and a variant of ElGamal encryption, which will generate a lot of time cost. VerSA [9] still designs a verification scheme based on the double mask, while VSAF uses a linear homomorphic hash function to construct a lightweight verification scheme and outsources the aggregation process of verification proofs to a third party that does not collude with CS, thereby reducing the verification cost.

Defending Against Gradient Reconstruction Attacks As shown in Fig.7, both our VSAF scheme and the FedAvg scheme [17] were trained via federated learning on the MNIST and CIFAR-10 datasets, respectively. During training, we applied a Gradient Reconstruction Attack to each method. The results indicate that under this attack, the FedAvg scheme was able to recover image information after approximately 70 rounds on MNIST and around 210 rounds on CIFAR-10. In contrast, our VSAF scheme did not leak any image information.

8. Concluding Remarks and Future Work

In this work, we propose VSAF, a verifiable and secure aggregation scheme designed for federated learning in edge computing. VSAF employs a combination of single masking with Bloom filtering for lightweight, dropout-tolerant privacy protection of user gradients. A linear homomorphic hash function is used to design a verification algorithm that ensures correct aggregation while minimizing verification overhead. Security analysis confirms the high security and correctness of VSAF, supported by comprehensive theoretical and experimental results.

In future work, we aim to reduce computational and communication overheads whilst ensuring robustness to user dropouts. In addition, we seek to develop methods to correct erroneous aggregated results while verifying their correctness.

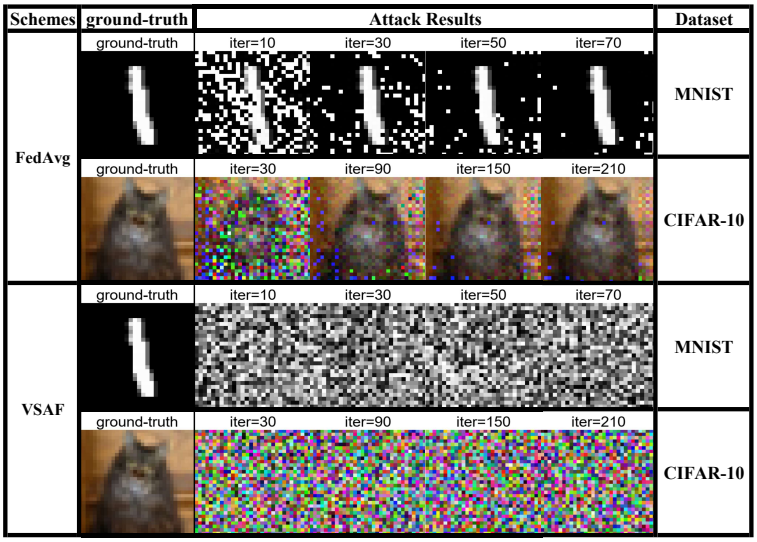


Fig. 7. Defending against gradient reconstruction attacks

Acknowledgments. This work is supported in part by the Scientific Research Fund of Hunan Provincial Education Department (No. 24A0337), the Natural Science Foundation of Hunan Province (No. 2025JJ50348), and the Natural Science Foundation of Fujian Province (No. 2022J05106). The authors appreciate the editor and anonymous reviewers for their invaluable feedback.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. Bell, J.H., Bonawitz, K.A., Gascón, A., Lepoint, T., Raykova, M.: Secure single-server aggregation with (poly) logarithmic overhead. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. pp. 1253–1269 (2020)
3. Bloom, B.H.: Space/time trade-offs in hash coding with allowable errors. Communications of the ACM 13(7), 422–426 (1970)
4. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. pp. 1175–1191 (2017)
5. Chen, Y., Qin, X., Wang, J., Yu, C., Gao, W.: Fedhealth: A federated transfer learning framework for wearable healthcare. IEEE Intelligent Systems 35(4), 83–93 (2020)
6. Cisco: Cisco annual internet report (2018–2023) (Mar 2020), https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf
7. Filho, C.P., Marques Jr, E., Chang, V., Dos Santos, L., Bernardini, F., Pires, P.F., Ochi, L., Delicato, F.C.: A systematic literature review on distributed machine learning in edge computing. Sensors 22(7), 2665 (2022)

8. Guo, X., Liu, Z., Li, J., Gao, J., Hou, B., Dong, C., Baker, T.: Verifl: Communication-efficient and fast verifiable aggregation for federated learning. IEEE Transactions on Information Forensics and Security 16, 1736–1751 (2020)
9. Hahn, C., Kim, H., Kim, M., Hur, J.: Versa: Verifiable secure aggregation for cross-device federated learning. IEEE Transactions on Dependable and Secure Computing (2021)
10. Han, G., Zhang, T., Zhang, Y., Xu, G., Sun, J., Cao, J.: Verifiable and privacy preserving federated learning without fully trusted centers. Journal of Ambient Intelligence and Humanized Computing pp. 1–11 (2022)
11. Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., Ramage, D.: Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604 (2018)
12. Hitaj, B., Ateniese, G., Perez-Cruz, F.: Deep models under the gan: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. pp. 603–618 (2017)
13. Jiang, C., Xu, C., Zhang, Y.: Pflm: Privacy-preserving federated learning with membership proof. Information Sciences 576, 288–311 (2021)
14. van Leeuwen, J. (ed.): Computer Science Today. Recent Trends and Developments, Lecture Notes in Computer Science, vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995)
15. Li, T., Li, J., Chen, X., Liu, Z., Lou, W., Hou, Y.T.: Npmml: A framework for non-interactive privacy-preserving multi-party machine learning. IEEE Transactions on Dependable and Secure Computing 18(6), 2969–2982 (2020)
16. Lin, L., Zhang, X.: Ppverifier: A privacy-preserving and verifiable federated learning method in cloud-edge collaborative computing environment. IEEE Internet of Things Journal 10(10), 8878–8892 (2022)
17. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh, A., Zhu, J. (eds.) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282. PMLR (20–22 Apr 2017), https://proceedings.mlr.press/v54/mcmahan17a.html
18. Melis, L., Song, C., De Cristofaro, E., Shmatikov, V.: Inference attacks against collaborative learning. arXiv preprint arXiv:1805.04049 13 (2018)
19. Mou, W., Fu, C., Lei, Y., Hu, C.: A verifiable federated learning scheme based on secure multiparty computation. In: International Conference on Wireless Algorithms, Systems, and Applications. pp. 198–209. Springer (2021)
20. Phong, L.T., Aono, Y., Hayashi, T., Wang, L., Moriai, S.: Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security 13(5), 1333–1345 (2018)
21. Ribičre, M., Charlton, P.: Ontology overview. Motorola Labs, Paris (2002), [Online]. Available: http://www.fipa.org/docs/input/f-in-00045/f-in-00045.pdf (current October 2003)
22. Shamir, A.: How to share a secret. Commun. ACM 22(11), 612–613 (nov 1979), https://doi.org/10.1145/359168.359176
23. Shokri, R., Shmatikov, V.: Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. pp. 1310–1321 (2015)
24. Song, C., Ristenpart, T., Shmatikov, V.: Machine learning models that remember too much. In: Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security. pp. 587–601 (2017)
25. Treleaven, P., Smietanka, M., Pithadia, H.: Federated learning: the pioneering distributed machine learning and privacy-preserving data technology. Computer 55(4), 20–29 (2022)
26. Vailshery, L.S.: Internet of things (iot)—statistics and facts. https://www.statista.com/study/27915/internet-of-things-iot-statista-dossier/ (2021)

27. Wang, F., He, Y., Guo, Y., Li, P., Wei, X.: Privacy-preserving robust federated learning with distributed differential privacy. In: 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). pp. 598–605 (2022)
28. Wang, X., Bettini, C., Brodsky, A., Jajoida, S.: Logical design for temporal databases with multiple granularities. ACM Transactions on Database Systems 22(2), 115–170 (1997)
29. Xu, G., Li, H., Liu, S., Yang, K., Lin, X.: Verifynet: Secure and verifiable federated learning. IEEE Transactions on Information Forensics and Security 15, 911–926 (2019)
30. Xu, G., Li, H., Zhang, Y., Xu, S., Ning, J., Deng, R.H.: Privacy-preserving federated deep learning with irregular users. IEEE Transactions on Dependable and Secure Computing 19(2), 1364–1381 (2020)
31. Yan, Y., Kamel, M.B., Zoltay, M., Gál, M., Hollós, R., Jin, Y., Péter, L., Tényi, Á.: Fedlabx: a practical and privacy-preserving framework for federated learning. Complex & Intelligent Systems pp. 1–14 (2023)
32. Zhang, J., Liu, Y., Wu, D., Lou, S., Chen, B., Yu, S.: Vpfl: A verifiable privacy-preserving federated learning scheme for edge computing systems. Digital Communications and Networks 9(4), 981–989 (2023)
33. Zhang, S., He, J., Liang, W., Li, K.: Mmds: A secure and verifiable multimedia data search scheme for cloud-assisted edge computing. Future Generation Computer Systems 151, 32–44 (2024), https://www.sciencedirect.com/science/article/pii/S0167739X23003564
34. Zhang, S., Hu, B., Liang, W., Li, K.C., Gupta, B.B.: A caching-based dual k-anonymous location privacy-preserving scheme for edge computing. IEEE Internet of Things Journal 10(11), 9768–9781 (2023)
35. Zhang, S., Hu, B., Liang, W., Li, K.C., Pathan, A.S.K.: A trajectory privacy-preserving scheme based on transition matrix and caching for iiot. IEEE Internet of Things Journal pp. 1–1 (2023)
36. Zhang, S., Yan, Z., Liang, W., Li, K.C., Dobre, C.: Baka: Biometric authentication and key agreement scheme based on fuzzy extractor for wireless body area networks. IEEE Internet of Things Journal pp. 1–1 (2023)
37. Zhang, S., Yang, Y., Liang, W., Sandor, V.K.A., Xie, G., Raymond, K.K.: Mkss: An effective multi-authority keyword search scheme for edge–cloud collaboration. Journal of Systems Architecture 144, 102998 (2023), https://www.sciencedirect.com/science/article/pii/S1383762123001777
38. Zhang, Y., Yu, H.: Towards verifiable federated learning. arXiv preprint arXiv:2202.08310 (2022)
39. Zhao, J., Zhu, H., Wang, F., Lu, R., Liu, Z., Li, H.: Pvd-fl: A privacy-preserving and verifiable decentralized federated learning framework. IEEE Transactions on Information Forensics and Security 17, 2059–2073 (2022)
40. Zhou, H., Yang, G., Huang, Y., Dai, H., Xiang, Y.: Privacy-preserving and verifiable federated learning framework for edge computing. IEEE Transactions on Information Forensics and Security 18, 565–580 (2023)
41. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. Advances in neural information processing systems 32 (2019)

Shiwen Zhang received his Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, in 2016. He is currently an associate professor at the School of Computer Science and Engineering, Hunan University of Science and Technology. He is a member of IEEE and CCF. His research interests include identity authentication, security and privacy issues in Wireless Body Area Networks (WBAN), cloud computing, privacy protection, and information security.
Email: shiwenzhang@hnust.edu.cn

Feixiang Ren received a B.S. degree from Henan University of Technology, and currently pursuing an M.S. degree from the school of Computer Science and Engineering at Hunan University of Science and Technology. His research interests include privacy-preserving and verifiable federated learning. Email: rfx.point@mail.hnust.edu.cn

Wei Liang received a Ph.D. degree in computer science and technology from Hunan University in 2013. He was a Post-Doctoral Scholar at Lehigh University from 2014 to 2016. He is currently a Professor and the Dean of the School of Computer Science and Engineering at Hunan University of Science and Technology, China. He has authored or co-authored more than 140 journal and conference papers. His research interests include identity authentication in WBAN and security management in wireless sensor networks (WSN). Email: weiliang@hnu.edu.cn

Kuanching Li is a Professor at the School of Computer Science and Engineering, Hunan University of Science and Technology. Dr. Li has co-authored over 150 conference and journal papers, holds several patents, and serves as an associate and guest editor for various scientific journals. He has also held chair positions at several prestigious international conferences. His research interests include cloud and edge computing, big data, and blockchain technologies. Dr. Li is a Fellow of the Institution of Engineering and Technology (IET). Email: aliric@hnust.edu.cn

Al-Sakib Khan Pathan received a B.Sc. degree in computer science and information technology from the Islamic University of Technology (IUT), Bangladesh, in 2003 and a Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2009. He is a Professor at the Department of Computer Science and Engineering, United International University (UIU), Bangladesh. He has served as General Chair, Organizing Committee Member, and TPC member in top-ranked conferences such as INFOCOM, CCGRID, GLOBECOM, ICC, LCN, GreenCom, and IRI. He received the IEEE Outstanding Leadership Award (GreenCom'13) and two IEEE Outstanding Service Awards (IRI'20, IRI'21), and has co-edited/co-authored 34 books. Email: sakib.pathan@gmail.com

Received: February 21, 2025; Accepted: April 26, 2025.

Computer Science and Information Systems 22(4):1433–1456 https://doi.org/10.2298/CSIS241017055T

Augmented Reality Mobile Application as a Support in Presentation of Orthodox Iconography

Dušan Tatić[1], Radomir Stanković[2], Detelin Luchev[3], Maxim Goynov[4], and Desislava Paneva-Marinova[5]

[1] Mathematical Institute of the Serbian Academy of Sciences and Arts
11000 Belgrade, Serbia
dusan.tatic@mi.sanu.ac.rs
[2] Mathematical Institute of the Serbian Academy of Sciences and Arts
11000 Belgrade, Serbia
radomir.stankovic@gmail.com
[3] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Sofia, Bulgaria
dml@math.bas.bg
[4] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Sofia, Bulgaria
m.goynov@math.bas.bg
[5] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Sofia, Bulgaria
d.paneva@math.bas.bg

Abstract. Orthodox Christian icons are a very valuable part of the national historical heritage of Bulgaria. To preserve and properly present data about icons, it was created a web-based platform Virtual Encyclopedia of Bulgarian Iconography (BIDL). It contains icon descriptions and context explanations as well as information about icon painters, drawing methods, and painting techniques used, for a large number of icons all over Bulgaria. Observation and contemplation about the icons can be motivated by religious feelings, in which case icons are viewed as an aid in worship, or by interest in their artistic features. In both cases, it is important and useful to get information about an observed icon at the site of observation and at the time while observing. In this work, we presented a mobile application based on augmented reality technology that facilitates and speeds up access to iconographic content stored on the BIDL platform. The main goal, based on image recognition by a specially designed augmented reality module, is providing instantaneous and on-site information about the concrete icon observed by the visitors of churches and monasteries, or museums and galleries. The classical search over a large database, requiring keywords such as geographical location, name of the church, and similar, to access the proper information is avoided, since the icons are immediately recognized.

Keywords: Augmented reality, Mobile applications, Cultural heritage, Orthodox iconography.

1. Introduction

Icons are pictures on wood, fabric, glass, stones, or other materials, representing God's images, the Lord Jesus Christ, the Mother of God, Saints, holy persons, and biblical scenes

and events. They are expressions of Orthodox Faith, teaching, and worship. Icons are treasured in churches, monasteries, and other holy places, but also in homes, museums, and art galleries. To properly present these immeasurably valued items of Bulgarian national heritage, it was developed a web-based platform called the Virtual Encyclopedia of Bulgarian Iconography (BIDL) [22]. This Encyclopedia contains information about hundreds of Bulgarian icons from the 9th to the 19th century located in many churches, monasteries, museums, and other places.

The primary purpose of icons is to help in worship by aiding the observer to focus on the divine things while observing an icon. At the same time, icons are very valuable artistic artifacts from ancient times to date. For an observer, motivated by either religious or artistic interest, it is important to get more detailed information about a concrete icon, the subject of his interest. It is reasonable to assume that observers desire to get this information on the site and at the time when an icon is in focus of their interest. In that respect, searching over BIDL appears as a time-consuming option, uncomfortable to be performed on-site, and which also requires providing additional data to find the icon in the database. This data necessary for a BIDL search might not be available for an observer which can also misinterpret similar icons. Further, a kind of disappointment would be receiving the answer that the related information is unavailable in BIDL after spending some time on the search. Therefore, in this paper, we present an augmented reality (AR) technology-based application for mobile devices that overcomes these disadvantages.

The main idea is that the AR technology-based application for mobile devices recognizes an icon by the camera of the mobile device and immediately leads to the corresponding link with the desired information or, in rare, but still possible cases, issues the answer that such data is not available. The recognition is performed by a specially designed Augmented Reality Icon Recognition (ARIR) mobile application. This application is realized for both Android and iOS supported devices and it will be freely accessible on the corresponding markets Google Play and App Store, respectively. The ARIR application is implemented as an upgrade to the BIDL, and, therefore, the complete system is called AR-BIDL.

The remaining sections of the paper are structured as follows. In section 2, we review the literature that concerns the usage of augmented reality for the presentation improvement of cultural and religious heritage. In section 3 we present the motivation for this work and identify the main goals. Based on the main goals we define the system architecture and provide the implementation details in section 4. In section 5 we explain how the system is used and identify the potential users of the system. Experimental testing and verification of the system are given in section 6. The scalability of the system of potential growth is discussed in section 7 while section 8 presents the conclusion.

2. Related work – AR at religious places

Augmented reality, as one of the emerging trends and technologies in libraries, galleries, museums, and archaeological sites helps increase interest and knowledge about cultural heritage, especially among the younger generation [18], [11], [3]. Various types of AR solutions have been explored in order to immerse visitors in cultural heritage content [24], [5], [12]. Literature reviews on the application of augmented reality in cultural heritage show that augmented reality mobile applications have a positive impact on the immersion

and engagement of visitors to cultural sites [26]. Such mobile applications can provide guided tours to enhance visitor's experience at cultural heritage sites [15]. Also, applications can provide interactive storytelling to visitors by recognizing the monuments or their parts [28]. AR applications can be used to show educational content about artworks in galleries or outdoors, as for instance pictures on ancient rocks [23], [1], [30].

Special types of applications concern the usage of augmented reality at religious places for digital storytelling about historical and spiritual content. AR in religious places is used to provide an innovative interactive way that enhances visitors' understanding and meaning of the religious content. Moreover, augmented reality technologies can be used to preserve, and display virtual reconstructions of religious places to visitors.

A mobile application based on augmented reality is developed to better present the cultural heritage of the religious place of Piazza dei Miracoli in Pisa [6]. Augmented reality technology is used to provide historical information about important landmarks in the open space of this location. The application guides visitors to nearby places through different periods and shows information about important stages using augmented reality. The visitor chooses between nearby landmarks that he wants to explore and visualize the important data for that selection by using augmented reality. A virtual timeline is embedded to explore data from different epochs and the chosen period overlays 3D models or images over the current state. The study's findings showed that this engaging application can improve visitors' experiences in exploring the information at the location [6].

To preserve and better present religious heritage the AR is used for the virtual outdoor reconstruction of the demolished Reformed Church from Brasov City [2]. The application developed for mobile devices was used to recognize the place where the church existed. The recognition of the current state is based on image recognition and uses photos captured from different angles and stored in the local database. The application shows virtual reconstruction based on the visitor's location and recognition of the place where the church existed. The old photos overlap the current state which is the field of view of the visitor camera. The results of the visitor's survey showed that the AR application for the reconstruction of religious and cultural heritage is easy, interesting, and enjoyable to use.

The paper [14] presents how a 3D scanned model was used to reconstruct the Exeter Cathedral west front. This reconstruction is used to recreate the colors of the façade stonework that existed in the past. The main goal was to optimize the 3D colored model for augmented reality visualization. In this way, visitors can use augmented reality to recognize the current façade and see the colorized reconstruction of stone parts.

The photogrammetry technique for 3D model generation can be used to preserve virtual religious architecture. In [17], photogrammetry is used to create 3D models of the most representative altars of the Cathedral of San Pedro in Guayaquil, Ecuador. Using augmented reality, the photogrammetric generated 3D altar models are used to present religious heritage and increase learning about religious content. The survey results showed that combining AR with photogrammetric technology is effective and improves knowledge about cathedral heritage [17].

Virtual reconstruction of the Ayazini Virgin Mary Church interior using augmented reality is realized with the aim of better presentation of old religious heritage [27]. The interior elements of the church are ruined over time and reconstruction is complex and expensive. Therefore, the demolished church elements such as columns are modeled in 3D based on expert opinions. The mobile application is used to recognize QR codes placed

at the exact places of demolished columns in the interior of the church. When applications capture QR codes inside the church virtual reconstruction of columns is visible. In this way, visitors can better explore and sense the church space.

Illustrated pages of promotional or learning material can be used for AR recognition and immersive presentations for multimedia education about cultural and religious heritage. Interactive brochures created for guiding through the Temple of Debod in Madrid are used for AR recognition [8]. AR is implemented to provide interactive storytelling about important parts of the site. Pictures of the eight most significant engravings from the temple walls were used in the brochure. AR recognition of temple engravings is overlaid with multimedia historical content and enables visitors to acquire more knowledge about them. Similarly, AR applications have been created to attract students to learn better Malang temple history and understand temple relief art [10]. Images of temples and temple reliefs have been used as AR markers printed on flyers and textbooks. Recognition of AR markers shows virtual overlays as reconstructions of temples and colorization of reliefs.

A guide based on augmented reality technology was created to recognize religious artworks in Museo Diocesano of Milan [7]. This is a religious museum that has a collection of sacred artworks. Augmented reality is used to establish interaction with the museum exhibition and provide a deeper meaning of religious artworks. Five paintings have been chosen for recognition with religious scenes and different meanings. Also, the depicted scenes are not understandable to regular visitors. Recognizing the paintings and the depicted scenes with AR technology multimedia content is used to provide detailed explanation and interpretations.

Augmented reality may be used for the recognition of the relics in order to provide interactive storytelling. As an example of such applications, we point out the application designed for the Basilica of Saint Catherina of Alexandria in Galatina as an aid for the enhancement and understanding of religious and cultural heritage [4]. The AR is implemented to recognize the most famous frescoes located on the interior walls of the Basilica. An image tracking solution is used with eleven frescos that are stored as image markers. Recognizing the frescos, storytelling is realized by overlaying relevant multimedia content such as audio interpretation or image reconstructions. The survey results show that this AR application is an effective and attractive tool for interpretation and learning about frescos in this basilica as well as other places with frescos from related periods or artistic styles.

In the related literature, AR technology was used to present, explain, and improve the presentation of cultural and religious heritage. These applications concern the usage of AR technology for solving specific problems on concrete indoors and outdoors in particular religious places. In such situations, the number of AR tracking objects is small, and related multimedia content is limited in quantity and easy to handle with contemporary digital devices and their memory capacities. Thus, all necessary data is stored locally in the memory of the mobile device as a part of the application. Therefore, the authors didn't consider solutions for the storage problems and scalability of the systems.

AR system scalable architecture for large areas that uses large numbers of multimedia objects for cultural heritage is given in [25], [16]. These systems consider outdoor location-based AR tracking technology for showing historical information to visitors

throughout the city. The systems are made for big city areas and discuss the techniques of optimization for content that is used in the application.

In our work, we consider image tracking solutions for a wide area that considers cultural and religious heritage focused on indoor usage. The BIDL platform contains information about many places such as churches and monasteries stored in the database and each of these places typically has a large number of icons. Thus, the number of AR targets is large, as well as the multimedia content associated with them. However, as a visitor can be at only one place at a time, locally stored image markers for other locations can cause unnecessary usage of the local memory of mobile devices. Further, the BIDL platform enables adding new content or updating information in the server database. If icon image markers are stored locally, each content modification in the BIDL, will necessarily require an update of the mobile application AR-BIDL. Accordingly, our solution is created to store the icon image markers on the BIDL server system. This enables scaling the system and overcoming local storage problems. Also, the AR recognition system enables quick information retrieval from a remote server about nearby icons.

3. Motivation

The motivation for realizing this work is based on twofold goals:

- using AR technology to speed up information retrieval about icons in visitors' surroundings,
- to provide optimized AR recognition content concerning storage of icon image markers for the visitor's current location.

3.1. Searching problem

The BIDL platform stores a wide spectrum of places such as churches or monasteries with descriptions of hundreds of iconographic objects stored on the server. That data is available to visitors using web technologies. Concrete information about icons is available using keywords or in a predefined list of icons ordered by title, author, scene, etc. This approach has limitations whenever a visitor is at the exact location and wants to find information about the observed icon. Searching for icon information might be time-consuming if the visitor has no prior knowledge about icons that are of his current interest or hasn't prepared in advance for visitation to that place. Also, the search can be long if information about the observed icon isn't stored in the database.

In this work, we extend the current structure of the BIDL platform and improve search data about concrete icons at the location of the visitor. We created a mobile application based on augmented reality technology. This technology is used to provide a virtual signal or element and notify the visitor about the availability of information on the observed icon. Also, basic information can be obtained by recognizing the icons of interest and more detailed information is provided by interacting with virtual elements. If the icon is not recognized, this can be a signal that information about the observed icon is unavailable.

3.2. Storage problem

To achieve the AR recognition effect, image targets of icons should be prepared for recognition. AR development tools usually have two possibilities to store tracking targets, by using local database storage or cloud recognition services.

When targets are stored locally on the device, the AR application can immediately start icon recognition. A static solution from the point of storage where image targets are embedded locally in the application is good for projects that do not have frequent updates. Using this kind of target organization is unsuitable for the BIDL database since it enables adding new iconographic content or information updates. Thus, this kind of AR application will require editing the project and updating the application each time new material is added to the BIDL. Also, there are local memory concerns as for all locations covered by in the BIDL icon targets should be stored locally although only those related to one location should be used at the time.

Cloud recognition is designed to work with huge amounts of tracking targets that can be stored in a remote database. The popular AR tools have their custom cloud system solutions that can accept about a hundred thousand image targets for recognition. In this way, storage concerns of the AR mobile application are resolved. This solution enables dynamic target editing and integration with other systems and services to store data. Also, cloud recognition solution assumes some costs based on the frequency of targets used per day. As the BIDL database has information about a few hundred and the potential to have thousands of icons this solution is insufficient for point-of-data amount and extra expenses in price terms.

Considering the storage problem, our solution provides only necessary icon image targets for AR recognition. The application dynamically receives image targets stored on the BIDL platform. As the database stores information about various churches and monastery icons, the mobile application receives only content for one location at the time. This is realized through a web service that responds depending on the location of the visitor. The service provides image tracking and icon information that is shown when the icon is recognized. Additionally, any new content added to BIDL will be automatically delivered to the mobile application without updating.

3.3. Modeling AR-BIDL system

The AR-BIDL model in Fig. 1 is proposed as a solution based on the key goals that the system has to address. The BIDL platform is on the server side with an implemented specialized web service. This service enables communication with the client side which is realized as an ARIR application. ARIR application combines several components to communicate with the web service and provide information about nearby icons. When a visitor is at the location of a church or monastery, the mobile application sends the GPS coordinates to the web service. The web service prepares data about icons based on the received location coordinates as a response. The ARIR application processes the response and provides data to the AR module. Then, the AR module loads image markers for tracking, and basic virtual information is provided upon recognition of the concrete icon. AR module enables interaction with virtual elements such as virtual buttons projected on screen. In this interaction, visitors receive detailed information about the icon of interest.

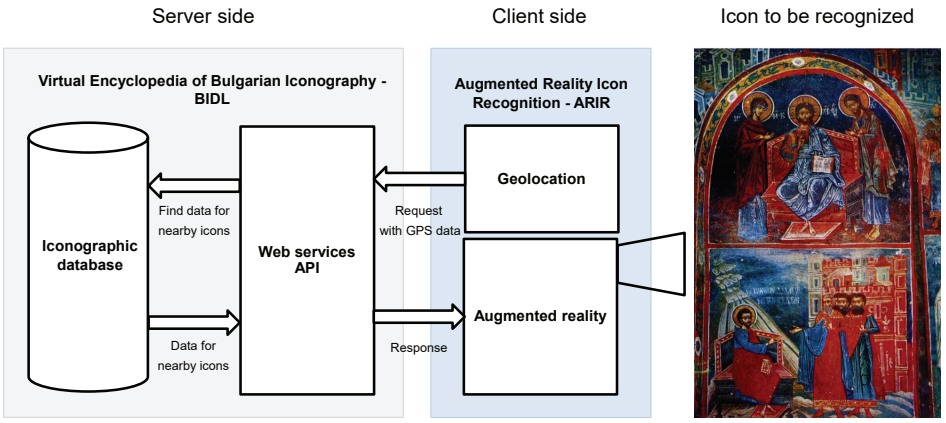


Fig. 1. AR-BIDL system model

4. Architecture of the system

Based on the proposed model, a client-server architecture is developed, and the system realized (Fig. 2). This section provides a detailed description of the BIDL system architecture with necessary web services for content delivery to the client side. Also, the ARIR mobile application as client-side architecture is described as well as the implementation of the entire system.

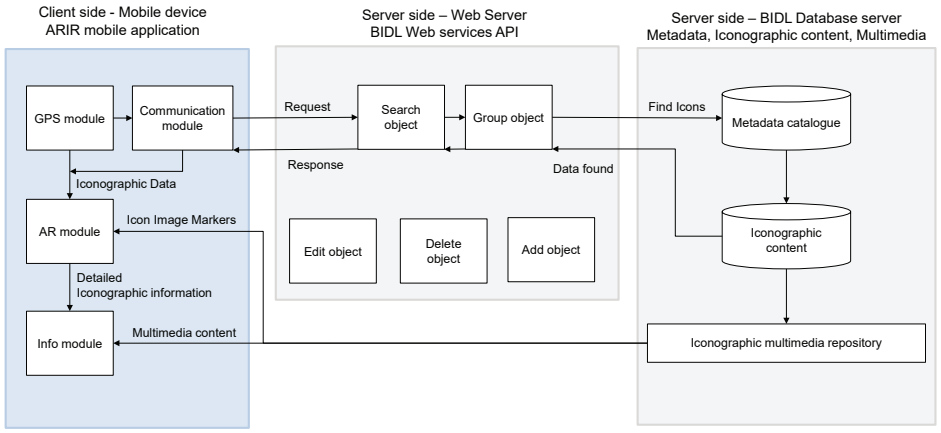


Fig. 2. AR-BIDL system architecture

4.1. BIDL as the Server Side

Virtual Encyclopedia of Bulgarian Iconography is based on the CultIS software platform [13]. CultIS is selected as a versatile, flexible, and highly adaptable digital content management solution developed by the IMI-BAS team [21]. With its active development using modern web technologies (Node JS, MongoDB, VueJS, Bootstrap, Sphinx search, etc.) it is a stable solution for digital libraries, virtual museums, galleries, archives, and other kinds of content management-based installments areas. It is viewed both as a suitable tool for experts in the field, as well as an appropriate learning tool for students.

BIDL database stores detailed information about iconographic objects which are used for semantic annotations and indexing. The iconographic object is described with a title of the icon, iconographical object type, the author or artist, iconographical school, the period when it is created, dimensions, location and source, identification notes, description, iconographical technique, and comments such as current state and restoration details (Fig. 3). Also, each iconographical object can have associated multimedia files such as images, audio, 3D, and video stored in a multimedia repository with metadata descriptions. Especially, for the system presented in this paper, added multimedia data concerns icon image marker information.

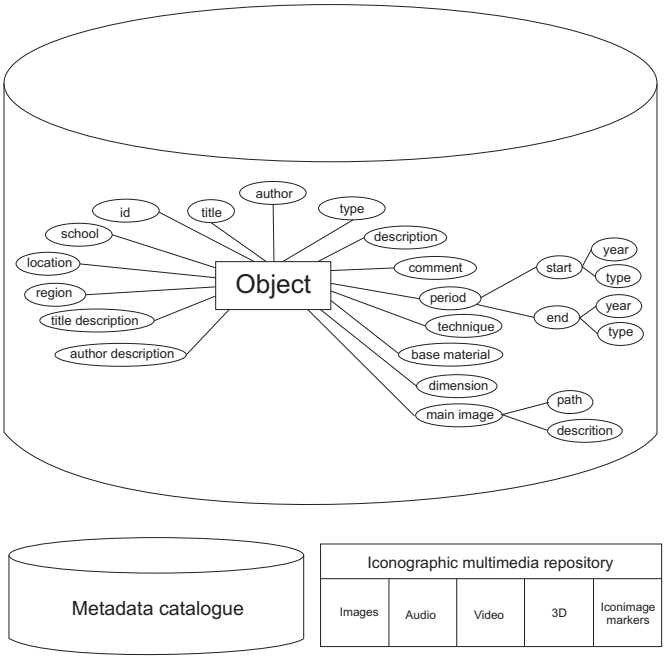


Fig. 3. Iconographic object and iconographic multimedia repository

Metadata structures in CultIS are managed using a dynamic model-building service discussed in Fig. 4, [20]. It provides the ability to maintain complex structures (including

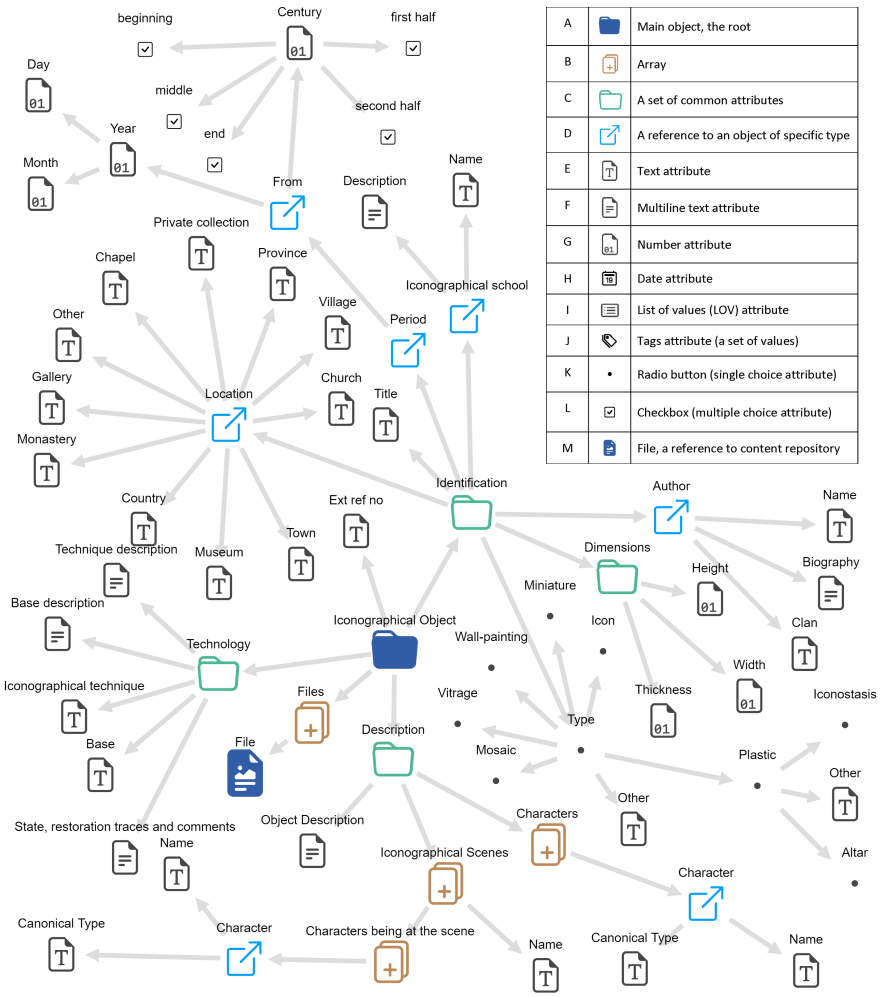


Fig. 4. BIDL metadata structure

arrays and recursive relations) and to extend them anytime. This is used for faster keyword searches of iconography content in the BIDL database.

The BIDL has implemented an application programming interface, an API-based backend, using the REST/JSON standard [19]. This API allows other parties (websites, mobile or desktop apps, etc.) to connect and communicate with the platform, request data using queries, and receive content according to their access level. These features are essential for easy and lightweight integrations with other local or cloud services and software products. Some of those services are:

- Add object service can be used to store iconographic objects in the database,
- Edit object service serves to modify information about infographic objects,
- Delete object service is created to remove information about infographic objects from the database,
- Group object service creates collections of iconographical objects,
- Search object service is used to achieve communication with the ARIR mobile application and provide information about icons.

A specialized search service on the BIDL platform has been developed to receive HTTP requests. This request has implemented the GPS location of the client-side mobile device. The service uses this GPS location to determine the nearest church or monastery. The service does calculations and checks if there is an orthodox object in the range of 1 km. Next, the database query delivers information about icons in the nearest church or monastery. This information is classified by group object service, prepared in JSON format, and sent to the client side. The information stores iconographical objects information and links to the multimedia content such as icon image markers for AR recognition.

4.2. ARIR Mobile Application Modules

The client-side is realized as the ARIR mobile application in the Unity engine. The development is done as a cross-platform application designed for different types of mobile devices with operating systems such as Android or iOS. ARIR consists of four modules realized as Unity scenes. These modules are the Communication module, the GPS module, the Info module, and the AR module as shown in Fig. 2.

GPS Module The GPS module is created to receive the geographical location provided by a location-based service on a mobile device. The received longitude and latitude are used for sending the request to the REST service in the BIDL using the Communication module. Additionally, this module provides notifications about the regularity of service response and whether to move forward with the AR module.

Communication module The Communication module exchanges data with the web service implemented on the BIDL platform. It sends requests about the current location in order to get information related to the iconographic objects in the nearest church or monastery. Next, this module receives as a response data about iconographic objects from the server. These data are parsed to provide textual information about iconographical objects. Also, data integrates links to multimedia content such as image targets for the AR module or photos, audio, video, and for the Info module visualization.

AR Module The AR module is developed by using the EasyAR SDK [29] to recognize and track icons at locations such as churches or monasteries. Photos of icons are stored in the multimedia repository on the server as AR image markers for recognition and tracking. The main components of the AR module and their relations are shown in Fig. 5. Information about the locations of AR image markers on the server is provided as a part of the response for each iconographic object. These data are provided to the special data structure named IconographicDataManager. Iconographic objects, besides descriptive elements about icons, have implemented links to the icon image markers.

The AR module starts when markers are dynamically loaded into the application using ARManager which loads marker links from IconographicDataManager and creates IconImageMarker. In this way, for each provided link, ARManager creates an icon image marker. Next, ARManager creates a virtual object as VRObject associated with the icon image marker. This virtual object will be shown during the recognition. The virtual object has an integrated virtual button, name, and unique identifiers (ID) of the icon.

When an icon is recognized, the corresponding virtual object overlays the icon with a virtual button and basic information such as its name. The other iconographical data can be displayed, if necessary, in this step. In interaction with the virtual objects the AR module links to the Info module. The ID is used to identify more detailed data about the recognized icon from IconographicDataManager and is shown in the Info module.

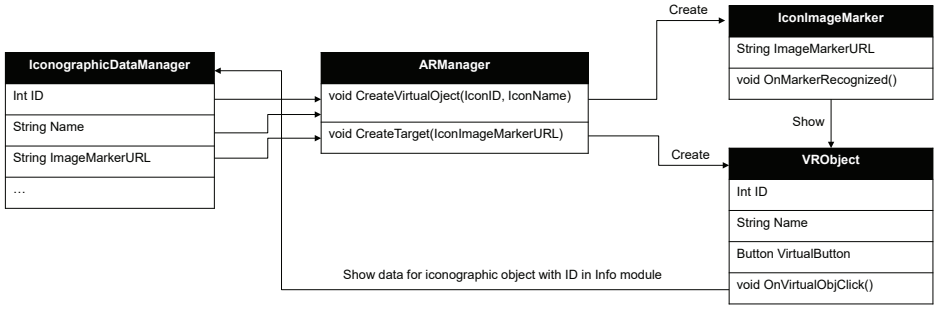


Fig. 5. Important components of the AR module

Info module The Info module is developed to show information about a recognized icon with the AR module. Based on the icon ID detailed information is displayed as textual information. Multimedia content such as audio, video, or 3D, is supported for visualization in this module and it is available via a link to the BIDL multimedia repository.

4.3. Realization of ARIR

The realization of the ARIR mobile application and usage of implemented modules are shown in Fig. 6. The application starts with the activation of a location-based service in the GPS module. This module loads the longitude and latitude of the visitor device to optimize the search for the icons that will be sent from the server. Using the Communication

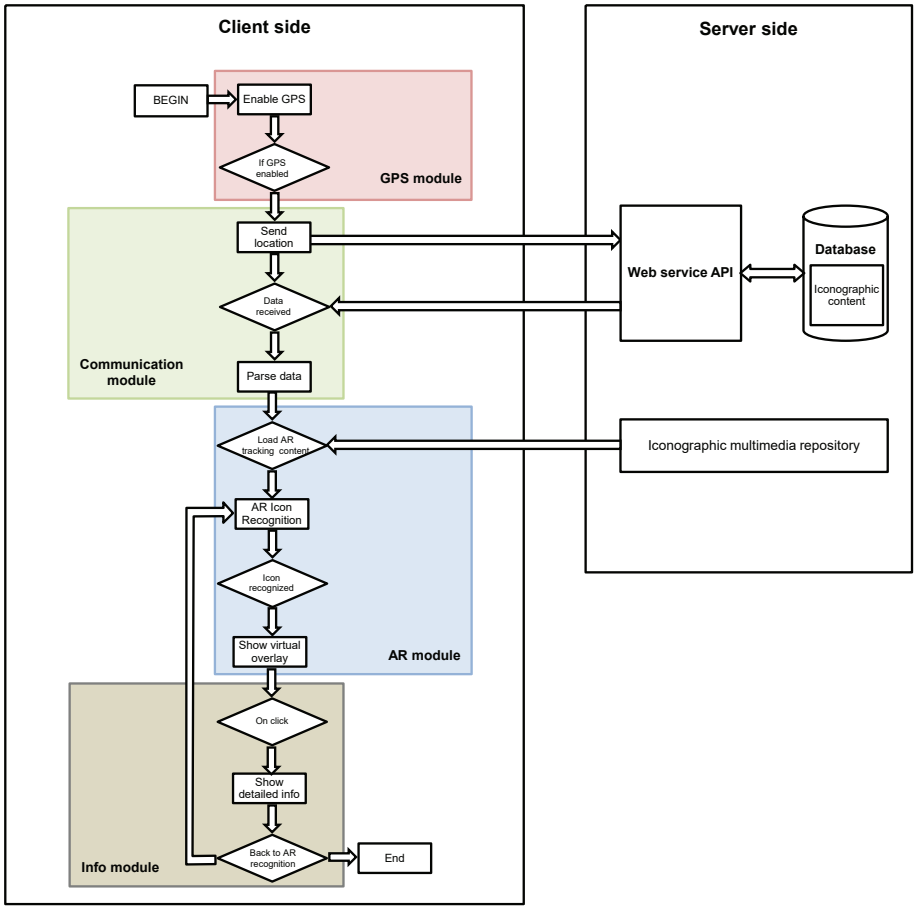


Fig. 6. Algorithm of ARIR application usage.

module these data are embedded as a part of the request and sent to the specialized search service. The search service processes this request by comparing this data with locations of places stored in the database such as churches or monasteries. As a response, iconographic content for the nearest place is prepared and sent in the form of JSON format to the ARIR mobile application.

The Communication module parses response data received from the server. These data have detailed information on iconographic objects and necessary data for the AR recognition. Data necessary for the AR module are the URL of the image markers and basic data of the icons such as ID and name. Based on the provided URL the AR module receives the image markers from the iconographical media repository.

AR module starts when data is successfully received from the search service and image markers are loaded from a media repository on the BIDL platform. Then, the AR tracking can begin and information about icons of interest can be quickly found at the location. When an icon is recognized, the basic information about the icon and interactive virtual object is displayed as an overlay during the tracking. Through the interaction with this virtual object, the AR module activates the Info module. The info module loads and displays detailed multimedia information about the recognized icon.

5. Usage of the ARIR

The ARIR application development concerns the creation of an interface design. As the application is aimed for the usage at religious places, a minimalistic interface design is used. Furthermore, the types of users are determined for the usage of the ARIR.

5.1. Interface Design

The user interface is designed according to the modules functionalities used for data visualization. Fig. 7 shows screenshots of the design of the application user interface. The first screenshot represents the home screen design. The button on the menu activates the procedure for the GPS module. If the location-based service is deactivated, the user has to enable it to proceed with the application. This is presented in the second screenshot. When a visitor enables a geolocation service background process sends a request for iconographic data of the location. The third screenshot shows that data was successfully received from the server, and it provides a brief explanation about the usage of AR. The activation of the AR module, which enables the device's camera and icon recognition, is given in the fourth screenshot. The last screenshot represents the visualization of the Info module, which displays more information about the identified icon.

5.2. Usage of AR-BIDL

This section presents a usage scenario for the AR-BIDL in situations usually met in practice. It is assumed that a visitor to an Orthodox church may be interested in learning more about the interior icons. At the location, he is informed that the application AR-BIDL is freely available at Google Play or AppStore and can be downloaded by scanning the corresponding QR codes printed on an info table provided at the site, or on a flyer, or shown

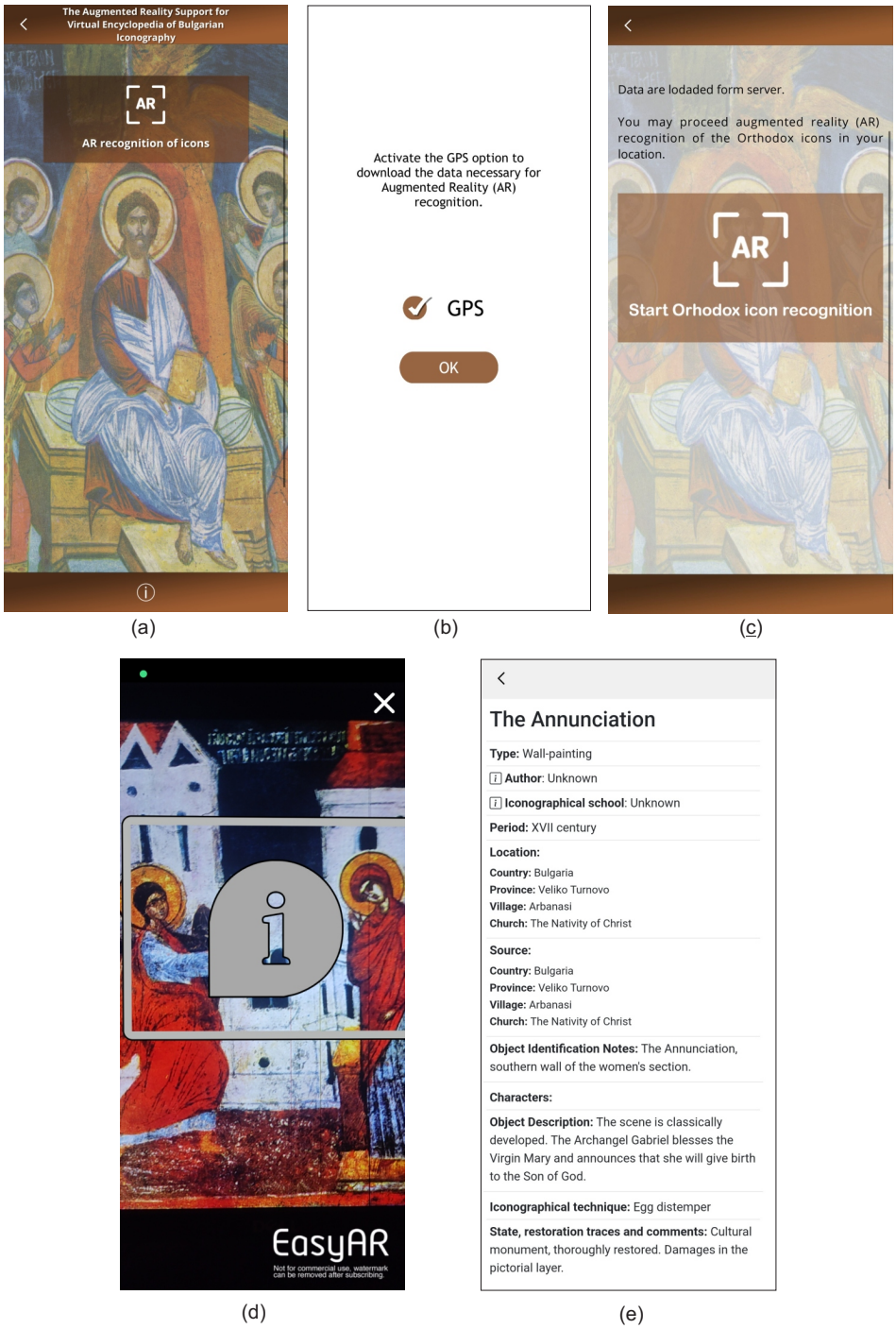


Fig. 7. (a) Home screen of the application, (b) Activation of the GPS module, (c) Data download from a server, (d) Activation of AR module, (e) Visualization of Info module.

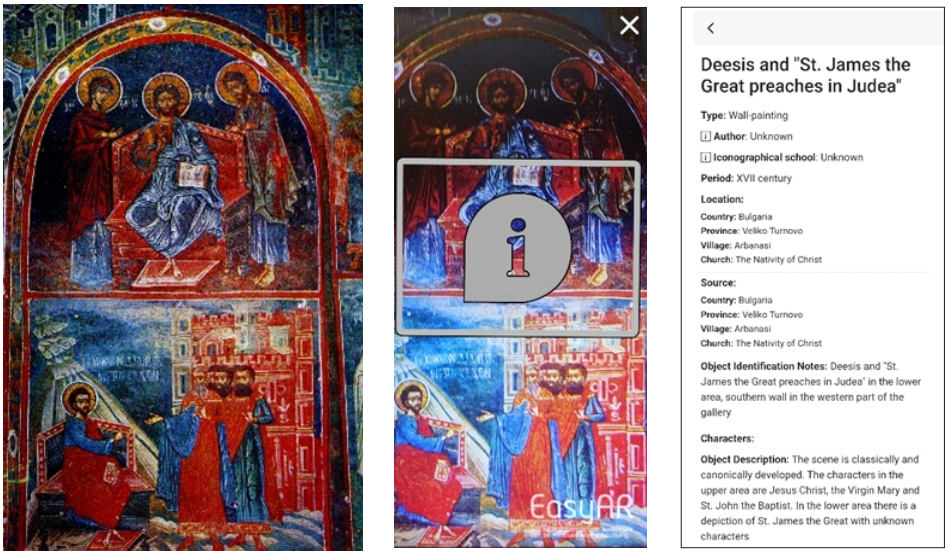
Icon image target			
Name	The Annunciation	Deesis and "St. James the Great preaches in Judea"	All angel-kind was amazed
Type	Wall-painting	Wall-painting	Wall-painting
Author	Unknown	Unknown	Unknown
Iconographical school	Unknown	Unknown	Unknown
Period	XVII century	XVII century	XVII century
Location:			
Country	Bulgaria	Bulgaria	Bulgaria
Province	Veliko Turnovo	Veliko Turnovo	Veliko Turnovo
Village	Arbanasi	Arbanasi	Arbanasi
Church	The Nativity of Christ	The Nativity of Christ	The Nativity of Christ
Source:			
Country	Bulgaria	Bulgaria	Bulgaria
Province	Veliko Turnovo	Veliko Turnovo	Veliko Turnovo
Village	Arbanasi	Arbanasi	Arbanasi
Church	The Nativity of Christ	The Nativity of Christ	The Nativity of Christ
Object Identification Notes	The Annunciation, southern wall of the women's section.	Deesis and "St. James the Great preaches in Judea" in the lower area, southern wall in the western part of the gallery	"All angel-kind was amazed", the southern wall of the women's section.
Object Description	The scene is classically developed. The Archangel Gabriel blesses the Virgin Mary and announces that she will give birth to the Son of God.	The scene is classically and canonically developed. The characters in the upper area are Jesus Christ, the Virgin Mary and St. John the Baptist. In the lower area there is a depiction of St. James the Great with unknown characters	The scene is classically and canonically developed.
Iconographical technique	Egg distemper	Egg distemper	Egg distemper
State, restoration traces and comments:	Cultural monument, thoroughly restored. Damages in the pictorial layer.	Cultural monument, thoroughly restored. Damages in the pictorial layer.	Cultural monument, thoroughly restored. Damages in the pictorial layer.

Fig. 8. Some of the iconographical objects of the Nativity of Christ Church

in some other suitable way. As an example, table at Fig. 8 shows information about three icons in the Nativity of Christ Church.

After downloading and installing, the visitor starts the ARIR application at the location. The GPS module receives the visitors' location when the AR button is pressed and then obtains information about nearby icons from the BIDL server.

The augmented reality scene opens when parameters for recognition and tracking icons are loaded from the BIDL server. The camera on the mobile device is activated and the visitor points the mobile device toward a specific icon as shown in Fig. 9 (a). The image captured by the camera appears on the screen of the mobile device. When the icon Deesis and "St. James the Great preaches in Judea" captured by the camera is recognized, the virtual button appears as an overlay (Fig. 9 (b)). This is also a signal to the visitor that information about the observed icon is available.



(a) (b) (c)

Fig. 9. (a) Icon image target - Deesis and "St. James the Great preaches in Judea", (b) Augmented reality tracking Orthodox icon, (c) The Application displays the detailed information received from the BIDL platform.

Interaction with this button provides a link to the Info scene where more detailed information about the recognized icon is provided. This information given in table at Fig. 8 is displayed in the Info scene in the format shown in Fig. 9 (c).

In this way, through the ARIR, the information about an icon is provided instantaneously at the site while observing the icon. The search over the large BIDL database, with necessary requests for additional information about the geographical location or other data to locate the icon, is avoided. Further, possible misinterpretations due to human errors in recognizing similar icons are eliminated.

5.3. Users of the application

The ARIR application is aimed at contemporary visitors who want to be informed about icons at the location in a fast and comfortable way. Augmented reality enables digital overlays over the icons that could provide information according to visitor's interest. These digital overlays unlock the artistic features and religious meanings of observed icons. Also, the application may deliver interactive storytelling about saints or biblical events, which is supposed to be of a particular interest for the younger users. This can be done by creating more animated material that can be shown inside a mobile application. At the same time, including specific multimedia material can improve presentation for visitors with disabilities. Recognizing the icons of interest audio storytelling can provide significant assistance to visitors with visual disabilities. Embedding video material with sign language can improve the interpretation of religious and artistic meaning for visitors with hearing problems. Moreover, specialized guided tours about iconographical content could be organized during the visitation to one or more places such as churches and monasteries. This enriches the traditional visitations, making them more informative and interactive. In this respect, the use of AR should preserve the spiritual significance of holy places while improving the religious and cultural experience.

This system can be useful as a remote learning tool for students. The AR application used directly on the site will provide educational content quickly upon icon recognition. Especially students of arts and theology can get educational material as interactive learning content to discover information about the observed icons in situ. This simplified interaction can help students to more effectively interpret depicted icons with their spiritual meaning. Also, it can help a deeper understanding of the iconographic techniques, period of realization, and chosen materials for iconography development. Moreover, this system enables professors to organize interactive learning experiences at exact locations. This can be performed as learning tours at remote classrooms where students have tasks to scan and explore the concrete icons at given locations.

This system can be used not just for visitors or students, but also for experts in the field of iconography such as conservators, iconographers, topologists, etc. In this case, the system can be used to share digital documentation stored in the BIDL database. Also, stored global knowledge can serve for analysis and collaboration work for processes such as conservation or restoration. This includes important metadata that can be shown during recognition.

6. Experimental testing and verification of the application

As it is customary practice, we performed experimental testing and verification of the developed ARIR application. Experiments were directed towards checking the functionality and usability of the application and were performed by following recommendations for testing AR-based applications presented in [25], [16], [9]. It should be taken into account that these recommendations are primarily intended for AR applications where the primary tasks are generating 3D models and their correct positioning and good visualization. In the case of the present application, the most important issue is to test the application response when the user is at an appropriate proximity to the location where iconographic objects of interest are exhibited. The application was installed on mobile devices with

various performances typical for ordinary users. Testing was performed at five different religious institutions in Bulgaria, two monasteries and three churches, at locations covered by different telecommunication networks offered by national providers. We also analyzed the response of the application to a request for presenting information content related to particular icons.

6.1. Functionality testing

To test the functionality of the application with respect to the speed of the response to the users' requests, we performed experiments with 5 experienced users familiar with this particular application and also various other similar applications. It was assumed that the user is in proximity to the object of interest. We measured the delay between the moment of issuing the request to the web service for the data until augmented reality was ready for recognition. The obtained average response was no more than 20 seconds. It however can be concluded that the response time primarily depends on the size of the communication package exploited by the user of the application but also the speed of the Internet on the location. After augmented reality recognition of an icon, when interaction with a virtual object is achieved, the application response was within no more than 5 seconds showing the icon information. The application was functional in all cases.

6.2. Usability testing

The usability test is performed with users who haven't used this application previously. A total of 10 users tested the application at three different locations. At each location, it was installed a 100x70 cm table with QR codes for downloading the application and a brief explanation of its purpose. Despite that, a conclusion is that at the beginning, it was necessary to first provide a clear and precise explanation about the way of using it. Later, the demands for assistance were considerably reduced.

7. Scaling of the system

The AR-BIDL is a system that can provide fast information about the hundreds of Bulgarian Orthodox icons at the location. This system has the scaling potential to be extended by including more icons for AR recognition and inserting information about them into the database. A good feature of the proposed approach is that no update is required after expansion of the database by including icons from other locations. This expansion can further concern adding new icons from Serbian churches and monasteries since this system is realized as a joint research project "Development of Software Tools and Multimedia Technologies for Digital Presentation, Preservation and Management of Cultural Heritage" between the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences and the Mathematical Institute of the Serbian Academy of Sciences and Art. This project can be extended to other orthodox nearby countries to serve as a regional database for learning and education about iconography.

The system's limitation is that the AR recognizes only icons whose locations are known in advance in places such as churches and monasteries. If we suppose to add in the BIDL database icons that do not depend on location then it is required to reorganize the

AR BIDL system's structure. This can include home icons placed at traditional Orthodox prayer corners. Also, incorporates icons that are part of museum collections that can be displayed in different locations. Accordingly, the new structure should compare the images captured by the camera with all icon image targets stored in the database. This can be realized using the cloud recognition solution where the database can store and compare up to 100 thousand targets.

In the case of the AR tool that we already used, EasyAR SDK enables cloud recognition service. This cloud solution enables integrated Web services API intended to manage image targets and provide communication with other services. The BIDL solution also enables integration with cloud services and therefore the proposed solution as Augmented Reality Cloud BIDL (ARC-BIDL) is given in Fig. 10.

The specialized service for adding new iconographical content on the BIDL platform should be extended for communication with cloud service applications. This service should send image targets to the cloud application instead of storing them on the BIDL media database. In the BIDL database, only the image target ID is stored as a response from the cloud recognition service.

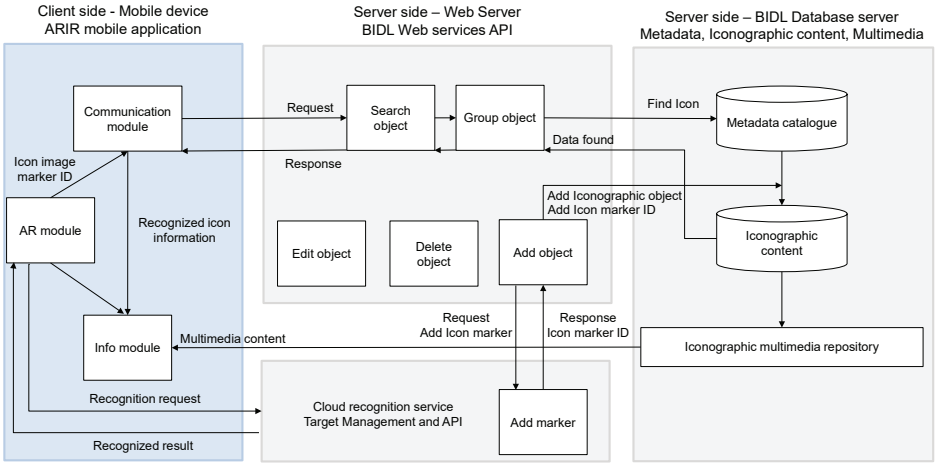


Fig. 10. Proposed ARC-BIDL system architecture.

The proposed ARIR mobile application solution starts and activates the AR module for recognition of Orthodox icons in visitors' surroundings. Then, the AR module sends a request for recognition of the image captured by the camera. As a response, cloud recognition service provides image target information. Overlay in the form of the virtual button appears on the screen of mobile device during the tracking of the recognized icon. Through the interaction with this virtual object, the ARIR loads the Info module. Also, as a background process, the Communication module is activated to send the request for the related content to the BIDL server. The request implements the icon image target ID of the recognized icon and sends it to the BIDL search service. The BIDL service responds

to the request and sends back the content about the recognized icon. This information will be parsed by the Communication module and sent for visualization in the Info module.

8. Conclusion

Virtual Encyclopedia of Bulgarian Iconography (BIDL) contains information about hundreds of Bulgarian iconographical artifacts including data about the holy person or other contents inscribed on the icon, as well as information about the painters and artistic techniques, type of colors, and other related data. As is the case with any web-based platform, the general problem in exploiting BIDL, in terms of speed and comfort, is finding information about a concrete icon on the site and at the time when a visitor observes the icon in a church, monastery, museum, or gallery. Typically for the usage of databases, it is necessary to provide keywords. This is resolved by developing the ARIR mobile application based on augmented reality technology to facilitate and speed up the search for related data over the BIDL platform.

Using the AR module of the ARIR application, the icon is recognized which enables a quicker search for information about the icon of interest at the exact location compared to the classical search using keywords. Manual retrieval results obtained from the database of BIDL by comparing images can be time-consuming and potentially lead to mismatching which is prevented by the AR module where the icon is immediately recognized. Furthermore, if the icon is not in the database, manual searching may cause a time loss, which the AR system prevents. AR overlays the icon with the virtual object as the signal to the visitor that information about the observed icon is available in the BIDL. The related data are projected after clicking on it. The system was experimentally verified by performing usability and functionality testing at several locations. Our further work under the above-mentioned bilateral project will involve different analyses and ways of informing potential users about the availability of the ARIR application and its features.

Acknowledgments. This research work was carried out and is supported by the joint research project "Development of Software Tools and Multimedia Technologies for Digital Presentation, Preservation and Management of Cultural Heritage" between the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences and the Mathematical Institute of the Serbian Academy of Sciences and Arts (2023-2025). Also, this work was supported by the Serbian Ministry of Science, Technological Development, and Innovation through the Mathematical Institute of the Serbian Academy of Sciences and Arts. The authors are grateful to the reviewers whose constructive comments were useful in improving the presentation of the paper.

References

1. Blanco-Pons, S., Carrión-Ruiz, B., Luis Lerma, J., Villaverde, V.: Design and implementation of an augmented reality application for rock art visualization in Cova dels Cavalls (Spain). Journal of Cultural Heritage 39, 177–185 (2019), https://doi.org/10.1016/j.culher.2019.03.014
2. Boboc, R.G., Gîrbacia, F., Duguleană, M., Tavčar, A.: A handheld augmented reality to revive a demolished Reformed Church from Braşov. In: Proceedings of the Virtual Reality International Conference - Laval Virtual 2017. pp. 1–4. VRIC '17, Association for Computing Machinery (2017), https://doi.org/10.1145/3110292.3110311

3. Capecchi, I., Bernetti, I., Borghini, T., Caporali, A., Saragosa, C.: Augmented reality and serious game to engage the alpha generation in urban cultural heritage. Journal of Cultural Heritage 66, 523–535 (2024), https://doi.org/10.1016/j.culher.2024.01.004
4. De Paolis, L.T., Gatto, C., Corchia, L., De Luca, V.: Usability, user experience and mental workload in a mobile augmented reality application for digital storytelling in cultural heritage. Virtual Reality pp. 1434–9957 (2022), https://doi.org/10.1007/s10055-022-00712-9
5. Díaz, P., Bellucci, A., Yuan, C.W., Aedo, I.: Augmented experiences in cultural spaces through social participation. Journal on Computing and Cultural Heritage (JOCCH) 11(4), 1–18 (2018), https://doi.org/10.1145/3230675
6. Duguleana, M., Brodi, R., Girbacia, F., Postelnicu, C., Machidon, O., Carrozzino, M.: Time-travelling with mobile augmented reality: A case study on the Piazza dei Miracoli. In: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. pp. 902–912. Lecture Notes in Computer Science, Springer International Publishing (2016), https://doi.org/10.1007/978-3-319-48496-9_73
7. Greci, L.: An augmented reality guide for religious museum. In: Augmented Reality, Virtual Reality, and Computer Graphics. pp. 280–289. Lecture Notes in Computer Science, Springer International Publishing (2016), https://doi.org/10.1007/978-3-319-40651-0_23
8. Gutierrez, J.M., Molinero, M.A., Soto-Martín, O., Medina, C.R.: Augmented reality technology spreads information about historical graffiti in Temple of Debod. Procedia Computer Science 75, 390–397 (2015), https://doi.org/10.1016/j.procs.2015.12.262
9. Hammady, R., Ma, M., Powell, A.: User experience of markerless augmented reality applications in cultural heritage museums: 'MuseumEye' as a case study. In: Augmented Reality, Virtual Reality, and Computer Graphics. pp. 349–369. Lecture Notes in Computer Science, Springer International Publishing (2018), https://doi.org/10.1007/978-3-319-95282-6_26
10. Hariyanto, L.S., Anggriani, S.: Studying the history of temple relief in Malang using textbooks based on augmented reality. ILEARNed 2(1), 34–43 (2021)
11. He, Z., Wu, L., Li, X.R.: When art meets tech: The role of augmented reality in enhancing museum experiences and purchase intentions. Tourism Management 68, 127–139 (2018), https://doi.org/10.1016/j.tourman.2018.03.003
12. Ibiş, A., Çakici Alp, N.: Augmented reality and wearable technology for cultural heritage preservation. Sustainability 16(10), 4007 (2024), https://doi.org/10.3390/su16104007
13. Institute of Mathematics and Informatics, Bulgarian Academy of Sciences: CultIS, https://cultis.math.bas.bg/en
14. Khosravi, M., Zia, R., Chang, J.: Exeter Cathedral: A colour reconstruction for use in augmented reality devices. In: Proceedings of EVA London 2022. pp. 28–33. BCS Learning & Development (2022), http://dx.doi.org/10.14236/ewic/EVA2022.6
15. Koo, S., Kim, J., Kim, C., Kim, J., Cha, H.S.: Development of an augmented reality tour guide for a cultural heritage site. Journal on Computing and Cultural Heritage 12(4), 1–24 (2020), https://doi.org/10.1145/3317552
16. Lee, G.A., Dünser, A., Kim, S., Billinghurst, M.: CityViewAR: A mobile outdoor ar application for city visualization. In: 2012 IEEE International Symposium on Mixed and Augmented Reality - Arts, Media, and Humanities (ISMAR-AMH). pp. 57–64 (2012), https://doi.org/10.1109/ISMAR-AMH.2012.6483989
17. Llerena-Izquierdo, J., Cedeño-Gonzabay, L.: Photogrammetry and augmented reality to promote the religious cultural heritage of San Pedro Cathedral in Guayaquil, Ecuador. In: Applied Technologies. pp. 593–606. Communications in Computer and Information Science, Springer International Publishing (2019), https://doi.org/10.1007/978-3-030-42520-3_47

18. Nepali, S., Tamang, R.: A review on emerging trends and technologies in library. American Journal of Information Science and Technology 6(1), 8–15 (2022), https://doi.org/10.11648/j.ajist.20220601.12
19. Paneva-Marinova, D., Goynov, M., Luchev, D.: Towards wider sharing of iconographical art content. In: Proceedings of the Fourth International Conference Digital Presentation and Preservation of Cultural and Scientific Heritage (DiPP2014). vol. 4, pp. 127–134 (2014), https://doi.org/10.55630/dipp.2014.4.14
20. Paneva-Marinova, D., Goynov, M., Luchev, D.: Multimedia Digital Library: Constructive Block in Ecosystems for Digital Cultural Assets. Basic Functionality and Services. LAP LAMBERT Academic Publishing, Berlin, Germany (2017)
21. Paneva-Marinova, D., Goynov, M., Zhelev, Y., Monova-Zheleva, M., Mitreva, E., Luchev, D., Pavlov, R., Pavlova, L.: Full-fledged access and usability of content in digital cultural heritage library: Approaches, paradigms and implementation. Journal on Computing and Cultural Heritage (JOCCH) pp. 1–12 (2023), https://doi.org/10.1145/3631135
22. Pavlova-Draganova, L., Georgiev, V., Draganov, L.: Virtual encyclopaedia of the Bulgarian iconography. In: Proceedings of the International Conference Modern (e-) Learning (ICMEL). pp. 165–170 (2006)
23. Pierdicca, R., Frontoni, E., Zingaretti, P., Sturari, M., Clini, P., Quattrini, R.: Advanced interaction with paintings by augmented reality and high resolution visualization: A real case exhibition. In: Augmented and Virtual Reality. pp. 38–50. Lecture Notes in Computer Science, Springer International Publishing (2015), https://doi.org/10.1007/978-3-319-22888-4_4
24. Ridel, B., Reuter, P., Laviole, J., Mellado, N., Couture, N., Granier, X.: The revealing flashlight: Interactive spatial augmented reality for detail exploration of cultural heritage artifacts. Journal on Computing and Cultural Heritage (JOCCH) 7(2), 1–18 (2014), https://doi.org/10.1145/2611376
25. Sánchez Berriel, I., Pérez Nava, F., Albertos, P.T.: LagunAR: A city-scale mobile outdoor augmented reality application for heritage dissemination. Sensors 23(21), 8905 (2023), https://doi.org/10.3390/s23218905
26. Silva, C., Zagalo, N., Vairinhos, M.: Towards participatory activities with augmented reality for cultural heritage: A literature review. Computers & Education: X Reality 3, 100044 (2023), https://doi.org/10.1016/j.cexr.2023.100044
27. Süvari, A., Okuyucu, Ş.E., Çoban, G., Eren Tarakci, E.: Virtual reconstruction with the augmented reality technology of the cultural heritage components that have disappeared: The Ayazini Virgin Mary Church. Journal on Computing and Cultural Heritage 16(1), 1–16 (2023), https://doi.org/10.1145/3579361
28. Tatić, D., Stanković, R., Jovanović, M., Stojanović, J., Andrejević, D., Arsić, N.: Monument to the liberators of Niš relates to us the history of the city. Review of the National Center for Digitization 39, 64–73 (2021)
29. VisionStar Information Technology: EasyAR: Augmented Reality SDK, https://www.easyar.com/
30. Westin, J., Råmark, A., Horn, C.: Augmenting the stone: Rock art and augmented reality in a Nordic climate. Conservation and Management of Archaeological Sites 23, 258–271 (2021), https://doi.org/10.1080/13505033.2023.2232416

Dušan Tatić received his M.Sc. and Ph.D. degrees in Electrical Engineering and Computer Science from the Faculty of Electronic Engineering, University of Niš. He was a research assistant at the Faculty of Electronic Engineering in Niš from 2012 to 2021. From 2021 to present, he works as a research associate at the Mathematical Institute

of the Serbian Academy of Sciences and Arts. His research interests primarily concern augmented reality technologies and their application in an industrial environment. Also, he does research concerning the applications of augmented reality and multimedia technologies in the presentation of the national historical, cultural, technical, and scientific heritage of Serbia. As a member of the ARhiMedia group, he was involved in developing various multimedia projects for cultural institutions in Serbia, including over 50 mobile applications for the presentation of national and cultural heritage.

Radomir S. Stanković received his B. Sc degree in Automatic and Informatic from the Faculty of Electronics in Niš, University of Niš in 1976, and M. Sc. and PhD degrees in Applied Mathematics from the Faculty of Electrical Engineering in Belgrade, University of Belgrade in 1984 and 1986, respectively. He was a professor at the "Mija Stanimirović" School of Electrical Engineering in Niš from November 1976 to April 1987, and from then until July 2017, he worked at the Computer Science Department of the Faculty of Electronics in Niš, going through elective titles, assistant professor until 1992, associate professor until 1997, and a full professor since 1997, in which title he retired in 2017. From July 1, 2017, until his retirement on July 1, 2019, he worked as a research professor at the Mathematics Institute of the Serbian Academy of Sciences in Belgrade. He was awarded in 1997 a Kyushu Institute of Technology Fellowship and worked at the Department of Computer Science and Electronics, Kyushu Institute of Technology, Iizuka, Fukuoka, Japan. In 2000, he was awarded the Nokia Professorship award by the Nokia company in Finland. From August 1999 to April 2017, he worked half-time at the International Signal Processing Center at the Department of Signal Processing, Tampere University in Tampere, Finland, first as a visiting researcher and from April 2009 to April 2017 as a visiting professor. Radomir Stanković's area of interest include switching theory, multi-valued logic, spectral techniques and signal processing. His particular area of interest is the history of computing. From 2012 until now, he also works towards applications of information technologies in the digitization of national heritage as the founder and head of the ARhiMedia group of the Faculty of Electronics in Niš and the Mathematics Institute of the Serbian Academy of Sciences in Belgrade. So far, he has participated in the development or managed the development of 45 applications for mobile devices in the field of presenting national heritage.

Detelin Luchev is an Assoc. Professor at IMI-BAS, Scientific Secretary of the Mathematical Linguistics Department at IMI—BAS (2012 - 2024) and Chair of the Mathematical Linguistics Department at IMI—BAS (2024 – at present). He holds a MA in History, a Master's degree in Computer Science with a specialization in Language and Multimedia Technologies and PhD in Ethnology. He has strong background in Theory and Development of Digital Libraries, Technology-enhanced Learning, Knowledge Presentation for Folklore Digital Library. He has experience in development of applications for Serious Games and Storytelling, Digital Libraries, and Virtual Museums. He is a key participant in 25 national and international research projects. He has more than 95 publications in research journals and conference proceedings, 1 book, and more than 110 citations. He is a key researcher in the development of 4 digital content management systems for cultural heritage and 2 serious games.

Maxim Goynov is an Assist. Professor at IMI-BAS. He has strong background in Design and Software Development of Digital Libraries and Serious Games. He is a key participant of 15 national research projects. He has more than 45 publications in research journals and conference proceedings, 1 book, and more than 70 citations. He is a key researcher in the software development of 6 digital content management systems for cultural heritage and 2 serious games.

Desislava Paneva-Marinova is a Professor and a Senior Researcher at IMI—BAS, Chair of the Mathematical Linguistics Department of IMI—BAS (2012 - 2024), Secretary of the Scientific Council of IMI—BAS (2018 – at present). She holds a Bachelor's degree in Mathematics and Computer Science, a Master's degree in Computer Science with a specialization in Language and Multimedia Technologies. Doctoral thesis "Semantic-oriented Architecture and Models for Personalized and Adaptive Access to the Knowledge in Multimedia Digital Library", successfully defended before the Specialised Academic Council for Informatics and Mathematical Modelling in 2008. She has strong research background in Theory and Development of Digital Libraries, Technology-enhanced Learning, Knowledge and Semantic Web Technologies, Human-Computer Interaction, Personalization and Content Adaptation, etc. She has experience in development of applications for Digital Libraries, Virtual Museums, Digital Ecosystems with Cultural Assets, Serious Games and Storytelling. She is a key participant and a site leader of 25 national and international research projects, advisor of 7 PhD students, etc. She has more than 150 publications in research journals and conference proceedings, 2 books, and more than 450 citations. She is a key researcher in the development of 8 digital content management systems for cultural heritage and 2 serious games. She is teaching at 3 Bulgarian universities in the mentioned above academic fields.

Received: October 17, 2024; Accepted: April 26, 2025.

Computer Science and Information Systems 22(4):1457–1482 https://doi.org/10.2298/CSIS241119056W

Federated Learning with Committee Mechanism for Class Imbalance

Lang Wu and Yi Dong

School of Applied Science Beijing Information Science and Technology University, 102206
Changping District, Beijing, China
wulang@bistu.edu.cn
19862140675@163.com

Abstract. Federated learning is a collaborative machine learning approach where multiple clients train a global model without sharing raw data. Federated learning has high application value in the fields of IoT, healthcare, and others due to its decentralized data processing and privacy protection features. Despite its advantages, the classic federated learning algorithm, Federated Averaging (FedAvg), faces some limitations that affect its optimization speed and compromise system security. This paper introduces FedCCSM, a federated learning framework designed to address class imbalance and malicious client behavior. Firstly, to accelerate model optimization, a client selection mechanism is introduced based on specific criteria, ensuring a high-quality data or powerful computational clients participate in the aggregation process. This speeds up optimization and improving overall efficiency. Secondly, the adoption of a committee mechanism involves selecting a client committee to screen the model before aggregation, enhancing system security. This committee serves as a precautionary measure to prevent malicious clients from conducting adversarial attacks by intentionally providing inaccurate updates or compromising the integrity of the global model integrity. By doing so, the security and reliability of the global model are ensured throughout the collaborative learning process. Thirdly, by simulating mechanisms for unbalanced clients, the algorithm's practical application effectiveness is strengthen. Experiments on MNIST and CIFAR-10 datasets demonstrate that FedCCSM improves accuracy on imbalanced datasets by 3% compared to FedAvg and reduces the influence of malicious clients by 5%. These results highlight the potential of FedCCSM in enhancing federated learning robustness and fairness in security-sensitive applications.

Keywords: Federal Learning, Committee, Imbalance DateSet, Transcendence Coefficient, Reliability Value Criterion.

1. Introduction

As machine learning and artificial intelligence continue to shape industries worldwide, the evolution towards intelligence is evident. Traditional centralized machine learning algorithms[1][27] necessitate users to upload their local data to a central server in exchange for high-quality machine models. However, relinquishing control over data raises concerns regarding security and privacy, potentially violating user personal interests and information security. The General Data Protection Regulation (GDPR), implemented by the European Union in 2018, underscore the importance of standardized information technology practices and the establishment of a secure cyberspace environment to cornerstone

big data development and implementation. In the machine learning domain, a distributed learning framework, not requiring users to disclose private data, is targeted; yet, it can still achieve model training, known as Federated Learning (FL)[28].

FL is a decentralized machine learning approach. In Fig1, a single round of FL mainly follows the following four steps:

- First, the central server initializes the global model and sends it to the participants;
- Second, the participants train their local models using their local data, producing local model parameter updates;
- Third, the participants send the parameter updates to the central server;
- Fourth, the central server aggregates the parameter updates from all participants to generate the new parameters of the global model, and sends them back to participants.

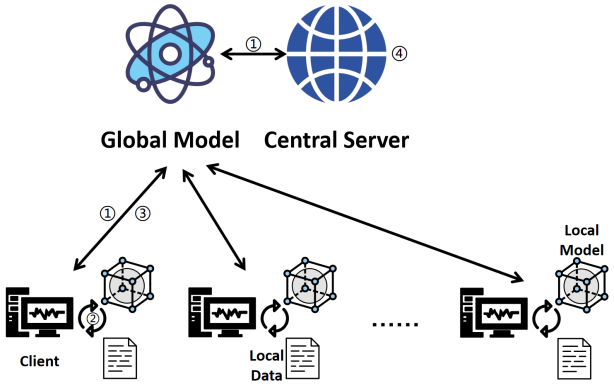


Fig. 1. Federal Learning Workflow Diagram

This process iterates continuously, allowing the global model to be optimized and enhanced without exposing individual data.

Federated learning, as a decentralized machine learning approach, offers the advantage of enabling model training across devices and organizations without centralizing data, thus safeguarding data privacy. This decentralization approach also mitigates data transfer and communication costs. Additionally, federated learning integrates data from diverse sources, enhancing the model's generalization capabilities, and facilitates local model updates, thereby improving training efficiency. Compared to centralized model training, federated learning better accommodates dispersed data sources and privacy protection requirements in real-world scenarios, presenting wide-range application.

As of now, there have been significant advancements and innovations in federated learning. Federated Averaging Algorithm (FedAvg), proposed by McMahan, is a practical method for joint learning of deep networks through iterative model averaging[28]. Zhou introduced FedGAM, an innovative federated learning algorithm designed to address client drift issues. It does so by introducing gradient norm perception minimization to achieve a locally flat loss function shape and utilizes control variables to correct local

updates, effectively solving the global flatness problem[37]. FedProx, as proposed by Li et al., addresses heterogeneity issues in federated networks by serving as a generalization and re-parameterization of FedAvg[24]. Moreover, WFB is a watermarking-based copyright protection framework for federated learning models that leverages blockchain technology to ensure model ownership and prevent unauthorized usage [33]. In addition, FedBN, introduced by Li et al, employs local batch normalization to alleviate feature shift before model averaging, thereby accelerating convergence compared to FedAvg[25]. FedFast, proposed by Muhammad et al, aims to accelerate distributed learning, achieving good accuracy for all users early in the training process and benefiting from reduced communication costs and improved model accurate[29]. Zhou et al proposed a new hierarchical FL framework RoPPFL, a robust aggregation edge FL framework tailored for computing applications. It supports privacy-preserving hierarchical FL and is resistant to poisoning attacks[40]. Moreover, Wei et al introduced a zeroth-order stochastic FL method based on Nesterov's zeroth-order (gradient-free) technique, considering both constant and diminishing step size strategies[36]. In addition, Pedrycz et al advocated for expressing Machine Learning (ML) construction results' credibility in terms of information granularity, extending the scope of FL evaluations[31]. Nergiz et al converted several classic DL methods, including the Big Transfer model, into federated versions[30]. Du et al proposed a Network Intrusion Detection algorithm (NIDS-FLGDP) based on Gaussian differential privacy federated learning[10]. Added to that, Zhang et al developed a platform architecture for a blockchain-based Industrial Internet of Things (IIoT) fault detection FL system, along with a novel Centroid Distance Weighted FedAvg (CDW_FedAvg) algorithm[9]. These advancements enhance federated learning 's efficacy and applicability.

Currently, the primary obstacle hindering the practical deployment of FL systems are their susceptibility to attacks from malicious clients[15] and impacts by imbalance data. In such systems, the central server lacks control over clients' behavior and access to their private data. Therefore, malicious clients can deceive the server by sending modified and harmful model updates, launching adversarial attacks on the global model[2]. Two types of adversarial attacks are prevalent: non-targeted attacks [6] and targeted attacks. The former aims to degrade overall model performance, causing the model to produce incorrect predictions without specifying a specific target category. This type of attack is considered as a Byzantine attack, leading to deteriorating model performance or training failure[22]. As for targeted attacks[9][6][39], they are specific, aiming to modify the model's behavior on particular data instances chosen by the attacker, such as misclassifying an image of a cat as a dog while keeping the model's performance unaffected on other data instances. Therefore, this attack requires defining a specific target category to misclassify the input as the specified target category. Both types of attacks can have catastrophic consequences, underscoring the importance of promptly detecting malicious attackers and remove their models from the FL algorithm. Defense against Byzantine attacks has been extensively researched in distributed ML. For instance, Chen proposed a variant of the classic gradient descent method based on geometric median averaging of gradients. Firstly, the parameter server groups the received gradients into non-overlapping batches to increase the similarity of non-Byzantine batches and then applied the median of batch gradients to mitigate the impact of Byzantine machines[8]. Moreover, Blanchard et al. introduced Krum, formulating the tolerance properties of aggregation rules; it is the first provably Byzantine-fault-tolerant distributed SGD algorithm[4]. In addition, Fung et

al. describes a novel defense method called FoolsGold, which is used to identify poisoning sybils based on the diversity of client updates in the distributed learning process. This system does not limit the expected number of attackers, requires no auxiliary information outside of the learning process, and makes fewer assumptions about clients and their data [11]. Furthermore, Han proposed three new robust aggregation rules for distributed synchronous Stochastic Gradient Descent (SGD) under a general Byzantine failure model, where attackers can randomly manipulate the data transferred between the servers and the workers within the Parameter Server (PS) architecture[13]. Finally, Yin developed a median-based distributed learning algorithm, achieving optimal statistical performance, better communication efficiency, and provable robustness while requiring just one communication round[38]. Obviously, the impact of malicious attacks is severe. The above is basically based on the improvement of the algorithm, yet it cannot completely block malicious attacks.

In previous research on federated learning, there are limitations and challenges in addressing class imbalance and client data integrity. Some studies focus on addressing the challenges posed by class imbalance but fall short in considering client data integrity and privacy protection. Other studies concentrate on client data integrity but often overlook the impact of class imbalance on model performance. Therefore, there are still gaps and deficiencies in research on addressing class imbalance and client data integrity in federated learning. In this context, integrating defense against malicious attack and addressing data imbalance in FL is a logical step forward. To this end, the proposal of Federated Learning with Committee Consensus and Selection Mechanism (FedCCSM) is significant[19]. The inclusion of a committee mechanism offers an effective means to detect malicious clients, thereby bolstering the security and resilience of the FL system. By implementing the committee mechanism, decisions regarding the acceptance of model updates from a participant can be made through methods like voting, effectively preventing malicious clients from impacting the system. According to the data characteristics of the client data set, the weight is weighted to enhance the anti-unbalance performance of the model. Moreover, the committee mechanism can incorporate security checks and validation mechanisms, such as data authenticity verification and model parameter legitimacy, further fortifying the system's security. Therefore, FedCCSM stands poised to effectively identify and counteract malicious client behavior while protecting participant data privacy, thereby enhancing the overall security and trustworthiness of the FL system. This approach holds significant application value across various fields such as healthcare, finance, smartphones, and IoT devices, paving the way for widespread adoption of FL in real-world scenarios.

Through studying the above questions, we will introduce new mechanisms into federated learning for improvement. As a result, the major contributions of this work can be summarized as follows:

- Improved global model training speed is achieved through the introduction of a boxplot coefficient screening mechanism. This method involves selecting high-accuracy client models for aggregation into the global model in the subsequent round of FL. By excluding relatively low-accuracy models, this approach accelerates the convergence speed of the global model;
- Incorporation of a committee consensus mechanism to detect and exclude malicious client models. Committee members assess and validate the model parameters sub-

mitted by clients to ensure they are non-malicious and capable of enhancing test accuracy. Only models meeting the criteria are allowed to participate in model aggregation, effectively screening out malicious clients;
- Implementation of a committee member election mechanism aimed at ensuring the integrity of elected committee members. Through predefined election criteria, clients failing to meet the specified standard are identified as malicious and barred from participating in client elections. This empowers the committee to effectively distinguish and exclude malicious models from participation;
- Development of a simulation for imbalanced datasets to emulate real-world client behavior. Since real client datasets frequently exhibit imbalances, which can impede training progress and undermine model efficacy, our simulation employs imbalanced datasets to assess the performance of federated learning algorithms accurately.

2. Related Work

2.1. Committee Mechanism

The committee mechanism[30] refers to a method of integrating and coordinating multiple independent ML models to improve overall predictive performance. In the committee mechanism, each model is trained independently, and the final prediction is derived from the combined voting or weighted average of all models. This integration method helps overcome the limitations of individual models, thereby enhancing prediction accuracy and robustness. The committee mechanism finds widespread application across diverse fields including financial risk assessment, medical diagnosis, natural language processing. Its advantages lie in its capability to leverage the strengths of multiple models, reduce overfitting risks, and improve generalization, demonstrating a strong adaptability in handling complex and high-dimensional data.

Therefore, the committee mechanism plays a pivotal role in enhancing the performance of ML models across various application scenarios, especially in FL[7]. Referring to Algorithm 1, in a FL training session, the committee mechanism orchestrates the collaborative process. Initially, the steps involve initializing the global model, followed by participants downloading the global model and training their local models using their local data to generate parameter updates. Subsequently, these updates are then transmitted to the committee mechanism, which aggregates them to derive new parameters for the global model. Finally, the committee mechanism distributes the aggregated global model parameters to participants for updating their local models. This iterative process facilitates FL, allowing for the optimization and enhancement of the global model while safeguarding data privacy.

2.2. Consensus Mechanism in Blockchain

The consensus mechanism[20], originated from the Byzantine Generals' Problem, describes a trust and consistency problem in a distributed system. This problem involves ten small countries surrounding a large country, with at least more than half of the small countries requiring to participate in the siege to achieve victory. However, if betrayal happens during the attack, the invaders may be annihilated. Therefore, each small country

Algorithm 1 Federated Learning Committee Mechanism

```
procedure (Global Model Initialization)
    global_model = initialize_model()
    for each round of training do
        for each participant do
            train_local_model = global_model
            local_model = train_local_model(local_data)
            local_parameters = get_parameters(local_model)
            send_parameters_to_committee(local_parameters)
        end for
        Committee Aggregation:
        global_parameters = aggregate_parameters_from_committee()
        global_model.update_parameters(global_parameters)
        Model Distribution:
        send_global_model_to_participants(global_model)
    end for
end procedure
```

does not trust the others. This example is similar to the need for nodes in a distributed system to reach a consensus regarding a decision. However, if there is a possibility of unfaithful behavior among the nodes in the system (i.e., betrayal), a consensus mechanism is required to ensure that a consensus decision can still be reached even in the presence of such unfaithful nodes.

Within the consensus mechanism, nodes can be divided into block-producing nodes, validating nodes, and accounting nodes (in Fig2). Moreover, the nodes responsible for proposing blocks are called block-producing nodes, also known as block producers, accountants, leaders, master nodes, or proposers. However, the nodes responsible for validating blocks are known as validating nodes, also called validators or backup nodes. Validating nodes must verify the legitimacy of the block producers and the blocks, as well as the correctness of the signatures. Finally, the nodes responsible for maintaining the blockchain database are called accounting nodes. Such nodes must store all blocks and verify them. Block-producing nodes, validating nodes, and accounting nodes are collectively referred to as consensus nodes. Therefore, the consensus mechanism main process includes electing block producers, proposing blocks, validating blocks, and updating the blockchain[26]. In each round, firstly a new block producer is elected. Then, the block producer proposes a block (packaging legitimate transactions from the network into a new block). Subsequently, validators verify the legitimacy of the new block. Finally, the accounting node prescribes the newly agreed block into the local database end to update the blockchain.

In the current research on federated learning, FedAvg, as a commonly used optimization method, is widely applied in the model aggregation process. However, FedAvg has certain shortcomings in terms of model security. Specifically, due to the use of a simple average aggregation method, FedAvg poses risks of privacy leakage and model tampering, which could potentially threaten the overall security of federated learning systems. To address the security deficiencies of FedAvg, the method proposed in this study combines committee mechanisms and consensus mechanisms. By introducing committee mecha-

nisms, each participant forms an independent committee during the model update process, where committee members supervise and verify each other, enhancing the reliability and security of model updates. Additionally, through the introduction of a consensus mechanism, participants must reach a consensus before submitting the model update results to the central server, ensuring the consistency and trustworthiness of model updates. This federated learning approach that combines committee and consensus mechanisms not only enhances model security but also effectively improves model performance and convergence speed, bringing new insights and opportunities for the development of federated learning systems.

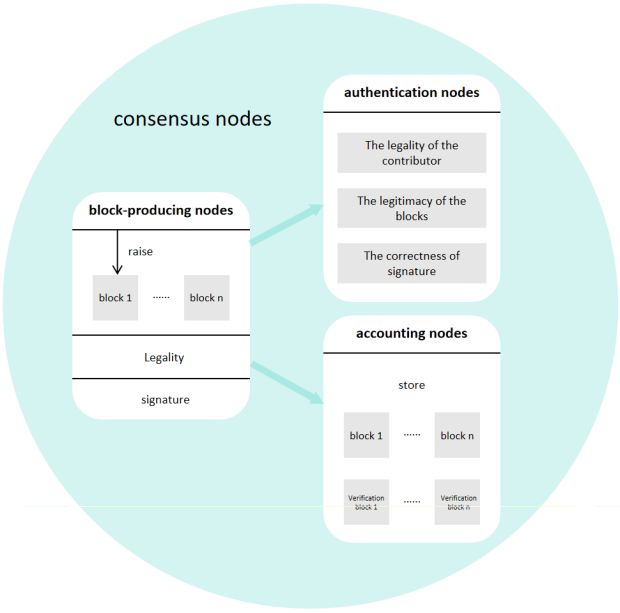


Fig. 2. Consensus Mechanism

2.3. Federal Learning and Imbalanced Dataset

Federated learning[28] is a decentralized machine learning approach that enables multiple edge devices or clients to collaboratively train a shared global model without exchanging raw data. Instead, model updates are computed locally on each device using its own data, and only the encrypted or aggregated updates are sent to a central server for aggregation. This privacy-preserving technique allows for efficient model training while protecting the privacy and security of sensitive data, making federated learning ideal for applications in healthcare, finance, and other industries where data privacy is a top priority.

Data imbalance is a common and real challenge in federated learning, like FedIBD[14]. Due to the potential increase in computational resource requirements and time costs associated with asynchronous learning, FedIBD may face challenges in performance stability

and model generalization when dealing with imbalanced data. Additionally, data privacy concerns and deployment complexities are limitations of FedIBD that require further research and improvement to enhance the system's reliability and scalability. To address the issue, we need to simulate real-world data. In the simulation test, it is a key to grasp the data imbalance. A class-imbalanced dataset[3] refers to a dataset where the number of samples in each class differs significantly, leading to a classification problem. In such datasets, the number of samples in certain classes may be much larger than in others, resulting in an uneven data distribution. Based on the total sample size, representing the classification standard, class-imbalanced datasets are divided into globally-balanced locally-imbalanced and globally-imbalanced locally-imbalanced datasets. Referring to Fig3(b), the total data quantity for each client is balanced. However, data distribution for different clients highlights a locally imbalanced state, representing the globally-balanced locally-imbalanced type. Referring to Fig3(c), the total quantity of data for each client is imbalanced, and their distribution across the different data categories is also imbalanced. Even in cases where the overall data is of a minority class, it may be the majority class locally; for instance, data k has the most data on client A, but is in the middle overall, while data j is in the majority class overall, but does not appear for client A.

The data class imbalance phenomenon is common in several real-world applications, such as rare diseases in medical diagnosis and in financial fraud detection. Moreover, the existence of class-imbalanced datasets can affect the training and performance of ML models, as they tend to predict the classes with more samples, while neglecting those with fewer samples. Therefore, the unbalance of data sets should be emphasized in federated learning.

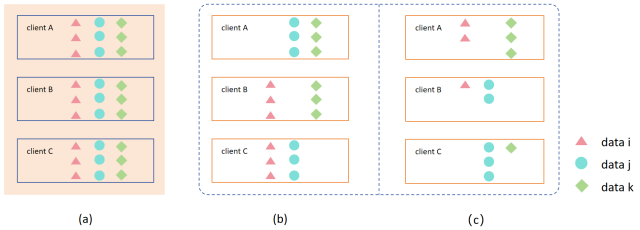


Fig. 3. Balanced and Imbalanced datasets

2.4. Malicious Client Detection

Malicious Client Detection[5][16] is a critical issue in Federated Learning, aiming to ensure that the global model is not compromised by malicious clients in a collaborative multi-party environment. Several studies have focused on proposing robust detection and defense mechanisms to address the potential impact of malicious clients on model training. Li et al. provided a review of the challenges and future directions of Federated Learning, discussing how to address the interference from malicious clients[23]. Chen and Zhou proposed a robust Federated Learning algorithm that identifies and mitigates the impact of malicious clients to improve model performance[22]. Konečný et al. proposed a

Federated Learning framework for distributed optimization, pointing out the impact of malicious clients on the training process[18]. Zhu and Han provided a detailed review of attacks and defenses in Federated Learning, proposing defense strategies based on attack types[34]. Xie et al. proposed a method to enhance the robustness of Federated Learning through malicious client detection and defense mechanisms[21]. Overall, although many existing methods provide certain robustness in theory and can handle some malicious client behaviors, their detection accuracy in practical applications still has room for improvement. These issues significantly limit the applicability of existing methods in real-world scenarios, especially in tasks with high requirements for accuracy.

3. Federated Learning Based on Committee Consensus and Selection Mechanism

The classic FedAvg learning algorithm is susceptible to contamination by malicious clients [17] due to its indiscriminate aggregation of client models. Moreover, the average aggregation method can hinder overall model training speed. FedCCSM addresses these issues by integration a committee mechanism to select and evaluate clients. In such a defensive algorithm, two criteria are employed to select committee members from the client pool, asked with filtering well-trained client models. Committee members must possess the capability to score models, enabling them to control the client models participating in the aggregation rather than aggregating all models blindly. By ensuring the integrity of committee members, the probability of malicious clients disrupting the global model training process is reduced significantly. To guarantee the honesty of committee members and facilitate secure aggregation, a new committee mechanism has been devised, encompassing a scoring system, selection strategy, and election strategy. Meanwhile, to simulate the actual client dataset distribution, random sharding is utilized to simulate an imbalanced dataset, achieving the most realistic effect.

Therefore, this section will provide a detailed introduction to the proposed framework and mechanism. In this case, we assume there are C clients forming a client group $\{C_i\}_{i=1}^{C}$, and the dataset for each client is denoted as M_{C_i}. The meanings of each variable notation are in Table 1.

3.1. Allocate Client Datasets

Referring to Fig4, step I refers to allocating the training dataset and testing the dataset for each client. For a dataset M containing m samples with a total of k classes, each class containing N samples; therefore, the dataset can be represented as $M = \{N_i\}_{i=1}^{K}$. To construct an imbalanced dataset, it is required to generate first the imbalanced parameters. Typically, for a collection of data containingk classes, the number of data classes in an imbalanced dataset is randomly selected between 1 and k. Therefore, a random array S containing C elements is generated, where each element is a random number between 1 and k. This array serves as the random shard array, where each element denotes the number of shards for the corresponding client.

To create a globally balanced and locally imbalanced train dataset, the data volume for each client has been determined during the simulation. To maximize the use of dataset M, the data volume for each client is set as m/C. Therefore, each shard size for the i^{th}

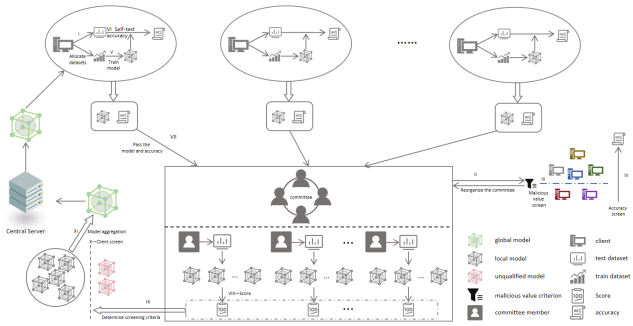


Fig. 4. FedCCSM Framework Diagram

client is m/CS_i. As a result, the entire dataset can be divided into CS_i shards. By using the np.random.permutation function to randomly permute the indices of all shards, client N_i takes the first S_i indices corresponding to the shards, ensuring data randomness. This approach leads to a more realistic simulation for data situation for several clients.

For a globally unbalanced training dataset, the construction process only requires a modification to one step of the process used for creating a globally balanced and locally imbalanced dataset. Specifically, the shard size for the entire dataset is not determined by each client but is instead uniformly defined. We denote this uniform shard size as s, which must fall between 1 and K (here, s can also serve as an indicator of the imbalance rate). Considering the scenario with the maximum data volume, the maximum data volume for each client remains m/C. Hence, the size of each shard becomes m/Cs. The subsequent steps, where each client acquires a certain number of shards, are still dictated by the array S, implying that client N_i's dataset comprises S_i different types of data, totaling mS_i/sC of data.

The algorithm also searches for malicious clients; thus, in addition to the imbalanced dataset, it is required to simulate a portion of malicious clients. To construct malicious clients, an additional step is added to the algorithm after creating the imbalanced dataset, where a portion of the information in each shard is modified to become incorrect. This study uses the Modified National Institute of Standards and Technology (MNIST) handwritten digit dataset as the basis for detection. For example, the handwritten digit data corresponding to a number x is changed to correspond to $9-x$; thus, the handwritten digit for 6 becomes 3. A client with such erroneous training data is considered a malicious client. The trained models by these clients will have significant different performances compared to those trained by normal clients, leading to the evaluation of the exact model effectiveness.

When simulating the testing dataset, it is crucial to ensure an adequate representation of each class of data. Therefore random shuffling of shards is not permissible. Instead, a portion of data from each class is sequentially selected to form the testing dataset. This step concludes the simulation of the clients.

3.2. Electing the Committee for Generation

Next, step II "reorganize the committee" is proposed. The details shown in Fig5, it mainly consists of selecting clients to form a committee G, consisting of a central server G_0 as fixed member and a number of members, where the number of members set as g. As the objective consists of detecting clients throughout the committee, the committee members should meet certain criteria. Step III involves determining the reliability of clients, while step IV focuses on sorting and selection.

This algorithm defines the reliability scoring criterion e to filter unreliable clients. Specifically, after the federated learning training is conducted by each client, each client tests and obtains a local prediction accuracy rate. During the initial selection process, the local model and the locally measured accuracy rate are transferred to the committee. The fixed member detects each client's local model to obtain the prediction accuracy rate. The absolute value of the difference between the local accuracy rate of each local model and the accuracy rate measured by the fixed member is taken as the reliability of this round. According to the magnitude of the reliability from low to high, g temporary members are selected. In subsequent rounds, the g temporary members selected in the previous round are responsible for testing the local models of all other clients in this round. Thus, each local model acquires g prediction accuracy rates. The absolute values of the differences between each of these accuracy rates and the local accuracy rate of the model are taken and then averaged to obtain the reliability f_i. According to the magnitude of the reliability from low to high, g temporary members are selected for the next round. The g temporary members selected in each round, along with the fixed member, form a $g+1$ person federated learning committee, which determines the selection criteria for the local models of each client participating in the global model aggregation.

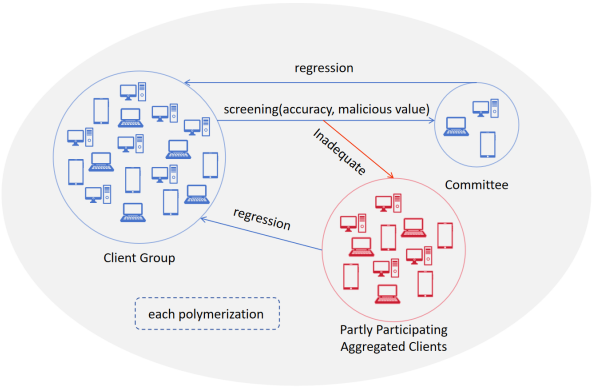


Fig. 5. The Interaction Between Clients and Committee

The remaining clients that do not meet the criteria participate in the subsequent local training steps.

Algorithm 2 Electing the Committee for Generation

Server executes:
Initialize G^0
for each round i = 0, 1, 2,... **do**
  if i mod v = 0 **then**
    for each round j = 0, 1, 2,... **do**
      $C_j \rightarrow f_j$
      if $f_j < e$ **then**
        $C_j \rightarrow G^i$
      end if
    end for
  end if
  $G_i \rightarrow C$
end for

3.3. Model Training

The focus is on non-convex neural network objectives. For machine learning problems, the typical formulation is $f_i(w) = l(x_i, y_i; \omega)$, representing the loss incurred by using model parameter ω to predict on example (x_i, y_i). It is assumed that there are K clients participating in this round of training, where P_i is the set of indices of data points on client C_i, and $n_i = |P_i|$. Therefore, the algorithm considered applies to the following form of finite-sum objective[22]:

$$\min_{\omega \in R^d} \mathrm{f}(\omega) \quad where \quad \mathrm{f}(\omega) = \sum_{i=1}^{K} \frac{n_i}{n} \mathrm{F_i}(\omega) \quad where \quad \mathrm{F_i}(\omega) = \frac{1}{n_i} \sum_{j \in P_j} \mathrm{f_j}(\omega). \tag{1}$$

Consequently, we proceed to step V, involving model training. For each client C_i, the parameters of the global model $\omega^{(l)}$(the l^{th} round) are first loaded into the client's model. Then, the client's data loader is created based on the client's own training dataset where each client undergoes d rounds of local training. In each round, a pair of data and labels is retrieved from the training set and iterated upon. Firstly, data and labels are moved to the device for computation; in such case, the CPU is deployed. The data is then fed into the neural network.

Secondly, Refer to Fig6, a Convolutional Neural Network (CNN) is established with two 5*5 convolutional layers[32], with 32 and 64 channels, respectively. Each layer is followed by a 2*2 max-pooling layer. This is connected to a fully connected layer with 512 units and ReLU activation, and a final softmax output layer (totaling 1,663,370 parameters).

Next, the reshaping of the input data into a tensor of shape (-1, 1, 28, 28) is performed. The process expression for the first convolutional layer is as follows:

$$h^{(l)} = \sigma(\omega^{(l)} * h^{(l-1)} + b^{(l)}). \tag{2}$$

Then, applying the ReLU the activation function $f(x) = max(0, x)$ is performed. Following this process, a pooling layer is applied, followed by another convolutional layer, an activation function, and another pooling layer, resulting in the tensor being flattened

into a one-dimensional shape having a size of (-1, 7*7*64). Consequently, the flattened tensor is passed through the first fully connected layer, followed by another ReLU activation function. Then, the result is passed through the second fully connected layer to generate the final output. As a result, the output consists the predicted result.

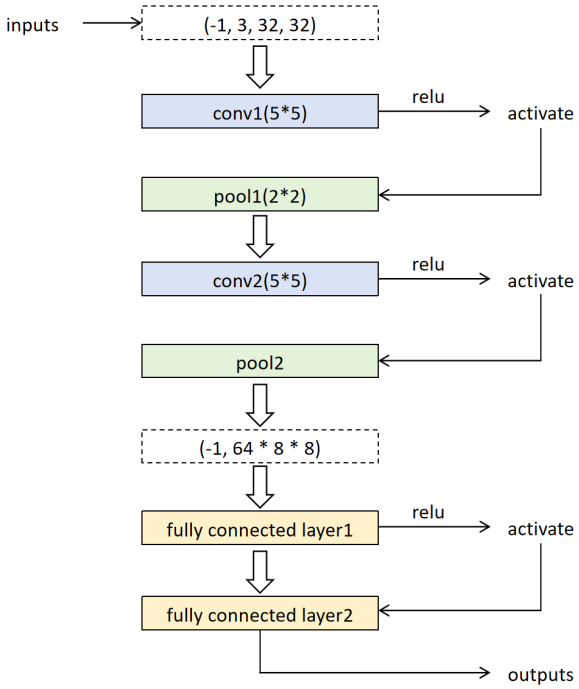


Fig. 6. The steps of Convolutional Neural Network

Thirdly, the loss function is calculated between the predicted results and the true values using the cross-entropy loss function. As the MNIST dataset essentially represents a ten-classification problem, with labels being arrays of length ten where the i^{th} element is equal to the unit and the rest are null, we use the cross-entropy loss function to calculate the loss. If i represents the sample, y denotes the actual label, a indicates the predicted output, and n highlights the total number of samples; then, the loss value is written as follows:

$$loss = -\frac{1}{n}\sum_{i} y_i \ln p_i \,. \tag{3}$$

Fourthly, the backpropagation is performed to calculate the gradients. Fifthly, the Stochastic Gradient Descent (SGD) algorithm is utilized as the optimizer to update the model's parameters. For each i, the algorithm expression is defined as follows:

$$\theta_j := \theta_j + \alpha(y^{(i)} - \mathrm{h}_\theta(x^{(i)}))x_j^{(i)} \quad where \quad \mathrm{h}_\theta(x) = \theta_0 + \sum_{i=1}^{n} \theta_i x_i \,. \tag{4}$$

Finally, after the local training is completed, the client model is updated with the latest model, and the accuracy of the model is measured with the local test data set in Step VI.

3.4. Model Selection and Aggregation

After each client C_i completes his local training in this round, he obtains the latest model ω_i^l and evaluates its accuracy on the local test set, denoted as a_i.

Step VII involves transmitting the trained local updated model parameters and accuracy to the committee. Each committee member G_i and central server G_0 receives all updated model parameters and their accuracies to prepare for the next step.

Step VIII contains the committee members scoring for each client. The members of the committee test the local models of the other clients based on their local test datasets. Each local model of the client will obtain g test values.

In step IX, by differentiating these test values with the accuracy of the client's own measurement and averaging the absolute value of the difference, the reliability e_j of the client's local model can be determined. Each model $\omega_j^{(l)}$ corresponds to a score set include reliability and accuracy. To prevent unreliable clients from infiltrating the committee and maliciously changing the scores given to the models, all values are analyzed in a unified centralized analysis to determine the reliability criteria e. (described in section 4.1).

Then, the accuracy rates of the selected local models of the clients are sorted, and the upper and lower quartiles and the interquartile range are calculated to determine the accuracy standard line k, as follows:

$$k = q_1 - t * iqr \,. \tag{5}$$

where t is a tunable multiple of iqr, known as exceedance coefficient.

Step X is represented in Figure 3. Exclude clients with accuracy below k and reliability above e[12]. The models' parameters that meet both criteria are aggregated, corresponding to step XI.

$$\theta_i^{(l+1)} = \frac{1}{n+1} \sum_{i=0}^{n} \theta_i^{(l)} \,. \tag{6}$$

Finally, in step XII, the aggregated global model $w^{(l+1)}$ is transmitted to the Central Server. In this way, a complete round of federated learning process is achieved.

4. Experimental Results

The experimental testing is divided into two phases based on the imbalanced dataset, namely, all normal simulation and the introduction of malicious clients. Before starting the experiment, it is important to determine the key coefficients.

Table 1. Notation Explanation

Notation	Description
C_i	Client with the index i
M_{C_i}	Dataset of client i
N_i	Data of the i^{th} category
Dr	Train dataSet
De	Test dataSet
S	Randomly partitioned array
S_i	The i^{th} shard number
s	Uniform shard number(also an indicator for the imbalance rate)
K	The number of clients participating in each round of training
G_i	The i^{th} committee member
G_0	fixed central server
H_i	The i^{th} candidate member
P_k	The set of indices of data on client k
t	Exceedance coefficient
ω	General model parameters
$\omega^{(l)}$	The l^{th} round global model
$W^{(l)}$	filter
$\omega_i^{(l)}$	The local updated model of the i^{th} client in the l^{th} round
a_i	The accuracy of client i's self-assessment
d	epoches
e	malicious standard
a_i^j	The accuracy of client i in detecting $\omega_i^{(l)}$
f_j	The maliciousness of client j
θ_j	The parameters
k	Accuracy standard
t	Threshold multiplier
iqr	Interquartile range
$q_{1/3}$	Lower/upper quartile
p_i	predicted output
m	maliciousness criterion
u	committee number
v	model validation frequency of communications

Algorithm 3 Federated Learning Based on Committee Consensus and Selection Mechanism

Server executes:
Initialize ω^0
for each round i = 1, 2,... **do**
    W $\leftarrow C_i(e_i > e)$
    $C_F \leftarrow$ (random set of m clients)
    for each client j $\in C_F$ **do**
        $w_j^{(l)}$ = ClientLocalUpdate(j, $\omega_j^{(l-1)}$)
        $a_j = TestClientAccuracy(C_De, \omega_j^{(l)})$
    end for
    for each Committee member in W_i **do**
        $a_i^j = TestClientAccuracy(\omega_j^{(l)})$
    end for
    $f_j = \sum_{k=1}^{w} a_i^j$
    q1, q3, iqr $\leftarrow$ boxplots(a_i)
    k = q1 - t * iqr
    $\omega^{(l+1)} \leftarrow Aggregation(k, e, \omega_{C_F}^l)$
end for

ClientLocalUpdate: // Run on client k
$\beta \leftarrow$ (split M_k into batches of size B)
for each local epoch i from 1 to e **do**
    for each batch b $\in \beta$ **do**
        $\omega_j^{(l+1)} \leftarrow \omega_j^{(l)}$ - $\eta(\omega_j^{(l)}$; b)
    end for
end for
return $w_j^{(l+1)}$ to server.

4.1. Coefficient Adjustment and Determination

This algorithm involves two parameters: the transcendence coefficient t and the reliability value criterion, both playing a crucial role in the client screening step. The transcendence coefficient t determines the standard line for screening accuracy, i.e., the lower accuracy bound. Experiments are conducted with t=0.3,0.4,0.5,0.6 and 0.7 as the imbalance rates to generate the accuracy of the global model while keeping other parameters constant.

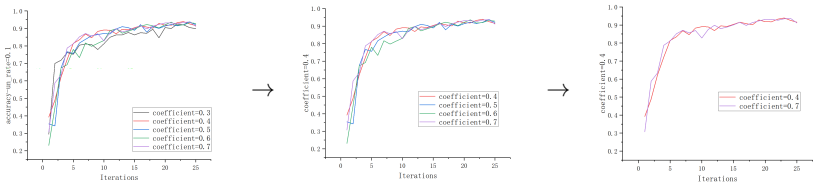
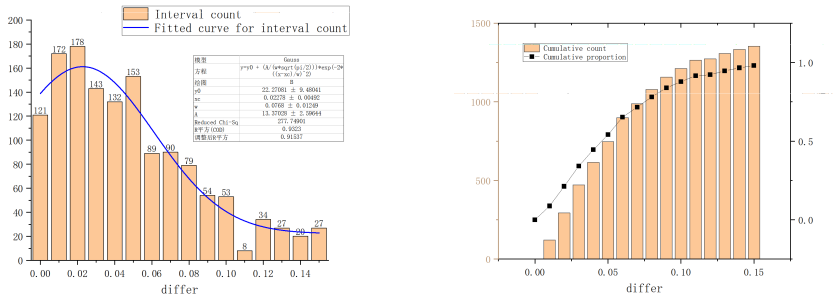


Fig. 7. Experimental Results with Different Transcendence Coefficient

Referring to Fig7, the algorithm performs best when t=0.4 referring to the speed of accuracy improvement and the stability during the improvement process. The reliability value criterion also applies experimental data as a reference. A large dataset is constructed by scoring clients with no malice where the normal range and values of scores are analyzed. This analysis helps determining the standard for reliability values. During the experiment, approximately 600 scores are collected from 30 rounds for two randomly selecting clients. Based on Fig8, 91.59% of clients have reliability scores less than 0.11. Therefore, the standard for screening unreliability clients in the model is set accordingly.



(a) Distribution of Differences **(b)** cumulative difference statistics

Fig. 8. Client Malice Score Distribution Chart

4.2. Global Imbalance

Our goal is to develop outstanding models to enhance the performance of devices in image classification and language modeling tasks, thereby improving user experience and device usability.

In the experiment of global balance and local imbalance, MNIST and CIFAR-10 datasets are deployed as the experimental datasets to test the effectiveness of the algorithm.

In the experiments concerning local imbalance in the dataset, a series of experiments were undertaken to investigate the model's ability to manage imbalanced classes. In this section, the dataset was partitioned into multiple subsets according to class combinations. Models were trained and assessed for each subset to analyze their performance in the context of local imbalance. Specifically, the study juxtaposed the accuracy, precision, and other metrics of the models across various subsets, along with their efficacy in identifying imbalanced classes.

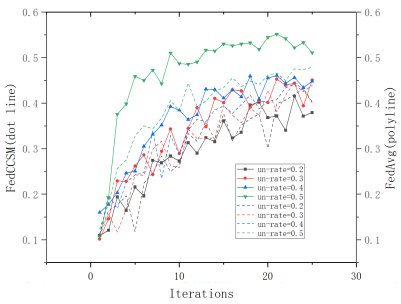
We conducted a significance t-test[35] on the accuracy of multiple experiments under the same parameters to ensure the stability of the results.

Ensuring all parameters are consistent, we obtained ten accuracy results from ten experiments: 0.9779, 0.9794, 0.9805, 0.9796, 0.9803, 0.9593, 0.9736, 0.9728, 0.9735, 0.9682. The null hypothesis (H_0) is "These values fluctuate around a mean with a small amplitude, and the average remains stable around 0.9751." The alternative hypothesis (H_1) is "These values fluctuate around a mean with a large amplitude, and the average is not stable around 0.9751." Firstly, the calculated average of these values is 0.9751. Secondly, the sample standard deviation is calculated to be 0.0065. Thirdly, the t-value is computed as

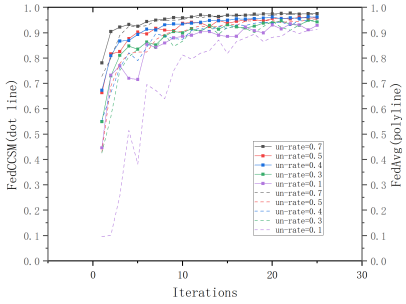
$$t = \frac{0.9751 - 0.9751}{0.0065/\sqrt{10}} = 0. \tag{7}$$

In the fourth step, the critical value for a t-distribution with 9 degrees of freedom is determined by referring to the t-distribution table. For a two-tailed test, the critical value is ±2.262. Finally, based on the calculated t-value of 0, which is less than the critical value of 2.262, we fail to reject the null hypothesis. Therefore, we accept the null hypothesis, indicating that the stability of these values around the mean of 0.9751 is good, with a small fluctuation amplitude.

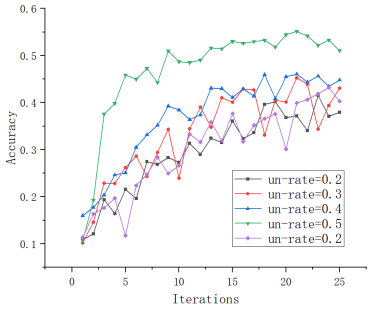
In the experimental analysis, a noticeable trend emerged wherein the improved FedCCSM model showed a swifter improvement in global model accuracy in comparison to FedAvg. Fig9(a) and (b) illustrates the phenomenon, showcasing an increase in training rounds and the accuracy of the improved model speed, while FedAvg displayed a comparatively slower progress. It is evident that on the Cifar-10 dataset, especially when the un-rate is 0.5, FedCCSM shows an overall accuracy improvement of around 5% compared to FedAvg throughout the entire federated learning process. Additionally, there are improvements of 1-2% on other un-rates as well. Furthermore, the improved model surpassed FedAvg in overall accuracy, reflecting superior precision. Additionally, the smoother curve of the improved model, characterized by minimal fluctuations, indicates a more stable response to changes in the dataset. These findings underscore the superior performance and enhanced stability of the improved model in managing imbalanced datasets, providing



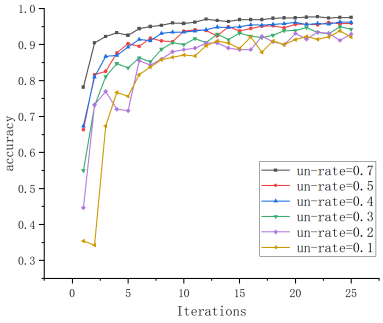
(a) FedCCSM and FedAvg in Cifar-10



(b) FedCCSM and FedAvg in Mnist



(c) FedCCSM along in Cifar-10



(d) FedCCSM along in Mnist

Fig. 9. Differ Global Imbalance Accuracy Results between FedAvg and FedCCSM in Mnist and Cifar-10

more reliable results for practical applications. Currently, the experimental results support the superior performance of the improved model over FedAvg.

Moreover, Fig9(c) and (d) corroborate the aforementioned observations, depicting a rapid improvement in accuracy with an increase in training rounds for the improved model. Despite the varying resistance encountered during model training due to different imbalance rates, as the imbalance rate increases, a trend of increased rounds is applied to reach the peak; yet, good results can be achieved.

Table 2. Comparison of methods in different hyperparameters on MNIST

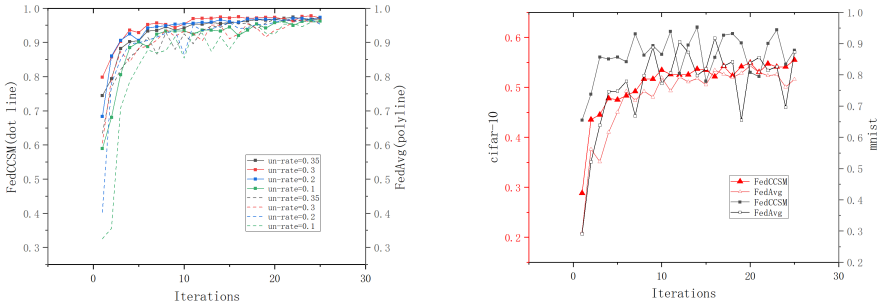
Hyperparameters value		**FedAvg**	**FedCCSM**
batchsize	$b = 8$	0.9463	0.9363
	$b = 10$ (used)	**0.9593**	**0.9499**
	$b = 15$	0.9298	0.9041
	$b = 20$	0.9090	0.8975
learning_rate	$lr = 0.001$	0.8297	0.7887
	$lr = 0.005$	0.9218	0.9052
	$lr = 0.01$ (used)	**0.9593**	**0.9499**
	$lr = 0.05$	0.9493	0.9408

Based on the experimental results and sensitivity analysis of hyperparameters, we found that in this study, the FedCCSM method demonstrates better robustness and stability. Particularly, under the conditions of batch size = 0.01 and learning rate = 10, its performance significantly outperforms the FedAvg method. Therefore, in this experiment, we used these parameters for overall experimentation and comparison. This indicates that the FedCCSM method exhibits better adaptability and stability in federated learning tasks, serving as an effective optimization method to enhance model performance and convergence speed. Hence, it is recommended to prioritize the use of the FedCCSM method in practical applications to achieve better results.

4.3. Global Balance

In the global balance experiment section, this study conducted unified training and evaluation on the entire dataset, incorporating malicious clients into the simulated client group. From the experimental data, it is evident that our proposed federated learning model, FedCCSM, demonstrates accelerated initial performance enhancement compared to the traditional FedAvg. Furthermore, in terms of the change in global model accuracy, the balance between local imbalance and global imbalance is closer to Independent and Identically Distributed (IID), as evidenced by the relatively higher accuracy.

According to Fig10(a), FedCCSM exhibits higher accuracy and a steeper slope, indicating faster convergence and superior performance during the early stages of model training. Additionally, as the training progresses, the performance improvement of FedCCSM gradually stabilizes and surpasses FedAvg, suggesting that FedCCSM can maintain stable performance in the later stages of training. Compared to FedAvg, FedCCSM demonstrates clear advantages in terms of convergence speed and performance stability.



(a) FedCCSM and FedAvg in Honest Global Balance

(b) FedCCSM and FedAvg in Malicious Device Detection

Fig. 10. Results about Global Balance in Honest and Malicious

The following experiment involved manipulating the training set based on imbalance, simulating malicious clients by introducing incorrect training sets. This was conducted to evaluate the algorithm's resistance to malicious behavior. When applied to MNIST and CIFAR-10, we intentionally disrupted the correspondence between the data and labels, resulting in mislabeled data. This adversely affected the overall training process.

The results in Fig10(b) demonstrate that the improved code outperforms FedAvg in overall accuracy, indicating greater resistance to malicious behavior. Furthermore, the curve of the enhanced code exhibits smoother trajectories with smaller fluctuations, suggesting that it can maintain accuracy more consistently when faced with malicious data. This provides more reliable results for practical applications.

Table 3. Comparison of methods on the MNIST dataset under global balance and malicious

Hyperparameters	**value**	**Methods**			
		FedAvg	**FedProx**	**MOON**	**FedCCSM(this paper)**
Reliable	un-rate=0.1	0.9599	0.9393	0.9526	0.9652
	un-rate=0.2	0.9533	0.9592	0.9621	0.9735
	un-rate=0.3	0.9601	0.9526	0.9592	0.9717
	un-rate=0.35	0.9719	0.9622	0.9637	0.9709
Malicious	Malic-ratio=0.2	0.8728	0.8841	0.8775	0.9160

The experimental results demonstrate that the FedCCSM model shows improvement in federated learning, with higher accuracy compared to traditional FedAvg, FedProx, and MOON models, regardless of the imbalance rate. In this experiment, the highest accuracy improvement was observed at an imbalance rate of 0.2, with an increase of 2.02%. When facing malicious clients with a simulated proportion of 0.2, FedCCSM showed improvements of 4.32%, 3.19%, and 3.85% compared to FedAvg, FedProx, and MOON, respec-

tively. Clearly, when dealing with a higher proportion of malicious devices, FedCCSM can better maintain model accuracy, demonstrating stronger robustness and generalization capabilities.

The aim of this study, as demonstrated through the above experiments, is to thoroughly analyze the performance of the model on malicious clients with imbalanced class datasets. Additionally, the study seeks to explore effective training strategies for addressing imbalanced class datasets and enhance the model's ability to resist malicious attacks. The results consistently demonstrate that FedCCSM has the capability to handle imbalanced datasets and resist malicious attacks.

4.4. All Balance

In order to test the algorithm's defense effectiveness against malicious attacks and compare its advantages and disadvantages with other similar algorithms, this part is based on the MNIST dataset, comparing the performance of FedCCSM with three methods, MUD-HoG, Foolgolds, and Fedavg, evaluating the performance indicators. Malicious attacks are categorized into non-targeted poisoning attacks[39] and targeted poisoning attacks, with targeted poisoning attacks including single-label flip attacks[8], multi-label flip attacks, and backdoor attacks[4][11]. We selected backdoor attacks as the type of attack to test the algorithm's performance. Backdoor attacks involve attackers implanting trigger patterns (backdoor triggers) in certain training/testing data to inject a backdoor, causing the model to make incorrect judgments on data with specific features. The experiment initializes the number of clients to 40, including 50% malicious clients. A federated learning model is trained over 200 communication rounds with 10 local training rounds.

Table 4. Comparison of methods on the MNIST dataset under iid distribution

Methods	Backdoor Accuracy	Model Testing Accuracy
Fedavg	99.3	97.9
Foolsgold	99.6	98.3
MUD-HoG	82.5	98.4
FedCCSM(this paper)	11.6	98.7

The experiment evaluated the effectiveness of defense methods from two aspects: model testing accuracy and backdoor accuracy. Model testing accuracy refers to the accuracy of the global model on the test set. Backdoor accuracy assesses adversarial backdoor training, measuring the number of injected samples classified as the attacker's target label. If a client's sample with a backdoor attack is predicted as the malicious target, it is considered successful in identifying and defending against the attack on that client. Depending on the exclusion of clients with backdoor attacks, a decrease in backdoor accuracy implies a reduction in the number of clients carrying backdoor attacks in the client group. Therefore, as the algorithm trains the model towards the later stages, a low value of backdoor accuracy indicates the exclusion of more backdoor attacks, suggesting that the algorithm provides the best defense against backdoor attacks.

Table 2 shows the performance results on the MNIST dataset. Compared to other defense methods, FedCCSM provides the best protection, being able to effectively eliminate

backdoor attacks to the greatest extent, demonstrating the strongest defense capability. Additionally, it improves model testing accuracy while minimizing the impact on model training.

In this study, we delve into the theoretical implications of federated learning and deep learning in the context of imbalanced datasets and privacy security. Our research findings reveal the performance differences of different algorithms in handling imbalanced datasets, providing theoretical guidance for adjusting algorithms to improve model accuracy and robustness. Additionally, our study explores the effectiveness and limitations of privacy protection technologies in federated learning and deep learning, laying a theoretical foundation for designing more secure federated learning frameworks. These discoveries offer important theoretical support for addressing challenges related to imbalanced datasets and privacy security, and provide valuable insights for future research and practical applications.

However, it is worth noting that this study also has some limitations that need further exploration and resolution. Firstly, our experimental datasets are limited to MNIST and Cifar-10, without studying more datasets. Additionally, our research is primarily focused on theoretical analysis and simulated experiments, and the practical application effects in real-world scenarios still need further validation. Therefore, ongoing attention and updates are necessary.

5. Conclusion

This paper introduces FedCCSM, a federated learning framework that addresses class imbalance and malicious client behavior through a committee-based consensus mechanism. Experiments on MNIST and CIFAR-10 datasets demonstrate significant improvements in accuracy and robustness compared to baseline methods, highlighting the effectiveness of FedCCSM. These findings contribute to advancing federated learning by improving fairness and reliability in distributed systems.

Future research should focus on exploring adaptive committee selection strategies, applying FedCCSM to diverse datasets, and assessing its scalability in real-world applications. Research can be conducted on how to make federated learning systems resilient to a wider range of malicious attacks, including data poisoning, model manipulation, and privacy breaches. This may involve the development of new security and privacy protection techniques to ensure the security and robustness of federated learning systems. Additionally, new evaluation metrics can be developed to assess performance on imbalanced datasets.

Acknowledgments. This work is supported by the Beijing Information Science and Technology University young backbone teacher program project (YBT 202447), the Future Blockchain and Privacy Computing Advanced Center Project Funding (NO.202203).

References

1. Athilakshmi, R., Jacob, S.G., Rajavel, R.: Automatic detection of biomarker genes through deep learning techniques: A research perspective. Stud. Informatics Control, Vol.32(No.2), 51–61. (2023)

2. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V.: How to backdoor federated learning. In: International conference on artificial intelligence and statistics. pp. 2938–2948. PMLR (2020)
3. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter 6(1), 20–29 (2004)
4. Blanchard, P., El Mhamdi, E.M., Guerraoui, R., Stainer, J.: Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in neural information processing systems 30 (2017)
5. Calik Bayazit, E., Koray Sahingoz, O., Dogan, B.: Deep learning based malware detection for android systems: A comparative analysis. Tehnički vjesnik 30(3), 787–796 (2023)
6. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
7. Che, C., Li, X., Chen, C., He, X., Zheng, Z.: A decentralized federated learning framework via committee mechanism with convergence guarantee. IEEE Transactions on Parallel and Distributed Systems 33(12), 4783–4800 (2022)
8. Chen, Y., Su, L., Xu, J.: Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. Proceedings of the ACM on Measurement and Analysis of Computing Systems 1(2), 1–25 (2017)
9. Dai, J., Chen, C., Li, Y.: A backdoor attack against lstm-based text classification systems. IEEE Access 7, 138872–138878 (2019)
10. Du, J., Yang, K.: Nids-flgdp: Network intrusion detection algorithm based on gaussian differential privacy federated learning. Journal of Circuits, Systems and Computers 33(03), 2450048 (2024)
11. Fung, C., Yoon, C.J., Beschastnikh, I.: Mitigating sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866 (2018)
12. Gupta, A., Luo, T., Ngo, M.V., Das, S.K.: Long-short history of gradients is all you need: Detecting malicious and unreliable clients in federated learning. In: European Symposium on Research in Computer Security. pp. 445–465. Springer (2022)
13. Han, Y., Zhang, X.: Robust federated training via collaborative machine teaching using trusted instances. arXiv preprint arXiv:1905.02941 (2019)
14. Hou, Y., Li, H., Guo, Z., Wu, W., Liu, R., You, L.: Fedibd: a federated learning framework in asynchronous mode for imbalanced data. Applied Intelligence 55(2), 1–17 (2025)
15. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al.: Advances and open problems in federated learning. Foundations and trends® in machine learning 14(1–2), 1–210 (2021)
16. Kavitha, S., Uma Maheswari, N., Venkatesh, R.: Intelligent intrusion detection system using enhanced arithmetic optimization algorithm with deep learning model. Tehnički vjesnik 30(4), 1217–1224 (2023)
17. Kolasa, D., Pilch, K., Mazurczyk, W.: Federated learning secure model: A framework for malicious clients detection. SoftwareX 27, 101765 (2024)
18. Konečnỳ, J., McMahan, B., Ramage, D.: Federated optimization: Distributed optimization beyond the datacenter. arXiv preprint arXiv:1511.03575 (2015)
19. Kumar, P., Kumar, R., Kumar, A., Franklin, A.A., Garg, S., Singh, S.: Blockchain and deep learning for secure communication in digital twin empowered industrial iot network. IEEE Transactions on Network Science and Engineering 10(5), 2802–2813 (2022)
20. Lashkari, B., Musilek, P.: A comprehensive review of blockchain consensus mechanisms. IEEE access 9, 43620–43652 (2021)
21. Lewis, C., Varadharajan, V., Noman, N.: Attacks against federated learning defense systems and their mitigation. Journal of Machine Learning Research 24(30), 1–50 (2023)
22. Li, S., Cheng, Y., Wang, W., Liu, Y., Chen, T.: Learning to detect malicious clients for robust federated learning. arXiv preprint arXiv:2002.00211 (2020)

23. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE signal processing magazine 37(3), 50–60 (2020)
24. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems 2, 429–450 (2020)
25. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv:2102.07623 (2021)
26. Li, Y., Chen, C., Liu, N., Huang, H., Zheng, Z., Yan, Q.: A blockchain-based decentralized federated learning framework with committee consensus. IEEE Network 35(1), 234–241 (2020)
27. MAIER, M.I., Czibula, G., DELEAN, L.R.: Using unsupervised learning for mining behavioural patterns from data. a case study for the baccalaureate exam in romania. Studies in Informatics and Control 32(2), 73–84 (2023)
28. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
29. Muhammad, K., Wang, Q., O'Reilly-Morgan, D., Tragos, E., Smyth, B., Hurley, N., Geraci, J., Lawlor, A.: Fedfast: Going beyond average for faster training of federated recommender systems. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1234–1242 (2020)
30. Nergiz, M.: Federated learning-based colorectal cancer classification by convolutional neural networks and general visual representation learning. International Journal of Imaging Systems and Technology 33(3), 951–964 (2023)
31. Pedrycz, W.: Advancing federated learning with granular computing. Fuzzy Information and Engineering 15(1), 1–13 (2023)
32. Qian, Y., Bi, M., Tan, T., Yu, K.: Very deep convolutional neural networks for noise robust speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24(12), 2263–2276 (2016)
33. Shao, S., Wang, Y., Yang, C., Liu, Y., Chen, X., Qi, F.: Wfb: watermarking-based copyright protection framework for federated learning model via blockchain. Scientific Reports 14(1), 19453 (2024)
34. Sikandar, H.S., Waheed, H., Tahir, S., Malik, S.U., Rafique, W.: A detailed survey on federated learning attacks and defenses. Electronics 12(2), 260 (2023)
35. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 623–632 (2007)
36. Wei, L.Y., Yu, Z., Zhou, D.X.: Federated learning for minimizing nonsmooth convex loss functions. Mathematical Foundations of Computing 6(4), 753–770 (2023)
37. Xu, Y., Ma, W., Dai, C., Wu, Y., Zhou, H.: Generalized federated learning via gradient norm-aware minimization and control variables. Mathematics 12(17), 2644 (2024)
38. Yin, D., Chen, Y., Kannan, R., Bartlett, P.: Byzantine-robust distributed learning: Towards optimal statistical rates. In: International conference on machine learning. pp. 5650–5659. Pmlr (2018)
39. Zhang, W., Lu, Q., Yu, Q., Li, Z., Liu, Y., Lo, S.K., Chen, S., Xu, X., Zhu, L.: Blockchain-based federated learning for device failure detection in industrial iot. IEEE Internet of Things Journal 8(7), 5926–5937 (2020)
40. Zhou, H., Zheng, Y., Jia, X.: Towards robust and privacy-preserving federated learning in edge computing. Computer Networks 243, 110321 (2024)

Lang Wu Ph.D., is a lecturer at Beijing Information Science and Technology University. She received her Doctor of Science degree from Harbin Institute of Technology in

2016. Her research focuses on the application of artificial intelligence in complex systems, with particular emphasis on meta-learning, federated learning, privacy-preserving computation, and personalized model optimization. She is dedicated to addressing uncertainty modeling and intelligent decision-making in high-dimensional data analysis. Dr. Wu built a solid foundation in mathematical theory through early work on high-precision numerical solutions of partial differential equations. Later, she shifted her focus to the deep integration of machine learning with real-world industrial applications, proposing various optimization methods that have significantly enhanced the practicality and robustness of models in critical fields such as finance and healthcare. In recent years, she has published more than ten papers in domestic and international academic journals, and some of her research outcomes have been successfully applied in real-world systems. She possesses strong research capabilities and a solid foundation in engineering practice.

Yi Dong is a master's student at Beijing Information Science and Technology University, primarily engaged in research on federated learning. Her work focuses on data privacy protection and distributed modeling techniques, aiming to enhance model generalization and robustness in multi-source heterogeneous data environments. Under the guidance of her supervisor, she actively participates in research projects, demonstrates strong programming skills and academic competence, and is exploring the application potential of federated learning in real-world scenarios such as finance and healthcare.

Received: November 19, 2024; Accepted: May 10, 2025.

Computer Science and Information Systems 22(4):1483–1508 https://doi.org/10.2298/CSIS240925058D

Data-Driven Traffic Management: Enhancing Road Safety through Integrated Digital Twin Technology ⋆

Miloš Durković[1], Petar Lukovac[1], Demir Hažić[2], Dušan Barać[1], and Zorica Bogdanović[1]

[1] University of Belgrade - Faculty of Organizational Sciences, Jove Ilića 154
11000 Belgrade, Serbia
mdurkovic127@yahoo.co.uk
[2] Petroleum Development Oman, Mina Al Fahal, Qurum
Muscat, Oman
demir.hadzic@gmail.com

Abstract. This paper proposes a data-driven approach to enhancing traffic safety through the integration of digital twins, in-vehicle monitoring system and machine learning. The main goal is to contribute to solving problems related to driver behavior, inadequate road signage infrastructure, and delayed maintenance by developing a digital twin model that leverages real-time data for predictive analysis, coaching, and maintenance. Using the Prophet algorithm, the model predicts compliance with traffic regulations, identifies frequent driver violations, and highlights deficiencies in road signage, enabling timely interventions. The innovation of this solution lies in its ability to synchronize real-time data from drivers, vehicles and road infrastructure and provide predictive insights, creating a scalable and adaptable framework for traffic management. The proposed model is tested in a proof-of-concept scenario, where it demonstrated significant improvements in road safety.

Keywords: smart mobility, digital twins, machine learning, IVMS, road safety.

1. Introduction

Road safety has been recognized as a critical component of UN Sustainable Development Goals (SDGs). SDG Goal 3 emphasizes ensuring healthy lives and promoting well-being for all, with Target 3.6 specifically aiming to halve the global number of deaths and injuries caused by road traffic accidents by 2030 [35]. Similarly, SDG Goal 11 focuses on making cities and human settlements inclusive, safe, resilient, and sustainable, with Target 11.2 promoting access to safe, affordable, accessible, and sustainable transport systems, with a particular emphasis on enhancing road safety. In addition to these goals, the UN has established five key pillars to further promote road safety. Pillar 1 focuses on Road Safety Management, Pillar 2 emphasizes Safer Vehicles, Pillar 3 targets Safer Road Users, Pillar 4 addresses Post-Crash Response, and Pillar 5 aims to create a Safer Driving Environment [34]. These pillars serve as a comprehensive framework for reducing global road traffic fatalities and injuries worldwide.

⋆ This is an extended version of a conference paper Digital Twins of Road Signage: Leveraging AI and RFID for Improved Road Safety, XIX International Symposium Unlocking the Hidden Potentials of Organization Through Merging of Humans and Digitals, SYMORG 2024

Recent advancements in driving behavior analysis and traffic management have leveraged machine learning (ML), deep learning (DL), Internet of Things (IoT) sensors, and data-driven methods to improve safety and efficiency. Existing research on driving behavior analysis has primarily focused on fuel efficiency [11,28], machine learning classification [30], and behavioral profiling, yet critical road safety factors remain underexplored. Studies on eco-driving have successfully optimized fuel consumption [10], but lack emphasis on broader traffic safety concerns. Similarly, ML/DL-based classification models have achieved high accuracy in distinguishing driving behaviors, yet they face scalability challenges and often fail to integrate real-world environmental data and traffic infrastructure [27]. While these studies demonstrate the potential of ML, IoT, and DL in traffic management, they often lack a holistic approach that integrates real-time data from drivers, vehicles, and road infrastructure into a unified framework. Furthermore, the absence of predictive capabilities for compliance with traffic rules and proactive safety interventions highlights the need for a more comprehensive, scalable solution that addresses both driver behavior and infrastructure management.

To address these gaps, this research introduces a data-driven framework that integrates Digital Twin technology, In-vehicle monitoring system IVMS and ML. The proposed model is designed with the aim to improve compliance with traffic rules, analyze road signage infrastructure, and predict unsafe driving behaviors.

For these purposes, we have posed the following research questions:

- RQ1: How can the DT of road safety identify and mitigate traffic sign non-compliance violations?
- RQ2: How can driving patterns of non-compliance with traffic signage be identified and used to prevent future motor vehicle incidents?
- RQ3: What insights can the DT of Road safety provide for road signage maintenance and infrastructure improvement?

The structure of this paper is as follows. Section 2 provides definitions, background, and related studies relevant to the research problem. Section 3 presents the Road Safety Digital Twin model and its components. Section 4 outlines the experimental research, including data collection and the application of analytical tools and machine learning techniques for predictive modeling. Section 5 discusses the findings in relation to the stated research questions. Finally, conclusions are presented in Section 6.

2. Related Work

This work is based on DT technology, which obtains real-world data through the IVMS system. In this section, we analyze the relevant concepts of digital twins in mobility and IVMSs, and present the analysis of relevant studies that analyze vehicle and driving behavior.

2.1. Digital Twins in Mobility

A Mobility Digital Twin (MDT) framework [36] has been introduced by Wang, characterized as an artificial intelligence (AI)-based data driven cloud–edge–device framework for mobility services.

The MDT framework operates across three distinct planes, as we can see in Figure 1:

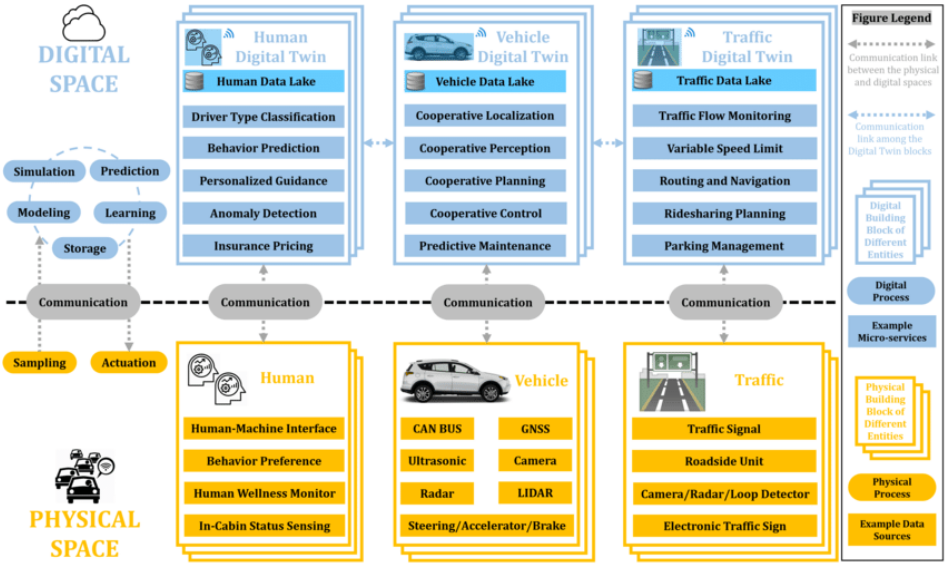


Fig. 1. Illustration of the Mobility Digital Twin framework[36]

1. the physical space, encompassing humans, vehicles, and traffic infrastructure,
2. the digital space, which contains digital counterparts of these physical entities and
3. the communication plane, which facilitates interaction between the physical and digital spaces[36].

This concept is highly significant, enabling functions like monitoring, management, maintenance, optimization, and forecasting. To achieve these goals, physical assets are equipped with RFID and smart sensors, which simplify data capture across the entire product life cycle and transmit real-time data to edge or cloud servers. The more sensors deployed, the more precise insights are obtained. This data is then analyzed and presented to users in an accessible format.

DTs leverage advanced technologies such as IoT for data acquisition, 5G for transmission, and AI/ML for data analysis and prediction. Additionally, tools like Virtual Reality (VR) and Augmented Reality (AR) enhance data representation and provide immersive experiences [18,26].

The Digital Twin collects real-time data from sensors [4] and correlates it with historical data previously obtained from the same vehicle, enabling it to make informed decisions and issue alerts for any unsafe conditions.

However, the effective deployment of DTs depends heavily on the advancement of these technologies, which are still maturing. This ongoing evolution limits the ability of DTs to fully realize their benefits. In addition, the implementation of DTs is often impeded by high costs, the complexity of managing numerous interconnected systems, and the necessity for continuous investment to keep up with rapid technological progress. Advancements in simulation and modeling tools, IoT device connectivity, expanded bandwidth,

and improved computing architectures [8] will be the key to enabling DTs to become a predominant tool for both companies and governments.

2.2. In-Vehicle Monitoring Systems (IVMS)

IVMS technology offers a comprehensive approach to improving driver safety by leveraging both hardware and software to capture critical vehicle and driving data. Real-time feedback mechanisms, including in-cab warning lights, and auditory alerts, provide drivers with immediate notifications when specific parameters, such as speed limit, are exceeded. This fosters the development of safer driving habits. This data is transmitted via various networks (cellular, Wi-Fi, or satellite) to remote servers, where it is stored for retrospective analysis, coaching, and reporting on driver behavior and vehicle performance. Crucially, IVMS can operate without constant connectivity, synchronizing data once the connection is restored, ensuring that drivers and managers receive consistent feedback [22,25].

IVMS track and analyze vehicle activity to improve road safety and efficiency [12]. These systems are becoming increasingly common, even mandatory in some areas. IVMS, which is accessible for both private and commercial use, can provide valuable insights for young drivers, helping to pinpoint areas where their driving competencies can be enhanced.

In-vehicle monitoring systems consist of five essential layers: the object layer, the sensing layer, the network layer, the data layer, and the application layer, as shown in Figure 2. These layers work together to provide comprehensive monitoring and management capabilities within vehicles [21].

- Object layer: This layer moves beyond traditional license plates by creating a digital information source for each vehicle and driver. This source acts like a digital ID card, containing identification and regulatory data for both machine and human use (think barcode vs. readable text).
- Sensing layer: This layer uses advanced technologies like RFID, GPS, and cameras to collect real-world data about vehicles and drivers. It essentially translates the physical world into a digital one, creating a rich pool of information for analysis.
- Network layer: This layer ensures data collected from various sensors gets transmitted across different regions. It acts like a digital highway, carrying information through wired networks, wireless connections, or even satellites.
- Data layer: This layer acts as the information hub. It stores sensor data, manages system information, and provides tools to analyze and extract insights from the collected data. It can also control sensor devices and provide basic functions like data retrieval to the application layer.
- Application layer: This layer puts the collected information to use. It offers functionalities through various interfaces (large screens, mobile apps, etc.) to manage vehicles and drivers in the real world. This layer essentially builds the foundation for a modern and intelligent transportation system.

Shell (the second-largest investor-owned oil and gas company) presents significant benefits from IVMS, including a 60% reduction in speeding and a major drop in accidents [32]. The system also encourages safer habits like seat belt use and discourages

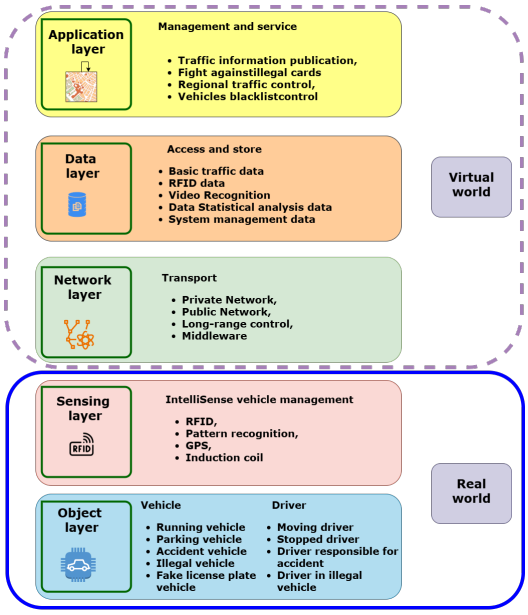


Fig. 2. IVMS 5-layers model [21]

harsh driving behaviors. Beyond safety, IVMS offers cost savings through reduced fuel consumption, less wear and tear on vehicles, and potentially lower insurance premiums.

The International Association of Oil & Gas Producers (IOGP) analyzes studies [22] related to the general usage of IVMS in commercial vehicles. Research has shown that IVMS implementation, combined with tailored driver coaching, leads to significant reductions in risky behaviors like speeding, harsh braking, and cornering, as well as a decrease in motor vehicle incidents (MVIs). IVMS also enhances journey management by providing insights into driving patterns, identifying "hot spots" with frequent incidents, and optimizing routes to avoid high risk areas.

Furthermore, IVMS offers multiple layers of security, including real-time GPS tracking for emergency response, vehicle recovery, and protection against false claims in MVI investigations. As the science of IVMS evolves, best practices continue to be refined, contributing to better driver compliance, improved driving skills, and overall road safety [6,22].

2.3. Related Studies

This research aims to address a critical gap in the existing literature, which is the absence of a DT model capable of integrating data from the all three fundamental entities in traffic and transportation (driver, vehicle and road) to generate actionable insights for enhancing road safety. To bridge this gap, we identified relevant studies that analyze vehicle and driving behavior using onboard diagnostics (OBD) data, apply ML and DL techniques to study vehicle and driver behavior, and leverage RFID technology for road-related applications. Based on our analysis, we categorized the relevant studies into the following

three groups: Studies conducted by using Onboard Diagnostics dataset, Studies applying ML and DL techniques, and Leveraging technology for road-related applications.

2.3.1. Studies Conducted Using Onboard Diagnostics Dataset

Existing research has primarily focused on utilizing OBD data and advanced computational models to assess driving patterns and promote eco-friendly driving. Some studies have developed methods to classify safe and unsafe driving behaviors by analyzing key vehicle parameters such as speed, engine RPM, throttle position, and engine load, achieving high classification accuracy through machine learning techniques like the AdaBoost algorithm [11]. Others have proposed eco-driving evaluation models that identify critical fuel-related driving events using statistical approaches such as principal component analysis and multiple linear regression, achieving predictive accuracy of up to 96.72 % [10]. Additionally, some research has explored the integration of IoT vehicle sensors with gamification strategies to encourage fuel-efficient driving, where real-time feedback based on throttle position and engine RPM helps drivers adopt safer and more sustainable driving behaviors [28]. Further advancements include the use of type-2 fuzzy logic models to assess eco-driving skills, incorporating factors such as engine speed, acceleration, and pedal position to evaluate driving style and its impact on fuel consumption, demonstrating the potential for significant fuel savings [38]. While these studies provide valuable insights into vehicle performance and driver behavior, they lack an integrated DT framework that combines real-time vehicle data with road and environmental factors to enhance predictive modeling and adaptive driving recommendations.

2.3.2. Studies Applying ML and DL Techniques

These analysis have explored various methodologies, including knowledge-based approaches, classical machine learning, and deep learning techniques, with sensor fusion emerging as a key factor in detecting aggressive, inattentive, and intoxicated driving [1]. Studies utilizing clustering methods have examined driver behavior based on open traffic data, focusing on factors such as the use of safety systems and mobile phone distractions [7]. Other research efforts have leveraged cloud-based machine learning and deep learning systems to classify driving behavior, integrating big data management techniques and clustering algorithms to distinguish between eco-friendly and aggressive driving styles [30]. Additionally, deep learning models have been applied to naturalistic driving data to assess compliance with traffic regulations, demonstrating high accuracy in real-time driver monitoring [2]. These studies provide valuable insights into vehicle and driver behavior analysis, they primarily focus on data processing and classification. The lack of a unified DT framework limits the potential for predictive modeling and real-time interventions, underscoring the need for a more holistic approach to road safety.

2.3.3. Leveraging Technology for Road-Related Applications

The traffic sign detection has explored different approaches, including RFID-based methods and deep learning techniques. Studies have investigated the use of active and passive RFID tags for traffic sign detection, where active tags provide stable detection at speeds exceeding 100 km/h and distances up to 30 m but come with high costs and the need for

battery charging, while passive tags offer a more affordable alternative with a limited 5 m range, though they may fail in scenarios like overtaking or roadwork avoidance [27]. Other approaches have focused on computer vision and deep learning, utilizing hierarchical classification models combined with object detection algorithms such as YOLOv5 to enhance traffic sign localization [37]. While these studies contribute significantly to traffic sign recognition, they primarily focus on detection and classification without integrating a DT approach that would connect traffic signage data with real-time road and vehicle conditions. This gap highlights the need for a more holistic DT framework that enhances road safety through predictive analytics and adaptive decision-making.

Since this study builds upon our previous conference paper, "Digital Twins of Road Signage: Leveraging AI and RFID for Improved Road Safety,"[14] it serves as the foundation for our ongoing research in this field and the expansion of our work. In our previous study, we focused on RFID technology and its potential for recognizing road signage, where RFID-enabled traffic signs collected real-time data on road conditions, which were then compared with values obtained from IVMS reports. Additionally, we introduced a DT model of road signage, which has been further developed for this study. In this paper, we extend our research by utilizing real-world data and incorporating a predictive component for driving behavior analysis.

3. Digital Twin Model of Road safety

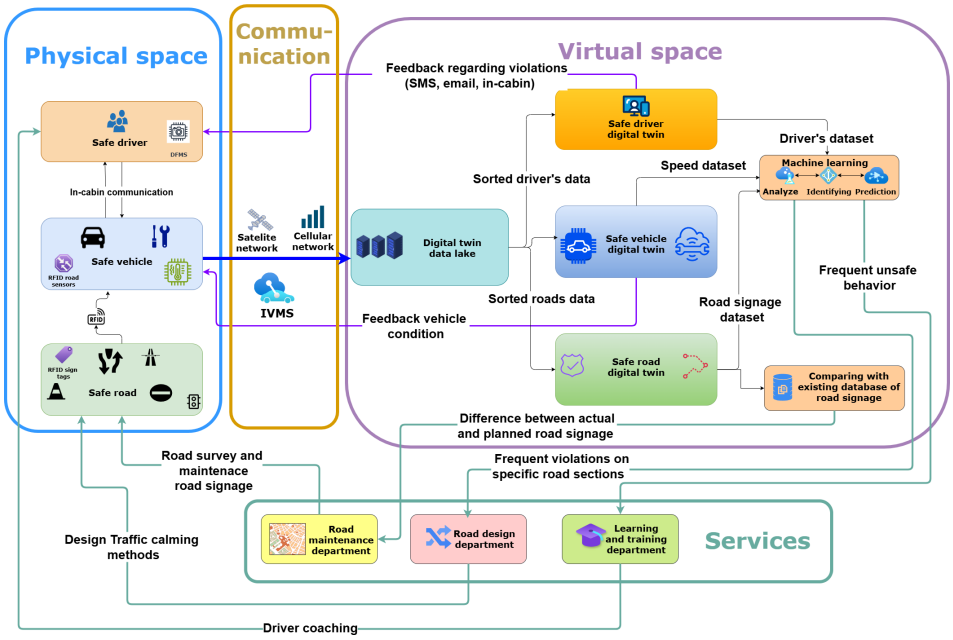


Fig. 3. Digital twin of Road safety

The role of the DT extends beyond simple simulation as it actively interacts with the physical system and adjusts to evolving external circumstances. DT technology heavily relies on data intake and correlation analysis, closely intertwined with data-intensive modeling, driven by advanced Machine Learning and Deep Learning, as well as big data analytics [13]. Our DT model of Road Safety is based on enhancing IVMS using RFID technology and artificial intelligence, as illustrated in Figure 3. In the following sections, we discuss in detail each component of the model we developed.

3.1. Physical Space

The physical space of this model is structured around the three fundamental pillars of road safety: Safe Driver, Safe Vehicle and Safe Road. At the core of this system is the vehicle, which serves as the central element connecting the Safe Road through an RFID-based link and the Safe Driver via in-cabin connectivity. The connection between the Safe Road and the Safe Vehicle is unidirectional, meaning that the vehicle only receives data from the road infrastructure without sending feedback. However, the connection with the Safe Driver is bidirectional, enabling certain driver-related parameters to be monitored and analyzed. This includes data from Driver Fatigue Management systems, as well as vehicle sensor readings influenced by driver behavior, such as speed, engine RPM, harsh braking, and harsh acceleration.

Beyond driver-related parameters, the Safe Vehicle system continuously transmits data regarding of the vehicle's overall condition and reliability. This includes predictive maintenance, utilizing sensors connected to the CANBUS system (Controller Area Network Bus), which facilitates real-time communication between the vehicle's electronic components. Additionally, other distributed sensors within the vehicle contribute to monitoring critical operational aspects, ensuring a comprehensive approach to road safety.

3.2. Virtual Space

In the virtual space, a data lake serves as the central repository where all incoming data is processed and categorized into three DTs entities: DT Safe Driver, DT Safe Vehicle, and DT Safe Road. Each DT is responsible for analyzing data within its specific domain of traffic safety, applying predefined parameters and filters to assess conditions, and either providing feedback or forwarding the information for further processing and analysis.

The DT Safe Driver can detect repeated behaviors through the Driver Fatigue Management System (DFMS) and initiate corrective actions to enhance road safety. These actions may include enforcing a mandatory 20 minute rest break, alerting a supervisor or responsible authority, or even temporarily disabling the vehicle to prevent further driving. Additionally, by combining driver behavior data with road signage information from DT Safe Road and utilizing Machine Learning (ML) algorithms, the system can identify driving patterns and predict potentially unsafe actions that may pose a risk to road safety.

The DT Safe Vehicle continuously monitors the condition of critical vehicle components, providing recommendations for corrective and predictive maintenance based on real-time sensor data collected from the vehicle's internal systems.

The DT Safe Road plays a key role in speed regulation and traffic compliance. By leveraging geolocation data of speed limit signs and stop signs (where a stop sign is technically equivalent to a speed limit of 0 km/h), it can compute speed predictions and alert

drivers to reduce speed accordingly. RFID readings of traffic signs act as triggers for these calculations, ensuring that sign data remains accessible even when a vehicle is unable to connect to GPS or GSM servers. Additionally, the RFID reader facilitates real-time traffic sign detection and transmits this information to the DT Safe Road, helping to maintain an updated database of road signage assets and improving overall traffic management.

3.3. Communication

The In-Vehicle Monitoring System serves as a key communicator between the physical and virtual spaces within the DT framework for road safety. It continuously monitors and collects data on vehicle performance and driver behavior, including speed, engine RPM, harsh braking, harsh acceleration, and geolocation, while integrating with RFID technology and CANBUS sensors to support predictive maintenance and traffic sign detection. This data is transmitted to the virtual space, where DT Safe Driver, DT Safe Vehicle, and DT Safe Road process the information and provide feedback for interventions, such as mandatory rest breaks, maintenance alerts, and speed regulation compliance. Each DT entity analyzes the data based on predefined safety parameters, ensuring appropriate real-time corrective measures. By enabling a continuous flow of information, IVMS enhances road safety, vehicle reliability, and overall traffic efficiency within an intelligent transportation ecosystem. Transmission is conducted through cellular and satellite networks.

3.4. Services

Certain services have emerged as a necessary component of the DT framework, as they cannot be strictly classified within either the physical or virtual space. These services are developed based on the 5D Digital Twin model [20] and serve as an added value to the DT system. Their primary objective is to enhance the functionality of each digital entity, ultimately improving the entire road safety ecosystem centered around a specific driver, vehicle, and road network. We have identified three key services, each operating within a dedicated department: Road Maintenance Department, Road Design Department, and Learning & Training Department.

The Road Maintenance Department analyzes data by comparing vehicle records with the existing road infrastructure database and identifying discrepancies based on real-time vehicle inputs. If inconsistencies are detected, the system triggers the Road Survey and Maintenance Road Signage service to assess and address road safety deficiencies. Following this, the Road Design Department utilizes machine learning algorithms to predict driver behavior patterns and detect frequent violations of traffic regulations. This analysis serves as a trigger for implementing traffic-calming measures, aimed at reducing risks on specific road sections. Additionally, the Learning & Training Department focuses on driver coaching, offering targeted safety training in areas where drivers exhibit unsafe behaviors, contributing to a proactive approach in road safety improvement.

4. Experimental Research

The primary goal of the experimental research is to evaluate the DT model in improving road safety. The implementation of the DT of Road safety is based on a layered system

architecture comprising edge-level data acquisition and centralized cloud-based processing through the infrastructure of the IVMS service provider. Data collection is carried out using the FMS Fusion 300 IVMS device, which integrates GNSS modules (GPS, GLONASS, BeiDou) and CANBUS connectivity for vehicle telemetry acquisition. Data transmitted via cellular or satellite networks are stored and processed within a centralized Digital Twin data lake hosted on web servers managed by the IVMS provider, where analytics are executed using Python-based geospatial libraries and the Meta Prophet forecasting algorithm. This analytical framework enables the deployment of intelligent DT services, including predictive road maintenance, road signage evaluation, and adaptive driver coaching.

4.1. Context of the Experimental Research

The Digital Twin of Road Safety model served as the foundation for this research. This study involved data collection through IVMS, where historical reports on vehicle movements and the corresponding drivers were generated by the IVMS provider's server. Compared to the model, the research did not implement the RFID component used to confirm the presence of specific traffic signs, but this does not affect the research as a whole. Geospatial data analysis plays a crucial role in traffic monitoring systems, which utilizes

Fig. 4. The FMS Fusion device for collecting data

Python libraries like GeoPandas and Shapely [5] to analyze vehicle behavior around road signs. The central concept behind this script is to evaluate whether vehicles complies with speed limits within specific zones of influence, created around the road signs based on their direction. Using GPS data from both road signs and vehicles, the script transforms this raw input into a GeoDataFrame [5], enabling spatial operations such as determining whether a vehicle falls within the influence zone of a sign. This technique leverages geospatial theory by treating locations as geometric points and constructing directional buffer zones to model the area affected by each road sign. The code further employs spatial joins to analyze the interaction between vehicles and road signs, analyzing whether a vehicle's location is within a sign's zone of influence. Once this spatial link is established, the script compares vehicle speed to the road sign's speed limit to detect potential violations. The theoretical framework here is grounded in spatial data theory, particularly

in defining directional influence zones, which are critical for ensuring that only vehicles moving in the direction of the sign are considered. Finally, the system evaluates speed violations within these zones, providing a basis for traffic enforcement and highlighting how geospatial analytics can enhance road safety by ensuring that vehicles adhere to speed limits. The Meta Prophet model (Figure 4.) [29] is a time series forecasting tool that relies on an additive approach to decompose a time series into four primary components: trend, seasonality, holiday effects, and noise. The trend captures the long-term changes in the series, seasonality reflects periodic fluctuations, holiday effects account for the impact of special dates or periods, and noise represents unpredictable random variations [15].

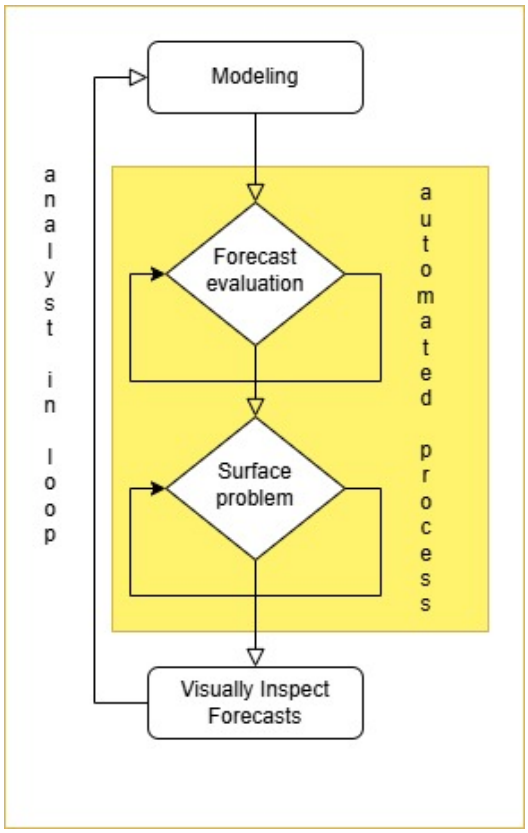


Fig. 5. The Meta Prophet model

The decomposition process in the Prophet model is divided into two parts: trend decomposition and seasonality decomposition. Trend decomposition utilizes a segmented linear function to capture both linear and non-linear components of the trend. Seasonality decomposition employs a Fourier series to break down seasonal patterns into multiple cycles. During parameter learning, the Prophet model estimates regression coefficients using least squares [33].

The overall forecast at any given time t is represented by the equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

In this formula, $y(t)$ denotes the predicted value at time t, with each term capturing a different aspect of the time series. The trend, seasonality, and holiday effects are represented by the following equations:

$$g(t) = k + at + \sum_{i=1}^{G} Ci \cdot sigmoid(t - ti)$$

$$s(t) = \sum_{n=1}^{N} (a_n \sin(\frac{2\pi nt}{P}) + b_n \sin(\frac{2\pi nt}{P}))$$

$$h(t) = \sum_{j} KjI$$

The trend component $g(t)$ models long-term changes and can be represented either by a piecewise linear function or a logistic growth model. K is the offset term indicating the overall average, a is the slope of the linear trend, t denotes time, G is the number of inflection points in the non-linear trend, Ci is the growth rate at each inflection point $(t - t_i)$, N is the order of the seasonal pattern, P is the period of the seasonality, and K_j is the effect associated with the $j - th$ holiday. The indicator function I equals 1 if time t is during the $j - th$ holiday and 0 otherwise [3].

The Meta Prophet model is well-suited for this research due to its advanced time-series forecasting, anomaly detection, and predictive analytics capabilities, which align with the study's goals of improving road safety, traffic compliance, and infrastructure management. Its ability to analyze IVMS data enables accurate predictions of traffic sign violations, identifying high-risk areas and driver non-compliance patterns. By handling complex, multi-variable traffic data, it detects behavioral trends and key compliance factors, enhancing driver monitoring and intervention strategies.

4.2. Data Collection

For the purposes of this study, data was provided by the IVMS provider with the consent of the participating company, ensuring that no personal identifiers were disclosed. The data collection period spanned from January 2023 to May 2024. Additionally, a field survey was conducted to compile a list of traffic signs, including geolocation data and speed limits, covering a total of 25 signs, among which were speed limit signs and stop signs. The research also involved the analysis of data from 22 drivers and two vehicles.

The FMS Fusion 300, presented in Figure 5., is an advanced IVMS device designed for precise data collection and processing related to vehicle and driver activities. It is equipped with an ARM (Advanced RISC Machine) Cortex-A53 Octa-core processor, 16GB eMMC (embedded MultiMediaCard) storage, and 2GB LPDDR3 (Low Power Double Data Rate 3) RAM. The device utilizes GNSS (Global Navigation Satellite System) capabilities, including GPS (Global Positioning System), GLONASS (Global Navigation Satellite System, Russia), and BeiDou (Chinese Satellite Navigation System) to

accurately capture and record real-time location data, vehicle speed, and trip duration [16]. IVMS uses cellular or satellite networks to transmit data. This research was conducted in an area without natural or artificial obstacles (such as tunnels or mobile coverage gaps). However, existing studies [9] [23] indicate that cellular signals can be maintained even inside tunnels. After data collection, the IVMS device transmits the information to IVMS provider servers, which, within the DT model, are referred to as the DT data lake. For our research, we extract various types of data from the DT data lake to construct DT Safe Driver, DT Safe Vehicle, and DT Safe Road. Later, this classified data, combined with additional collected inputs, is processed using the Meta Prophet algorithm to generate predictive results. These predictions help in triggering specific DT Services, enhancing road safety and traffic management.

4.3. Results of Experimental Research

This subsection presents the key findings derived from the experimental implementation of the proposed DT of Road Safety. Following data collection within the DT Safe Driver, DT Safe Vehicle, and DT Safer Road components, as an integral part of this research, the focus in the subsequent subsections is placed on the application of analytical tools and machine learning techniques for predictive modeling and dataset comparison, in accordance with the defined research objectives.

4.3.1. Identification of Violations Related to Traffic Sign Non-Compliance

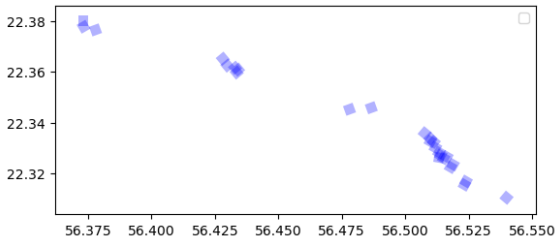


Fig. 6. The position and influence zone of traffic signs

To detect violations associated with traffic sign non-compliance, the analysis began with the integration of a geospatial dataset containing the precise coordinates of all traffic signs. A directional influence zone was then established for each sign by generating a directional buffer aligned with the sign's orientation. The direction, initially provided in degrees (with 0° indicating North), is converted into radians to facilitate the calculation of horizontal and vertical displacements based on trigonometric functions. These displacements outline a line extending from the sign's location in the specified direction.

To represent the influence zone, a buffer is generated around this line, with its width set to half the specified distance, ensuring that the zone accurately captures the sign's directional impact along the road. Additionally, square ends are produced for the buffer

to simulate the area affected by the sign. This method is applied to each road sign in a GeoDataFrame, resulting in a new column that contains the directional influence zone for each sign.

The approach ensures that the influence zone is not isotropic but instead follows the direction in which the sign impacts traffic in Figure 6. This method allows for the precise delineation of areas influenced by road signs, improving the accuracy of our model. Afterward, a spatial join is conducted between the driver data and the road sign influence zones to link each driver to the respective zone they are located within. This spatial join matches the driver's position with the influence zone of road signs based on their spatial relationship as shown in Listing 1.

```
# Spatial join to associate driver data with the road sign influence zones
joined_gdf = gpd.sjoin(driver_data_gdf,
road_signs_gdf.set_geometry('influence_zone'),
how='left', op='within')

# Ensure the geometry column is correctly set
# joined_gdf = gpd.GeoDataFrame(joined_gdf,
geometry=driver_data_gdf.geometry.name,
crs=driver_data_gdf.crs)

joined_gdf['geometry'] = joined_gdf['geometry_left']
# or another appropriate column
joined_df = joined_gdf
joined_gdf = gpd.GeoDataFrame(joined_gdf,
geometry='geometry', crs=driver_data_gdf.crs)
# def is_moving_with_sign(vehicle_direction,
sign_direction, tolerance=30):
#     Check if the vehicle's direction is aligned with the road sign's
    direction.
#     Allow some tolerance for slight deviations.
#     diff = abs(vehicle_direction - sign_direction)
% 360
#     return diff <= tolerance or diff >=
(360 - tolerance)

# # Filter based on direction
# joined_gdf['correct_direction'] = joined_gdf.apply(
# lambda row: is_moving_with_sign
(row['vehicle_direction'],row['direction']), axis=1
# )
# valid_vehicles_gdf = joined_gdf
[joined_gdf['correct_direction']]

valid_vehicles_gdf = joined_gdf
#driver_data_gdf = gpd.GeoDataFrame
(driver_data, geometry='geometry')
```

Listing 1.1. Code to associate driver data with sign influence areas

The outcome is GeoDataFrame that initially contains data from both drivers and road signs. To ensure the correct heading (direction) is maintained, the geometry column is explicitly set to the driver's original heading. The final GeoDataFrame is then validated to confirm it has the correct coordinate reference system. Although the commented-out portion of the code provides a method for checking if a vehicle's direction aligns with the road sign's direction ($is\_moving\_with\_sign function$), this function calculates the

angular difference between the vehicle's and the sign's directions, allowing for a tolerance range to account for slight deviations. If needed, the function could be applied to filter out vehicles moving in directions that do not align with the sign's influence, creating a subset of vehicles (($valid\_vehicles\_gdf$) that are correctly oriented relative to the signs. However, in this instance, the script concludes by setting ($valid\_vehicles\_gdf$) to the result of the initial spatial join, effectively retaining all joined records without further directional filtering.

Table 1. Output table with drivers violations

TachoDate (Date)	TachoDate (Time)	Driver	Speed	Speed limit	Violation
1/2/23	8:09:48	Driver 2	46	50	FALSE
1/2/23	8:09:58	Driver 2	27	50	FALSE
1/2/23	8:10:18	Driver 2	40	30	TRUE
1/2/23	8:10:19	Driver 2	40	30	TRUE
1/2/23	8:10:23	Driver 2	31	30	TRUE
1/2/23	8:10:23	Driver 2	31	30	TRUE
1/2/23	8:10:28	Driver 2	22	30	FALSE
1/2/23	8:10:28	Driver 2	22	30	FALSE
1/2/23	8:10:30	Driver 2	20	30	FALSE
1/2/23	8:10:30	Driver 2	20	30	FALSE
1/2/23	8:10:31	Driver 2	20	30	FALSE
1/2/23	8:10:31	Driver 2	20	30	FALSE
1/2/23	8:10:34	Driver 2	20	30	FALSE

In this research, we identify instances where vehicles exceed the speed limit and count the number of violations in Table 1. A new column, "violation", is created in the dataset by comparing each vehicle's recorded speed against the speed limit specified by the corresponding road sign. This comparison determines whether the vehicle's speed exceeds the limit, with the result indicating a violation if the speed is above the limit. After identifying these violations, the total number is calculated by summing the number of instances where violations occurred. This count represents the total number of vehicles that exceeded the speed limit according to the road signs they encountered. Finally, the total number of speed violations detected within the dataset is displayed, highlighted in red in Figure 7.

4.3.2. Predictions of Driver Behavior

If we examine compliance with traffic regulations, particularly speed limits, we can distinguish between intentional and unintentional driving behavior[31]. Intentional non-compliance may arise due to a lack of knowledge, unawareness of consequences, absence of enforcement or monitoring ("nobody sees me"), or reduced concentration. While occasional instances of such behavior can happen to any driver, the concern arises when these actions become repetitive patterns. Identifying such patterns enables us to anticipate future traffic violations, which could ultimately lead to motor vehicle incidents. This research focuses on predicting potential driver behavior and drawing data-driven conclusions based on these predictions to enhance road safety interventions.

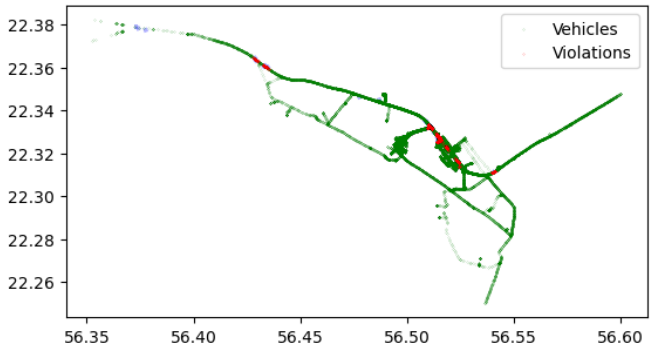


Fig. 7. Spatial join to associate driver violations with the road sign influence zones

Table 2. Output a table with predictions of drivers' behavior

ds	yhat_lower	yhat_upper	trend_lower	trend_upper	yhat	Driver
2024.10.20	-0.142796082	0.832923817	0.115616094	0.115616157	0.341534197	Drv1
2024.10.27	-0.15413158	0.781563299	0.112826379	0.112826445	0.338744484	Drv1
2024.11.03	-0.120619894	0.790254016	0.110036663	0.110036735	0.335954771	Drv1
2024.11.10	-0.093254485	0.806663005	0.107246948	0.107247024	0.333165057	Drv1
2024.11.17	-0.177004874	0.790667688	0.104457232	0.104457312	0.330375344	Drv1
2024.11.24	-0.137346225	0.760633958	0.101667516	0.101667601	0.327585631	Drv1
2024.12.01	-0.138367901	0.805363476	0.098877801	0.09887789	0.324795917	Drv1
2024.12.08	-0.173509915	0.783826341	0.096088085	0.096088179	0.322006204	Drv1
2024.12.15	-0.133032178	0.791374811	0.093298369	0.093298469	0.319216491	Drv1
2024.12.22	-0.134742919	0.783052654	0.090508654	0.090508758	0.316426777	Drv1
2024.12.29	-0.132399781	0.796272074	0.087718938	0.087719047	0.313637064	Drv1
2025.01.05	-0.137950205	0.771677584	0.084929223	0.084929337	0.31084735	Drv1
2025.01.12	-0.165706826	0.790620813	0.082139506	0.082139626	0.308057637	Drv1
2025.01.19	-0.166182664	0.75937172	0.079349791	0.079349916	0.305267924	Drv1
2025.01.26	-0.171092742	0.798246192	0.076560075	0.076560205	0.30247821	Drv1
2025.02.02	-0.168234231	0.784721728	0.073770358	0.073770495	0.299688497	Drv1
2025.02.09	-0.195392163	0.743006299	0.070980644	0.070980783	0.296898784	Drv1
2025.02.16	-0.172924502	0.756590949	0.068190928	0.068191072	0.29410907	Drv1
2025.02.23	-0.16626537	0.755384227	0.065401212	0.065401361	0.291319357	Drv1
2025.03.02	-0.1781599	0.764092547	0.062611495	0.06261165	0.288529644	Drv1
2025.03.09	-0.166079517	0.753248377	0.059821779	0.059821939	0.28573993	Drv1
2025.03.16	-0.159474004	0.772116024	0.057032063	0.05703223	0.282950217	Drv1
2025.03.23	-0.210509577	0.744856374	0.054242347	0.054242519	0.280160504	Drv1
2025.03.30	-0.194617814	0.735365529	0.051452631	0.05145281	0.27737079	Drv1
2025.04.06	-0.186525308	0.727976744	0.048662915	0.0486631	0.274581077	Drv1
2025.04.13	-0.190258675	0.745741923	0.045873199	0.04587339	0.271791363	Drv1

The output dataset shown in Table 2. presents predictions for driver behavior based on time series analysis, with each row corresponding to a specific time-stamped prediction. Key variables in the dataset include the timestamp (ds), the underlying trend component

(trend), and the prediction intervals ($yhat\_lower$ and $yhat\_upper$), which indicate the range within which the actual driver behavior is expected to fall. The dataset also provides confidence intervals for the trend component ($trend\_lower$ and $trend\_upper$), as well as additive terms that account for daily and weekly seasonal patterns, along with other influencing factors. The impact of these seasonal patterns is further detailed in the daily and weekly components, while the multiplicative terms are set to zero, indicating no multiplicative effects were considered in this dataset. The $yhat$ value represents the predicted driver behavior at each timestamp, and the $Driver$ column specifies the driver associated with each prediction [29].

The results indicate that the trend component for each driver remains consistent over time, suggesting a stable underlying pattern in their behavior. The prediction intervals, represented by $yhat\_lower$ and $yhat\_upper$, are relatively narrow, which implies a high level of confidence in the predictions. Seasonal effects are evident in the daily and weekly components, with the data indicating variations in driver activity or compliance at different times of the day and across different days of the week [19].

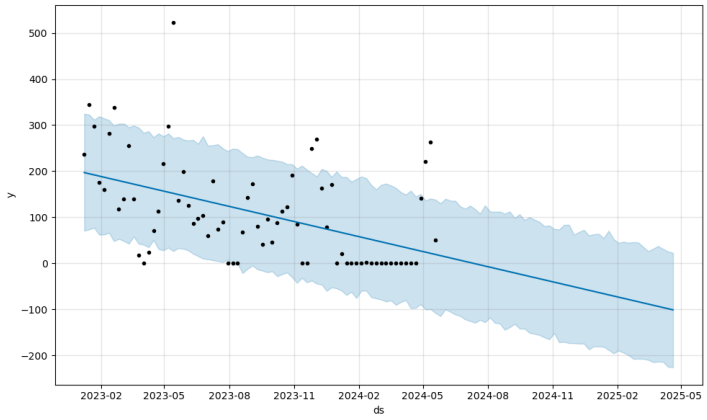


Fig. 8. Forecast trend of traffic rules violations

The plot in Figure 8. shows the predicted trend of traffic rule violations associated with posted traffic signs from 2023 until mid-2024. It suggests that traffic rule violations are expected to decrease significantly over time. This could be due to various factors, such as better enforcement, increased driver awareness, or changes in traffic management policies. However, the confidence interval indicates some uncertainty, and the model's performance should be closely monitored to ensure it continues to provide reliable forecasts. In our research, we analyzed historical driving data to forecast future behaviors, focusing on the likelihood of violations and changes in habits. This methodology enables the identification of key behavioral trends, seasonal variations, and anomalies at the individual driver level, thereby offering valuable insights for targeted safety improvement and performance management.

In Figure 9, we present forecasts for four drivers, each representing a typical driving behavior pattern identified among the 22 analyzed driver behaviors in this study. For

Driver 1, the forecast indicates a slight decline in violations, suggesting a potential improvement in adherence to traffic regulations over the analyzed period. However, the wide prediction interval suggests significant uncertainty, possibly due to inconsistent driving patterns. The data shows noticeable seasonal patterns, with some recurring peaks and dips, and violations clustered around certain values. The forecast for Driver 2 shows a relatively stable trend with a slight upward movement towards the end of the forecast period, suggesting a steady but slightly increasing rate of violations. The prediction interval is moderately wide, indicating some uncertainty but less variability compared to Driver 1. The pattern of violations is more consistent, with fewer fluctuations, indicating more stable driving behavior. The forecast for Driver 3 indicates a slight upward trend in violations, pointing to a possible increase over time. The prediction interval is wide, especially in the middle and towards the end of the forecast, reflecting high uncertainty and potentially erratic driving behavior. The pattern shows pronounced spikes and dips, suggesting more irregular behavior or external influences. The forecast for Driver 4 (driving during maintenance) exemplifies how predictions should ideally appear for all drivers, reflecting stable driving patterns with minimal fluctuations and promoting good driving behavior. In contrast, Driver 1 shows a downward trend but with high variability, Driver 2 exhibits a more stable pattern with slight increases, and Driver 3 displays an upward trend with considerable uncertainty.

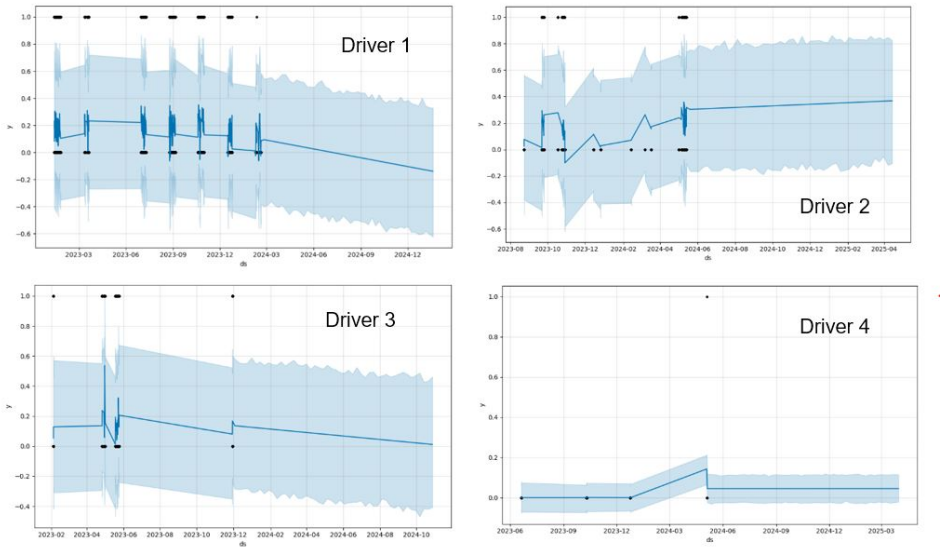


Fig. 9. Forecast for driving behaviors of selected drivers 1, 2, 3 and 4

4.3.3. Forecasting Compliance with Road Signage as Part of Maintenance and Infrastructure Improvements

The following analysis presents forecasts of driver compliance with various traffic signs, including speed limits and stop signs on both main and side roads, as illustrated in Figure 10. The results identify distinct behavioral patterns, ranging from consistent adherence to fluctuating violations, emphasizing critical areas for traffic management intervention. The forecast for the ”Speed Limit 40 km/h” sign on the main road indicates a generally

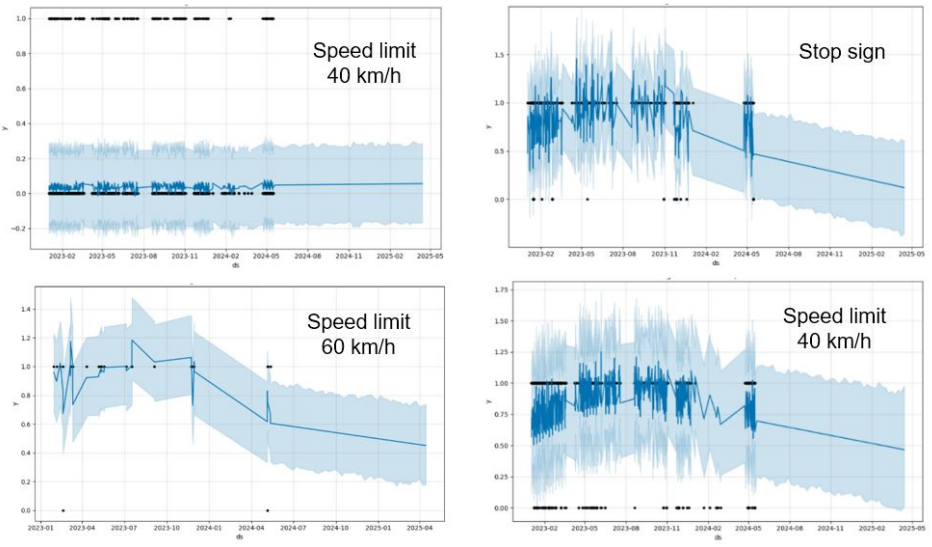


Fig. 10. Prediction of compliance with traffic signs

stable trend, with minimal fluctuations around the posted speed limit. The prediction interval remains narrow, signifying a high level of confidence in the projected values. This consistency suggests that drivers largely comply with the speed limit, with only occasional deviations observed. The stability in adherence implies that existing traffic control measures are effective in maintaining compliance at this location.

In contrast, the forecast for the ”Stop Sign” on the side road reveals significant variability in driver behavior, with widening fluctuations, particularly from mid-2023 onward. The prediction interval expands, reflecting greater uncertainty and an increased frequency of violations. A downward trend in compliance beginning in early 2024 raises road safety concerns, indicating a potential decline in driver awareness or enforcement effectiveness. This suggests the need for targeted traffic interventions to enhance compliance at this intersection.

The forecast for the ”Speed Limit 60 km/h” sign on the main road exhibits moderate variability, with the prediction interval widening over time, reflecting increasing uncertainty in driver behavior. A downward trend in compliance is observed starting from early 2024, suggesting a decline in adherence to the speed limit. The growing volatility in driver

behavior may indicate inconsistent enforcement or changing driving patterns, necessitating further analysis and possible intervention to ensure traffic safety at this location.

Similarly, the forecast for the sign ”Speed Limit 40 km/h” on the main road highlights high variability in driver compliance, characterized by frequent fluctuations in the number of violations. The confidence interval remains wide throughout the forecast period, indicating substantial uncertainty in future trends. Although a slight decrease in violations is projected over time, irregular driving patterns persist, emphasizing ongoing challenges in maintaining compliance with posted speed limits. These findings suggest a need for further monitoring and potential traffic calming measures to improve speed regulation and overall road safety.

4.3.4. Potential Validity Issues

While it may appear that the volume of collected data is a limitation for a comprehensive experimental analysis, our DT of Road safety, although tested on a limited dataset, is inherently scalable and can be effectively extended to larger geographic areas, encompassing a greater number of vehicles, road segments, and traffic signs. This aspect will be explored in future research.

5. Discussion

This section provides a comprehensive discussion of the principal findings derived from the experimental investigation, aligning them with the previously formulated research questions. Each part is dedicated to one question, drawing connections between the results and the research aims. By combining insights from data analytics, user behavior patterns, and interactive vehicle system responses modeled through the Digital Twin environment, the discussion explores the DT of Road Safety capacity to support data-driven road safety improvements.

5.1. How Can the Digital Twin of Road Safety Identify and Mitigate Traffic Sign Non-Compliance Violations?

The DT model, specifically the DT Safe Road, in this research utilizes real-world data collected from the physical environment combined with existing traffic signage records to identify all instances of traffic non-compliance. This serves as the initial step toward the ultimate goal of mitigating traffic sign violations and enhancing compliance.

Our finding provides an answer to RQ1 by establishing a method for identifying traffic sign violations and gaining an initial insight into potential black spots [24] within the road network. To address this issue, we identify two key factors influencing non-compliance. The first factor is driving behavior, particularly human errors, which can be intentional or unintentional [31]. The second factor relates to Safe Road infrastructure, where addressing this challenge involves implementing traffic calming measures to reduce speed and enhance compliance. This can be achieved by activating Road Maintenance Department services within the Digital Twin model, ensuring timely interventions and improvements in road safety. These measures will be explored in more detail through RQ2 and RQ3.

5.2. How Can Driving Patterns of Non-Compliance with Traffic Signage Be Identified and Used to Prevent Future Motor Vehicle Incidents?

If we examine compliance with traffic regulations, particularly speed limits, we can distinguish between intentional and unintentional driving behavior[31]. Intentional non compliance may arise due to a lack of knowledge, unawareness of consequences, absence of enforcement or monitoring (”nobody sees me”), or reduced concentration. While occasional instances of such behavior can happen to any driver, the concern arises when these actions become repetitive patterns. Identifying such patterns allows us to anticipate future traffic violations, which could ultimately lead to motor vehicle incidents. This research focuses on predicting potential driver behavior and drawing data-driven conclusions based on these predictions to enhance road safety interventions. By analyzing recurrent behaviors, the model aims to predict and mitigate potential violations before they escalate into safety-critical events.

This study reveals intentional non-compliance with speed-related traffic regulations and identifies behavioral patterns where drivers improve or deteriorate their driving habits over time. By utilizing predictive tools, we can anticipate that certain behaviors may lead to motor vehicle incidents, including asset damage and injuries to pedestrians and other road users. These risk patterns trigger one of the DT services within the Learning & Training Department, which then delivers targeted awareness programs and knowledge-based interventions to enhance driver safety and compliance. When discussing unintentional driving behavior, we will explore this aspect further in response to RQ3 in the next chapter.

5.3. What Insights Can the Digital Twin of Road Safety Provide for Road Signage Maintenance and Infrastructure Improvement?

The DT Safe Road is designed as a key component of road infrastructure, aimed at enhancing road signage maintenance and infrastructure improvements. Communication within road networks, encompassing interactions among vehicles and between vehicles and infrastructure, has begun to garner interest within the traffic engineering community [17].

Fig. 11. Road signage survey after analysis

In our post-analysis of certain graphs, we gained key insights that significantly impacted the graphical representation, underscoring the importance of this type of analysis. In particular, we found that the stop sign had no recorded values despite being located on a secondary road, where violations should have been more frequent. Interviews with drivers and data from the road maintenance department revealed that the stop sign had been damaged (knocked down), and one of the 40 km/h signs had been rotated by the wind, showing the speed limit in the opposite direction, as shown in Figure 11. These issues were rectified before our field data collection, but it is important to note that two months passed between the initial occurrence and the repair. In the context of road safety, such delays are unacceptable given the potential consequences of these events. Furthermore, some traffic signs were incorrectly positioned or had faded to the point of being barely legible, a fact that we were aware of during data collection. Our graphical analysis confirmed that a portion of traffic violations could be attributed to these specific issues.

6. Conclusions

In conclusion, mitigating traffic incidents caused by sign disobedience requires a multifaceted approach that considers the interdependent relationship between drivers, roads, and vehicles. By synchronizing these elements and leveraging technological advancements, significant progress can be made toward enhancing road safety. The integration of IVMS and DT offers a transformative opportunity to complement traditional road safety strategies and foster a safer, more efficient transportation ecosystem.

This research proposes a DT model that integrates all three pillars of traffic safety: Safe Driver, Safe Vehicle, and Safe Road, by combining IVMS data with machine learning techniques. The model presents a comprehensive traffic safety approach, facilitating detailed vehicle data collection and enabling real-time traffic sign recognition. By analyzing this data, the system can predict driver behavior, monitor compliance with speed limits, and provide proactive warnings, thereby promoting adherence to traffic regulations.

The research findings highlight the ability to identify drivers in need of additional coaching, evaluate the effectiveness of safety interventions, and ensure long-term improvements in driving behavior. The focus remains on promoting consistent, safe driving habits, with forecasting models ideally showing stable, low-violation patterns. Additionally, classifying violations by traffic sign type and predicting trends in non-compliance rates allow for targeted road surveys and the implementation of traffic-calming measures to enhance safety.

Future research should focus on scaling the DT road safety model across broader geographical areas and a diverse range of vehicles. Expanding its application would enable better validation of its effectiveness across various road networks and traffic environments, providing critical insights into cost-effectiveness, scalability, and the potential for widespread adoption. Furthermore, integrating additional sensor technologies, such as LiDAR (Light Detection and Ranging) and camera-based systems, alongside IVMS would enhance traffic sign detection accuracy and data collection reliability under diverse road conditions.

Advancing machine learning techniques remains crucial for improving predictive capabilities related to driver behavior and traffic management. Incorporating external factors such as weather conditions, time of day, and traffic density could increase forecasting

accuracy and enable more dynamic safety interventions. Additionally, future research should focus on real-time applications of the model, offering immediate feedback to drivers and ensuring proactive traffic management. As the adoption of DT technology in smart city traffic systems continues to grow, this model holds significant potential to contribute to data-driven traffic management solutions and enhanced road safety outcomes.

References

1. Adhikari, B.: Using visual and vehicular sensors for driver behavior analysis: A survey. Arxiv Cornell University (2023), dOI 10.48550/arXiv.2308.13406
2. Al-Hussein, W.A., Por, L.Y., Kiah, M.L.M., Zaidan, B.B.: Driver behavior profiling and recognition using deep-learning methods: In accordance with traffic regulations and experts guidelines. International Journal of Environmental Research and Public Health vol.19 (2022), dOI 10.3390/ijerph19031470
3. Almazrouee, A., Almeshal, A., Almutairi, A., Alenezi, M., Alhajeri, S.: Long-term forecasting of electrical loads in kuwait using prophet and holt-winters models. Applied Sciences vol.10 (2020), dOI 10.3390/app10165627
4. Ashfaq, N., Khan, S., Niaz, F., Usman, S.M., Niaz, I., Yanbing, J.: Smart city iot application for road infrastructure safety and monitoring by using digital twin. In: 2022 International Conference on IT and Industrial Technologies (ICIT). pp. 1–6 (2022)
5. Belhaj Ali, A.: Spatial Statistics with Python: Theories, Techniques and Applications. Independently published (09/ 2024), iSBN:979-8338972335
6. Bell, J.L., Taylor, M.A., Chen, G.X., Kirk, R.D., Leatherman, E.R.: Evaluation of an in-vehicle monitoring system (ivms) to reduce risky driving behaviors in commercial drivers: Comparison of in-cab warning lights and supervisory coaching with videos of driving behavior. Journal of Safety Research vol.60, pp.125–136 (2017), dOI 10.1016/j.jsr.2016.12.008
7. Blagojević, M., Šošić, S.: Application of cluster analysis in the behaviour of traffic participants relating to the use of safety systems and mobile phones. Facta Universitatis, Series: Electronics and Energetics 33(4), 655–668 (2020), dOI 10.2298/FUEE2004655B
8. Botin, D., Lozoya-Reyes, J., Vargas-Maldonado, R., Rodríguez-Hernández, K., Lozoya-Santos, J., Ramírez Moreno, M.A., Ramirez-Mendoza, R.: Digital twin for urban spaces: an application. In: Proceedings of the 6th North American Conference on Industrial Engineering Operations Management. Monterrey, Mexico (2021)
9. Broz, J., Tichy, T.: Road tunnel positioning: Enabling location-based services in gnss-denied environments. IEEE Access vol.12, pp.156694–156701 (2024), dOI 10.1109/ACCESS.2024.3479278
10. Chen, C., Zhao, X., Yao, Y., Zhang, Y., Rong, J., Liu, X.: Driver's eco-driving behavior evaluation modeling based on driving events. Journal of Advanced Transportation 2018 (1/ 2018), dOI 10.1155/2018/9530470
11. Chen, S.H., Pan, J.S., Lu, K.: Driving behavior analysis based on vehicle obd information and adaboost algorithms. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. vol. 1, pp. 102–106. Hong Kong (3/ 2015), dOI 10.1155/2018/9530470
12. Chepuru, A., Maddireddy, G., Rawoof, H.A., Rajarapu, P., Shivani, S.: Integrated vehicle monitoring system. International Journal for Research in Applied Science and Engineering Technology vol.8, pp. 1663–1669 (5/ 2020), dOI 10.22214/ijraset.2020.5271
13. Durković, M., Hadžić, D., Barać, D., Despotović-Zrakić, M., Bogdanović, Z., Radenković, B.: Digital twin city congestion management model as part of smart mobility in gcc countries. Marketing and Smart Technologies: Proceedings of ICMarkTech 2023 pp. 341–352 (2023), dOI 10.1007/978-981-97-3698-024

14. Durković, M., Hadžić, D., Lukovac, P.: Digital twins of road signage: Leveraging ai and rfid for improved road safety. In: Kostić-Stanković, M.P., Mijatović, I.P., Krivokapić, J.P. (eds.) SYMORG 2024, XIX International Symposium Unlocking the Hidden Potentials of Organization Through Merging of Humans and Digitals. University of Belgrade – Faculty of Organizational Sciences (2024)
15. Ertürk, M.A.: Comparison of time series forecasting for intelligent transportation systems in digital twins. Electrica 2024 vol.24, pp. 375–384 (2024), dOI 10.5152/electrica.2024.23200
16. FMSTech, O.: Driver fatigue management solution in adipec (2023), [Online]. Available: https://fms-tech.com/fms-tech-to-showcase-driver-fatigue-management-solution-in-adipec-2023/ (current July 2025)
17. Gobbi, H.U., dos Santos, G.D., Bazzan, A.L.C.: Comparing reinforcement learning algorithms for a trip building task: a multi-objective approach using non-local information. Computer Science and Information Systems, ComSIS, Novi Sad, Srbija 21(1), 291–308 (2023), dOI 10.2298/CSIS221210072G
18. Guo, J., Bilal, M., Qiu, Y., Qian, C., Xu, X., Choo, K.K.R.: Survey on digital twins for internet of vehicles: Fundamentals, challenges, and opportunities. Digital Communications and Networks vol. 10, pp. 237–247 (4/ 2024), dOI 10.1016/J.DCAN.2022.05.023
19. Gupta, R., Yadav, A.K., Jha, S.K., Pathak, P.K.: Long term estimation of global horizontal irradiance using machine learning algorithms. Optik vol.283 (7/ 2023), dOI 10.1016/J.IJLEO.2023.170873
20. Hassan, M., Svadling, M., Björsell, N.: Experience from implementing digital twins for maintenance in industrial processes. Journal of Intelligent Manufacturing vol.35, pp. 1–10 (02/ 2023), dOI 10.1007/s10845-023-02078-4
21. Hu, L., Li, H., Xu, X., Li, J.: An intelligent vehicle monitoring system based on internet of things. In: Proceedings of 7th International Conference on Computational Intelligence and Security, CIS 2011. pp. 231–233 (2011), dOI 10.1109/CIS.2011.59
22. IOGP: Implementing and sustaining an in-vehicle monitoring system programme (5/ 2024), [Online]. Available: https://www.iogp.org/bookstore/product/implementing-an-in-vehicle-monitoring-program-a-guide-for-the-oil-and-gas-extraction-industry/ (current July 2025)
23. Jiang, S., Xu, Q., Wang, W., Peng, P., Li, J.: Vehicle positioning systems in tunnel environments: a review. Complex & Intelligent Systems vol.11/64, pp.1–34 (2025), dOI 10.1007/s40747-024-01744-1
24. Karamanlis, I., Nikiforiadis, A., Botzoris, G., Kokkalis, A., Basbas, S.: Towards sustainable transportation: The role of black spot analysis in improving road safety. Sustainability vol.15(19) (2023), dOI 10.3390/su151914478
25. Krum, A., Miller, A., Soccolich, S.: Evaluation of an in-vehicle monitoring system among an oil and gas well servicing fleet. National Surface Transportation Safety Center for Excellence (NSTSCE),Blacksburg, Virginia (5/ 2020)
26. Marai, O.E., Taleb, T., Song, J.: Roads infrastructure digital twin: A step toward smarter cities realization. IEEE Network vol. 35, pp. 136–143 (2021), dOI 10.1109/MNET.011.2000398
27. Mariut, F., Fosalau, C., Zet, C., Petrisor, D.: Experimental traffic sign detection using i2v communication. In: Proceedings of 35th International Conference on Telecommunications and Signal Processing (TSP). pp. 141–145 (5/ 2012), dOI 10.1109/TSP.2012.6256269
28. Massoud, R., Bellotti, F., Berta, R., Gloria, A.D., Poslad, S.: Eco-driving profiling and behavioral shifts using iot vehicular sensors combined with serious games. In: 2019 IEEE Conference on Games (CoG). pp. 1–8 (2019)
29. Meta: Prophet: Forecasting at scale (2017), [Online]. Available: https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale (current July 2025)
30. Peppes, N., Alexakis, T., Adamopoulou, E., Demestichas, K.: Driving behaviour analysis using machine and deep learning methods for continuous streams of vehicular data. Sensors vol.21 (2021), iSSN 1424-8220, DOI 10.3390/s21144704

31. Rahman, M., Islam, M., Al-Shayeb, A., Arifuzzaman, M.: Towards sustainable road safety in saudi arabia: Exploring traffic accident causes associated with driving behavior using a bayesian belief network. Sustainability vol.14, pp.6315 (05/ 2022), dOI 10.3390/su14106315
32. Shell: In-vehicle monitoring systems improve driving skills (2024), [Online]. Available: https://www.shell.com/business-customers/shell-fleet-solutions/health-security-safety-and-the-environment/in-vehicle-monitoring-systems-can-help-everyone-to-improve-their-driving-skills.html (current July 2025)
33. Taylor, S.J., Letham, B.: Forecasting at scale. The American Statistician vol.72, pp.37–45 (2018), dOI 10.1080/00031305.2017.1380080
34. UN, N.: Un road safety strategy 5 pillars, [Online]. Available: https://www.un.org/sites/un2.un.org/files/2021/04/undssun_road_safety_strategy_5_pillars.png (current July 2025)
35. UN, N.: Un sustainable development goal 3.6, [Online]. Available: https://www.un.org/sustainabledevelopment/health/ (current July 2025)
36. Wang, Z., Gupta, R., Han, K., Wang, H., Ganlath, A., Ammar, N., Tiwari, P.: Mobility digital twin: Concept, architecture, case study, and future challenges. IEEE Internet of Things Journal vol.9, pp. 17452–17467 (9/ 2022), dOI 10.1109/JIOT.2022.3156028
37. Xu, J., Huang, Y., Ying, D.: Traffic sign detection and recognition using multi-frame embedding of video-log images. Remote Sensing vol.15 (06/ 2023), dOI 10.3390/rs15122959
38. Zdravković, S., Vujanović, D., Stokic, M., Pamucar, D.: Evaluation of professional driver's eco-driving skills based on type-2 fuzzy logic model. Neural Computing and Applications vol. 33 (09/ 2021), dOI 10.1007/s00521-021-05823-z

Miloš Durković is currently pursuing a Ph.D. in Software Engineering and E-Business at the Faculty of Organizational Sciences, University of Belgrade. His dissertation focuses on smart mobility and innovative road safety technologies. His research interests include artificial intelligence, internet technologies, advanced road safety systems, and the Internet of Things (IoT). He is currently employed by Dareen Global, working on a road safety standards project with Petroleum Development Oman. He brings extensive transportation and logistics experience from his prior service in the Serbian Armed Forces.

Petar Lukovac is a teaching assistant at the Faculty of Organizational Sciences, University of Belgrade. He is a PhD student in Software engineering and e-business, pursuing a PhD dissertation in the field of smart environments and blockchain technologies. His research interests include web3, internet technologies, e-business, and IoT.

Demir Hadžić, PhD, has over 24 years of full-time professional experience, including key roles with the Government of Serbia and Petroleum Development Oman. He specializes in transportation, logistics, and HSSE, with a career focused on developing and implementing strategic solutions that promote sustainability, safety, and operational efficiency. He is recognized for his ability to lead multicultural teams across both government and private sectors, leveraging his integrated expertise in safety, business, and sustainability to drive long-term success.

Zorica Bogdanović, PhD, is a full professor and the Head of the IoT Center at University of Belgrade - Faculty of Organizational Sciences. Her research interests include internet of things, smart environments, and internet technologies.

Dušan Barać, PhD, is a full professor and the Vice-Dean for digital development at University of Belgrade - Faculty of Organizational Sciences. His research interests include digital transformation, e-commerce and software engineering.

Received: September 25, 2024; Accepted: June 25, 2025.

Computer Science and Information Systems 22(4):1509–1532 https://doi.org/10.2298/CSIS241225059Z

Fire Detection Models Based on Attention Mechanisms and Multiscale Features

Shunxiang Zhang[1,⋆], Meng Chen[1], Kuan-Ching Li[2], Hua Wen[1], and Liang Sun[1]

[1] School of Computer Science and Engineering, Anhui University of Science & Technology, 232001 Huainan, China
sxzhang@aust.edu.cn
2260662967@qq.com
1762636707@qq.com
2582132681@qq.com

[2] Department of Computer Science and Information Engineering (CSIE), 13 Providence University, 43301 Taizhong, Taiwan
kuancli@pu.edu.tw

Abstract. Fire detection is critical in applications such as fire management and building safety, but dispersion and blurring of flame and smoke boundaries can present challenges. Multiple upsampling and downsampling operations can blur the localisation signals, thus reducing accuracy and efficiency. To address this problem, we propose the AMMF(Attention Mechanisms and Multiscale Features) detection model, which integrates an attention mechanism and multi-scale feature fusion to improve accuracy and real-time performance. The model incorporates a dynamic sparse attention mechanism in the backbone network to enhance feature capture and restructures the neck network using CepBlock and MPFusion modules for better feature fusion. MDPIoU loss and Slideloss are then utilised to reduce the bounding box regression error and address the sample imbalance problem respectively. In addition, parameters are shared by merging 3×3 convolutional branches, which optimises the detection head and improves computational efficiency. The experimental results show that AMMF-Detection can significantly improve the detection speed and accuracy on the public dataset.

Keywords: Fire detection,YOLO,Feature fusion,ynamic sparse attention,Multi-scale features

1. Introduction

Detecting objects is essential for analyzing and understanding flame and smoke images, with its main objective being to precisely identify and pinpoint components like the smoke and fire source.The complexity and diversity of fire scenes, the irregular shapes of target objects, and the presence of numerous interfering elements in the images lead to low detection accuracy. Additionally, when processing the original image, up-sampling introduces additional pixels, increasing the sparsity of the original features, while down-sampling causes a loss of localization information. This results in the gradual blurring or disappearance of small flame and smoke details, which negatively impacts detection accuracy. Moreover, the current computational efficiency remains a significant challenge, as

⋆ Corresponding author

fire detection must meet real-time requirements. Consequently, fire detection models must possess strong abstraction and generalization capabilities to handle the complex and dynamic nature of fire scenarios. These demands further complicate target detection, making fire detection a highly challenging task.

Fire detection algorithms are generally divided into three main types: methods based on classifiers, model compression techniques, and deep learning approaches [1,2].Traditional classifier-based techniques rely on manually designed feature extractors to derive image features [3], and then use algorithms like SVM, ID3, or BP neural networks to identify fire and smoke. While model compression methods enhance detection speed, they often face challenges in achieving a balance between accuracy and efficiency. Conversely, fire detection models based on deep learning possess the ability to autonomously extract image features, enabling them to recognize intricate patterns and finer details with greater precision, which leads to enhanced detection accuracy. Mainstream deep learning-based object detection algorithms include SSD [4], YOLO [5,6,7,8], and Transformer-based RT-DETR [9]. Despite advancements in object detection, these methods still face several challenges. First, target features in the image are often scattered with fuzzy boundaries, causing the original features to become more dispersed and sparse during feature fusion, which increases the model's complexity in processing up-sampled features and degrades its performance on certain features. Second, reducing image resolution through down-sampling operations causes the loss of fine details in flame and smoke images. This loss adversely affects the model's capacity to detect small objects or identify localized features. Moreover, increasing model complexity presents a challenge for real-time inference, particularly in resource-constrained environments. Finally, issues such as sample imbalance, poor data quality, and insufficient training data further affect model performance, potentially leading to misdetection or detection failures.

Taking the above considerations into account, this paper introduces the AMMF-Detection model, designed to achieve a balance between detection accuracy, processing speed, and computational efficiency. As illustrated in Figure 1, the model's overall structure comprises three key components: the backbone network, the neck network, and the detection head.The MPfusion module, designed for feature fusion, combines feature maps from three different scales. The CepBlock, a feature extraction module, employs distinct structures during training and inference, ensuring high accuracy during training and fast inference speed. The detection part introduces a novel detection head that integrates seamlessly with the original convolutional block without compromising model performance. This paper aims to enhance the accuracy of fire detection and the speed of inference, all while minimizing the complexity of the model. This goal is accomplished by comprehensively extracting edge characteristics, including color, shape, and texture, from images of flames and smoke to improve the accuracy of fire detection. The model aims to better serve applications in fire safety and emergency response, including fire management, warehousing and logistics, building safety, and other monitoring and detection scenarios.

Our contributions are summarized in the following three aspects.

(1)We propose a method to optimize the backbone network using dynamic sparse attention. This approach aims to enhance the model's feature memory and recognition capabilities, enabling it to focus more effectively on feature selection. By addressing the challenges posed by complex backgrounds, this method effectively mitigates issues of target misdetection and omission.

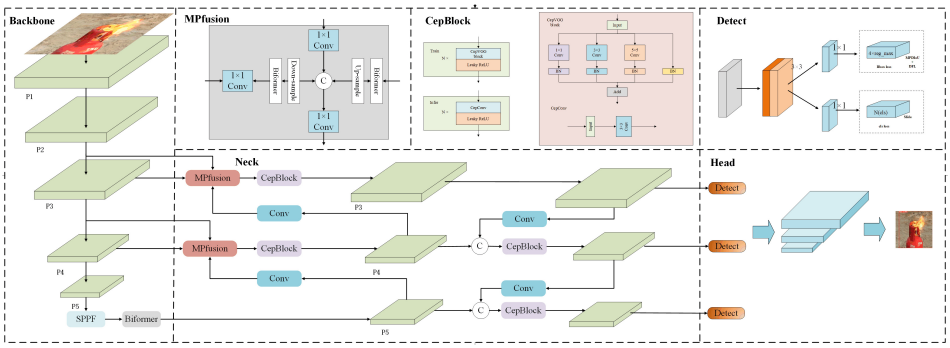


Fig. 1. Overall framework of proposed AMMF

(2)We propose a method that integrates the parameter reconfiguration of the CepBlock module with the MPFusion module to redesign the neck network. This redesign enables the fusion of feature maps across multiple receptive fields, facilitating deeper exploration of features at different levels and improving the overall feature representation.

(3)We propose a novel lightweight detection head design that simplifies the model structure by merging the 3×3 convolutions in the original branches, achieving parameter sharing. This design enhances computational efficiency and makes the model more suitable for deployment on resource-constrained devices.

The structure of the paper is outlined as follows: Section 2 reviews related research on fire detection. Section 3 details the proposed improved methodology. Section 4 explains the experimental setup, training approach, and results. and Section 5 summarizes the key findings and conclusions.

2. Related work

Current fire detection techniques can be categorized into three main approaches: traditional methods relying on image features, target detection algorithms utilizing model.

2.1. Conventional fire detection methods based on image features

Traditional fire detection methods predominantly relied on handcrafted feature extractors, emphasizing characteristics like color, luminance, texture, and edges in images. For instance, Chen et al. [10] proposed an approach based on RGB chromaticity to detect flame and smoke pixels without the need for physical measurements. Similarly, Binti Zaidi et al. [11] leveraged RGB and YCbCr color components, analyzing their specific values to identify fire. Vipin et al. [12] introduced a rule-based model that classified flame pixels by separating luminance and chrominance within the RGB and YCbCr color spaces.

To further enhance detection accuracy, researchers have also investigated texture feature extraction. Dimitropoulos et al. [13] employed background subtraction and color analysis to identify potential ignition regions, followed by modeling fire behavior using spatiotemporal features and dynamic texture analysis. Ye et al. [14] developed a dynamic

texture descriptor based on surface waveform transformations combined with a Hidden Markov Tree model, which was employed for smoke detection in video sequences.

Other approaches have combined color and motion features. For example, Chunyu et al. [15] applied an optical flow algorithm to calculate flame motion features, which were then integrated with color features for video-based fire detection. Li et al. [16] introduced a framework that integrates flame color, dynamic motion patterns, and flicker properties. Although these approaches have enhanced the reliability and precision of fire detection, the complexity inherent in fire scenarios frequently constrains their effectiveness. Handcrafted feature extractors struggle to comprehensively represent object features in such scenarios, leading to a decline in feature extraction precision.

2.2. Fire detection models based on model compression

The primary approaches for model compression and acceleration include network pruning and sparsification [17], lightweight model design [18,19,20], knowledge distillation [21], and compact network architecture development. Techniques such as pruning and sparsification simplify the model by detecting and eliminating redundant parameters or connections, which effectively reduces the computational load and parameter count in neural networks. These methods are often applied to pre-trained models. Alternatively, a compact network architecture can be selected during the initial model design phase.

For instance, C. Szegedy et al. [22] introduced a model architecture grounded in the Hebbian principle and multiscale processing for target detection tasks. Similarly, N. Ma et al. [23] developed ShuffleNetv2, which leverages the ChannelShuffle operation and pointwise grouped convolution to enable efficient feature extraction and information exchange. This design achieves remarkable performance and computational efficiency across multiple computer vision tasks. Furthermore, A. Howard et al. [24] proposed MobileNetv3, which optimizes feature extraction and model compression by employing techniques such as candidate network structure search and network tilting.

These lightweight architectures primarily address the challenges of reducing computational requirements and parameter counts. However, they often come with a trade-off in accuracy, particularly when compared to models like YOLO. While these models excel in specific scenarios, they may struggle to achieve YOLO's level of precision in more complex environments.

2.3. Deep learning based fire detection methods

In recent years, fire detection techniques based on deep learning have primarily employed either single-stage or two-stage strategies. Among these, single-stage methods—such as SSD [4], SPPNet [25], YOLOv3 [26], and YOLOv4 [27] are widely favored due to their ability to quickly and directly predict target categories and locations from input images.In contrast, two-stage methods, including R-CNN [28], Fast R-CNN [29], Faster R-CNN [30], and Mask R-CNN [31], offer higher accuracy but are generally slower, making them widely adopted in target detection tasks. Despite significant progress, these methods face challenges related to high storage and computational resource requirements. To address these limitations, the YOLO series has demonstrated superior performance through continuous iterations and enhancements, particularly in fire detection tasks.

J. Miao et al. [32] introduced an enhanced real-time fire detection algorithm built upon YOLOv5s. This approach incorporates sensory field enhancement along with channel attention mechanisms, aiming to improve both the efficiency and precision of recognizing flames and smoke. Similarly, M. Luo et al. [33] improved YOLOv5s for fire detection by replacing the SPP module with the WASP module and introducing attention mechanisms along with a small-target detection layer, effectively enhancing the detection of small-scale forest fires. Additionally, Li, Pu, and Li, Songbin et al. [34,35] developed fire detection algorithms leveraging target detection CNNs and implicit depth supervision mechanisms, respectively, which addressed the trade-offs between accuracy, model size, and processing speed.Majid et al. [36] Combining EfficientNetB0 with an attention mechanism to propose a fire detection model, real-world fire image dataset achieved good results.Pincott et al. [37] developed a computer vision-based indoor fire and smoke detection system using the Faster R-CNN Inception V2 and SSD MobileNet V2 models, which was initially evaluated with a small training dataset and achieved some results. These approaches addressed prevalent challenges in fire detection algorithms, such as insufficient accuracy and significant latency.

These advancements highlight the potential of deep learning in enhancing fire detection accuracy and reducing false-negative rates. However, existing models still exhibit deficiencies, such as inadequate key feature extraction, limited feature map representation capabilities, suboptimal target loss calculations, and high model complexity. To address these shortcomings, this study adopts YOLOv8n as the benchmark model and aims to improve its capability in detecting flame and smoke boundaries. Key improvements include refining the loss function, incorporating attention mechanisms, optimizing feature fusion, and enhancing feature selection. Simultaneously, efforts are directed toward reducing computational complexity and storage requirements to enable practical deployment and application.

3. Improved methodologies

3.1. Backbone network improve

The backbone network of YOLOv8 uses convolutional and inverse convolutional layers to extract features, using residual connectivity and bottleneck structure to optimise network size and performance. The C2f (Convolution to Fuly Connecied) module is used as the basic building block, but feature redundancy exists after SPPF (Spatial Pyramid Pooling - Fast).

To boost detection accuracy on the fire dataset and refine feature extraction, the dynamic sparse attention mechanism from dynamic sparse attention [38] is applied. Integrated into layer 11 of the backbone network, this mechanism efficiently calculates attention by isolating irrelevant key-value pairs and focusing on the most relevant ones. Leveraging the query-adapted input feature map, the model focuses more on essential key information, reduces the impact of background noise, lowers computational and storage demands, enhances its understanding of the input, and ultimately improves detection accuracy. dynamic sparse attention employs an attention mechanism to capture global feature relationships, offering a superior global perception capability compared to traditional local CNN models. This mechanism functions by encoding the input data sequence,

computing and normalizing the dot product between queries and keys, and then applying weighted summation. The attention formula is presented in Equation (1):where $\sqrt{d_k}$ is the scaling factor to prevent concentration of weights and gradient vanishing.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

The attention mechanism proves useful for conducting a global analysis of images, enabling the extraction of features related to small-scale flames and smoke.However, this approach also increases computational complexity and consequently raises computational costs. The CBAM introduced by Woo et al., employs a dual attention mechanism focusing on spatial and channel dimensions. While it demonstrates strong performance, it suffers from significant computational overhead, making it less efficient and not lightweight. In contrast, the ECA (Efficient Channel Attention) module introduced by Wang et al. minimizes model complexity. However, it demonstrates lower effectiveness in fire detection tasks because of its restricted ability to facilitate channel interactions.To address the challenges of high computational complexity and memory usage associated with conventional attention modules, fire detection platforms constrained by resource limitations cannot afford to integrate these modules. To mitigate these issues, sparse queries are proposed as a resource-efficient alternative to global queries. This concept has inspired research into dynamic sparse attention mechanisms, such as Bi-Level Routing Attention. This approach partitions the input feature map into distinct, non-overlapping regions and uses linear mapping to produce the query, key, and value. An adjacency matrix representing region-to-region affinities is computed by multiplying the region-level query with the transposed region-level key through matrix operations. The routing index matrix, which preserves the top-k connections for each region, is utilized to achieve fine-grained token-to-token attention. The dynamic sparse attention module is ultimately integrated into the backbone network at its 11th layer. This integration includes combining two feature vectors, applying depthwise separable convolution, performing layer normalization, and conducting multilayer perceptron computations. In this context, Q , K,and V refer to the query, key, and value, respectively.$W^q, W^k, W^v \in \mathbb{R}^{(C\times C)}$ The projection weights for the query, key, and value are represented accordingly. The corresponding calculation is provided in Equation (2):

$$Q = X^r W^q, K = X^r W^k, V = X^r W^\nu \tag{2}$$

Constructing a directed graph to determine the regions that should be concerned with each given region, we first derive the region-level queries Q^r and K^r by applying Q and the K to the mean value of each region ,which have dimension $\mathbb{R}^{(S^2\times S^2)}$. Then, by computing the matrix multiplication between Q^r and the transposed K^r , we obtain the adjacency matrix A^r of the region-to-region affinity graph with dimension $\mathbb{R}^{(S^2\times S^2)}$.The computation of the adjacency matrix for inter-region correlation can be expressed as shown in Equation (3):

$$A^r = Q^r (K^r)^T \tag{3}$$

In the neighbourhood matrix, the entry A^r is used to measure how semantically related two regions are. The indexes of the top-k connections are retained row by row using the

routing index matrix $I^r \in \mathbb{N}^{S^2 \times k}$ Using the region-to-region routing index matrix I^r, fine-grained token-to-token attention can be computed. The ith row of matrix I^r contains the indexes of the first k most relevant regions of the ith region, which is calculated as shown in Equation (4):

$$I^r = topkindex(A^r) \tag{4}$$

Using the region-to-region routing index matrix I^r, fine-grained token-to-token attention can be computed. For each query token in region i, the key-value pairs in the concatenation of all k routing regions located in the index set $I^r_{(i,1)}, I^r_{(i,2)}, \ldots, I^r_{(i,k)}$ are processed Since these routing regions are scattered over the entire feature graph, in order to implement this step efficiently, the tensor of the keys and values needs to be collected first and computed as shown in Equations (5) and (6):

$$K^g = gather(K, I^T) \tag{5}$$

$$V^g = gather(V, I^T) \tag{6}$$

The above formulation uses an attention operation on the collected (gather) key-value pairs and introduces a local context augmentation term LCE(V), where LCE(V) is parameterised using a depth-separable convolution with a convolution kernel size of 5. The computation is shown in Equation (7):

$$O = Attention(Q, K^g, V^g) + LCE(V) \tag{7}$$

The module employs a two-level routing attention mechanism that is incorporated into Layer 11 of the backbone network to enhance the model's attention to critical target details, thereby improving detection accuracy.

3.2. Optimisation of neck network

Drawing on the EfficientCepBiPAN idea of YOLOV6 [39], a new neck network is designed, which includes two key components: the MPFusion module and the CepBlock module.The MPFusion module is optimised for the up- and down-sampling part of the original model, which incorporates an attention mechanism before the up- and down-sampling, and the three adjacent layers in a cascade operation to fuse the low-level features in the trunk to the high-level features in the neck, so that more accurate position signals are retained in the process of feature fusion, and efficient fusion of multi-scale feature maps and large and small target information is achieved. This process not only strengthens the model's capacity to detect targets but also improves its comprehension of image content.The design of the MPFusion module helps to cope with the scale differences of diverse targets in the real scene, so as to capture the target's features in a more comprehensive way. Inspired by RepBlock, the CepBlock module is designed, which is mainly optimised for the large perceptual field of the model. During the training phase, a 5×5 convolutional kernel is introduced to expand the perceptual field while maintaining the network's depth. The four branches in this phase are employed separately for feature extraction, with each branch undergoing a distinct reparameterization process. Specifically, the 1×1 convolution is reconfigured using padding with 3×3 and 5×5 convolutional

kernels, with multiple instances of the 5×5 kernel applied. convolution kernel, 5 × 5 convolution kernel through the weighted average of neighbouring weights compressed into a 3 × 3 convolution, there is no convolution kernel of the residual channel to construct a class of convolutional input and output, that is, multiply a unit matrix can be, after the convolution layer and the BN layer fusion of the addition operation and then the output.CepConv is a 3 × 3 convolutional and the LeakyReLU activation function of the stack, Leaky ReLU can effectively solve the 0-gradient problem in the case of negative input, compared to the ordinary convolution block, less BN layer, the core idea is the fusion of Conv2d and depth-separable convolution, and finally directly add the parameters of these three convolutional layers to fuse them into an equivalent 3×3. Since 3×3 convolution has a high degree of optimisation on mainstream GPUs and CPUs, and has a high computational density, this design can greatly accelerate the inference speed. In the inference stage, the CepVGG block is transformed into CepConv, which can effectively accelerate the inference process using single branching. The MPFusion and CepBlock modules collaborate effectively, complementing each other to enhance the accuracy of information localization during the neck network's feature fusion process. The feature representation capability is enhanced by strengthening feature interactions and filtering out irrelevant information, allowing the model to better capture and understand target features. This improvement boosts the model's target discrimination ability. Additionally, the inference speed is increased without compromising accuracy, offering a practical approach for real-time target detection tasks. The designs of the two modules are illustrated in Figure 2, parts (a) and (b).

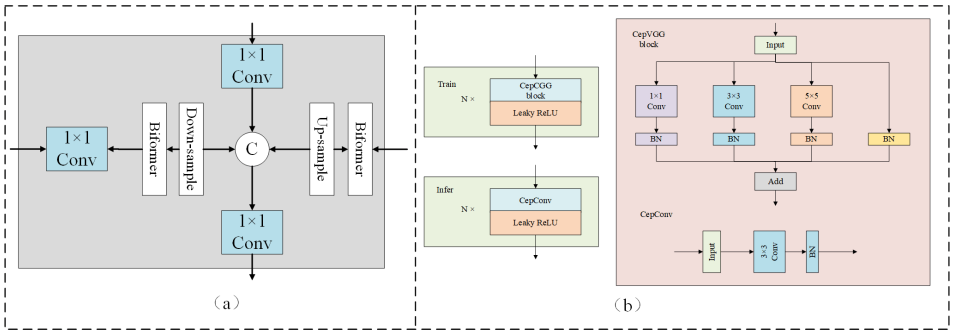


Fig. 2. The structure of MPFusion and CepBlock

3.3. Lightweight detection head reconfiguration

In the redesigned YOLOv8, the detection head has been optimized with a focus on lightweighting, aiming to overcome challenges related to model storage and execution. The original YOLOv8's detection head adopts a decoupled head structure and has been changed from Anchor-Based to Anchor-Free by removing the objectness branch and adopting two parallel branches, which are responsible for the extraction of category features and positional features respectively. However, this results in an increased parameter

count, which significantly raises the demand for storage and computational resources. Our new detection head design maintains high accuracy while keeping lightweight. Specifically, we have designed the original 3×3 convolution of the two branches to merge and share parameters, while employing a layer of 1x1 convolution for both classification and localisation tasks. This design helps decrease the model's parameter count, which in turn lowers storage demands, making the model more efficient to deploy and operate. By employing parameter sharing and applying optimization techniques, we effectively improve the overall performance of the model. The new detection head structure is shown in Figure 3, with two parallel branches performing classification and localisation tasks through a layer of 1x1 convolution, which is designed to remain lightweight while still being able to quickly process edge feature information and extract classification features. Compared to the original detection head of YOLOv8, our design reduces the storage footprint while still ensuring high detection accuracy. This improvement not only makes the model more suitable for resource-limited environments, but also accelerates the training and inference speed and improves the overall performance.

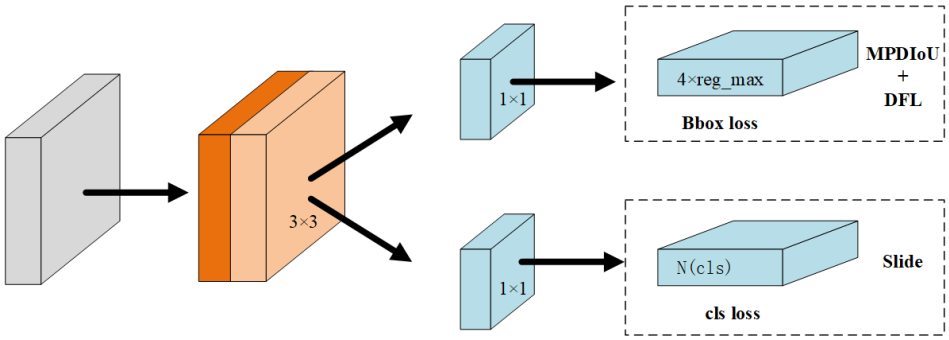


Fig. 3. The structure of the lightweight detection head

3.4. Optimisation of loss function

YOLOv8 adopts an anchorless design with a significant change in the loss function compared to the YOLOv5 series. The optimization goal is divided into two primary aspects: regression and classification. The classification component utilizes the sample weighting function (Slide Loss), whereas the regression process relies on the Distributional Keypoint Loss (DFL) along with the bounding box regression loss (MDPIoU). The complete loss function calculation is shown in Equation (8):

$$F_{loss} = \alpha_1 F_{Slideloss} + \alpha_2 F_{DFL} + \alpha_3 F_{MPDIoU} \tag{8}$$

To tackle the problem of sample imbalance in target detection tasks, Slide Loss has been introduced. Slide Loss primarily aims to balance samples with different difficulty levels by dynamically modifying their weights. The difficulty for each sample is evaluated based on the IoU values calculated between the predicted bounding boxes and the ground

truth. To minimize the inclusion of additional hyperparameters, the average IoU value across all bounding boxes is used as the threshold, denoted as μ with IoU values below μ are classified as negative, while those with IoU values above μ are categorized as positive. The calculation process is detailed in Equation (9):

$$\mathrm{F(x)} = \begin{cases} 1, x \leq \mu - 0.1 \\ e^{1-\mu}, \mu < x < \mu - 0.1 \leftarrow \\ e^{1-x}, x \geq \mu \end{cases} \tag{9}$$

To fully utilize samples with ambiguous classifications and those located near decision boundaries, Slide Loss is introduced. This method addresses challenging samples by categorizing them as positive or negative based on the parameter μ. Additionally, the Slide weighting function assigns greater importance to boundary samples by giving them higher weights, thereby enhancing the model's attention to classification-challenging cases.

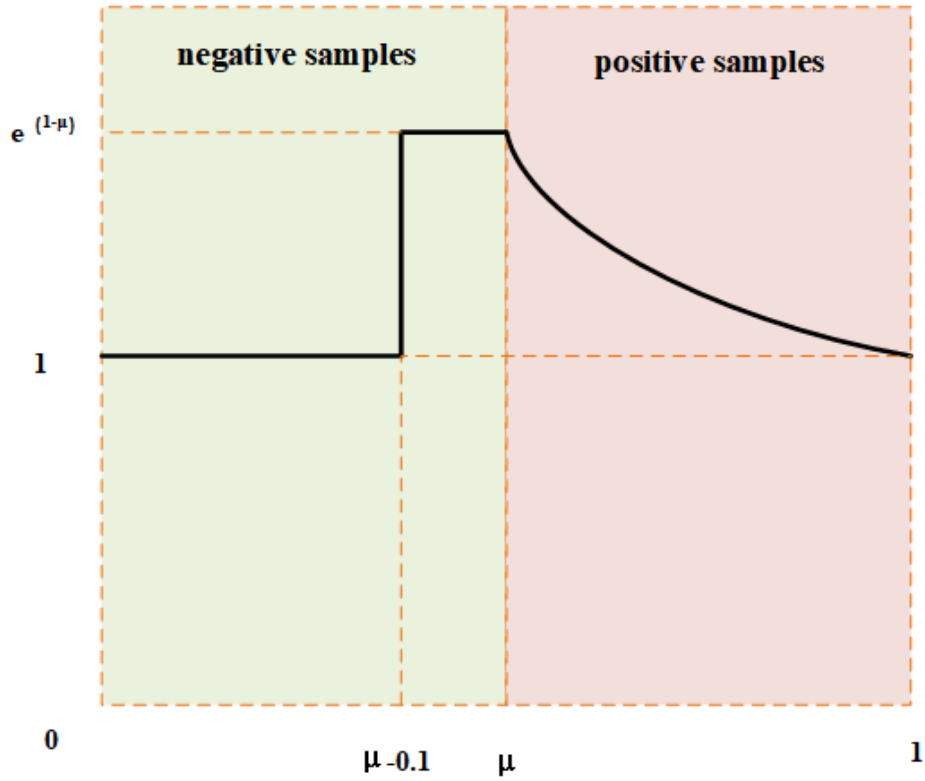


Fig. 4. Slide weighting function image

Slideloss [40], illustrated in Figure 4, represents a sliding loss function that adaptively determines the threshold parameters: μ for positive samples and μ for negative samples.By setting higher weights around μ the loss of difficult, incorrectly categorised examples

can be increased,which approach significantly enhances the model's classification performance, particularly for boundary cases and challenging samples.

There are many targets in the dataset with the same aspect ratio but inconsistent scaling. To solve this problem, this paper introduces MDPIoU [41] as an optimisation method for bounding box regression loss. For any convex shapes A and B, the widths and heights are denoted as w and h . The coordinates represent the upper-left and lower-right corner points of shapes A and B, respectively. $(x_1^A, y_1^A), (x_2^A, y_2^A)$ and $(x_1^B, y_1^B), (x_2^B, y_2^B)$ respectively,The derivation process of MDPIoU is shown in Equations (10) to (12):

$$d_1^2 = (x_1^B - x_1^A)^2 + (y_1^B - y_1^A)^2 \tag{10}$$

$$d_2^2 = (x_2^B - x_2^A)^2 + (y_2^B - y_2^A)^2 \tag{11}$$

$$MDPIoU = \frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \tag{12}$$

In the training phase, the set of predicted values for each bounding box predicted by the model is forced to approximate the set of true bounding boxes by minimising the loss function, hence for MPDIoU based loss function is defined as shown in Equation (13):

$$L_{MDPIoU} = 1 - MDPIoU \tag{13}$$

The coordinates of the four points can be used to derive all components of the current bounding box regression loss function. The transformation steps are outlined in Equations (14) to (18):

$$\begin{aligned} C_x &= \max(x_2^{gt}, x_2^{prd}) - \min(x_1^{gt}, x_1^{prd}) \\ C_y &= \max(y_2^{gt}, y_2^{prd}) - \min(y_1^{gt}, y_1^{prd}) \\ |C| &= C_x \cdot C_y \end{aligned} \tag{14}$$

$$x_c^{gt} = \frac{x_1^{gt} + x_2^{gt}}{2}, \quad y_c^{gt} = \frac{y_1^{gt} + y_2^{gt}}{2} \tag{15}$$

$$x_c^{prd} = \frac{x_1^{prd} + x_2^{prd}}{2}, \quad y_c^{prd} = \frac{y_1^{prd} + y_2^{prd}}{2} \tag{16}$$

$$w^{gt} = x_2^{gt} - x_1^{gt}, \quad h^{gt} = y_2^{gt} - y_1^{gt} \tag{17}$$

$$w^{prd} = x_2^{prd} - x_1^{prd}, \quad h^{prd} = y_2^{prd} - y_1^{prd} \tag{18}$$

Here, $|C|$ denotes the area of the smallest rectangle that encloses both B_{gt} and P_{rd} , $\left(x_c^{gt}, y_c^{gt}\right)$ and (x_c^{prd}, y_c^{prd}) represent the coordinates of the centre points of the groundtruth bounding box and the prediction bounding box, respectively, ${}_wgt$ and ${}_hgt$ represent the width and height of the groundtruth bounding box, ${}_Wprd$ and ${}_hprd$ represent the width and height of the prediction bounding box.According to From the coordinates of the upper-left and lower-right points, all factors present in existing loss functions—such as non-overlapping areas, distances between centroids, and variations in width and height—can be derived, as shown in Equations (16) to (18).This demonstrates that the MDPIoU utilized in our approach is both thoughtfully designed and computationally efficient. Figure 5 illustrates the loudness factor.

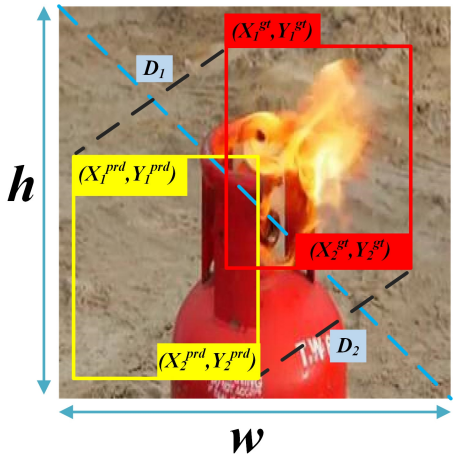


Fig. 5. Factors affecting MDPIoU

4. Experiments

4.1. Experimental environment

The environment and hardware platform parameters for the training phase of the experiment are shown in Table 1:

Table 1. Experimental environment configuration

parameter	configure
CPU	Intel Xeon Silver 4214R
GPU	NVIDIA GeForce RTX 3080 Ti
operating system	Ubuntu 18.04.5
Architecture	torch-1.9.0+cu111

During the training process, we set the key parameters according to Table 2.

4.2. Introduction to datasets

To comprehensively assess the performance of our enhanced algorithm, we chose two publicly accessible datasets: one from the Roboflow platform and the D-Fire fire detection dataset. Both datasets offer extensive annotations for flames and smoke across various scenarios, enabling a robust assessment of our algorithm's performance in complex environments.

The D-Fire dataset consists of 21,527 images featuring flames and smoke, averaging 2.52 bounding boxes per image. Conversely, in the categories labeled as "Smoke" and

Table 2. Experimental parameter configuration

parameter	settings
Epochs	150
Momentum	0.937
Initial learning rate	0.01
Final learning rate	0.01
Weight decay	0.005
Input image size	640 × 640
Optimizer	SGD
Data enhancement	Mosaic
Box Loss decay	7.5
Cls Loss decay	0.5
Batch size	32

”Fire and Smoke,” the average number of smoke-labeled frames is 1.13 bounding boxes per image. Altogether, the dataset comprises 26,557 bounding boxes, with 11,865 annotated as smoke and 14,692 identified as fire.

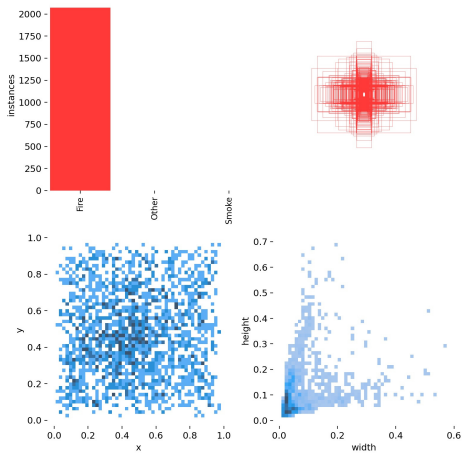


Fig. 6. Label distribution of the dataset

The dataset from the Roboflow platform features images of fire smoke captured in diverse environments, including both indoor and outdoor scenarios. The dataset is split randomly into three subsets: training, validation, and testing, following a 7:2:1 ratio. Specifically, the training set consists of 4,620 images, the test set includes 1,320 images, and the validation set comprises 660 images. There are three distinct types of annotations, as illustrated in Figure 6. The first subfigure highlights the quantity of various fire-related objects. The second subfigure presents the bounding box sizes, with all their center points aligned at a single location, suggesting a prevalence of small object regions within the

dataset. The third subfigure depicts the distribution of bounding box center coordinates, revealing that the majority of center points are clustered around the central region of the image.Finally, the fourth subfigure presents a scatter plot showing the widths and heights of bounding boxes. The darkest area in the lower-left corner highlights that the dataset primarily consists of small objects.

From the analysis of the dataset, it can be concluded that it predominantly consists of numerous small objects with a dense yet uneven distribution. Compared to traditional datasets used in computer vision tasks, this dataset is significantly larger and includes a variety of scales, scenes, and angles, making it more challenging than standard computer vision datasets. To enhance the model's performance and refine its development, this paper employs data augmentation techniques such as cropping, scaling, and color perturbation to improve data quality and increase diversity. The YOLOv8n model served as the baseline, with several ablation experiments performed to assess how each improvement strategy affected its performance, leading to the identification of the best configuration. Moreover, mosaic data augmentation was utilized in the last 10 training epochs to enhance the speed of model convergence.

4.3. Evaluation indicators

This experiment uses both the model itself metrics and TIDE metrics to measure the performance of the model in this paper at the same time, calculated as shown in Equations (19) to (21):

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

$$AP = \int_0^1 \text{Precision(Recall)d(Recall)} \tag{21}$$

The mean Average Precision (mAP) across all categories is derived by calculating the weighted average of the AP values for each sample category. This metric evaluates the model's detection performance across all categories and is computed as illustrated in Equation (22):

$$mAP = \frac{1}{K} \sum_{i=1}^{K} AP_i \tag{22}$$

APi in Equation (22) denotes the relationship between the value and the value of the category index.K denotes the number of categories of the samples in the trained dataset, and the value in this paper is 3.

To compare model runtime, this paper adopts Frames Per Second (FPS) as the performance metric. FPS, which indicates the number of image or video frames the model can process each second, is used to evaluate runtime efficiency. In order to further capture the more valuable error distributions in mAP, all FPs and FNs are grouped into six types, and FPs and FNs can be paired in some cases, and IoUmax is used to denote the overlap of the maximum IoU of an FP with the ground truth of a given category, with the foreground

IoU threshold denoted as tf and the background threshold denoted as tb. The following are the definitions of the six types of errors and the rules of determining them .

Classification error (**Cls**), i.e., for misclassification, $\text{IoU}_{\text{max}} \geq t_f$ (i.e., localisation is correct but misclassified).

Localisation error (**Loc**), i.e., for correct classification, $t_b \leq \text{IoU}_{\text{max}} \leq t_f$ (i.e., classification is correct but localisation is incorrect).

Both classification and localisation error (**Cls+Loc**), i.e., for misclassification, $t_b \leq \text{IoU}_{\text{max}} \leq t_f$ (i.e., classification is incorrect and incorrectly localised).

Duplicate Detection Error (**Duplicate**), i.e., correctly classified and $\text{GTIoU}_{\text{max}} \geq t_f$, but another higher-scoring test has already matched the GT (i.e., would have been correct if not for the higher-scoring test).

Background Error (**Bkgd**), i.e., $\text{IoU}_{\text{max}} \leq t_b$ for all GTs (detecting the background as the foreground).

Undetected GT errors (**Missed**), i.e., all undetected ground truth (FN) not covered by classification or localisation errors.

Smaller values of the above six metrics represent smaller errors and superior model performance.

4.4. Experiment 1:experimental results and analysis

Optimal improved position experiments with dynamic sparse attention The first few layers of the backbone network usually have smaller feature map size and depth, and it may be relatively faster to apply the attention mechanism at these layers to reduce the impact on the overall inference speed. To achieve optimal performance and support subsequent ablation experiments, comparative experiments were carried out in this study. Specifically, by enhancing the neck network, the dynamic sparse attention module was integrated into different layers of the backbone network, with the resulting experimental data summarized in Table 3. Specifically, dynamic sparse attention modules were applied individually to layers 3, 5, 7, 9, and 11 of the backbone network. The data in Table 3 demonstrate that the dynamic sparse attention11 model achieves superior detection performance, exhibiting a notable accuracy improvement and the lowest scores across all six error type indices when compared to other models. As a result, dynamic sparse attention11 was chosen as part of this paper's proposed improvement strategy. The experimental results demonstrate that incorporating the dynamic sparse attention module allows the network to more effectively identify and highlight important features within the image.

Table 3. Comparison of YOLOv8 with different dynamic sparse attention configurations

Model	AP	AR	mAP50	mAP50-95	FPS/bs32	Size/MB	GFlops	Cls	Loc	Both	Dupe	[c]Bkg	[c]Miss
YOLOv8+dynamic sparse attention3	0.951	0.912	0.961	0.879	**419**	**2.86**	3.01	0.11	1.38	**0.02**	0.09	0.76	0.55
YOLOv8+dynamic sparse attention5	0.948	0.921	0.959	0.881	412	**2.86**	3.01	0.12	1.45	0.06	0.08	0.73	0.61
YOLOv8+dynamic sparse attention7	0.957	0.915	0.948	0.880	409	**2.86**	3.01	**0.09**	1.31	0.03	**0.06**	0.88	0.58
YOLOv8+dynamic sparse attention9	0.946	**0.935**	0.964	0.882	406	**2.86**	3.01	0.11	1.35	0.04	**0.06**	0.86	0.71
YOLOv8+dynamic sparse attention11	**0.971**	0.919	**0.967**	**0.888**	400	**2.86**	**3.01**	**0.09**	**1.29**	**0.02**	**0.06**	**0.67**	**0.49**

Analysing Table 3 shows that the performance of adding dynamic sparse attention to the backbone network is superior, with most of the metrics improved, the recall has de-

creased probably because the features extracted at layer 11 are not discriminative enough, and despite the decrease in recall, the overall increase in performance shows the positive impact of the addition of dynamic sparse attention to the 11th layer of the backbone network. The improvement in other metrics suggests that it has advantages in improving detection precision and accuracy.

Experiment 2: ablation experiment This paper further explores the impact of different improvement strategies to evaluate the enhancement of the improvement model by gradually adding each module. The specific programmes are as follows.

(1) YOLOv8n: original model for target detection.

(2) YOLOv8n+A: Replace the original BCElosss with Slideloss.

(3) YOLOv8n+B: Use MPDIoU to replace the original CIoU.

(4) YOLOv8n+D: Improve the neck network.

(5) YOLOv8n+C+D: Improve the backbone network by adding dynamic sparse attention to (4).

(6) YOLOv8n+C+D+E: using optimised lightweight detection head on top of (5).

(7) YOLOv8n+C+E: removing improvements to the neck network from (6).

(8) YOLOv8n+A+B+C+D+E: combining (2),(3),and (6).

The detection performance and model parameters of each model are listed in the Table 4, where YOLOv8n is the baseline model. The tables and images are briefly analysed below: YOLOv8n:The baseline model, with most of its metrics, is only ranked in the middle of the pack in the ablation experiments, which suggests that even if a model does not have a high number of model parameters in it, it still leads to a long inference time due to the redundant network layers. The final FPS index of the improved model reaches 434, which can ensure the real-time requirement in real deployment. YOLOv8n+A:Using the strategy of improvement A on top of the baseline model, the recall and map50 scores of the model are slightly improved to 0.975 and 0.944 respectively, which demonstrates the reasonableness of the improvement in solving the problem of imbalance between difficult and simple samples, while the the decrease in Cls indicates a substantial improvement in the classification error. YOLOv8n+B: Adopting the B improvement strategy based on the baseline model, the map50-95 , AR and AP of this model slightly decreased respectively, which may be due to the large differences between targets affecting the overall performance of recall and accuracy. However, map50 is improved, while its Loc and Miss metrics are decreased indicating the superiority of the frame loss improvement. YOLOv8n+D:The D improvement strategy was used on the basis of the baseline model, which fused feature representations of different layers, reduced the number of channels and convolutional kernel size of some layers, and increased the convolutional layers, with the number of parameters and the amount of computation reduced by 10% and 11%, respectively. All other indexes are improved, which indicates that the model after the improved neck network makes a substantial improvement in the detection performance and running speed of the model. YOLOv8n+A+B+C+D+E: This model has improved all indexes compared with YOLOv8n, and the error rate is obviously reduced, which indicates that there is a large improvement in the detection performance, and the overall performance is higher, therefore, we identified this model as the optimal improved model, which shows that the improvement of the baseline model is feasible and effective and reasonable considering the accuracy and speed of special equipment and detection scenarios.

Table 4. Comparison of ablation experiment indicators

Models	AP	AR	mAP50	mAP50-95	FPS/bs32	Size/MB	GFlops	Cls	Loc	Both	Dupe	Bkg	Miss
YOLOv8n	0.968	0.938	0.970	0.898	400	2.86	8.1	0.18	0.77	0.01	0.03	0.71	0.42
YOLOv8n+A	0.975	**0.944**	**0.977**	0.902	400	2.86	8.1	**0.03**	0.86	0.02	0.04	0.74	0.16
YOLOv8n+B	0.966	0.928	0.971	0.890	384	2.86	8.1	0.08	0.70	0.01	0.07	0.86	0.27
YOLOv8n+D	0.968	0.937	0.972	0.897	434	**2.57**	7.2	0.10	0.75	0.01	0.05	0.68	0.13
YOLOv8n+C+D	0.971	0.919	0.967	0.888	384	2.82	7.2	0.09	1.29	0.02	0.06	0.67	0.49
YOLOv8n+C+D+E	0.959	0.922	0.961	0.880	416	2.88	6.5	0.20	**0.72**	0.01	0.11	1.01	0.09
YOLOv8n+C+E	0.966	0.935	0.971	0.893	**454**	3.91	8.1	0.08	0.78	0.01	0.06	0.72	0.21
YOLOv8n+A+B+C+D+E	**0.981**	0.940	0.974	**0.907**	434	2.88	**6.5**	0.08	0.76	**0.01**	0.04	**0.49**	**0.08**

Experiment 3:comparative experiments n the field of target detection, deep learning methods are classified into level 1 and level 2 categories, distinguished by their anchor generation mechanisms. In real engineering scenarios, real-time processing of fire and smoke images is more in line with practical needs. Therefore, in order to combine both accuracy and hardware dependency considerations, it is more practical to choose the level 1 target detection method. In this experiment, we selected the YOLO series as the object of comparison test, including advanced and general models such as YOLOv5-s, YOLOv3, YOLOv3-tiny and YOLOv6. These models have been widely used in a variety of embedded scenarios and published in several papers. To emphasize the advantages of the models used in this experiment, we selected the enhanced versions developed in this study for comparison. Comparison experiments are conducted with YOLOv5-s, YOLOv3, YOLOv3-tiny and YOLOv6. To ensure fairness, no pre-training weights were used in all model training processes.The outcomes of the comparative experiments are presented in detail in Table 5 and Table 6. Review these tables to gain insights into the findings.

(1) YOLOv3 has the highest number of parameters, totaling 103,666,553 bytes. While its overall performance metrics are impressive, its real-time speed is limited to 66 FPS, making it unsuitable for the real-time detection tasks required in this study.

(2) The YOLOv5-s model is highly lightweight. however, its precision is low on this dataset, suggesting a high false detection rate. Despite this, it demonstrates better performance in terms of recall. The primary reason for these observations lies in the complexity of the dataset's background and the significant variation in target sizes. These factors negatively impact the precision, but the improved version of the model is better suited for the task at hand.

(3) YOLOv3-tiny is also a lightweight model, most of the indicators are low, compared to the performance of YOLOv5-s is still insufficient, which is also YOLO after several versions of iteration much led to make YOLOv5 performance has been greatly improved.

(4) YOLOv6 has the lowest recall rate, indicating that the model has a certain leakage rate, fire smoke detection task, not only requires high real-time, in the leakage rate requirements are also more stringent, the poor performance of this important indicator, making the model is not suitable for the task in this paper.

(5) Our model comprehensive comparison of other models in the same series, the overall performance of the best, especially in the FPS this indicator is excellent, while the model in basically does not change the number of far away model parameters, in the deployment difficulty and real-time and other convenient to meet the real-life engineering needs, while the model also has good robustness, accuracy and practicality.

Table 5. Comparison of experimental indexes of data sets of each model on roboflow

Models	AP	AR	mAP50	mAP50-95	FPS/bs32	Size/MB	GFlops	Cls	Loc	Both	Dupe	Bkg	Miss
YOLOv5-s	0.963	0.962	0.972	0.894	270	2.86	7.2	0.11	1.02	0.01	0.03	0.86	0.71
YOLOv3	0.968	0.961	0.981	0.920	66	98.86	282.2	0.44	0.83	0.23	0.19	0.65	0.28
YOLOv3-tiny	0.959	0.935	0.973	0.895	285	4.04	11.8	0.58	1.19	0.19	0.21	0.67	0.25
YOLOv6	0.955	0.917	0.960	0.853	277	11.56	18.9	0.16	1.35	0.10	0.13	0.66	0.32
TOOD	0.972	0.931	0.978	0.901	386	31.8	125.9	0.12	0.86	0.03	0.08	0.59	0.16
Our	0.981	0.938	0.974	0.907	434	2.88	6.5	0.08	0.83	0.01	0.06	0.49	0.08

Table 6. Comparison of the experimental metrics for each model on the D-Fire dataset

Models	AP	AR	mAP50	mAP50-95	FPS/bs32	Size/MB	GFlops	Cls	Loc	Both	Dupe	Bkg	Miss
YOLOv5-s	0.761	0.731	0.772	0.458	270	2.86	7.2	0.14	1.16	0.04	0.05	0.92	0.84
YOLOv3	0.777	0.759	0.781	0.481	66	98.86	282.2	0.49	0.95	0.29	0.24	0.74	0.36
YOLOv3-tiny	0.765	0.721	0.773	0.453	285	4.04	11.8	0.68	1.27	0.24	0.24	0.79	0.23
YOLOv6	0.753	0.711	0.762	0.414	277	11.56	18.9	0.19	1.44	0.14	0.17	0.73	0.42
TOOD	0.768	0.727	0.766	0.451	386	31.8	125.9	0.13	0.78	0.05	0.07	0.62	0.19
Our	0.786	0.728	0.789	0.467	434	2.88	6.5	0.12	0.93	0.03	0.09	0.55	0.17

Experiment 4:individual comparison with YOLOv8 To verify the impact of the enhanced model on detection performance, we carried out a comparative experiment between the improved model and the baseline model, YOLOv8n. Figure 7 shows the trend analysis of precision, recall, mAP50 and mAP50-95 for AMMF-Detection (orange curve) and YOLOv8n (blue curve) on the validation dataset. The metrics show rapid improvement throughout the iterations and gradually approach stable values, and AMMF-Detection ends up with higher convergence values.

Experiment 5:visualisation and analysis The interpretability of deep learning models is a key issue limiting their application and development, thus becoming a research hotspot in artificial intelligence. We evaluated the model performance in terms of confusion matrix, feature map visualisation and inference experiments through comparative experiments. The confusion matrix intuitively reflects the classification accuracy, the feature map visualisation demonstrates the distribution of the model's attention to the target, and the inference experiment verifies the model's generalisation ability and robustness on a new fire image dataset. These methods comprehensively reveal the strengths and limitations of the model.

As illustrated in Figure 8, the diagonal indicator region within the confusion matrix for AMMF-Detection is notably higher compared to YOLOv8n. This suggests an improved capability of our model in accurately classifying target categories. Also, the proportion of objects whose backgrounds are judged to be flames has been reduced, which means that the improved model reduces the miss detection rate for this category, but the accuracy for smoke has been reduced, which is due to the complexity and polygonal shape of the smoke itself. A comparison of the heat maps reveals that both models exhibit a certain degree of false detection, with the flame frequently being misclassified as background. To address this, we selected three representative images to visualize their feature maps.

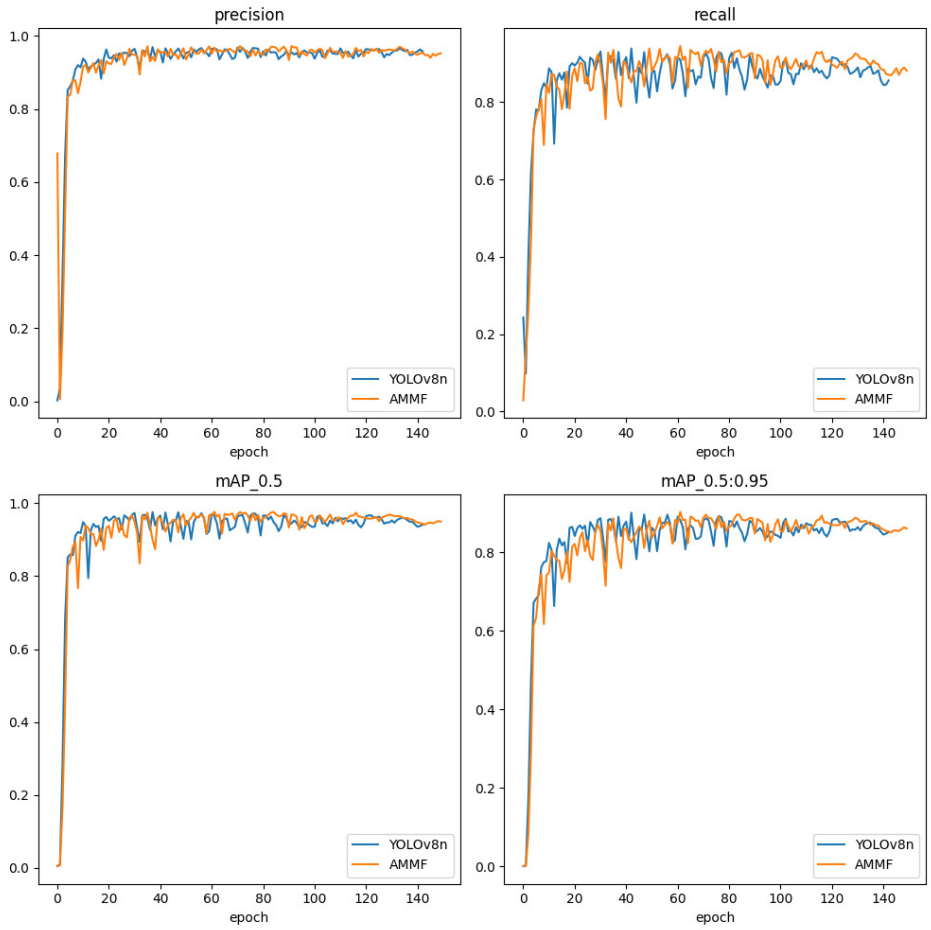


Fig. 7. Trend analysis of the indicators of the model training process

This visualization offers a clearer and more intuitive way to observe the model's focus, which is highly beneficial for improving the overall detection performance. In the same image, it is evident that AMMF-Detection focuses on a broader range, including similar target objects. The original model, however, demonstrates a notable false detection rate and insufficient confidence levels for the detected targets. In contrast, the improved model exhibits a more precise focus, reduces the error-prone regions, and achieves higher confidence scores. A detailed comparison of the feature maps is presented in Figure 9.

To evaluate the generalization ability of the proposed method, a diverse dataset of images was collected, and inference experiments were performed. These images feature numerous small objects, posing significant challenges for detection. The inference experiment results, presented in Figure 10, include detection outcomes for indoor and outdoor targets of varying sizes. As shown in Figure 10, our method achieves high-quality detection performance across diverse and complex environments. The model demonstrates

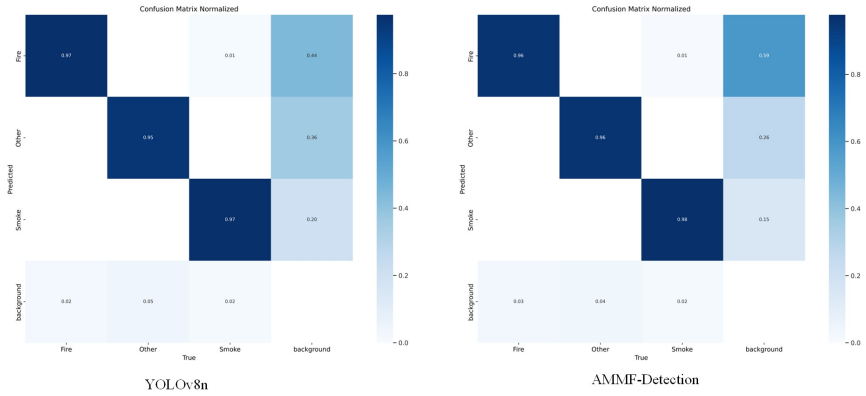


Fig. 8. Comparison of confusion matrices

Fig. 9. Comparison of heatmap visualizations

minimal instances of missed detections, effectively showcasing the adaptability and robustness of the proposed approach across different scenarios.

From the experimental results, it can be concluded that the proposed approach successfully identifies target objects in various and challenging environments, encompassing both indoor and outdoor scenarios. Additionally, it demonstrates the capability to handle target objects of various sizes. These findings highlight the method's strong adaptability and generalization capabilities, making it suitable for application in numerous real-world scenarios.

5. Conclusion

In this paper, we introduce an effective and streamlined fire detection model, termed AMMF-Detection, which is an optimization based on YOLOv8. This model addresses the challenges of bounding box optimization and sample imbalance in fire detection tasks by incorporating the MPDIoU bounding box distance metric and the SlideLoss classification loss function. Moreover, the integration of the dynamic sparse attention mechanism

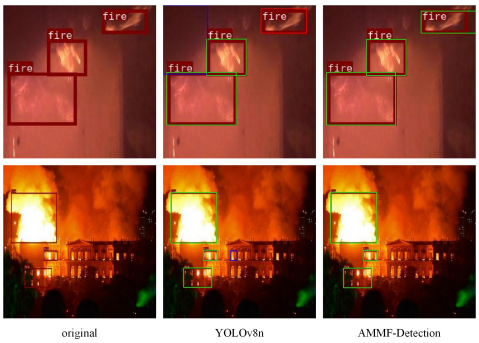


Fig. 10. Inference experiment results

improves the model's ability to capture global contextual information and understand image content. Additionally, the neck network is redesigned by incorporating the CepBlock module and the MPFusion module, further refining the overall architecture. Finally, the detection head is restructured to achieve a lightweight design, reducing the model's computational complexity. Experimental findings reveal that the optimized model attains an average precision of 97.4% at a 50% recall rate and 90.7% across a recall range of 50% to 95%.Additionally, the frames per second (FPS) metric improves from 400 to 434. The fire detection model presented in this paper holds significant practical applications. Future studies can further optimize the model's performance and validate its application in various other domains and tasks.The reconstruction of the neck network in our improved model introduces complexity and a relatively high number of feature fusion steps, leading to variations in model size and inference time. There is still potential to optimize the model's computational consumption. Future work will prioritize investigating distillation and pruning techniques to compress the model's parameters and structure, aiming to strike a balance between complexity and performance, thereby improving its efficiency and overall performance.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant NO.62076006), the Opening Foundation of the State Key Laboratory of Cognitive Intelligence, iFLYTEK (Grant NO.COGOS-2023HE02), and the University Synergy Innovation Program of Anhui Province (Grant NO.GXXT-2021-008).

References

1. Y. Wang, Y. Han, Z. Tang, et al. A Fast Video Fire Detection of Irregular Burning Feature in Fire-Flame Using in Indoor Fire Sensing Robots. *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.
2. L. Zhang, C. Lu, H. Xu, et al. MMFNet: Forest fire smoke detection using multiscale convergence coordinated pyramid network with mixed attention and fast-robust NMS. *IEEE Internet of Things Journal*, 2023.
3. M. Mueller, P. Karasev, I. Kolesov, et al. Optical flow estimation for flame detection in videos. *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2786–2797, 2013.

4. W. Liu, D. Anguelov, D. Erhan, et al. SSD: Single shot multibox detector. In *Computer Vision–ECCV*, Springer, pp. 21–37, 2016.
5. J. Redmon, S. Divvala, R. Girshick, et al. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.
6. J. Redmon and A. Farhadi. YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263–7271, 2017.
7. C. Y. Wang, I. H. Yeh, H. Y. Mark Liao.Yolov9: Learning what you want to learn using programmable gradient information.In *European Conference on Computer Vision*. Springer, Cham, 2025, pp. 1-21.
8. A. Wang, H. Chen, L. Liu, et al. Yolov10: Real-time end-to-end object detection.*arXiv preprint arXiv:2405.14458*, 2024.
9. W. Lv, S. Xu, Y. Zhao, et al. DETRs beat YOLOs on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2023.
10. T. H. Chen, P. H. Wu, Y. C. Chiou. An early fire-detection method based on image processing. In *Proceedings of the 2004 International Conference on Image Processing (ICIP)*, vol. 3, pp. 1707–1710, IEEE, 2004.
11. N. I. binti Zaidi, N. A. A. binti Lokman, M. R. bin Daud, et al. Fire recognition using RGB and YCbCr color space. *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 21, pp. 9786–9790, 2015.
12. V. Vipin. Image processing based forest fire detection. *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 2, pp. 87–95, 2012.
13. K. Dimitropoulos, P. Barmpoutis, N. Grammalidis. Spatio-temporal flame modeling and dynamic texture analysis for automatic video-based fire detection. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 2, pp. 339–351, 2014.
14. W. Ye, J. Zhao, S. Wang, et al. Dynamic texture based smoke detection using Surfacelet transform and HMT model. *Fire Safety Journal*, vol. 73, pp. 91–101, 2015.
15. Y. Chunyu, F. Jun, W. Jinjun, et al. Video fire smoke detection using motion and color features. *Fire Technology*, vol. 46, pp. 651–663, 2010.
16. Z. Li, L. S. Mihaylova, O. Isupova, et al. Autonomous flame detection in videos with a Dirichlet process Gaussian mixture color model. *IEEE Transactions on Industrial Informatics*, vol. 14, no. 3, pp. 1146–1154, 2017.
17. S. Han, H. Mao, W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.
18. M. Courbariaux, Y. Bengio, J. P. David. BinaryConnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.
19. M. Courbariaux, I. Hubara, D. Soudry, et al. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. *arXiv preprint arXiv:1602.02830*, 2016.
20. I. Hubara, M. Courbariaux, D. Soudry, et al. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research*, vol. 18, no. 187, pp. 1–30, 2018.
21. J. Yim, D. Joo, J. Bae, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4133–4141, 2017.
22. C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.
23. A. G. Howard, M. Zhu, B. Chen, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04825*, 2017.
24. A. Howard, M. Sandler, G. Chu, et al. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.

25. K. He, X. Zhang, S. Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
26. J. Redmon and A. Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
27. A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.
28. R. Girshick, J. Donahue, T. Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580-587, 2014.
29. R. Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440-1448, 2015.
30. S. Ren, K. He, R. Girshick, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137-1149, 2016.
31. K. He, G. Gkioxari, P. Dollár, et al. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961-2969, 2017.
32. J. Miao, G. Zhao, Y. Gao, et al. Fire detection algorithm based on improved YOLOv5. In *2021 International Conference on Control, Automation and Information Sciences (ICCAIS)*, IEEE, pages 776-781, 2021.
33. M. Luo, J. Huang, X. Sun, et al. Small Target Forest Fire Recognition Method based on Deep Learning. In *2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, IEEE, volume 3, pages 593-597, 2023.
34. P. Li and W. Zhao. Image fire detection algorithms based on convolutional neural networks. *Case Studies in Thermal Engineering*, 19:100625, 2020.
35. S. Li, Q. Yan, and P. Liu. An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism. *IEEE Transactions on Image Processing*, 29:8467-8477, 2020.
36. S. Majid, F. Alenezi, S. Masood, M. Ahmad, E. S. Gündüz, and K. Polat. Attention based CNN model for fire detection and localization in real-world images. *Expert Systems with Applications*, 189:116114, 2022.
37. J. Pincott, P. W. Tien, S. Wei, and J. K. Calautit. Indoor fire detection utilizing computer vision-based strategies. *Journal of Building Engineering*, 61:105154, 2022.
38. L. Zhu, X. Wang, Z. Ke, et al. dynamic sparse attention: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10323-10333, 2023.
39. C. Li, L. Li, Y. Geng, et al. YOLOv6 v3.0: A full-scale reloading. *arXiv preprint arXiv:2301.05586*, 2023.
40. Z. Yu, H. Huang, W. Chen, et al. YOLO-Facev2: A scale and occlusion aware face detector. *arXiv preprint arXiv:2208.02019*, 2022.
41. M. Siliang and X. Yong. MPDIoU: A loss for efficient and accurate bounding box regression. *arXiv preprint arXiv:2307.07662*, 2023.

Shunxiang Zhang received the Ph.D. degree from the School of Computing Engineering and Science, Shanghai University, Shanghai, China, in 2012. He is a Professor at the Anhui University of Science and Technology, Huainan, China. His current research interests include web mining, semantic search, and complex network

Meng Chen received her Bachelor's degree from Fuyang Normal University in 2023. She is currently a master's student in the School of Computer Science and Engineering,

Anhui University of Science and Technology. Her current research interests include computer vision, image segmentation, object detection, 3D object detection, and multimodal networks.

Kuan-Ching Li is currently a Distinguished Professor at Providence University, where he also serves as the Director of the High-Performance Computing and Networking Center. He published more than 450 scientific papers and articles and is the coauthor or coeditor of more than 50 books published by well-known publishers. He is the Editor-in-Chief of IJCSE and IJES and serves as an associate editor for several leading journals. His research interests include parallel and distributed computing, big data, and emerging technologies. He is a fellow of the IET and a Senior Member of the IEEE.

Hua Wen received his Bachelor's degree from Huainan Normal College in 2023. He is currently a master's student in the School of Computer Science and Engineering, Anhui University of Science and Technology, China. His current research interests include multimodal sentiment analysis, sentiment extraction in multiple scenarios, sarcasm detection, and multimodal networks.

Liang Sun received the Bachelor's degree from Anhui University of Technology in 2024. He is currently a master's student in the School of Computer Science and Engineering, Anhui University of Science and Technology, China. His current research interests include multimodal sentiment analysis, large language modeling, graphic dialog generation, machine translation.

Received: December 25, 2024; Accepted: June 27, 2025.

Computer Science and Information Systems 22(4):1533–1553 https://doi.org/10.2298/CSIS250131060G

Defining the Attractiveness Concept for Cyber Incidents Forecasting

Javier García-Ochoa, Alberto Fernández-Isabel, Clara Contreras, Rubén R. Fernández, Isaac Martín de Diego and Marta Beltrán

Rey Juan Carlos University
Department of Computing, ETSII
C/ Tulipán, s/n, 28933, Móstoles, Madrid (Spain)
{javier.garciaochoa, alberto.fernandez.isabel, clara.contreras, ruben.rodriguez, isaac.martin, marta.beltran}@urjc.es

Abstract. Cyber incident forecasting has several applications within the security field, such as attack projection, intention recognition, attack prediction, or situational awareness. One of the main challenges of these issues lies in analysing the proneness of an entity to be attacked by an adversary evaluating the relevance of different target features or behaviours. This paper presents a methodology that defines the *Attractiveness* concept to address this issue. *Attractiveness* is the possession of features or the exhibition of behaviours in entities that raise interest for potential adversaries. Thus, the more significant the *Attractiveness* value is, the greater the proneness of attacking could be considered. The concept is decomposed into three main branches: *basal attractiveness* (relevance of the entity in the world), *online reputation* (the opinion of the individuals and the reach of the entity), and *potential victimisation* (the interest that the entity arouses for potential attackers). Machine Learning (ML) methods in combination with Information Retrieval (IR) and text mining techniques have been proposed to gather relevant information and identify hidden patterns and relations in past security incidents. With this approach, potential targets could reduce their *Attractiveness*, focusing on those aspects that can be remedied. Alternatively, future risky situations could be predicted to better prepare for proactive protection, detection, and response. The proposal has been validated through several experiments.

Keywords: Attractiveness, Cyber incidents, Victimisation, Online reputation, Forecasting.

1. Introduction

In the current digital era, entities (companies, associations, public organizations, and organisations) are constantly in the spotlight for possible malicious intentions. Thus, they must be able to identify risky situations in the scope of cybersecurity to face potential threats coming from the Internet. This strengthens their defences to protect their information and systems from attackers if these risky situations materialise [15].

Risky situations are defined as those where a potential threat could materialise, resulting from a cybersecurity failure. This idea is closely related to incident forecasting, a widely addressed topic [6].

In the case of cyber-incident forecasting, the methods defined in this domain are beneficial for organisations to protect their digital assets and ensure business continuity. By

providing a quantifiable approach to cyber risk assessment, these methods enable these entities to estimate potential threats and vulnerabilities with a certain degree of confidence. Although forecasting does not imply absolute certainty, it offers valuable insights that support the development of a proactive security strategy, allowing for better strategic planning and resource allocation. This approach helps managers and decision-makers at different levels to understand the overall security posture, take preventive actions, identify weak points to reinforce them before they are exploited and prioritise countermeasures [23].

The cyber incident risk varies according to several variables to consider. For instance, specific sectors like energy, financial services, manufacturing, technology, or pharmaceuticals are usually more targeted by current threat actors [3]. Small to medium-sized businesses are also often targeted due to their lack of resources to defend themselves [11]. Moreover, entities holding valuable data (e.g. financial or personal) usually awake more interest, and those with a weak cybersecurity posture could become easier targets [25].

Consequently, the research presented here proposes a novel methodology to analyse the proneness of an entity to be attacked by an adversary using an evolution of the *Attractiveness* concept (firstly introduced in [5]). Thus, the main novelty of the proposal lies in considering static (i.e. firmographic features such as the entity sector, size, or revenue) and dynamic factors (i.e. reputation and dynamic factors such as the value of its information, the number of visible vulnerabilities, or the potential impact of an incident). A higher *Attractiveness* value indicates a higher probability of being targeted by adversaries. It is important to note that this analysis focuses on scenarios where attackers do not have a predefined target and look for easy targets to attack. In this point, it is important to remark that *Attractiveness* is an estimation, not an exact measure.

The proposed methodology uses three different branches to build the *Attractiveness* concept: *basal attractiveness* (relevance of the entity in the world), reputation on the Internet (the opinion of the individuals and the reach of the entity), and *potential victimisation* (the interest that the entity arouses for potential attackers and the possible Common Vulnerabilities and Exposures (CVE) detected). The main contribution of the proposal is the definition and combination of these branches to produce the final *Attractiveness* estimation.

Several data were collected for the experimental setup to validate the proposal. These data consist of information from cyber incidents reported by entities from multiple sectors. The dataset is completed with the static and dynamic features from these entities. DeNexus Inc. in the frame of the DICYME project (Ref: CPP2021-009025), provided support for this issue.

It is important to highlight that the only feasible approach relies on confirmed incidents and organisations that have publicly disclosed them. Notice that attempts cannot be quantified, as they are not publicly reported, making it impossible to collect and aggregate them across different organisations for analysis. Similarly, incidents affecting organisations that have not disclosed them remain unknown and, therefore, cannot be considered. Ultimately, anything that is not known cannot be accounted for in the analysis, which poses a significant challenge for comprehensive risk assessment.

This approach does not address aspects of an organisation's security posture, which is undoubtedly crucial in cyber risk quantification. Instead, it focuses on the organisation's posture as an entity, considering its more static characteristics, such as its scope, presence,

and engagement across networks and social media platforms. By analysing these factors, the proposed approach provides insights into the organisation's external exposure and perceived attractiveness to potential attackers. Consequently, when *Attractiveness* is combined with other methods and quantification approaches that incorporate internal security measures, a more comprehensive and accurate cyber risk estimation can be achieved.

The rest of this paper is organised as follows. Section 2 overviews the related work categorising existing cyber incident forecasting and similar methods. Section 3 sets out the motivation for this work and the research questions addressed. Section 4 presents the estimator for *Attractiveness* and introduces the proposed method. Section 5 details the development process and the dataset features used in the proposal. Section 6 validates the proposal, while Section 7 discusses the results focusing on strengths and limitations. Finally, Section 8 concludes and proposes future research lines.

2. Related Work

Forecasting cyber incidents consists of predicting future risk situations produced by attackers evaluating a specific set of features gathered from entities analysed in the present.

It is a complex task and faces at least two main challenges [21]. The first is the inaccessibility of adequate data and observations about past incidents. When available, the challenge is to extract relevant and reliable signals to treat sporadic and seemingly random acts of adversaries. This involves dealing with imbalanced ground truth labels and unconventional signals [20] gathered from public sources (incident databases, news, social media). The second is the ever-changing threat landscape [3], making it difficult to keep up with the latest threats and adapt forecasting models accordingly.

Intrusion Detection Systems (IDS) play a crucial role in this context by monitoring network traffic for suspicious activity and known threats, providing real-time alerts to potential security incidents [28]. Despite their effectiveness, IDS data can be overwhelming and often contain a high rate of false positives, adding to the complexity of accurately forecasting cyber incidents.

Different previous research has attempted to overcome these challenges with several approaches. Table 1 summarises the most significant prior work in this area. The *Goal* column captures the kind of proposed forecasting and can be used to predict the next adversary's move (attack projection, AP), to infer the adversary's motivation and goals (intention recognition, IR) or to anticipate upcoming cyber attacks (attack prediction and risk quantification, RQ). The *Data signals* column specifies the kind of data signals used to perform the forecasting, therefore, on which aspects the prediction depends. Finally, the *Model* column summarises the kind of Artificial Intelligence (AI) approach selected to achieve the forecasting.

Some studies emphasise the unpredictable nature of cybersecurity threats, suggesting that certain information about an attack can be used to predict subsequent attacks. For instance, network attacks are analysed using dependency graphs and intrusion responses [17]. Other approaches utilise honeypot data combined with probabilistic models like Markov Chains to identify patterns in attack propagation and target areas [7].

Real-time attack intention recognition is another focus area, employing neural network models to analyse known attack patterns and network behaviour [2]. Time series

Table 1. Summary of previous work on forecasting cyber attacks or incidents

Ref.	Goal	Data signals	Model
[17]	RQ	IDS alerts, intrusion responses and dependency graphs	Graph
[7]	AP	Honeypot evidences	Markov chain
[1]	RQ	IDS alerts and logs	Time series
[2]	IR	Known attacks patterns and signatures	Neural network
[19]	AP + RQ	Asset graphs, vulnerabilities and IDS evidences	Bayesian model
[26], [27]	RQ	Incidents landscape, geopolitical context, social mentions and sentiment	Bayesian model
[16]	RQ	Honeypot evidences	Neural network
[30]	RQ	CVE and Twitter	Neural network
[29]	AP	Asset graphs, vulnerabilities, attacker location and capability	Graph
[8]	RQ	IDS alerts and logs	Neural network
[12]	RQ	IDS alerts and logs	Neural network
[22]	RQ	IDS alerts	Neural network

analysis and dynamic risk assessment are also applied to predict incidents in critical cloud infrastructures [1].

In industrial systems, AI models such as Bayesian networks are usually used to evaluate the cyber risk and physical impact of potential attacks [19]. ML models, such as Support Vector Machines, Multi-layer perceptron, and k-Nearest Neighbours, are leveraged to forecast various types of cyber incidents based on past data [26].

Deep Learning frameworks, including Recurrent Neural Networks, capture long-term dependencies and non-linearity in the data to predict attack rates [16]. In addition, big data from social networks and vulnerability databases is used to identify cyber risks, offering strategies to mitigate these risks in critical infrastructures [30].

Nevertheless, a structural limitation is common to most of these approaches: they rely on internal telemetry (e.g., IDS alerts, raw network flows, honeypot traces or sensors) that presuppose a mature monitoring infrastructure and a willingness to share data. Many organisations, especially small and medium-sized enterprises (SMEs) and peripheral entities in supply-chain ecosystems, lack such instrumentation. This reliance introduces several biases: partial coverage, because entities without telemetry are systematically excluded; high noise levels, as IDS outputs can contain false positives, rendering costly preprocessing indispensable before model training; and a short prediction horizon, since the models become effective only seconds or minutes before (or during) the intrusion, offering little value for strategic planning or proactive investment decisions.

Moreover, these methods tend to focus on technical signs of threats rather than the intrinsic nature of potential victims. They assess anomalies in network activity but rarely incorporate information about the entity, its representation, or its visibility and appeal to

an adversary [5]. As a result, they overlook critical contextual or firmographic variables that often shape attackers' target selection.

Furthermore, signature- or pattern-based models are highly vulnerable to concept drift as attackers adapt their tactics, leading to rapid performance degradation. However, while attacker behaviours may evolve quickly, their target selection patterns tend to be more stable over time [9]. These shortcomings motivate the need for a complementary perspective that leverages externally observable, universally available signals, remains interpretable, and is still applicable even in the absence of network instrumentation.

These limitations are the foundation for introducing the concept of *Attractiveness*, a generalist approach to evaluate the proneness of an entity to be attacked. A comprehensive methodology is developed to integrate various features collected from entities and generate the final estimation. ML models are used in this process, mainly considering descriptive variables of entities such as entity country, category, or financial worth into a comprehensive measure designed to assess this proneness.

3. Motivation and Research Questions

This section illustrates the motivation of the proposal. The approaches previously presented have put into the spotlight different methods and techniques that can be applied to predict or forecast cyber incidents, attacks, or events. They could be organised according to some data signals: those related to the targets (IDS alerts, logs, asset graphs, and vulnerabilities), and those related to the attackers and their known behaviour (honeypot pieces of evidence, attack patterns, and signatures).

There is agreement regarding the convenience of applying predictive mechanisms to the targets instead of applying them to the adversaries. This approach makes sense, assuming that targets are not actively being novel or creative, trying to evade security controls as the adversaries are. Therefore, it is more likely that predictions adjust to reality to a greater extent when they are made on the targets than when they are made on the adversaries (there is much less reliable data available on their techniques, motivations, or objectives).

Even so, it is observed that there is great difficulty in deciding which signals should be taken into account about these targets because they are the most significant for making predictions. As shown in the previous section, the largest body of work focuses on predicting whether an attack is imminent, which can be considered an early warning solution. This prediction is based on internal target data collected and stored in IDS-type systems or logging solutions. In recent years, neural networks have proven to be suitable tools for this type of prediction (deep neural networks, recurrent neural networks, and long short-term memory), based on learning associations between different alerts and contextual information.

However, the primary research question is different in this proposal. Which specific high-risk features (static) or behaviours (dynamic) allow predicting the occurrence of an incident? In a less imminent time frame, not because indicators are being observed (IDS alerts, logs) that would allow for early warning. Furthermore, how can this prediction be made once identified or selected? Is it possible to define the *Attractiveness* concept so that it can be used to forecast the future number of cyber incidents?

This research question (RQ) leads us to more specific ones:

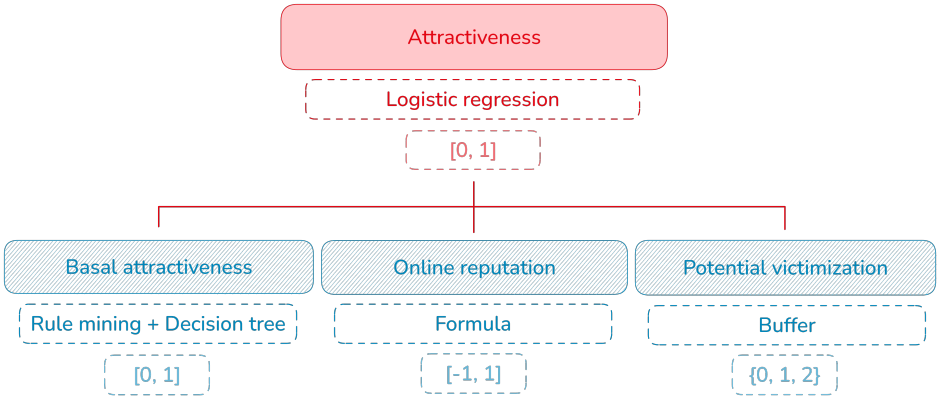


Fig. 1. Proposed high-level methodology

- RQ1: What specific features and behaviours enable the estimation of *Attractiveness*?
- RQ2: Can entities be grouped according to this *Attractiveness* to understand the obtained estimations? Can be the relationship between the entity *Attractiveness* or group *Attractiveness* and cybersecurity incidents identified?
- RQ3: How can this *Attractiveness* be used to forecast cybersecurity incidents? Are they inevitable, or is it possible to influence some of the aspects that influence *Attractiveness* to try to avoid them?
- RQ4: Are public data sources available to build the required data sets to work with this approach?

4. Estimation of Attractiveness

This proposal presents a methodology to estimate the *Attractiveness* value of each entity based on high-risk features (static) or behaviours (dynamic). This estimator considers the following critical aspects which can be evaluated in parallel: *basal attractiveness* (firmographic data), *online reputation*, and *potential victimisation*.

For each one of the aspects, the methodology gathers information about publicly confirmed cybersecurity incidents.

As a result of the methodology, a Logistic Regression algorithm joins the three aspects of the *Attractiveness*, producing a normalised value. This value represents the *Attractiveness* estimation for an entity being 0 the least attractive and 1 being the most attractive.

Figure 1 illustrates a general overview of this proposal. The next sections provide details about the estimation of the three considered aspects. Section 4.1 introduces the *basal attractiveness*, while Section 4.2 tackles the *online reputation*. Finally, Section 4.3 addresses the *potential victimisation*.

4.1. Basal attractiveness

The concept of *basal attractiveness* is based on the idea that certain entities are inherently more appealing to adversaries due to their static characteristics, commonly referred to as firmographic data (e.g., location, operational criticality, data sensitivity, and size).

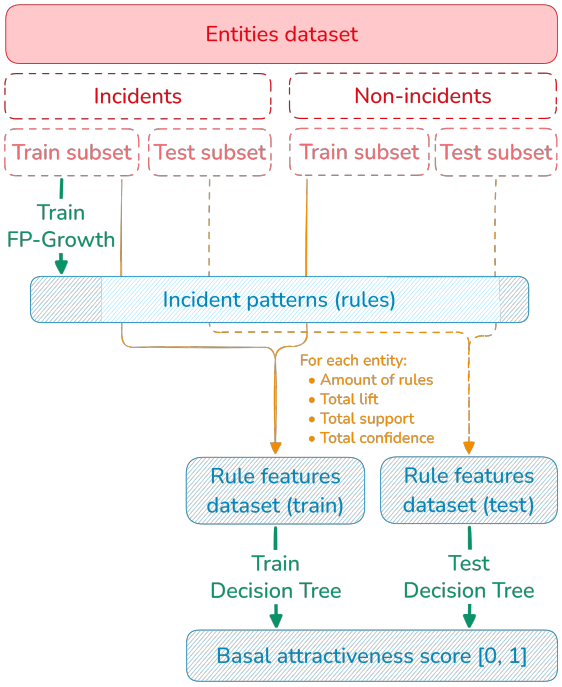


Fig. 2. Workflow of the *Basal attractiveness* model development

To model this, a dataset has been compiled containing both incidents and non-incidents involving various entities. Then, as shown in Figure 2, an association rule mining technique is used, specifically, the FP-Growth algorithm [32]. It is applied to uncover patterns within the incident data. For this, only a training subset of records corresponding to actual incidents is used, allowing the model to identify firmographic traits associated with increased risk.

Once the algorithm generates a set of association rules, these are used to evaluate both the incident and non-incident training subsets. For each observation, rule metrics such as support, confidence, lift, and the number of satisfied rules are computed and aggregated to build a new feature set for each entity.

This enriched dataset is then used to train a Decision Tree classifier, which outputs a binary prediction. The test subsets are subsequently applied to evaluate performance metrics on previously unseen entities.

The final output is a *basal attractiveness* score, ranging from 0 to 1. A score close to 0 indicates a low inherent risk of being targeted, while a score near 1 suggests a high baseline risk based on the entity's static attributes.

4.2. Online reputation

Online reputation tackles the acceptance and recognition that the entity has among the majority of people. The intuition behind the following proposition is explained by the

Table 2. Social media and networks considered for the *online reputation* estimation

Social media or network	
X (Twitter)	Tripadvisor
Facebook	Reddit
Instagram	Website domains
TikTok	Comments from website domains
YouTube	Forum domains

fact that an entity may be more attractive to an adversary based on its *online reputation* defined as the result of what users, customers, or employees write, communicate and share anywhere on the Internet based on their perceptions and experience in any moment of their relationship, direct or indirect, with the entity [26], [27], [30]. Thus, the positive and negative opinions or comments exchanged on social media and social networks can make an entity more attractive to certain attackers. This situation could change over time, finding an entity attractive during a period, and later unnoticed.

Social media is the dynamic content shared through corporate means (corporate website sections, corporate podcasts, corporate blogs). In these media, interaction is possible, but not expected or frequent. On the other hand, social networks allow the entity to share dynamic content using external platforms that are not controlled by the entity (e.g. Twitter, Reddit, or LinkedIn). Users may comment and chat about the shared content. This interaction is expected and more frequent.

Delving into reputation estimation, the Determ tool [13] has been used to gather specific information. Moreover, a formulation to produce the final indication has been built. In this sense, the *online reputation* (OR) indicator is based on the engagement (E), which can be defined as the ratio between the interaction (I) and reach (R) [10]. Reach represents the estimated number of people who read or mention a publication, while interaction (I) is the estimated number of people who answer it. Thus, if only one person has commented on a post with 100 views, the Engagement is $1/100$.

In this proposal, the *online reputation* is time-dependent because it represents a static picture in a specific moment, influenced by previous periods. Thus, OR^t is the temporal indicator computed per social media and network, where t is the current time. This indicator is built by using a weighted average of two other indicators: the entity engagement (EE^t) and the user engagement (UE^t):

$$OR^t = \alpha \cdot EE^t + (1 - \alpha) \cdot UE^t \,. \tag{1}$$

The EE^t indicator considers the engagement generated by the content published by the entity through its social media and networks. It is defined as:

$$EE^t = \frac{1}{E} \cdot \sum_{j=1}^{E} \left[\frac{EE_j^t}{\max\limits_{z=1\ldots t} EE_j^z} \right] , \tag{2}$$

where E is the number of social media and networks where the current entity interacts, and j represents each one of these media and networks. Therefore, EE_j^t means the entity engagement in the current time t in the social media or network j, and $\max\limits_{z=1\ldots t} EE_j^z$ is the maximum entity engagement considering the estimations previously achieved.

At the same time, EE_j^t can be defined as follows:

$$EE_j^t = \frac{1}{n_j^t} \cdot \sum_{i=1}^{n_j^t} \frac{IC_i^t}{RC_i^t}\,, \tag{3}$$

where n_j^t is the total number of mentions in the social media or network, while IC_i^t and RC_i^t are the interactions and reach in the social media j in each of the mentions i at the moment t.

The UE indicator considers the engagement generated by external mentions and comments (not generated or controlled by the entity) through social media and networks. It is defined as follows:

$$UE^t = \frac{1}{U} \cdot \sum_{k=1}^{U} \left[\frac{UE_k^t}{\max\limits_{z=1\ldots t} UE_k^z} \right], \tag{4}$$

where t is the current time, U is the number of mentions or interactions made by users related to the current entity, and k represents each one of these media and networks. Therefore, UE_k^t means the user engagement in the current time t in the social media or network k, and $\max\limits_{z=1\ldots t} UE_k^z$ is the maximum user engagement considering the estimations previously achieved.

At the same time, UE_k^t can be defined as follows:

$$UE_k^t = S_k^t \cdot \frac{1}{n_k^t} \cdot \sum_{p=1}^{n_k^t} \frac{IU_p^t}{RU_p^t}\,, \tag{5}$$

where n_k^t is the total number of mentions in the social media or network, while IU_p^t and RU_p^t are the interactions and reach in the social media k in each of the mentions p at the moment t. S_k^t estimates the sentiment value of each mention related to a social media or network k in a period t.

Then, the estimation of the sentiment value for each social media or network k is defined as follows:

$$S_k^t = \frac{positive_k^t - negative_k^t}{positive_k^t + negative_k^t}\,, \tag{6}$$

where $positive_k^t$ is the number of mentions labelled as positive in the social media or network k in the time t, and $negative_k^t$ is the number of mentions labelled as negative in the social media or network k in the time t. This formula produces a sentimental range between (-1) and $(+1)$.

Notice that the proposal does not consider the sentiment regarding entity engagement because it is associated with content published (and controlled) by the entity. Therefore, it can be assumed that it will always be eminently positive.

As a result, the *online reputation* estimator produces a value between (-1) and $(+1)$, where the lowest value represents a mediocre *online reputation* and the highest value indicates a good *online reputation* in social media and networks.

4.3. Potential victimisation

The intuition behind the *potential victimisation* (also a part of the dynamic behaviour) is explained by the fact that an entity may be more attractive to an adversary if it is often mentioned in underground forums or specific dark websites, or perceived as an approachable victim. Additionally, the entity may become more attractive if there are public data breaches.

Thus, forums, foreground and underground sites are monitored from the point of view of an adversary. There, knowledge and tools, data breaches, intelligence or business, the visibility of the entity infrastructure, and also a possible victim of an attack are analysed to detect if they are shared. These concepts are defined as follows:

- Risky visibility: number of direct mentions in monitored underground forums and dark web sites.
- Perceived ease of success: number of visible open ports and services, number of visible assets connected to the Internet, number of visible remote access protocols, number of visible third-party software dependencies, number of visible CVE, and number of visible wrong settings (i.e. default accounts, default or empty credentials, and valid leaked credentials).

These concepts are collected in two variables: Critic Info (number of mentions in dark web leaks) and Devices (number of devices connected to the Internet). These two variables are transformed into one with a buffer method in the following way:

- If both are 0, the *potential victimisation* value is 0.
- If one is 0 and the other is more than 0, the *potential victimisation* value is 1.
- If both are greater than 0, the *potential victimisation* value is 2.

As a result, the *potential victimisation* estimator produces a result between 0 and 2, where the lowest value indicates a very low *potential victimisation* and the highest value represents a very relevant *potential victimisation*.

5. Dataset development

This section details the development process of the dataset used in the experiments related to the proposal. Multiple primary and secondary data sources were considered, including public databases and data collection platforms, to generate this dataset. The quality and relevance of the data utilised are crucial to ensure the validity and reliability of the results obtained. Therefore, rigorous procedures have been implemented for cleaning, normalisation, and comparability of variables.

Delving into the set of firmographic variables related to *basal attractiveness*, they are defined as follows:

- Country: headquarters location based on country.
- Category: entity sector provided by RocketReach tool (see Table 3 for more details).
- Revenue: annual billing of the entity (USD).
- Earnings: annual profit of the entity (USD).
- Publicly traded: whether it is listed on the stock exchange (true/false).

Table 3. Sectors for entities in the *basal attractiveness* estimation

Sector	
Agriculture & Fishing	Media & Internet
Business Services	Metals & Mining
Chambers of Commerce	Organisations
Cities, Towns & Municipalities	Real Estate
Construction	Research & Technology
Consumer Services	Retail
Cultural	Software
Education	Telecommunications
Energy, Utilities & Waste Treatment	Trade, Supply Chain & Commerce
Finance	Transportation
Government	Healthcare
Hospitality	Insurance
Law Firms & Legal Services	Manufacturing

- Employees: size of the entity regarding the number of employees.
- Profitable: whether it is for-profit (true/false).

This information is gathered from RocketReach [31]. This tool allows consulting the names of entities. In this case, these names are those entities that appear in public databases where confirmed cyber incident victims are reported. These databases are the European Repository of Cyber Incidents (EuRepoC), Hackmageddon, Jam Cyber, TI Safe Incident Hub, KonBriefing, CISSM Cyber Attacks Database, and ICS STRIVE.

The dataset is completed with observations of similar entities that have not reported incidents. RocketReach is also used here to provide a set of these entities for each affected entity. Then, the category variable of the entities is checked and filtered. Finally, from the set of entities that share the category, one of them is randomly considered.

The generated dataset counts 675 observations, collected between September 2023 and May 2024. The proposed methods only consider incidents and non-incidents. Therefore, the data is transformed to a binary classification, obtaining 485 incidents and 190 non-incidents.

Regarding the distribution of incident types, they are organised as follows: 240 observations of ransomware, 120 observations of denial of service, 95 observations of data breach, 27 observations of destruction, and 189 observations of non-incident.

Concerning the properties of the variables, revenue has a wide range, indicating varied sizes of entities, with a mean of approximately $\$4.12$ billion. Notice that this mean is skewed by some entities, as evidenced by a maximum value of $\$271.57$ billion. Earnings also show a large spread and high variability, while most entities (87.7%) are not publicly traded. The mean number of employees is around $43,884$, the gain is highly skewed due to some large organisations. Most entities, however, are smaller as the median number of employees is 341.5, while more entities are profitable (52.6%).

Numerical variables have been discretised into four quartiles, resulting in the following ranges:

- Revenue: $[-2.72\text{e+}08, 6.79\text{e+}10)$, $[6.79\text{e+}10, 1.36\text{e+}11)$, $[1.36\text{e+}11, 2.04\text{e+}11)$ and $[2.04\text{e+}11, 2.72\text{e+}11]$

Table 4. Support, Confidence, and Lift for the top 10 Rules from the rule mining algorithm

Rule	Sup.	Conf.	Lift
Earnings=NaN → Employees=0.0	0.845	0.927	0.996
Employees=0.0 → Earnings=NaN	0.845	0.909	0.996
Earnings=NaN → Publicly traded=False	0.871	0.955	1.096
Publicly traded=False → Earnings=NaN	0.871	1.000	1.096
Publicly traded=False → Employees=0.0	0.807	0.926	0.995
Employees=0.0 → Publicly traded=False	0.807	0.867	0.995
Earnings=NaN, Publicly traded=False → Employees=0.0	0.807	0.926	0.995
Earnings=NaN, Employees=0.0 → Publicly traded=False	0.807	0.954	1.095
Publicly traded=False, Employees=0.0 → Earnings=NaN	0.807	1.000	1.096
Earnings=NaN → Publicly traded=False, Employees=0.0	0.807	0.884	1.096

- Earnings: $[-7.84\text{e+}09, -5.40\text{e+}08)$, $[-5.40\text{e+}08, 6.73\text{e+}09)$, $[6.73\text{e+}09, 1.40\text{e+}10)$ and $[1.40\text{e+}10, 2.13\text{e+}10]$
- Employees: $[2.50\text{e+}04, 6.25\text{e+}06)$, $[6.25\text{e+}06, 1.25\text{e+}07)$, $[1.25\text{e+}07, 1.875\text{e+}07)$ and $[1.875\text{e+}07, 2.50\text{e+}07]$

In the case of the *online reputation* case, its estimation involves a series of complex functions designed to extract and analyse data from social media mentions.

The data is pre-processed to segment time intervals into bi-weekly periods from the date the incident happened to the present. After that, the *online reputation* formula is applied to the data collected and included in the dataset.

The results of the formula range from -0.215 to 0.440 with a mean of 0.091, indicating generally positive reputations among entities. The variability (standard deviation of 0.103) suggests differences in how entities are perceived online.

Finally, two variables are included following the methodology for the *potential victimisation*. For devices, most entities do not report device-related incidents, with a 75th percentile value of 0. However, the maximum value of 100 indicates that some entities experience significant device-related incidents. For critical info, the values range from 0 to 314, with most entities (75th percentile) reporting at most 1 critical info incident, suggesting that breaches are not widespread.

6. Experiments

This section illustrates the viability of the proposal through various experiments. The three aspects of the methodology are considered. The relationships between *basal attractiveness*, *online reputation*, and their potential implications on incident occurrences are also explored. Each experiment is structured to validate theoretical assumptions through data analysis.

Delving into the experiments, a complete evaluation of the proposed methodology is presented. The following sections detail the specific experiments conducted to assess the three critical aspects: *basal attractiveness*, *online reputation*, and *potential victimisation*. Section 6.1 tackles the FP-Growth algorithm to estimate *basal attractiveness* based on specific entity features. Section 6.2 explores the relationship between entities regarding *online reputation* and their incidence of cyber incidents. Further analysis of the interaction

between *basal attractiveness* and *online reputation* with incident occurrences is presented in Section 6.3. Section 6.4 details the application of a Decision Tree classifier to predict incidents based on rule mining metrics. Finally, Section 6.5 integrates the insights from the previous experiments into a Logistic Regression model to estimate the combined effect of *basal attractiveness*, *online reputation*, and *potential victimisation* on predicting incident outcomes.

6.1. Rule mining algorithm for estimating basal attractiveness

The primary objective of using the FP-Growth algorithm [18] in this study is to estimate the *basal attractiveness* of entities. This method enables the discovery of underlying patterns within the dataset that contribute meaningfully to the perceived attractiveness, providing a structured approach to understanding the static factors associated with increased risk.

As a preliminary step, continuous variables in the dataset were discretised—i.e., converted into categorical intervals—to simplify the representation of the data and make it more suitable for pattern mining. The discretised data was then processed using the FP-Growth algorithm to identify frequent item sets. A minimum support threshold of 0.01 was used, ensuring that only patterns occurring in at least 1% of the transactions were considered. Other algorithm parameters were maintained at their default settings.

For mining frequent patterns, the dataset was split using an $80/20$ ratio, yielding 385 incident cases for training and a separate test set comprising 97 incident cases and 189 non-incident cases. It is important to highlight that the FP-Growth algorithm was trained exclusively on observations labelled as incidents.

Once extracted, the association rules were ranked by their *support* values, which indicate how frequently the associated item sets appear within the incident data. Rules with higher support are considered more representative and were analysed in detail to assess their contribution to the *basal attractiveness* estimation.

Table 4 presents the top 10 rules with the highest support. For instance, one of the most frequent rules reveals that entities experiencing cyber incidents tend to have few employees and lack available data on their annual earnings. This suggests that smaller organisations with limited financial transparency may be more vulnerable to attacks. Such findings can inform cybersecurity policies by encouraging targeted support for small enterprises that might otherwise lack the resources or visibility to manage cyber risks.

6.2. Relationship between online reputation and cyber incidents

In this experiment, the *online reputation* is estimated to assess the relationship between an entity's reputation in social media and social networks and the occurrence of cyber incidents. The α parameter of Equation 1 is fixed to 0.5 to provide neutral relevance to each part of the equation. The dataset was divided into two groups: those observations associated with incidents and those with non-incidents.

The non-parametric Wilcoxon rank-sum test is selected to assess whether there are statistically significant differences in *online reputation* scores between groups defined by their incident status [14]. This test compares the median of the *online reputation* scores of both groups. This approach is more robust to outliers and non-normal data distributions than mean comparisons used in t-tests.

A statistic W equals 49698 with a p-value of 0.06408 is provided. This result suggests no statistically significant difference in the distributions of *online reputation* scores between the non-incident and incident groups at the conventional 0.05 significance level. However, the p-value is close to the threshold, indicating a potential trend that could be significant with a larger dataset or different grouping methods.

The findings indicate that, while there is a visible difference in the median *online reputation* scores between the groups, it is not statistically significant under the current experimental setup. The marginally high p-value suggests a potential pattern in which *online reputation* could influence incident outcomes, albeit not strongly enough to be deemed significant.

6.3. Relationship between basal attractiveness and online reputation with incident occurrences

This analysis is focused on investigating the association between *basal attractiveness* and incident occurrences, as well as *online reputation* and incident occurrences, utilising the Chi-squared test of independence.

It is a non-parametric test to detect if two categorical variables are independent of each other across different groups [4]. It compares the observed frequencies in the data against the expected frequencies, which are calculated under the assumption that the variables are independent.

The data includes recorded incidents and non-incidents categorised by levels of *basal attractiveness* and *online reputation*. The data is grouped as follows:

- *Basal attractiveness*
  - Level 0: 119 non-incidents, 4 incidents.
  - Level 1: 9 non-incidents, 9 incidents.
  - Level 2: 1 non-incident, 57 incidents.
- *Online reputation*
  - Range $[-1, 0]$: 36 non-incidents, 29 incidents.
  - Range $(0, 1]$: 93 non-incidents, 41 incidents.

The Chi-squared test for *basal attractiveness* produced a statistic of $X^2 = 157.98$ with 2 degrees of freedom. The extremely low p-value ($< 2.2e - 16$) indicates a highly significant statistical association between *basal attractiveness* levels and incident occurrences. This suggests that *basal attractiveness* is not independent of incident status, with higher *Attractiveness* levels correlating with a higher frequency of incidents.

The Chi-squared test for *online reputation* produced a statistic of $X^2 = 3.1823$ with 1 degree of freedom. The p-value of 0.07444 suggests that while there is a notable trend, the association between *online reputation* ranges and incident occurrences does not reach conventional levels of statistical significance ($p < 0.05$). However, the p-value close to the threshold indicates a potential mild association that might become significant with a larger sample size or different categorisation of reputation scores.

Although the test for *online reputation* did not reach statistical significance independently, it is worth considering the potential interplay between *online reputation* and *basal attractiveness*. It is possible that *online reputation* could interact with *basal attractiveness* to influence incident occurrences in ways that are not captured when these variables are considered separately, this would be explored in further experiments.

6.4. Prediction of incidents based on rule mining metrics

This analysis consists of constructing a predictive model utilising a Decision Tree classifier. This model aims to forecast incidents by leveraging metrics obtained through rule-mining techniques.

A new dataset with 286 observations was prepared by labelling entities as 0 for non-incidents and 1 for incidents using the rule mining algorithm results for the test observations of the original dataset. The features extracted for the rule mining algorithm included counts of rules and their aggregated measures of support, confidence, and lift.

Once the new dataset is built, two steps are addressed: data preparation and splitting, and model training. The first splits the new dataset into training and testing groups with a 70-30 ratio, ensuring a balance between the learning and validation capabilities of the model. The second trains a Decision Tree classifier with specified hyper-parameters (10 as minimum samples split, 5 as minimum samples leaf, 5 as maximum depth, and 42 as random state). These values have been selected to prevent over-fitting issues while maintaining the model's generalisation ability.

The Decision Tree model achieved robust performance in the test with the following metrics: Precision: 0.87, Accuracy: 0.89, Recall: 0.74, and Kappa Statistic: 0.64.

Therefore, the model effectively classifies entities based on the rule mining metrics, with high Accuracy and a good balance between Precision and Recall. The high Precision rate indicates that the model is reliable in predicting incidents when it classifies an entity as such, while the Recall rate shows that it captures a significant proportion of actual incidents.

6.5. Prediction of incidents based on basal attractiveness, online reputation and potential victimisation

This last experiment aims to demonstrate that integrating the three critical aspects of the methodology improves the obtained results. A new dataset is created containing the three measures previously estimated as follows, with the attributes of the observations from the FP-Growth test:

- *Basal attractiveness*: output of the Decision Tree probability applied to the Decision Tree test dataset and trained with the output of the FP-Growth algorithm.
- *Online reputation*: output of Equation 1 with the α parameter fixed to 0.5.
- *Potential victimisation*: buffer applied to both columns of this category, getting a value in $\{0, 1, 2\}$.

A Logistic Regression model was built using the R package caret, which eases robust model building and evaluation. This package was selected due to its comprehensive array of functions that not only streamline model training but also provide extensive tools for tuning and evaluating model performance [24].

The experiment consists of three steps: data preparation, model training, and evaluation. Firstly, the dataset was read and processed to remove unnecessary entity identifiers and convert key variables into categorical forms suitable for analysis. Thus, the original dataset was discretised, while the information related to *potential victimisation* was transformed into a binary factor indicating the presence or absence of the two evaluated conditions. Then, the data were randomly split into a training set (80%) and a test

Table 5. Comparison of the metrics obtained by the Logistic Regression model for training and testing data

Metric	Training Data	Testing Data
Accuracy	0.931	0.923
Kappa Statistic	0.847	0.835
Recall	0.875	0.929
Precision	0.925	0.867

set (20%), using stratified sampling to maintain the proportion of incidents across these sets. In the next step, the Logistic Regression model was trained on the training set using cross-validation (5-fold) to optimise model parameters and prevent over-fitting. The model included the three outcomes of the aspects as predictors. The Logistic Regression model maps a linear combination of the predictors to a value between 0 and 1, representing the probability of an incident. A threshold is applied to the predicted probability of making a binary decision. If the predicted probability is greater than or equal to 0.5, the outcome is 1 (incident). If it is less than 0.5, it is classified as 0 (non-incident). In the last step, model performance was assessed on training and testing sets using confusion matrices and associated statistics to measure Accuracy, Recall, Kappa, and Precision.

Results are summarised in Table 5, showing key performance metrics for training and testing data. The model maintained high-performance metrics across training and testing datasets, with consistent metrics reported. The model exhibits strong accuracy, maintaining over 0.92 in both datasets. The Kappa statistic, which measures agreement beyond chance, indicates excellent model reliability with values of 0.847 and 0.835 for training and testing, respectively, suggesting that the model is consistent in its predictions across different data sets.

The increase in Recall from 0.875 in training to 0.928 in testing highlights the model's enhanced ability to identify positive cases in unseen data and generalise well without being overly fitted to the training data. This balance is crucial for practical applications where false positives and negatives carry significant implications.

The proposed model demonstrated high Accuracy and Recall. This high performance suggests that the discretisation of *basal attractiveness* and *online reputation*, including the binary variable of *potential victimisation*, provides a strong foundation for identifying patterns associated with incident occurrences.

7. Lessons Learned

This section synthesises the insights and key observations from the experiments designed to forecast possible cyber incidents using the *Attractiveness* concept. These findings contribute to a deeper understanding of the complex dynamics in cybersecurity threat assessment and the relevant features that make entities prone to suffering attacks.

Firstly, it is important to note that the developed methodology can be integrated into cybersecurity platforms to produce more robust predictive tools.

Regarding the selected ML, the FP-Growth algorithm to estimate the *basal attractiveness* revealed significant patterns correlated with cyber incident susceptibility in entities.

The discretisation data process proved invaluable in identifying these patterns, enhancing the understanding of key vulnerability factors.

The Decision Tree model achieved high Accuracy and Precision in classifying potential incident occurrences. This success illustrates the efficacy of ML approaches in extracting actionable intelligence from complex datasets.

The impact of *online reputation* on incident occurrences was evaluated using the Wilcoxon Rank-Sum test. Subtle, yet insightful, differences were found between affected and unaffected groups. This underscores the potential of nuanced statistical methods in identifying marginal trends. A statistically significant association between *basal attractiveness* and cyber incidents was confirmed when the Chi-square test of independence with *basal attractiveness* and *online reputation* was used. Thus, robust evidence is provided on how the physical infrastructure of attractiveness relates to the proneness of suffering possible attacks.

Integrating all the relevant variables into a Logistic Regression model enhanced predictive performance, affirming the value of synthesising multiple data sources and analytical perspectives.

Finally, despite its contributions, this proposal faces some limitations. In the scope of data and the generalisability of some findings, the data collected is limited and may not capture all the dimensions that influence cyber risk and increase or decrease the *Attractiveness*. Another limitation arises primarily because not all entities report cyber incidents when they occur. There can be various reasons for this lack of reporting, such as concerns about reputation damage or financial implications. Consequently, the data available for analysis might skew towards more transparent organisations or those mandated by regulation to disclose cybersecurity issues. Future studies could address this limitation by incorporating methods to estimate unreported incidents or using anonymised data contributions to encourage fuller disclosure from a wider range of entities.

8. Conclusions

This paper has introduced a novel methodology based on the *Attractiveness* concept. The proposed approach facilitates cyber incident forecasting by identifying entities prone to cyber attacks through three perspectives: the entity's relevance, its *online reputation*, and the interest it generates among potential attackers.

Regarding the first research question (RQ1), the study has identified specific static and dynamic features that enable the estimation of *Attractiveness*. Static features, such as firmographic data (e.g., sector, size, and revenue), establish a foundational risk profile by reflecting the inherent characteristics of an entity. Dynamic features, including *online reputation*, the number of visible vulnerabilities, and media exposure, provide a temporal dimension to the evaluation, capturing fluctuations in risk over time. The results indicate that combining these static and dynamic factors produces a more robust estimation of *Attractiveness*, offering insights into both baseline risk and evolving exposure.

For the second research question (RQ2), the findings confirm that entities can be grouped effectively based on their *Attractiveness* scores, revealing clear patterns between these groupings and the likelihood of experiencing cyber incidents. This analysis highlights a strong relationship between *Attractiveness* and cyber incidents, providing organi-

sations with actionable benchmarks for comparing their risk levels to similar entities and prioritising targeted interventions.

The third research question (RQ3) explored the use of *Attractiveness* for forecasting cyber incidents and its potential to influence risk mitigation. The results demonstrate that *Attractiveness* is a powerful predictor when used in conjunction with other data sources, such as threat intelligence and security posture. While certain factors contributing to *Attractiveness*, such as industry sector or size, are fixed and difficult to modify, others, such as reducing exposure by managing *online reputation* or addressing visible devices, can be proactively influenced to lower risk levels. This finding underscores the value of a proactive approach tailored to an entity's specific *Attractiveness* profile.

Finally, addressing the fourth research question (RQ4), the study confirmed the viability of using public data sources to build the required datasets for this methodology. Publicly available information, such as disclosed cybersecurity incidents, financial reports, social media activity, and vulnerability databases, provided the foundation for the analysis. However, the research acknowledges significant limitations due to the reliance on publicly reported data, as undisclosed incidents and unreported attack attempts remain inaccessible. Despite these constraints, the findings demonstrate that publicly available data, when combined with robust analytical techniques, offers a strong basis for cyber risk quantification and forecasting.

Future research could expand on this work by incorporating more observations into the dataset to verify the findings presented here. Furthermore, integrating real-time data streams could significantly enhance the model's performance. Additional efforts could focus on integrating intelligence from cyber threat actors, such as information from underground forums or dark web activities, to enhance understanding of adversarial behaviour and refine the *Attractiveness* concept further. The integration of AI-driven predictive analytics into real-time cyber defence systems could also be explored. These advancements would provide even greater value for organisations seeking to anticipate and mitigate cyber risks.

Acknowledgments. This work has been funded by the Spanish MICINN under the CPP program in the DICYME project (Ref: CPP2021-009025), partially funded by the XMIDAS project (PID2021-122640OB-I00), and supported by DeNexus Inc.

References

1. Abdlhamed, M., Kifayat, K., Shi, Q., Hurst, W.: A system for intrusion prediction in cloud computing. In: Proceedings of the International Conference on Internet of Things and Cloud Computing. pp. 1–9 (2016)
2. Ahmed, A.A., Mohammed, M.F.: Sairf: A similarity approach for attack intention recognition using fuzzy min-max neural network. Journal of Computational Science 25, 467–473 (2018)
3. Ardagna, C., Corbiaux, S., Van Impe, K., Sfakianakis, A.: ENISA ThreatLandscape 2022. European Agency for Cybersecurity (2022)
4. Argyrous, G., Argyrous, G.: The chi-square test for independence. Statistics for Social Research pp. 257–284 (1997)
5. Awan, M.S.K., Dahabiyeh, L.: Corporate attractiveness index: A measure for assessing the potential of a cyber attack. In: 2018 9th International Conference on Information and Communication Systems (ICICS). pp. 1–6. IEEE (2018)

6. Bakdash, J.Z., Hutchinson, S., Zaroukian, E.G., Marusich, L.R., Thirumuruganathan, S., Sample, C., Hoffman, B., Das, G.: Malware in the future? forecasting of analyst detection of cyber events. Journal of Cybersecurity 4(1), tyy007 (2018)
7. Bar, A., Shapira, B., Rokach, L., Unger, M.: Identifying attack propagation patterns in honeypots using markov chains modeling and complex networks analysis. In: 2016 IEEE international conference on software science, technology and engineering (SWSTE). pp. 28–36. IEEE (2016)
8. Ben Fredj, O., Mihoub, A., Krichen, M., Cheikhrouhou, O., Derhab, A.: Cybersecurity attack prediction: a deep learning approach. In: 13th international conference on security of information and networks. pp. 1–6 (2020)
9. Cho, J., Eling, M., Jung, K.: Spatial cyber loss clusters at county level and socioeconomic determinants of cyber risks. North American Actuarial Journal 29(2), 345–389 (2025)
10. Cioppi, M., Curina, I., Forlani, F., Pencarelli, T.: Online presence, visibility and reputation: a systematic literature review in management studies. Journal of Research in Interactive Marketing 13(4), 547–577 (2019)
11. CrowdStrike Holdings Inc: 2023 Global Threat Report. https://www.crowdstrike.com/global-threat-report/ (2023), online accessed: 2024-09-02
12. Dalal, S., Manoharan, P., Lilhore, U.K., Seth, B., Mohammed alsekait, D., Simaiya, S., Hamdi, M., Raahemifar, K.: Extremely boosted neural network for more accurate multi-stage cyber attack prediction in cloud computing environment. Journal of Cloud Computing 12(1), 14 (2023)
13. Determ d.o.o.: Determ - ai media monitoring and analytics software. https://www.determ.com/ (2024), online accesssed: 2024-09-02
14. Divine, G., Norton, H.J., Hunt, R., Dienemann, J.: A review of analysis and sample size calculation considerations for wilcoxon tests. Anesthesia & Analgesia 117(3), 699–710 (2013)
15. Djajasinga, N.D., Fatmawati, E., Syamsuddin, S., Sukomardojo, T., Sulistyo, A.B.: Risk management in the digital era addressing cybersecurity challenges in business. Branding: Jurnal Manajemen dan Bisnis 2(2) (2023)
16. Fang, X., Xu, M., Xu, S., Zhao, P.: A deep learning framework for predicting cyber attacks rates. EURASIP Journal on Information security 2019, 1–11 (2019)
17. GhasemiGol, M., Ghaemi-Bafghi, A., Takabi, H.: A comprehensive approach for network attack forecasting. Computers & Security 58, 83–105 (2016)
18. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data mining and knowledge discovery 8, 53–87 (2004)
19. Huang, K., Zhou, C., Tian, Y.C., Yang, S., Qin, Y.: Assessing the physical impact of cyberattacks on industrial cyber-physical systems. IEEE Transactions on Industrial Electronics 65(10), 8153–8162 (2018)
20. Husák, M., Kašpar, J.: Towards predicting cyber attacks using information exchange and data mining. In: 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC). pp. 536–541. IEEE (2018)
21. Husák, M., Komárková, J., Bou-Harb, E., Čeleda, P.: Survey of attack projection, prediction, and forecasting in cyber security. IEEE Communications Surveys & Tutorials 21(1), 640–660 (2018)
22. Jain, J.K., Waoo, A.A.: An artificial neural network technique for prediction of cyber-attack using intrusion detection system. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN) ISSN pp. 2799–1172 (2023)
23. van der Kleij, R., Schraagen, J.M., Cadet, B., Young, H.: Developing decision support for cybersecurity threat and incident managers. Computers & Security 113, 102535 (2022)
24. Kuhn, M.: Building predictive models in r using the caret package. Journal of statistical software 28, 1–26 (2008)
25. Mandiant: M-trends 2023. https://www.mandiant.com/m-trends/ (2023), online accessed: 2024-09-02

26. Okutan, A., Werner, G., Yang, S.J., McConky, K.: Forecasting cyberattacks with incomplete, imbalanced, and insignificant data. Cybersecurity 1, 1–16 (2018)
27. Okutan, A., Yang, S.J., McConky, K., Werner, G.: Capture: cyberattack forecasting using non-stationary features with time lags. In: 2019 IEEE Conference on Communications and Network Security (CNS). pp. 205–213. IEEE (2019)
28. Panigrahi, R., Borah, S., Bhoi, A.K., Mallick, P.K.: Intrusion detection systems (ids)—an overview with a generalized framework. Cognitive Informatics and Soft Computing: Proceeding of CISC 2019 pp. 107–117 (2020)
29. Polatidis, N., Pimenidis, E., Pavlidis, M., Papastergiou, S., Mouratidis, H.: From product recommendation to cyber-attack prediction: Generating attack graphs and predicting future attacks. Evolving Systems 11, 479–490 (2020)
30. Subroto, A., Apriyana, A.: Cyber risk prediction through social media big data analytics and statistical machine learning. Journal of Big Data 6(1), 50 (2019)
31. Szymoniak, S., Foks, K.: Open source intelligence opportunities and challenges–a review. Advances in Science and Technology. Research Journal 18(3) (2024)
32. Wang, K., Tang, L., Han, J., Liu, J.: Top down fp-growth for association rule mining. In: Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, PAKDD 2002 Taipei, Taiwan, May 6–8, 2002 Proceedings 6. pp. 334–340. Springer (2002)

Javier Sánchez García-Ochoa was born in Toledo, Spain, in 2000. He holds a bachelor's degree in Cybersecurity Engineering from Rey Juan Carlos University (URJC) and a master's degree in Cybersecurity and Privacy from the Open University of Catalonia (UOC). He worked in the private sector for over a year before joining Rey Juan Carlos University in 2023 as a research staff member. His main research activity is carried out within the public-private collaboration project DICYME ("Dynamic Industrial Cyber Risk Modelling based on Evidence"), focused on dynamic cyber risk quantification. He has also contributed to several research articles and conference proceedings. Additionally, he is involved in various other research tasks and projects related to cybersecurity and data science.

Alberto Fernández-Isabel was born in Toledo, Spain in 1984. He received a PhD in Computer Science from Complutense University of Madrid (UCM) in 2015. He obtained a scholarship at the Spanish National Research Council (CSIC) as a technical assistant. He has been working for several years on European and national projects as a predoctoral and postdoctoral researcher. Since 2019 he is Assistant Professor at the Higher Technical School of Computer Engineering (ETSII) at Rey Juan Carlos University (URJC). He has authored more than 30 scientific articles and books. He completes his background with a Master´s degree in Artificial Intelligence and a Master´s degree in Information Systems. His research interests include intelligent agents, machine learning, data visualization, and natural language processing in various application domains, including distributed programming, sentiment analysis, agent-based collaboration and negotiation, smart cities, and simulations.

Clara Contreras is as an Associate Professor at the Department of Computing, Universidad Rey Juan Carlos, Madrid (Spain) and Cybersecurity Engineer at Siemens. Her research interests are Cybersecurity and Artificial Intelligence.

Rubén Rodríguez Fernández was born in Bembibre, León, Spain in 1973. He received a PhD degree in Artificial Intelligence from Rey Juan Carlos University (URJC). He also received a master's degree in data science from URJC and a master's degree in Artificial Intelligence from the Technical University of Madrid (UPM). He is part of the Data Science Laboratory high performance research group and has been an Assistant Professor at URJC since 2024. His research interests include active learning, explainable machine learning, generative artificial intelligence, and applied machine learning.

Isaac Martín de Diego was born in Campaspero, Valladolid, Spain in 1973. He received a PhD degree in Mathematical Engineering from Carlos III de Madrid University in 2005 (Extraordinary Doctorate Award). Since 2023 he is a full professor at the Higher Technical School of Computer Engineering at Rey Juan Carlos University (Associate Professor from 2018). He is the co-founder of the Data Science Laboratory and Head of the Sports Analytics Master at Rey Juan Carlos University. He has been head of the ERICSSON Chair on Data Science applied to 5G. He is the author of more than 100 articles. His research interests include methods, processes, and tools for Data Science in various application domains: explainability, sampling, complexity, performance evaluation, visualization, recommendation systems and security with a special interest in Machine Learning algorithms and a combination of information methods.

Marta Beltrán received the master's degree in electrical engineering from Universidad Complutense of Madrid (Spain) in 2001, the master's degree in industrial physics from UNED (Spain) in 2003 and the PhD degree from the Department of Computing, Universidad Rey Juan Carlos, Madrid (Spain) in 2005. She is currently an Associate Professor at this department (on leave). She has published extensively in high-quality national and international journals and conference proceedings in the areas of parallel and distributed systems, cybersecurity and privacy. Her current research interests are Cloud computing, Edge/Fog Computing and Internet of Things, specifically, risk management, identity management and privacy-preserving mechanisms for these paradigms.

Received: January 31, 2024; Accepted: August 10, 2025.

Computer Science and Information Systems 22(4):1555–1575 https://doi.org/10.2298/CSIS240628061P

Hyperparameter optimisation in differential evolution using Summed Local Difference Strings, A Rugged But Easily Calculated Landscape For Combinatorial Search Problems

Husanbir Singh Pannu[1] and Douglas B. Kell[2]

[1] Department of Computer Science and Engineering, Thapar Institute
Patiala India 147004
hspannu@thapar.edu
[2] Research Chair in Systems Biology / Director of GeneMill
University of Liverpool UK L69 3BX
douglas.kell@liverpool.ac.uk

Abstract. We analyse the effectiveness of differential evolution hyperparameters in large-scale search problems, i.e. those with very many variables or vector elements, using a novel objective function that is easily calculated from the vector/string itself. The objective function is simply the sum of the differences between adjacent elements. For both binary and real-valued elements whose smallest and largest values are min and max in a vector of length N, the value of the objective function ranges between 0 and *(N-1)* × *(max-min)* and can thus easily be normalised if desired. String length, population size and generations for computational iterations have been studied. Finally, a neural network is trained by systematically varying three hyper-parameters, viz population (NP), mutation factor (F) and crossover rate (CR), and two output target variables are collected (a) median (b) maximum cost function values from 10-trial experiments and compared with SMAC3 and OPTUNA against grid and random search.
Keywords: Rugged landscape, differential evolution, neural networks, machine learning, optimization

1. Introduction

The tunably rugged fitness landscape reflects the intuition that combinatorial search problems can be seen in terms of a 'landscape' containing valleys and hills [26]. The NK model invented by Stuart Kauffman [27] can be adjusted by changing N (string length) and K to define the ruggedness level of the flexible landscape as shown in **Figure 1 .** To explain the search of most rugged string in the NK-landscape, a random three letter string has been considered (JIE) in Fig. 1(A). In the consecutive iterations, various single letter alterations have been considered in each iteration and cumulative distance of consecutive letters has been used for the cost function (Fig. 1(B)). Furthermore, the connected graph shows potential search paths in pursuit of extreme points (highest peak and deepest valley) within the combinatorial search space of string sequences. Darkness in colours in Fig. 1(C) signifies higher cost functions and finally in Fig. 1(D) this graph is one of the various tracks in complex NK-landscape showing the challenge of the optimization algorithm. Combinatorial search problems are common in both theoretical and

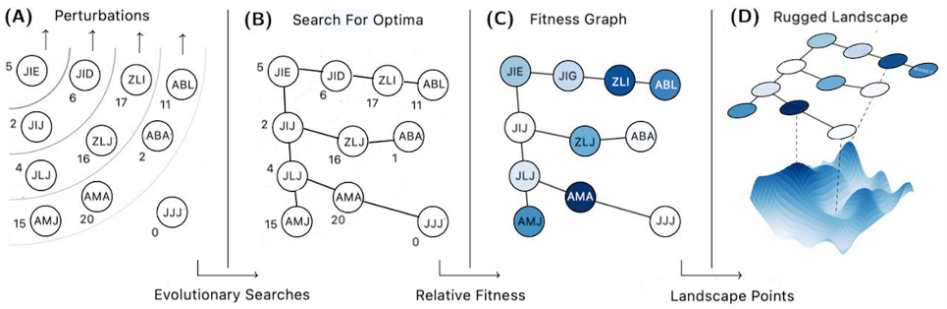


Fig. 1. (A) Combination space of various example string sequences of length 3. The digits represent the cumulative inter-letter distances of the string; the sequences along the arrows show the alterations of letters within consecutive strings to signify perturbations (B) Graphical rendering of potential paths to search for the global extrema (peak or valley) in the combinatorial string space (C) Relative colouring for visual comparison, darker means bigger cost function value i.e. higher cost function value (D) The graphs is one of the search possibility in the high NK-landscape showing the complexity of the search of extreme points through an algorithm

applied sciences such as optimisations, biology and complex evolutionary systems. As an example, in a business organisation, an agent may be searching a landscape of business opportunities. The valleys and hills represent the losses and profits. The journey through the landscape associates the decisions of the organisation whilst altering the structure of the organisation and modifying the products and services. All of the underlying processes interact in a complex evolutionary fashion and affect the cash flow and thus profit [29]. Most scientific problems can in fact be cast as combinatorial search or optimisation problems [28]. The kinds of optimisation in which we are here interested thus consist of a 'search space' or landscape in which a variety of inputs can be combined in potentially complex and nonlinear ways to lead to an output or objective function. Because the number of combinations always scales exponentially with the number of variables, the search spaces can easily be made to be far beyond any kind of exhaustive search (other than for small numbers of variables [41]), whether the search is computational or experimental. Heuristic methods, in which we seek to understand, simulate and navigate the landscape intelligently, are therefore appropriate. Among these, evolutionary algorithms of various kinds are pre-eminent [5]. Equally, because of the vastness of the search spaces, a variety of attempts have been made to create them *in silico*, using a more-or-less complex function of the inputs to calculate the output at that position in the search space. The idea of such strategies is that (notwithstanding that there is 'no free lunch' [11], [13], [40], [49], [48] an algorithm of interest may be assessed in competition with others [42], or its hyperparameters tuned to effect the most rapid searches. In evolutionary computing, three of these hyperparameters are the population size, the mutation rate, and the crossover rate [21]. In some cases, these hyperparameters can have very considerable effects on the efficiency of a search e.g. [12].

Well-known fitness functions used for creating landscapes of this type, often tuned to be 'deceptive' to evolutionary algorithms [22], [17], include NK [27],[21], [22] (and

variants such as NKp [6], OneMax/max-ones [45], [47], and the royal road [7], [43], [35]. Max-ones is especially easy to understand, since each variable is cast as an element of a (binary) string or vector of length N, and the fitness is simply the total number of elements containing a 1, with the maximum possible fitness obviously being N. That said, max-ones yields reasonably easily to evolutionary algorithms [23] as the crossover operator allows such algorithms more-or-less easily to combine building blocks (schemata [21]) successfully, not least because the 'fitness' of any element of a string is not context-sensitive.

We here develop and exploit a simple, related objective function in which the objective is not to maximise each element but to maximise *the sum of the differences between adjacent elements*. This is very easily calculated, allowing rapid assessment of different search algorithms. It provides for a landscape that is locally smooth but globally very rugged. Here the contribution to the overall fitness of any element of the string is absolutely context-sensitive. Thus, the problem is to find the most rugged string of length *n* using differential evolution which has three tuning parameters, viz. NP (population size of search particles), F (mutation factor), CR (crossover rate). The objective function for the hyperparameter optimisation for summed local difference strings has been defined with the following examples.

Example 1: String = "AAAA" has summed differences $|A-A|+|A-A|+|A-A| = 0$. String "ABCD" has summed differences $|A-B|+|B-C|+|C-D| = 3$. String "AZAZ" or "ZAZA" has $|A-Z|+|A-Z|+|A-Z| = 75$.

Example 2: The maximum ruggedness value of a string of length n is $(n-1)\times(U-L)$ where U and L are the maximum and minimum values of the alphabet values. In example 1 U = Z = 26 and L = A = 1 and n=4, thus $(4-1)\times(26-1) = 75$.

Even for binary strings this makes the problem much harder than max-ones. In addition, two very different (in fact maximally different) strings of length N have the same, maximum fitness, viz 101010...1010101 and 010101...0101010 of (N-1). We later also consider real-valued (integer) strings, such that the problem difficulty can be varied not only by varying the string length but by varying the number of allowable values in each position. Of the many variants of evolutionary algorithm, we focus on differential evolution, as originated by Storn and Price [38], [44] and reviewed e.g. in [9], [18], [36], [1], [16], [15], [46], as it seems to be highly effective in solving a wide range of problems.

Objective function and machine learning models need to be optimised according to the data distribution in order to find the best representative generalisation. Hyper-parameter tuning determines the best combination of free variables so that validation set yields the best performance. Hyperparameters have been tuned either by manual hit-and-trial, or through grid search, which involves systematically trying all possible combinations with a specified linear spacing. But both of these methods are limited to the human imagination or time duration to attempt grid search on a given level of granularity. Thus, an automatic machine learning based hyperparameter value optimisation has been proposed such as [30], [24], [10].

The paper is organised as follows: the first section is about the introduction of the rugged fitness landscape and local difference strings, Section 2 is a literature review, Section 3 covers the differential evolution and machine learning techniques used, Section 4 is about the results and discussion, Section 5 is the conclusion and future scope.

2. Literature Review

2.1. Differential Evolution Variants

In [31], a variant of DE using asymptotic termination based on the average differential of the cost function values has been proposed. The second modification is a new search for a critical parameter which helps to explore the search space. In [32], an ensemble of control parameters and mutation methods for DE has been proposed while considering the dynamic mutation strategies and set of values for control parameters. In [32], monkey king differential evolution using a multi-trial vector has been studied. The relation between exploitation and exploration depends upon control parameter and evolution strategy. To enhance the performance, multiple evolution strategies have been considered to generate multi-trial vectors. In [50], an ensemble of DE using multi-population approach and three distinct mutation methods, "rand/1", "current-to-rand/1", "current-to-pbest/1". CEC 2005 benchmark functions have been used for performance evaluation.

2.2. Hyper-parameter Tuning

In [2] an advanced hyper-parameter tuning technique 'Optuna' has been proposed. It allows API for dynamic user interaction, efficient search and pruning options, and easy to implement features. An automatic parameter tuner for sparse Bayesian learning has been proposed in [25]. The empirical auto-tuner has been used to address the neural network-based learning for performance comparison. In [20], a combination of stochastic differential equations and neural networks has been studied to extract the best combination of free parameters for the application of the economics dataset from Greater London using a Harris-Wilson model for a non-convex problem. In [25], a multi-label classification and complex regression problem has been addressed for auto-parameter tuning in deep learning models. In [8], ANNs have been used to predict the parameters for DE using 24 test problems from a Black-Box Optimisation Benchmarking dataset. In [34], parameter independent DE for analytic continuation has been studied using imaginary correlation functions of time. The parameters are embedded into the vectors which need to be optimised through the evolution. A study in [39] has proposed parameter optimisation for DE for the CED05 contest dataset that includes 25 complex mathematical functions with dimensions as high as 30.

3. Background

This section discusses the background techniques of differential evolution and artificial neural networks used in this research.

3.1. Differential Evolution (DE)

It is an efficient metaheuristic algorithm for numerical optimisation in which the output cannot be precisely defined from the input variables. For a given dimension of the data vector (string for example) it takes few input parameters such as number of points searching for the solution (population NP), mutation factor (F) and crossover rate (CR).

The algorithm has 4 phases of execution: initialisation of population particles, mutation, crossover and then selection [48] as shown in Fig. 2. This whole computation is repeated for a specified number of iterations (also known as generations) or a constrained time frame. Initialisation is usually done randomly from the normal distribution in the search domain followed by mutation which means finding the best parents to yield the child particle. Afterwards, crossover yields the child particle by varying proportions of parent particle attributes. The final step is selection, which means to update the current particle, how to use the new child along with other randomly selected particles with tuneable proportions. The idea is to maximise the randomness to avoid getting locked in local extrema. In our study, the simplest mutation formula (DE/rand/1) is used which is defined in (1) below:

$$Y_i = X1_i + F(X2_i - X3_i) \tag{1}$$

where *i* is the ith point computed in an iteration from X1, X2 and X3 random points out of the population. Next, to increase the diversity of muted vectors, the crossover operation is performed which is defined in (2). It mixes the target vector with another random vector in the population in an adjustable proportion using random probability function which can be defined using CR value. CR rate thus defines the ratio in which the new trial vector *Ui* inherits the values from mutation vector.

$$U_i = \{ \begin{array}{ll} Y_i & rand(0,1) \leq CR \\ X_i & otherwise \end{array} \tag{2}$$

For the selection phase, the cost function of this new trial vector is calculated after the crossover phase to compare with the fitness of the target vector Xi. The better among the two is selected to update the army of points in the population. Let *f(.)* be the fitness function then, selection is defined as in (3) below:

$$X(i+1) = \{ \begin{array}{ll} Y_i & f(Y_i) < f(X_i) \\ X_i & otherwise \end{array} \tag{3}$$

3.2. Artificial Neural Networks (ANN)

Artificial neural networks in data analysis got their origin from the behaviour of biological neurons on human brains. ANN consists of artificial nodes (neurons) which are interconnected through layers of other neurons to compute and refine the data using non-linear activation functions in the successive layers. Connection strengths among neurons is controlled by weight parameters (W). The simplest ANN in which we are interested is the multilayer perceptron (MLP), a three-layer structure is defined which contains the input layer, the hidden layer and the output. The standard structure of ANN is illustrated in **Figure 3** and details can be found in [3].

4. Results

Classical differential evolution uses three hyperparameters: the population size NP, a mutation factor F controlling the mutation rate, and a parameter CR that determines the extent

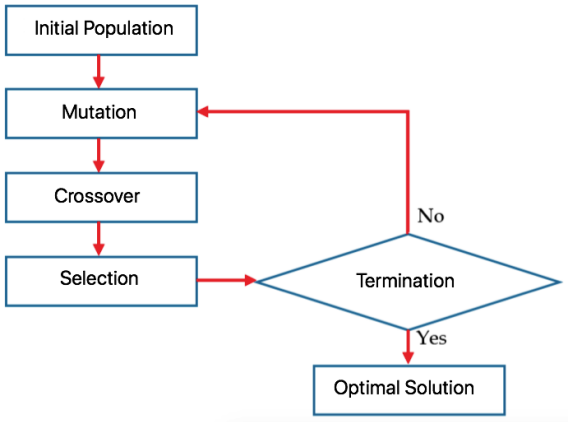


Fig. 2. Flow of differential evolution metaheuristic algorithm

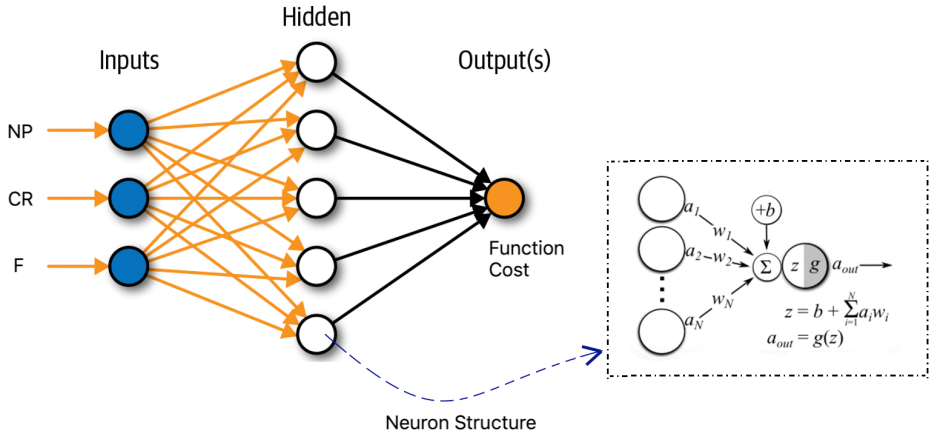


Fig. 3. An example artificial neural network with three inputs, one hidden layer with five neurons and the output layer . The inside structure of each neuron has been shown in the dotted box. A node or neurons is the weighted average of the input signals added with a constant (bias) value and travels through a function which is usually non-linear in nature such as a tanh, sigmoid or ReLU (Rectified Linear Unit) function

of the uniform crossover [45], [38] operator used. We start with standard and fixed (non-adaptive) values of NP = 50, F = 0.4, CR = 0.1 as recommended by Storn and Price. NP is usually scaled to the length of the vector (number of input variables) to be optimised, but not much is known for problems that have a great many input variables [37], for instance directed protein evolution [4],[14]; others are reviewed e.g. in [8].

4.1. Understanding the landscape ruggedness

To understand the nature of the landscape, it is convenient to create a random landscape in which, initially each element of the vector of length N=50 is set randomly to A to Z (**Table 1**). The random letters in each of the 10 vectors are the population candidates to search for the best rugged string of length 50.

Table 1. Table 1 Random landscape of 10 vectors (population size) with dimensions = 50 (string length) and initial cost function value (summed consecutive differences)

SR	Vector of length 50	Cost
1	GLPFFCVOBVANQGUQKHRRWBIDKURNCNYOCDBCQLQIUXHONOJGRK	384
2	UNQBHPJUCEZOADILXMBZVOVGFIKQTCZZVRCGSDXRTGZSBPRWSG	454
3	OOSKURLLTNYAORMWLJYLRQSRWNBKOYNAXRRNQJJJXSEIYXHNQQ	358
4	EJGFPREJIIHODQEXKTTNPBOEJGCNDRUFPDYVEOKXMWPWFTORLM	406
5	OBBMAREGDIBWCYMBEEBSPTIHNXOPSOMMXCPHJJFGDHRJEUNFGC	372
6	CPICSKHOLMUSGVJWJFMLUTMTRUZXQCBNZYCDVRFNWKDVGAXEJN	417
7	CYKGXKQWNTYTVYLPOCHLHOOPYPYUVZTEHKBRPCLADTIIOVJBGP	363
8	UYHYLMSVWLTLKWNKRKFBOUENTVSVCKJTJBMDYTAIIQLDDVBONY	408
9	AUMHUZPEHTEDLFCSBMMPFUZGGYSOJNUEMWAFCLSNZKMRJZDOIB	449
10	YIIBQZDVPNOABPVXDCALDDGXVTVPTUEKQVMIVBKWUCFESABITC	404

After 10 iterations of the DE algorithm, the best particle and cost are as follows. Particle = APGZ TTTA PZIZ HZAZ ZDAZ BAEA ZAAZ AVWZ ABAZ AZAZ AZZZ AZAZ AZ, Cost function = 773, Benchmark = $(26-1)\times(50-1)=1,225$. Elapsed time is 0.1955 seconds. The values of Dimensions = 50, Np = 10, Population Crossover Rate (CR) = 0.8 and Mutation Factor (F) = 0.85 for this example. **Figure 4** shows the consistent improvement in the cost function.

4.2. Varying the hyperparameters in standard differential evolution

All simulations are performed in MATLAB 2018b software with a system configuration of MAC Air (2017), 1.8 GHz Intel Core i5 processor, 8GB 1600 MHz DDR3 memory, HD Graphics 6000 1536 MB and macOS version 10.13.6. The experiment was repeated for 500-dimensional string, 10-trials and 50 generations with NP in [50,500], CR in [0.1,1], F in [0.4,0.9]. The stats were recorded for the median and maximum function values for 10-k trials listed in **Table 2**. A total of 600 values were collected for various combinations of NP, F and CR values suggested in [8]. Out of those 600 values, the best 5 were selected and they were calculated for extended iterations/generations as shown in **Table 3**. It infers that the cost function value generally tends to increase with iterations but not always. This

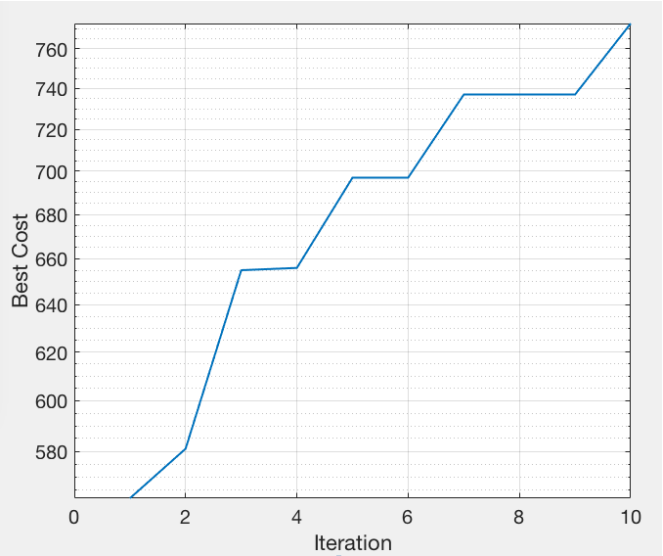


Fig. 4. Cost function values for 10 iterations for string length = 50. The values of Dimensions = 50, Np = 10, Population Crossover Rate (PCr) = 0.8 and Mutation Factor (F) = 0.85

is because at higher dimensions the problem is highly difficult to solve and not always yields the best solution during the metaheuristic search.

Table 2. Benchmark value for 500-dimensional string is $(26-1)\times(500-1) = 12,475. NP = \{50, 100, \ldots, 500\}, CR = \{0.1, 0.2, \ldots 1\}, F = \{0.4, 0.5, \ldots 0.9\}$ so total $10 \times 10 \times 6 = 600$ experiment values. The min and max values refer to the range of cost function values obtained through the experiments

	Range of function values	
10-trials	Minimum	Maximum
Median of cost function	5555.5	8870
	44.5%	71.1%
Maximum of cost function	5629	9226
	45.1%	73.96%

Table 4 and **Figure 5** illustrate the normalised function values for various combinations of population size and generations. For median of cost function values, it has been found that 9196 value is obtained with the NP=50 and Gen=500. For the maximum cost function values the value 9373 is found when NP=125 and Gen=200. Thus, it is not easy to say which of the NP and Generation values are the best ones to yield the best function cost but generally more generations with bigger population size is a good combination.

Table 3. Top 5 cases among 600 trials analysed for increased number of iterations from 100 to 1000. The winners are highlighted in bold text

CASE1	NP	F	CR	Iterations	100	200	300	500	1000
	500	0.9	1	Median	9632	10300	10375	10300	**10375**
				Maximum	9716	10550	10425	10375	**10625**
CASE2	NP	F	CR	Median	7935	8970	9269	9280	**9295**
	500	0.8	0.9	Maximum	8222	9106	9322	**9592**	9538
CASE3	NP	F	CR	Median	8801	9246	9159	9012	**9230**
	500	0.9	1	Maximum	8901	9362	9342	9181	**9328**
CASE4	NP	F	CR	Median	7002	9037	9218	**9536**	9242
	500	0.8	0.9	Maximum	7862	9152	9529	**9677**	9651
CASE5	NP	F	CR	Median	6075	8211	8500	**8675**	8450
	500	0.8	0.9	Maximum	6250	8451	8500	**8775**	8500

Table 4. Normalised comparison (NP*Gen constant) of cost function values (median and maximum over 10-trials) for F=CR=0.9, string length = 500. The values of VP and generations have been varied such that the product of NP and Gen is 25000. **Figure 5** is the visual illustration

Sr.	**NP**	**Generations**	**Median**	**Maximum**
1	500	50	8236	8272
2	250	100	9109	9394
3	125	200	9150	**9373**
4	50	500	**9196**	9304
5	10	2500	8137	8601

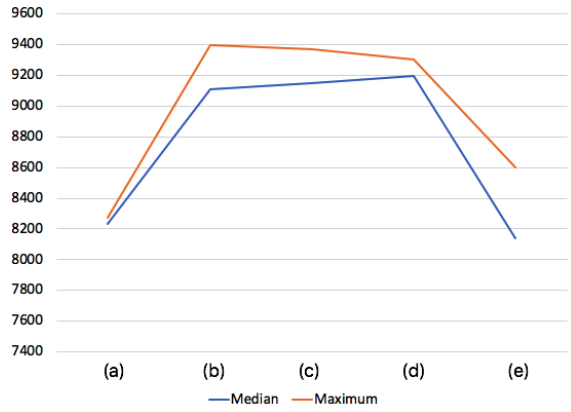


Fig. 5. Normalised comparison of various values of function cost for median and maximum using F=0.9, CR=0.9, string length (dimension)=500, 10-trials. (a) NP=500, Gen=50 (b) NP=250, Gen = 100 (c) NP=125, Gen=200 (d) NP=50, Gen=500 (e) NP=10, Gen=2500. The values of NP and Generations have been chosen such that the product remains constant (=25000) to analyse the effect of reducing population and increasing generations

Tables 5-7 illustrate time consumption analysis against variations in NP, string length and iterations of calculations. **Figures 6-8** are the graphical representations of these tables for the ease of visualisation. The time variations are almost linear giving the idea about the problem complexity in regard to the variable changes.

Table 5. For 50 iterations, string length 500 (data vector dimension), F=0.4, CR = 0.1 the time in seconds for various NP values has been calculated ranging from 50 to 500 for 10-trial experiments. **Figure 7** is the graphical rendering of this table

NP	50	100	150	200	250	300	350	400	450	500
Time	1.32	2.46	3.73	4.88	6.24	7.43	8.94	10.32	11.71	14.07

4.3. Training Data and ANN Training for Automatic Parameter Tuning

The neural network has been trained for 600 data points which includes three independent input variables: Population (NP), Mutation Factor (F) and Cross-mutation Rate (CR). The values range for NP is [50,500], F is [0.4,0.9], CR is [0.1,1] chosen according to our experiments and suggestion given in [50]. These 600 training data points have been collected by running 10-fold trials for each entry to find out median and maximum cost function values and is quite a time consuming task. This is due to finding the right range of these 3 hyperparameter values and then running the code to compile the data for few hours. So the differential evolution function is called 600×10 = 6,000 times for various values of these three hyperparameters. The training sample data and the computed median

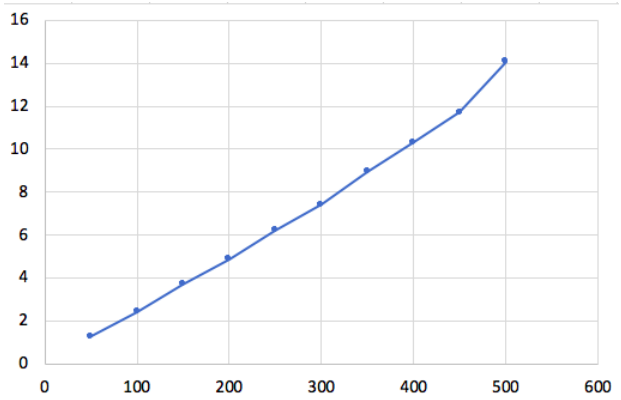


Fig. 6. Graph between NP (x-axis) and Time in seconds (y-axis) for fixed values of string length = 500, F = 0.4, CR = 0.1 and iterations = 50 for 10-trial experiments. The time consumption is nearly linear with the population size

Table 6. String length (Dimension) versus time consumption in seconds for fixed values of NP=100, F=0.4, CR=0.1, iterations = 50 and 10-trials for experiments to calculate median and mean function costs. The time consumption varies almost linearly with the string length. **Figure 8** shows the graphical representation

Dim	100	200	300	400	500	600	700	800	900	1000
Time	0.7992	1.1327	1.7634	2.1091	2.5229	2.8815	3.2637	3.5849	3.9694	4.2432

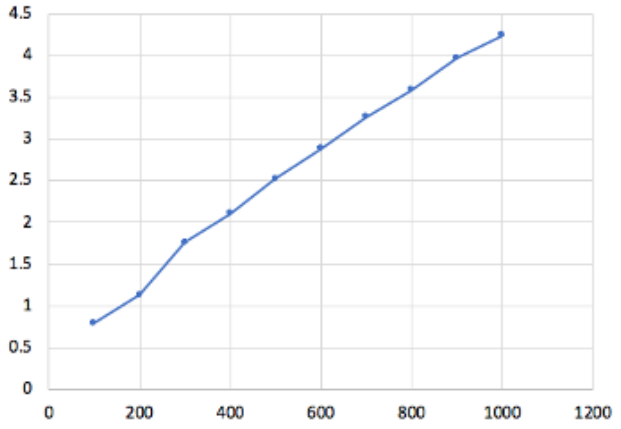


Fig. 7. Graph between string length (x-axis) and Time in seconds (y-axis) for fixed values o NP=100, F=0.4, CR=0.1, iterations = 50 and 10-trials for experiments to calculate median and mean. The time consumption varies almost linearly with the string length

Table 7. Iterations and time in seconds comparison for fixed values of string length = 500, NP=100, F=0.4, CR=0.1 and 10-trial experiments. The relation is almost linear and is rendered in **Figure 9**

Gen	50	100	150	200	250	300	350	400	450	500
Time	2.56	4.75	7.19	9.99	11.64	13.99	16.49	20.28	20.99	25.55

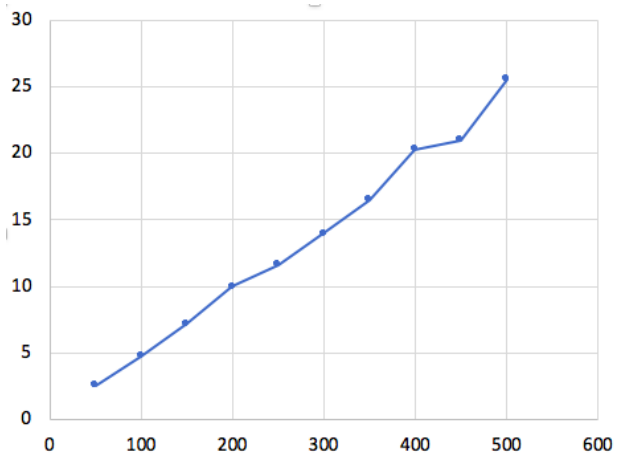


Fig. 8. Graph between iterations (x-axis) and time in seconds (y-axis) for fixed values of string length = 500, NP=100, F=0.4, CR=0.1 and 10-trial experiments. The relation is almost linear

and maximum function costs have been rendered in **Table 8**. Afterwards, Neural Network Fitting in MATLAB has been used for regression using three input variables and two target variables respectively. The default values of the model have been used, for example ratio of training: validation: testing is 70:15:15, number of hidden neurons = 10, Bayesian Regularization training algorithm for training. **Figures 9-10** demonstrate the frequency histogram of error values in relation to instances. After the training of ANN, the error of actual versus predicted values were collected for all training and tested values among 600 data points and plotted to see the quality of the trained model. Most of the error values are 0 which indicates that the model is well generalised and validated.

Table 8. Training data for ANN containing 3 input variables and two target variables. The $NP \in \{50, 100, \ldots, 500\}, F \in \{0.4, 0.5, \ldots, 0.9\}, CR \in \{0.1, 0.2, \ldots, 1\}$ so a total of $10 \times 6 \times 10 = 600$

Input variables				**Target var1**	**Target var2**
SR.	NP	F	CR	Median Fun Val.	Max Fun Val.
1	50	0.4	0.1	5559.5	5676
2	50	0.4	0.2	5705	5837
...	...	...	...	...	...
600	500	0.9	1	8870	9226

Tables 9-10 show the function costs on 10-trial average for the mean squared error (MSE) and regression coefficient (R) values as the number of hidden neurons increase during the ANN training.

Table 9. For the median cost function values (from 10-trials) the following are the MSE and R values for increasing number of neurons in ANN training

		Neurons	10	20	30	40	50	60	70	80
Median Function Cost	Train	MSE	3008.8	2102.2	1264.8	1472.8	1572.3	**1108.8**	1653	1540.5
		R	0.9975	0.9983	0.9987	0.9990	0.9985	0.9988	**0.9989**	**0.9989**
	Test	MSE	**2858**	3624	5883.7	5616.6	4622.5	7715.5	3162	5431.8
		R	0.9948	0.9959	0.9947	0.9962	**0.9974**	0.9946	0.9936	0.9957

Table 10. For the maximum cost function values (from 10-trials) the following are the MSE and R values for increasing number of neurons in ANN training

		Neurons	10	20	30	40	50	60	70	80
Max Fun. Cost	Train	MSE	8738	6682.7	6021.4	5308.3	5790.4	5791.7	5557.3	**5059.4**
		R	0.9937	0.9954	0.9958	0.9963	0.9958	0.9960	0.9961	**0.9964**
	Test	MSE	1148.3	**1192.8**	1308.1	1228.3	2106.7	1262.8	1689.5	1252.9
		R	**0.9934**	0.9916	0.9912	0.9914	0.9872	0.9908	0.9877	0.9926

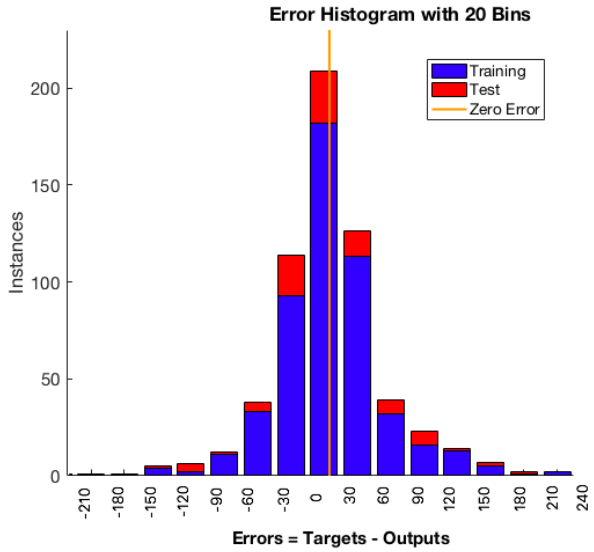


Fig. 9. Error histograms for training, testing and validation for 600 examples after ANN training for median cost function values. Ratio of train:validate:test used is 70:15:15. The target function used is the median cost function value from 10-trial experiments and input values to ANN were NP, F and CR

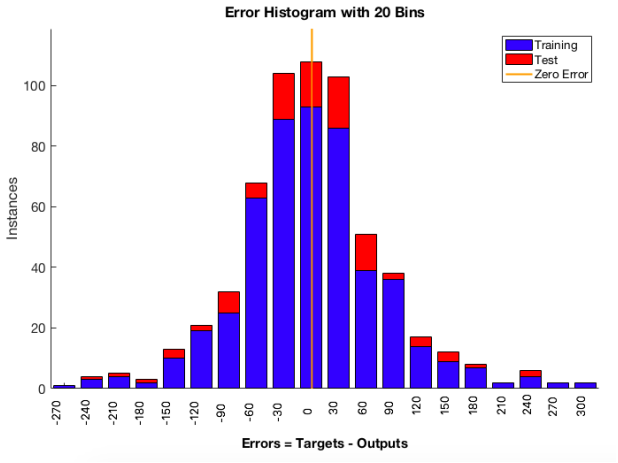


Fig. 10. Error histograms for training, testing and validation for 600 examples after ANN training for max cost function values. Ratio of train:validate:test used is 70:15:15. The target function used is the maximm cost function value from 10-trial experiments and input values to ANN were NP, F and CR

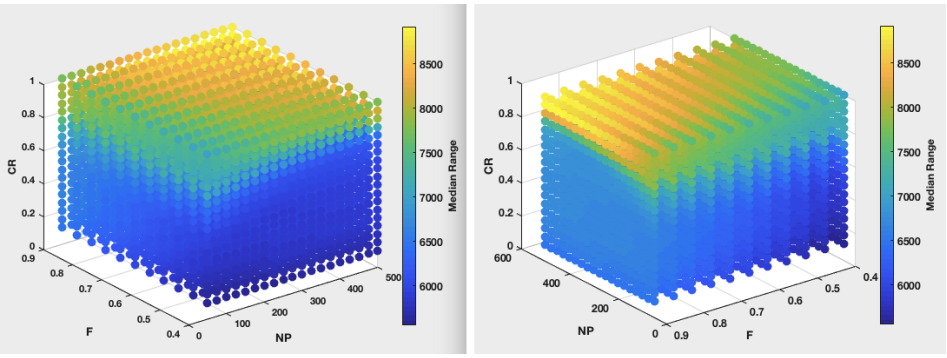


Fig. 11. The demonstration of 3,971 test points $(19 \times 11 \times 19)$ parameter combinations of $NP \in \{50, 75, \ldots, 500\}, F \in \{0.4, 0.45, ..., 0.9\}, CR \in \{0.1, 0.15, \ldots, 1\}$. The colour depicts the median of cost function values of 10-trial runs. The yellow shows higher (better) values and blue shows lower cost function values for those combinations of NP, F and CR. Left and right graphs are two perspective view of the same 3-d figure

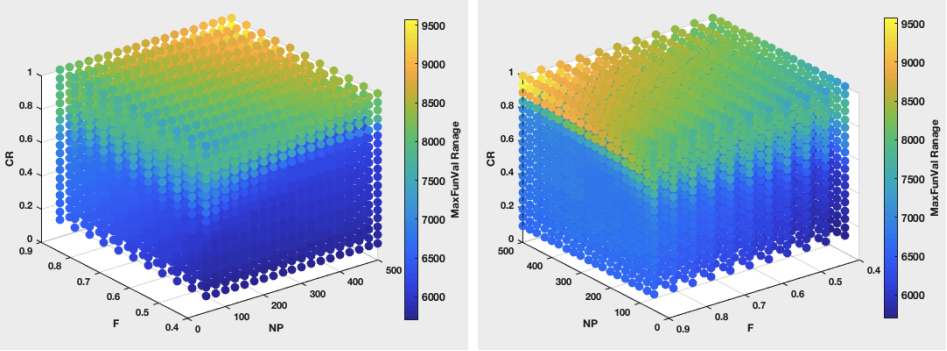


Fig. 12. The demonstration of 3,971 test points $(19 \times 11 \times 19)$ parameter combinations of $NP \in \{50, 75, \ldots, 500\}, F \in \{0.4, 0.45, \ldots, 0.9\}, CR \in \{0.1, 0.15, \ldots, 1\}$. The colour depicts the maximum value of cost function for 10-trial runs. The yellow shows higher (better) values and blue shows lower cost function values for those combinations of NP, F and CR. Left and right graphs are two perspective view of the same 3-d figure

Figures 11-12 show typical curves of the median and maximum fitness as a function of the number of evaluations for 10-trials. A 4-d representations of the test results obtained by ANN with 3,971 points has been shown, for a range of $NP \in \{50, 75, ..., 500\}, F \in \{0.4, 0.45, ..., 0.9\}, CR \in \{0.1, 0.15, ..., 1\}$ values. The colour illustrates the range of cost function values predicted by ANN. Yellow means higher (better) function value and blue means lower cost function value for that combination of NP, F and CR hyper-parameters. It can be observed that the best function values (both maximum and median values) are obtained from higher NP, CR and F values in general. It can be seen that the best function costs correspond to maximum values of NP, F and CR values for both figures.

4.4. Validation

Among the total tested values (3,971) we chose the best 10% values with maximum cost function values (yellow in **Figure 11-12**) to calculate the actual cost functions through differential evolution on the average of 10-trials. Afterwards, both the continuous variables (ANN suggested output and actual DE values) for (NP, F, CR) hyper-parameter combinations have been analysed using Pearson's correlation coefficients (ρ). For both cases of DE versus ANN test results $\rho_{\text{MEDIAN}} = 0.7$ and $\rho_{\text{MAXIMUM}} = 0.68$. It indicates a significant positive relationship among two variables [33]. Thus the hyperparameter optimisation in differential evolution with summed local difference strings can be performed efficiently using the neural network simulation.

4.5. Comparison with state-of-art

To analyse the effectiveness of ANN-based parameter tuning of DE, we compared the following techniques: (i) Grid Search, (ii) Random Search, (iii) Sequential Model-based Algorithm Configuration (SMAC) [25], (iv) Optuna [41](v) ANN in **Table 11** and **Figure 13**.

Grid Search suffers from the curse of dimensionality, resulting in an explosion in the number of possible evaluations, which is improved by Random Search. But random search is not well sorted and may miss the potential extremas. SMAC uses random forests (RF) and is a Bayesian optimiser in which RF helps in categorical variables to support large search space hyperparameter searches and is well scalable for increasing the number of training samples. It is available online https://github.com/automl/SMAC3. Optuna [2] is recent software used for hyperparameter optimisation using define-by-run API, pruning and search strategy implementation, versatile utility including distributed computing, scaling and interactive interface for users to modify the search space parameters dynamically. It is available online https://github.com/optuna/. Hyperparameter auto-tuning has also been performed for sparse Bayesian learning (SBL) in [19] using neural network-based learning, and has shown considerable improvement in recovery performance and convergence rate.

ANN performed better compared to Grid Search, Random Search, SMAC and Optuna (**Table 11 and Fig 13**.), but requires a considerable amount of time to generate the training data to simulate the behavior of the rigged function outcome for given values of NP, F and CR. Afterwards it is able to predict the new values of hyperparameters to search for the

Table 11. Hyperparameter optimisation for differential evolution algorithm to search for {NP,F,CR} by various algorithms. Benchmark value for string of length (=NP) of 500 = (26-1)×(500-1)=12,475 and 10 trial runs. Generations for DE=100 for all of them consistently and max of 10 trials experiments for the best cost calculations. Accuracy is the ratio of best cost and the benchmark (=12,475)

Methods	Details	Hyperpram Values	Best Cost	Accuracy
Grid Search	NP∈[50,500], F=CR=[0.40, 0.41,..1.0}	NP=390, F=0.95, CR=0.90	8616	0.6907
Random Search	NP∈[50,500], F, CR ∈ [0.4,1]	NP=450, F=0.91, CR=0.88	8572	0.6871
SMAC	NP ∈[50,500], F, CR ∈ [0.4,1]	NP=410, F=1, CR=0.98	8990	0.7206
Optuna	NP∈[50,500], F, CR ∈ [0.4,1]	NP=430, F=,0.87 CR=0.92	9244	0.7410
ANN	NP∈[50,500], F=CR=[0.40, 0.41,..1.0}	NP=500, F=0.90, CR=0.89	9438	0.7566

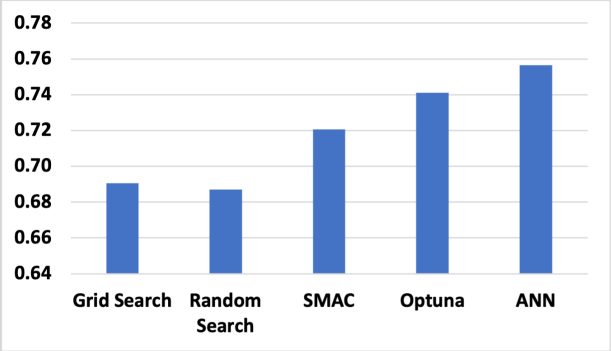


Fig. 13. Graphical representation of accuracy comparision of Table 11. Our ANN performed better than the other state-of-art hyperparameter techniques studied.

optimal combination of tuning parameters. ANN requires much less time as compared to DE to calculate the cost function once it is trained on a good sized set. For the sake of simplicity, only 600 examples have been used to train ANN for this case study, but a few thousand runs of DE would be better to yield the training data for ANN.

5. Conclusion

In this research, hyperparameter optimisation in differential evolution has been studied using Summed Local Difference Strings, which is a rugged but easily calculated landscape for combinatorial search problems with wide applicability in numerical optimisation, biological sciences, finance and organisational management. Differential evolution is a powerful numerical optimisation technique for non-differentiable and complex functions which cannot be nicely defined in mathematics, but it has three hyperparameters

(NP, F, CR) to be optimised. In this study a machine learning technique has been exploited to suggest the best possible combinations of hyper-parameters instead of a tedious grid search. The limitation of the technique is that training data collection is time consuming and needs a careful analysis of input variable ranges (NP, F, CR). Two output variables were recorded (median and maximum value) after 10-trial experiments of each combination of the hyperparameters. Finally, testing of the machine learning model has been employed on a bigger data (3,971) and the top 10% test results were compared with the actual DE results yielding a Pearson correlation coefficient of 0.7. Specifically, it was found that larger values of NP, F and CR hyper-parameters yield better outcomes.

In future, we plan to explore more bio-inspired algorithms and other ways to search for hyperparameters such as to embed the hyper-parameters in the very population being optimised. We will also explore binary strings and real valued strings with an extended range of values beyond 26 with advanced options of mutation and cross-over equations.

Acknowledgments. We thank the UK EPSRC and AkzoNobel for financial support via the SusCoRD project, grant EP/S004963/1. DBK thanks the Novo Nordisk Foundation for financial support (grant NNF20CC0035580).

References

1. Ahmad, M.F., Isa, N.A.M., Lim, W.H., Ang, K.M.: Differential evolution: A recent review based on state-of-the-art works. Alexandria Engineering Journal 61(5), 3831–3872 (2022)
2. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
3. Anderson, J.: An introduction to neural networks. mit press. Cambridge, MA., ISBN 10, 0262011441 (1995)
4. Arnold, F.H.: Directed evolution: bringing new chemistry to life. Angewandte Chemie (International Ed. in English) 57(16), 4143 (2017)
5. Bäck, T., Fogel, D.B., Michalewicz, Z.: Handbook of evolutionary computation. Release 97(1), B1 (1997)
6. Barnett, L., et al.: Ruggedness and neutrality-the nkp family of fitness landscapes. In: Artificial Life VI: Proceedings of the sixth international conference on Artificial life. pp. 18–27 (1998)
7. Belding, T.C.: Potholes on the royal road. arXiv preprint cs/0104011 (2001)
8. Centeno-Telleria, M., Zulueta, E., Fernandez-Gamiz, U., Teso-Fz-Betoño, D., Teso-Fz-Betoño, A.: Differential evolution optimal parameters tuning with artificial neural network. Mathematics 9(4), 427 (2021)
9. Chakraborty, U.K.: Advances in differential evolution, vol. 143. Springer Science & Business Media (2008)
10. Charilogis, V., Tsoulos, I.G., Tzallas, A., Karvounis, E.: Modifications for the differential evolution algorithm. Symmetry 14(3), 447 (2022)
11. Corne, D.W., Knowles, J.D.: No free lunch and free leftovers theorems for multiobjective optimisation problems. In: International Conference on Evolutionary Multi-Criterion Optimization. pp. 327–341. Springer (2003)
12. Corne, D., Oates, M., Kell, D.: On fitness distributions and expected fitness gains of parallelised mutation operators: Implications for high mutation rates and rate adaptation in parallel evolutionary algorithms. Parallel problem solving from nature-ppsn vii, Merelo Guervós, JJ; Adamidis, P pp. 132–141 (2002)

13. Culberson, J.C.: On the futility of blind search: An algorithmic view of "no free lunch". Evolutionary Computation 6(2), 109–127 (1998)
14. Currin, A., Swainston, N., Day, P.J., Kell, D.B.: Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. Chemical Society Reviews 44(5), 1172–1239 (2015)
15. Das, S., Mullick, S.S., Suganthan, P.N.: Recent advances in differential evolution–an updated survey. Swarm and evolutionary computation 27, 1–30 (2016)
16. Das, S., Suganthan, P.N.: Differential evolution: A survey of the state-of-the-art. IEEE transactions on evolutionary computation 15(1), 4–31 (2010)
17. Davidor, Y.: Epistasis variance: A viewpoint on representations, ga hardness, and deception. Complex systems 4(4), 1–20 (1990)
18. Feoktistov, V.: Differential evolution: in search of solutions. Springer (2006)
19. Gao, D., Guo, Q., Jin, M., Liao, G., Eldar, Y.C.: Hyper-parameter auto-tuning for sparse bayesian learning. arXiv preprint arXiv:2211.04847 (2022)
20. Gaskin, T., Pavliotis, G.A., Girolami, M.: Neural parameter calibration for large-scale multiagent models. Proceedings of the National Academy of Sciences 120(7), e2216415120 (2023)
21. Golberg, D.E.: Genetic algorithms in search, optimization, and machine learning. Addion wesley 1989(102), 36 (1989)
22. Goldberg, D.E.: Genetic algorithms and walsh functions: Part 2, deception and its analysis. Complex systems 3, 153–171 (1989)
23. He, J., Chen, T., Yao, X.: On the easiest and hardest fitness functions. IEEE Transactions on evolutionary computation 19(2), 295–305 (2014)
24. Hutter, F., Kotthoff, L., Vanschoren, J.: Automated machine learning: methods, systems, challenges. Springer Nature (2019)
25. Iliadis, D., Wever, M., De Baets, B., Waegeman, W.: Hyperparameter optimization in deep multi-target prediction. arXiv preprint arXiv:2211.04362 (2022)
26. Kauffman, S., Levin, S.: Towards a general theory of adaptive walks on rugged landscapes. Journal of theoretical Biology 128(1), 11–45 (1987)
27. Kauffman, S.A., Weinberger, E.D.: The nk model of rugged fitness landscapes and its application to maturation of the immune response. Journal of theoretical biology 141(2), 211–245 (1989)
28. Kell, D.B.: Scientific discovery as a combinatorial optimisation problem: how best to navigate the landscape of possible experiments? Bioessays 34(3), 236–244 (2012)
29. Levinthal, D.A.: Adaptation on rugged landscapes. Management science 43(7), 934–950 (1997)
30. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A.: Hyperband: A novel bandit-based approach to hyperparameter optimization. Journal of Machine Learning Research 18(185), 1–52 (2018)
31. Mallipeddi, R., Suganthan, P.N., Pan, Q.K., Tasgetiren, M.F.: Differential evolution algorithm with ensemble of parameters and mutation strategies. Applied soft computing 11(2), 1679–1696 (2011)
32. Nadimi-Shahraki, M.H., Taghian, S., Zamani, H., Mirjalili, S., Elaziz, M.A.: Mmke: Multi-trial vector-based monkey king evolution algorithm and its applications for engineering optimization problems. Plos one 18(1), e0280006 (2023)
33. Nettleton, D.: Selection of variables and factor derivation. Commercial data mining 48, 79–104 (2014)
34. Nichols, N.S., Sokol, P., Del Maestro, A.: Parameter-free differential evolution algorithm for the analytic continuation of imaginary time correlation functions. Physical Review E 106(2), 025312 (2022)
35. van Nimwegen, E., Crutchfield, J.P., Mitchell, M.: Statistical dynamics of the royal road genetic algorithm. Theoretical Computer Science 229(1-2), 41–102 (1999)

36. Onwubolu, G.C., Davendra, D.: Differential evolution: a handbook for global permutation-based combinatorial optimization, vol. 175. Springer (2008)
37. Piotrowski, A.P.: Review of differential evolution population size. Swarm and Evolutionary Computation 32, 1–24 (2017)
38. Price, K.V., Storn, R.M., Lampinen, J.A.: Differential evolution: a practical approach to global optimization. Springer (2005)
39. Ronkkonen, J., Kukkonen, S., Price, K.V.: Real-parameter optimization with differential evolution. In: 2005 IEEE congress on evolutionary computation. vol. 1, pp. 506–513. IEEE (2005)
40. Rowe, J.E., Vose, M.D., Wright, A.H.: Reinterpreting no free lunch. Evolutionary computation 17(1), 117–129 (2009)
41. Rowe, W., Platt, M., Wedge, D.C., Day, P.J., Kell, D.B., Knowles, J.: Analysis of a complete dna–protein affinity landscape. Journal of The Royal Society Interface 7(44), 397–408 (2010)
42. Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Design and analysis of computer experiments. Statistical science 4(4), 409–423 (1989)
43. Spirov, A.V., Myasnikova, E.M.: Heuristic algorithms in evolutionary computation and modular organization of biological macromolecules: Applications to in vitro evolution. Plos one 17(1), e0260497 (2022)
44. Storn, R., Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization 11(4), 341–359 (1997)
45. Syswerda, G., et al.: Uniform crossover in genetic algorithms. In: ICGA. vol. 3 (1989)
46. Tanabe, R., Fukunaga, A.: Reviewing and benchmarking parameter control methods in differential evolution. IEEE transactions on cybernetics 50(3), 1170–1184 (2019)
47. Tuson, A., Ross, P.: Adapting operator settings in genetic algorithms. Evolutionary computation 6(2), 161–184 (1998)
48. Wolpert, D.H.: What is important about the no free lunch theorems? In: Black box optimization, machine learning, and no-free lunch theorems, pp. 373–388. Springer (2021)
49. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE transactions on evolutionary computation 1(1), 67–82 (2002)
50. Wu, G., Mallipeddi, R., Suganthan, P.N., Wang, R., Chen, H.: Differential evolution with multi-population based ensemble of mutation strategies. Information Sciences 329, 329–345 (2016)

Husanbir Singh Pannu, PhD (https://sites.google.com/a/thapar.edu/hspannu/cv) is working as an Assistant Professor in the Computer Science and Engineering Department, at Thapar Institute of Engineering and Technology, Patiala, India. His research interests are Machine Learning, Image Processing, Data Analysis, Multimodal Learning systems and Numerical Optimizations. He was a postdoc research fellow at Trinity College Dublin Ireland, and University of Liverpool UK. He received his PhD from the University of North Texas USA, Master's from California State University Eastbay USA, MTech and BTech in Computer Science and Engineering. He has published about 50 research papers including journals, and international conferences. He is also an active peer reviewer of SCIE journals. He has finished supervision of 10 Master's and 3 PhD students.

Douglas Kell (CBE, DSc, MA, DPhil, FAAAS, FLSW, FSB, Research Chair in Systems Biology / Director of GeneMill, https://www.liverpool.ac.uk/people/douglas-kell) studied Biochemistry at Oxford University (including a Distinction in Chemical Pharmacology) followed by a D Phil at the same Institution. He spent many years at the University of Aberystwyth before moving to Manchester (UMIST) in 2002, where he led the Manchester Centre for Integrative Systems Biology. He came to Liverpool in December 2018.

He was a Founding Director of Aber Instruments Ltd (Queen's Award for Export Achievement 1998), and he has also served on secondment (0.8 FTE) as Chief Executive of the Biotechnology and Biological Sciences Research Council (2008-2013).

Received: June 28, 2024; Accepted: December 01, 2024.

Computer Science and Information Systems 22(4):1577–1598 https://doi.org/10.2298/CSIS250310062T

Digital Transformation in Public Accounting and Finance Management: a Clusters Literature Review

Ambrósio Teixeira[1], Xavier Martinez-Cobas[2], Alvaro Rocha[3], Maria José Gonçalves[4], Amélia Silva[5]

[1]Department of Accounting and Finance, CEOS.PP, ISAL, Funchal, Madeira, Portugal
ambrosio.teixeira@isal.pt
[2]Department of Accounting and Financial Economics, University of Vigo, Vigo, Galicia, Spain
xmcobas@uvigo.gal
[3]Department of Information Systems, ISEG, University of Lisbon, Lisboa, Portugal
amr@iseg.ulisboa.pt
[4]Department of Information Systems, CEOS.PP, ISCAP, Polytechnic of Porto, Porto, Portugal
mjose@iscap.ipp.pt
[5]Department of Accounting, CEOS.PP, ISCAP, Polytechnic of Porto, Porto, Portugal
acfs@iscap.ipp.pt

Abstract. This study investigates the literary corpus on the role and potential of digital transformation in public accounting and finance management. A total of 890 research papers was extracted from Scopus and Web of Science for bibliometric analysis, to investigate publishing trends, productive countries, and keyword analysis around the topic, and 24 relevant research publications, divided into two clusters, were selected for an in-depth analysis. The findings demonstrate that technologies have significantly transformed accounting and public finance by automating processes to reduce errors and save time, increasing transparency and accountability, preventing fraud with analytical tools, improving budget planning and monitoring, and integrating systems for a comprehensive financial view.

Keywords: accountability, digital transformation, public accounting, public finances, public management

1. Introduction

In today's interconnected global landscape, the expectations placed on Public Administration by both citizens and organisations have intensified significantly. The State is expected to optimise bureaucratic processes, modernise the administrative systems, and be transparent and accountable. Information technologies play a key role in this process, fostering communication between Public Administration and society, reshaping public services, and enabling governments to leverage technological innovation for enhanced public service delivery [1]. Technology makes it possible to identify new opportunities for Public Administration, particularly in the implementation of e-government, creating public value for the services provided [2], [3].

At the end of the day, it all begins and ends with new public financial management because it is all about how effective, efficient, and economical Public Administration is

in managing public resources. In other words, it concerns whether New Public Management can perform public responsibilities “as one of the basic responsibilities accepted by governments, with citizens as the major suppliers of financial resources for public sector institutions" [4].

Technology not only makes it possible to offer more effective and accessible public services to citizens, but it also enables strategic decision-making based on integrated and centralised data, collecting and making data available for analysis, as a basic product for implementing, monitoring, and developing predictive models to anticipate the future. This emphasis on data requires the adoption of new accounting practices and, consequently, new digital skills for the professionals who are in charge of this task [1]. The need for timely responses poses additional challenges for Public Administration [5], such as a flexible organisational structure, strategic planning and efficient and effective resource management, without jeopardising the sustainability of public finances. Moreover, the public governance paradigm involves knowledge sharing, coordination and collaboration between various state, market and social actors.

It is therefore evident that the implementation of a new accounting framework applied to Public Administration has been gaining importance as regards the management of public finances, with the three subsystems of public accounting. The need to increase the level of transparency, credibility and reliability of budgetary, financial and management information is crucial for assessing and disseminating the results achieved by public policies [7]. At the same time, these changes in the accounting framework for public administrations will also have an impact on the future of the public audit process, which is a vital activity in democratic countries to guarantee the relationship between government bodies and citizens and businesses. Auditing in the public sector also faces the same problems as the private sector, challenging the public interest [8] [9] [10].

Yet, while there is great evidence of the use and impact of new technologies in the private sector, “in the public sector, a similar indication does not exist” [11]. Indeed, the above points highlight a lack of research exploring the significance and potential impacts of digital transformation on public sector accountability and finance management. To address this gap and promote new research in this field, this paper assesses the current landscape of public finance management in the context of digital transformation, answering the following research question: How does digital transformation affect Financial Management and Accounting in the Public Sector? To answer this question, the paper presents a bibliometric analysis of the literature in the field.

In section 2, this article presents a theoretical approach to digital transformation in Public Administration and finance. It then describes the methodology used to carry out the literature review (section 3). Next, in section 4, the results are analysed and discussed. Finally, section 5 summarises the main changes in the areas under study and identifies topics for future research, as well as the study’s limitations.

2. Theoretical Background

Governments worldwide are embracing digital tools, social media platforms, algorithms, and artificial intelligence to not only revolutionise public services but also to promote

deeper engagement with citizens [12] [13]. The digital transformation is enhancing public services through technologies like AI, blockchain, and IoT. This transformative wave is expected to fortify various aspects of governance, including decision-making, transparency, accountability, and citizen-government relationships [14] [15].

The European Commission's goals for the digital decade include making all key public services available online, providing citizens with access to medical records, and having 80% of citizens use digital identity solutions. To achieve these, initiatives like GovTech collaborations, the Innovative Public Service Observatory, and funding programs are being implemented to support innovation, promote interoperability, and foster public-private partnerships. The technology readily available to entities has a positive effect on the transformation of public services, which allows governments to implement solutions such as e-government, harnessing information and communication technologies (ICT) to create public value [2], [3].

[16] argue that digital transformation in government represents a two-way street, in which public bodies and citizens actively participate through the co-production of public services. Enhancing citizens' adoption of electronic services results in improved quality, efficiency, and effectiveness of public services. Also, digital technology fosters transparency because it facilitates citizen access to information on resource management and promotes and assists the implementation of robust governance practices. Thus, new doors are opening for citizens' socio-political participation through digital technologies [17].

However, digitalisation has had considerable consequences on the labour market, with greater income disparities and reduced access to social security systems, which can be negative if not managed properly [18]. Furthermore, this digital revolution presents huge challenges regarding cybersecurity vulnerabilities and data integrity [18].

Literature in the field emphasises the role of digital technology in modernising accounting and accountability, contributing to automating repetitive tasks, freeing up time for strategic activities with higher value creation [19]; enhancing data quality, namely accuracy, reliability, and consistency in accounting data [20]; and combatting corruption by identifying irregularities and fraud through advanced data analytics.

In the information-oriented world, the challenges professionals face also need to be addressed. The types of skills, competencies and mindset required of finance professionals to perform at the level demanded for organisations has been stated by several authors [21]. This paradigm shift is characterised by the automation of routine tasks, enabling accounting professionals to redirect their focus towards strategic analyses [2], [22]. Digital accounting systems are instrumental in facilitating real-time reporting and data-driven decision-making, thereby bolstering transparency and accountability [18]. Notably, these changes are poised to reverberate across public sector auditing, ensuring accountability and transparency in the interactions between government entities and citizens and businesses [8] [10]. Cumulatively, the management of public resources is increasingly associated with the responsibility of managers to obtain better results with fewer public resources [23].

Digital technology's integration into Public Administration has far-reaching implications, transcending mere efficiency gains to encompass broader dimensions of accountability and integrity in governance. It also demands substantial changes, rather than just technological changes. Indeed, within the public sector, knowledge management is a powerful facilitator in the current push for greater efficiency in all areas as stated by [24], “to be transforming,

changes in technologies must be accompanied by changes in other organizational elements (like people or processes)", leadership models and organisational culture.

In essence, the amalgamation of digital technology and Public Administration underscores the transformative potential of digitalisation in reshaping governance paradigms and fostering accountability and transparency in the public sector. One notable aspect highlighted by the literature is the dearth of research on the nexus between public service digitalisation and accountability, particularly within accounting scholarship. This emphasises the importance of bridging this gap to fully understand the implications of digital transformation on governance and accountability [25].

3. Methodology

Our review can be categorised as a Thematic synthesis [26] because it examines the state of the literature of a specific topical area and uses all themes from all papers to create theme clusters. Therefore, our research question can be formulated as follows: How does digital transformation affect Financial Management and Accounting in the Public Sector?

Data collection took place in January 2024. We did not apply any chronological filter. In the first phase, we tried a separate search for each of the keywords. In Web of Science Core Collection (WOS), we applied the following strategy: search strategy (TITLE-ABS-KEY ("accountability" OR "accounting" OR "transparency") AND TITLE-ABS-KEY ("public finance" AND "management") AND TITLE-ABS-KEY ("digital transformation" AND "digitalization" AND "public sector")). In SCOPUS, we followed the same criteria.

In Figure 1, we summarize the research layout and results that led to the final set of articles. The literature search yielded 890 articles. Three authors reviewed and screened the titles and abstracts for inclusion and exclusion criteria. After applying the exclusion criteria — namely: (1) duplicates; (2) articles not available (3) do not focus on the subject under study, and inclusion, namely (1) papers that do not address digital transformation in public accounting and finance management and (b) articles mapping and reviewing the literature — 172 articles remained.

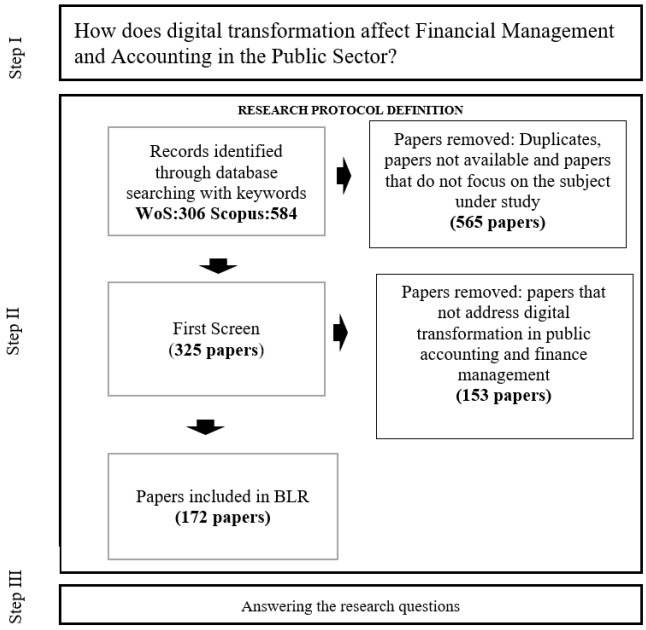


Fig. 1. Flowchart for literature selection

We used R Bibliometrix software [27] to perform bibliometric analysis and build data matrices for co-citation, clustering, scientific collaboration analysis and word analysis. Bibliometrics is increasingly applied across various disciplines to aid science mapping, addressing the growing volume and fragmentation of research driven by empirical contributions. For Network matrix creation, we used R Bibliometrix (http://www.bibliometrix.org). Based on the 172 manuscripts' database (title, abstract, keywords, authors, references), two clusters were created. The authors established a minimum of 30 citations in Web of Science or Scopus as inclusion criteria in the final clusters. Then, all titles, abstracts, literature reviews, and final considerations were read, and a document was created to contain the most relevant information extracted from those sections of each article. Finally, the contents were divided according to the main constructs to create a text that could explain the main theoretical approaches to each cluster, and the conclusions that had been drawn.

4. Results

4.1. Distribution of publications

Figure 2 shows an exponential increase in publications in the area under study from 2017 onwards. In 2022, there was an increase of 81 documents and in 2023 it reached 117 documents, which proves the novelty, relevance, and interest of the topic under investigation.

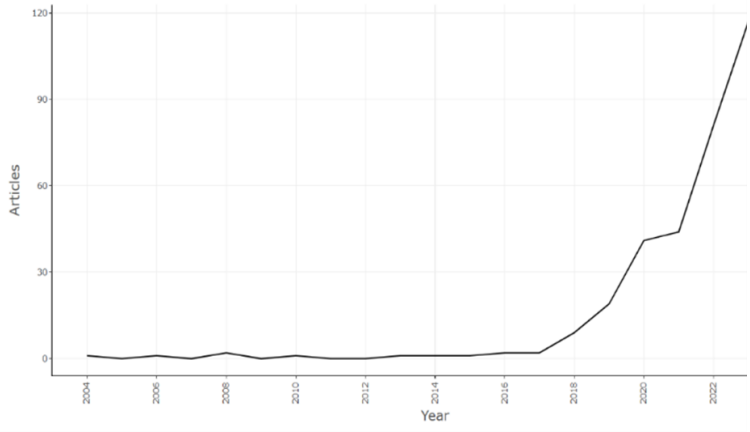


Fig. 2. Publication distribution per year

Regarding the number of publications by each author, if we use the lead author as a search criterion in this database sample, there are 10 authors with two or more articles published on the topic of digital transformation in public finance management. (Figure 3). Mergle and Stecollini stand out with 5 publications in the area, followed by Androniceanu with 4 publications.

As regards the journals with the highest visibility, Financial Accountability and Management and Government Information Quarterly is the journal with the highest impact measure H: 6, followed by Sustainability with 5 in the Impact measure H (see figure 4).

Figure 5 shows topics of interest over time. Advancements in technology have been pivotal in driving the digital transformation seen across the public sector. The COVID-19 pandemic has acted as a catalyst, accelerating this process. Digitalization began to have an impact from the year 2017, progressively intensifying. From 2022 onwards, the concepts that gained greater relevance were "digitalization/digitalisation," "digital transformation," and "public sector," followed by the term "public administration."

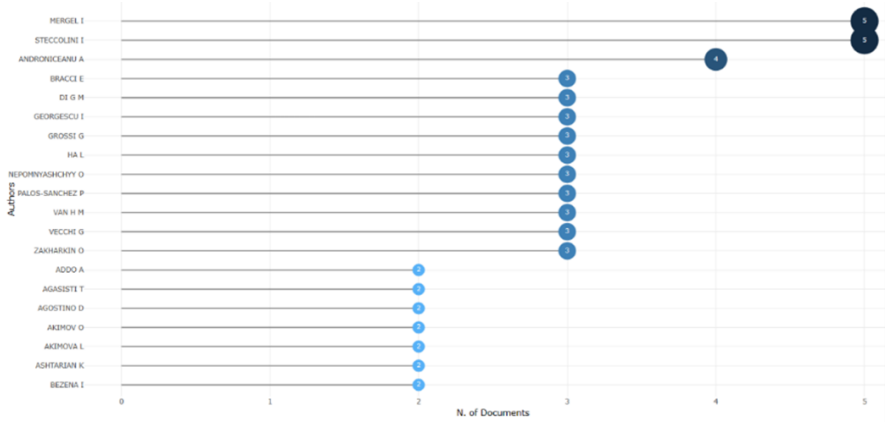


Fig. 3. Number of publications by author

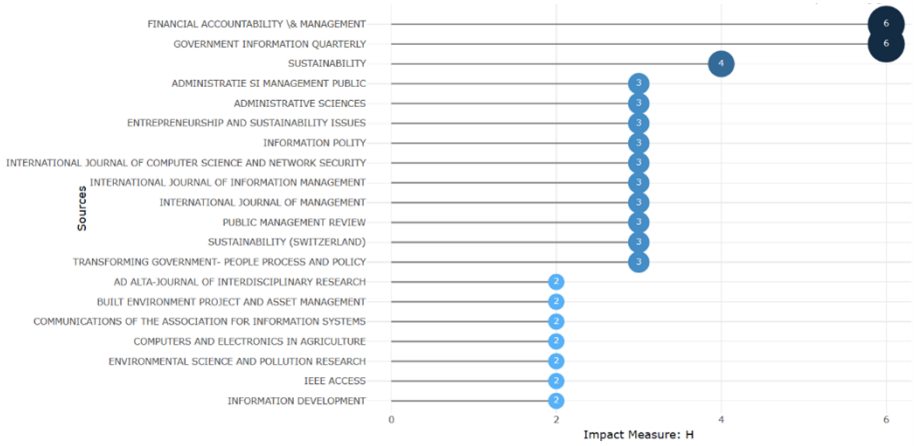


Fig. 4. Most relevant Sources

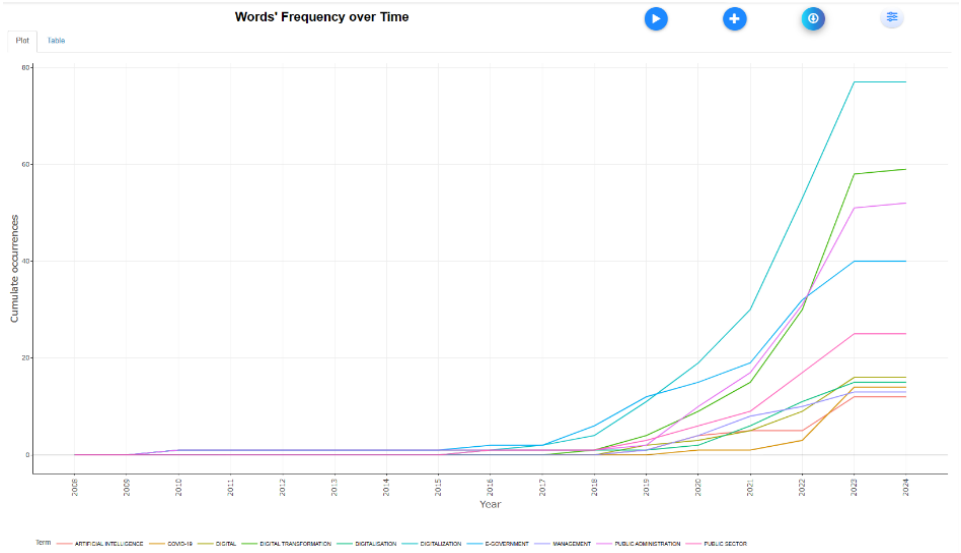


Fig. 5. Topics of interest over time

The Sankey diagram (Figure 6) illustrates relationships among authors, keywords, and countries. The size of each box is proportional to the frequency of occurrences of the respective theme. The flows connecting the boxes represent the thematic evolution, with thicker connecting lines indicating stronger associations between themes.[28].

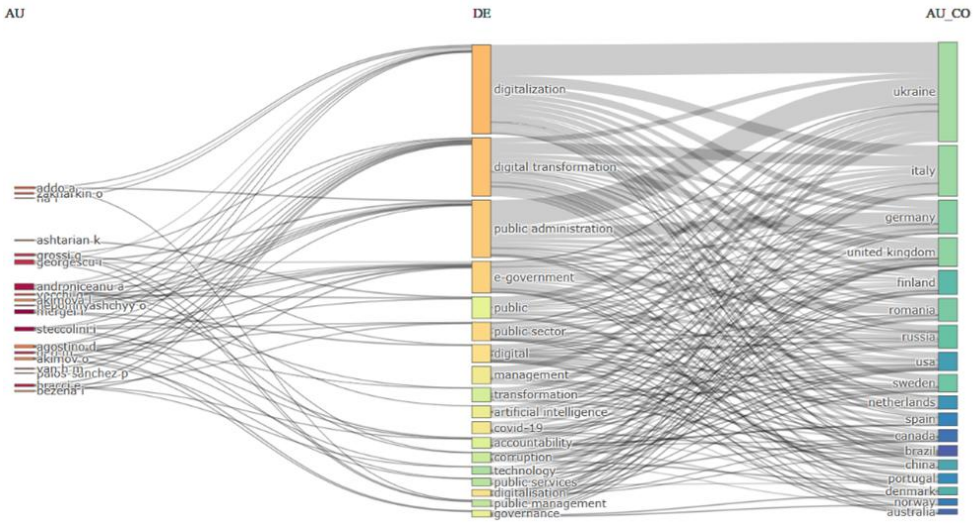


Fig. 6. Graph of three fields Authors (AU) - Keywords (DE) Countries (Au-CO)

Figure 6 shows that the most used keywords are "digitalisation", "digital transformation" and "Public Administration", with Androniceanu, Georgescu, Mergle, Steccolini and Agostino mentioning them the most. In terms of publications by country, the lead belongs to Ukraine, followed by Italy.

To understand how the contents of the selected articles converge in terms of the centrality of the study, correspondence factor analysis was used (see figure 7).

Figure 7 shows that the articles were correctly selected according to the main theme of digital transformation in the public sector, namely through digitization, big data, auditing, governance, and innovation.

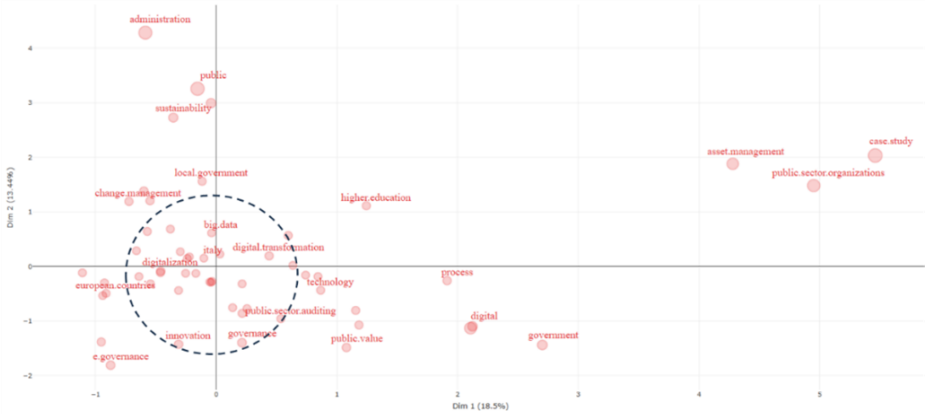


Fig. 7. Word Map

Also using factor analysis, figure 8 shows the dendrogram (grouping of similar items). The items with the greatest similarity are placed closer together and at the lowest level. There are 3 clusters and, within each cluster, the words that identify it.

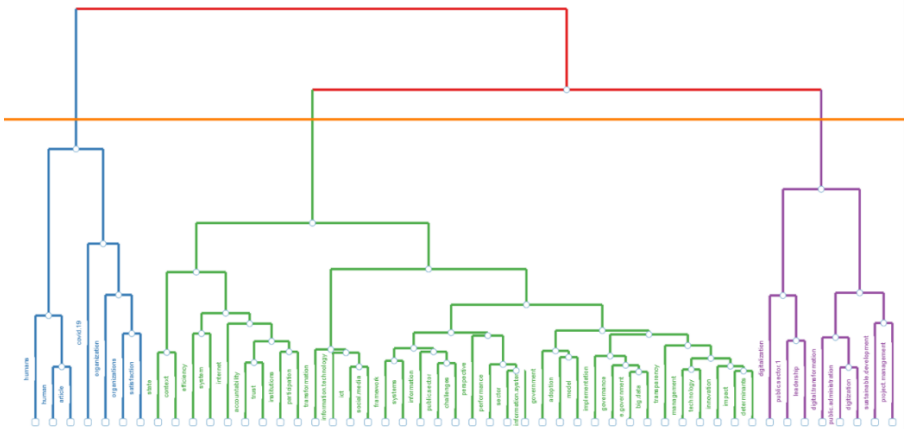


Fig. 8. Topic Dendogram

The Bibliometrix package [27] grouped the 172 studies into 3 clusters. A comprehensive analysis of each study was carried out, with the authors identifying the focus and main characteristic of each cluster, resulting in the following synthesis: Cluster 1 is defined as the cluster of technologies applied to the public sector and its tangible effects (identified in green in figure 10); while Cluster 2 is defined as the cluster of digital transformation of public administration and its intangible effects. Cluster 3 has only one article and was therefore excluded.

4.2. Structure and context of literature clusters

Cluster 1 encompasses one hundred and fifty-eight articles. Table 1 summarises the articles with at least 30 citations

Table 1. Cluster 1 articles

Authors	Title	Objectives	Methodology	Conclusions
[29]	Governance and innovation in public sector services: The case of the digital library.	Explore modes of governance and innovation in the public sector.	Longitudinal case study on modes of governance and innovation in the library sector.	Network Governance drives innovation by enhancing collaboration, knowledge sharing, and adaptability.
[30]	Digital transformation by SME entrepreneurs: A capability perspective	It assesses if SME entrepreneurs have led digital transformation despite scarce skills and resources.	Qualitative approach.	Successful digital transformation requires SMEs to upgrade skills in cognitive and social media management.
[31]	Digital service teams in government.	Understand the factors that lead to the initiation of digital service teams and identify the tasks of digital service teams in the public sector.	Qualitative interpretive approach comparative approach	EAGs improve digital service delivery through a hybrid IT governance model balancing centralisation and decentralisation.
[32]	Value positions viewed through the lens of automated decision-making: The case of social services.	State of the art of the public sector in e-governance.	Qualitative case study using interviews.	Greater accountability, lower costs and increased efficiency observed.
[33]	The study examines opportunities and challenges of blockchain in Japan's energy transition.	Responding to institutional challenges such as low renewable energy targets and grid interconnection.	Design research	It examined challenges and opportunities across technology, economy, society, environment, and institutions for Japan's blockchain-based microgrid.
[51]	Open innovation 4.0 as an enhancer of sustainable innovation ecosystems.	Identifying university links in sustainable innovation ecosystems. Proposing a package of policies for green governance.	Multiple case studies	Open innovation structures and university knowledge flows promote intelligent and responsible innovation cycles and funding
[34]	Digital transformation challenges: strategies emerging from a multi-stakeholder approach.	It outlines digital transformation strategies in the Tirol-Veneto region, highlighting challenges, actions, and the role of digital skills and culture.	Qualitative method with text mining and content analysis.	Digital transformation requires multifaceted strategic actions in three main pillars: Culture and Skills, Infrastructures and Technologies.
[35]	Digital transformation and knowledge management in the public sector.	It assessed how digital transformation affects knowledge management in	Quantitative method.	The success of digital government is strongly related to the quality of

		Portuguese public administration.		knowledge management in organisations.
[36]	Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities.	Incorporating Big Data and machine learning into energy efficiency in the public sector.	Data analysis, using deep neural networks, Rpart regression tree and random forest.	It proposes the MERIDA system that integrates predictive models for improving energy efficiency.
[37]	Functions of public management of the regional development in the conditions of digital transformation of economy.	Develop a methodological approach based on the system of differential equations.	Questionnaire survey	It highlights how decentralisation in Europe strengthens regional and local governance in Ukraine.
[38]	Digital government transformation: A structural equation modelling analysis of driving and impeding factors.	It quantitatively assesses factors enabling or hindering government digital transformation, emphasizing the need for change and collaboration.	Questionnaire survey	Urgency, collaboration, and management engagement drive TDG success.
[39]	It investigates RPA's opportunities, challenges, and implementation in supply management.	It analyses RPA's effects and implementation challenges in public and private procurement.	Multiple case study carried out to provide initial insights and generalisable propositions.	It examines RPA in Procurement and Supply Management, showing sector-specific advantages and challenges.
[40]	Knowledge management and digital transformation for Industry 4.0: a structured literature review,	Explore the interactions between knowledge management, digital transformation and Industry 4.0.	Literature review integrating qualitative and bibliometric analysis.	Identifies the link between CG, DT and the public sector. It stresses the crucial role of DT in the development of KM.
[41]	Co-production in digital transformation of public administration and public value creation: The case of Denmark.	It examines co-production in Denmark's digital strategy, highlighting citizen, economic, administrative, and social value.	Qualitative approach with interviews.	It explores co-production in Denmark's digital transformation, identifying four types of public value: citizen, economic, administrative, and social.
[42]	Government tax policy in the digital economy.	Identify the current problems, the changes required in tax policy in the context of the digital economy,	Discussion	International collaboration is essential to combat tax evasion, enhance transparency, and innovate tax systems.
[43]	It studies digital government units and their role in modernizing public management..	Kickstarting a public management research agenda focused on Digital Government Units	Qualitative approach. Multiple case studies, using interviews.	It outlines research questions on DGUs and the role of open standards and platforms in digital government transformation.
[44]	Digitalization, accounting and accountability: A literature review and reflections on future research in public services.	It examines research trends on digitalisation and accountability in public services.	Systematic literature review	It highlights accountability and inclusivity in digitalisation.
[45]	Drivers and outcomes of digital transformation: The case of public sector services.	Identify the expectations of public managers regarding the ongoing digital transformation projects	Case study approach with expert interviews	Long-term benefits need deep organisational change, hindered by training gaps and bureaucracy.

Cluster 1 emphasises the challenge of implementing new technologies, namely digitalisation, blockchain, big data, artificial intelligence, machine learning and cloud computing, to improve processes and interact with stakeholders in the digital economy. It

also highlights the risks associated with the digital economy, particularly in terms of security. Problems of interoperability and a lack of qualified training for human resources working in Public Administration have also been identified.

The cluster 2 encompasses six articles, as described below.

Table 2. Cluster 2 articles

Authors	Title	Objectives	Methodology	Conclusions
[53]	Formation of professional competences and soft skills of public administration employees for sustainable professional development.	Identify the skills required to perform the functions of public administration and identify the skills required.	Quantitative research with questionnaires	Digitalisation boosts public administration skills and service efficiency.
[46]	Socioeconomic and resource efficiency impacts of digital public services	It evaluates digitalisation's socio-economic, environmental, and welfare impacts in Europe.	It uses econometric techniques to estimate the impacts of digital public services.	Digital public services enhance the economy, society, and resource efficiency.
[47]	Public administration of planning for the sustainable development of the region in the context of total digitalization.	It develops a planning framework and evaluates public administration's role in sustainable regional development.	The functional graphic construction method portrays the public administration's mechanism for sustainable development.	The study highlights state management functions and their importance for regional development.
[48]	C-suite Leadership of Digital Government.	It outlines digital government leadership and a framework for research.	Literature review	It highlights the need for inclusive approaches in leading digitalisation. It introduces a conceptual framework with leadership roles for digital transformation.
[49]	Technology and digital transformation for the structural reform of the sports industry: building the roadmap	Design and create a tool to understand the digital structure. Develop a consultation tool for the digitisation needs of sports organisations.	Questionnaire survey	The creation of a consultation tool is crucial for the digital transformation. The tool's design makes it easy to collect data to understand the sports industry
[50]	An operational framework for the implementation of digital systems in public administration processes in the design phase	Define an operational framework that supports PAs to implement and check the digitalisation of their workflows for the design phase	Literature Review, interviews, conceptual discussion.	Proposal of an operational framework for implementation of digitalisation o

4.3. Analysis and discussion of results

The global rise in governmental adoption of digital tools and technologies finds substantial support in numerous studies, reflecting a concerted drive to modernise governance structures and bolster citizen engagement [12] [13]. Notably, a significant uptick in publications on this subject emerges from 2017 onward, particularly evident in 2022 (81 documents) and 2023 (117 documents), underscoring the burgeoning interest, relevance, and innovation within this domain. Leading the pack of impactful authors are Mergle and Stecollini, each boasting five publications, closely trailed by Androniceanu

with four. Cumulatively, the ten most cited articles have amassed a total of 2,611 citations, indicative of their considerable impact within the research sphere. Journals such as Financial Accountability and Management and Government Information Quarterly stand out with commendable H-index scores, attesting to their prominence in this field. Ukraine takes the lead in publications by country, closely followed by Italy. Key terms such as "Digitalisation," "Digital transformation," and "Public Administration" emerge as recurrent keywords, with Androniceanu, Georguscu, Mergle, Steccolini, and Agostino notably leveraging them.

The domain of qualitative approaches is well-defined, encompassing single case studies [32] [45], multiple case studies [34] [31] [43] [41] [39] and longitudinal case studies. Concerning quantitative studies, a variety of methods are employed, including econometric estimations [51] [52]), the development of machine learning-based models, correlation analysis [53], differential equations [37], and structural equation modelling [38]. These studies are predominantly exploratory, although a few incorporate international comparative data [31] [46]. Additionally, there are studies utilising systematic literature reviews [40] [35] [44] [48], as well as combinations of different methodologies [35] [42]. Furthermore, there are theoretical discussions [42], analytic models and frameworks [47] [49].

The clusters identified in the Bibliometric analysis demonstrate the multifaceted nature of digital transformation within Public Administration. Cluster 1 focuses on technologies applied to the public sector and their tangible effects; while Cluster 2 delves into the digital transformation of public administration and its intangible effects.

Cluster 1 focuses on technologies applied in Public Administration, highlighting the impact of digital transformation on performance; e-governance, decision-making processes, citizen engagement; and the role of knowledge management as an important facilitator for efficiency and quality in public services. Public Administration needs rigour and transparency in the management of public finances, creating "new forms of dialogic accountability with stakeholders" [44]. In addition to digitisation, other emerging technologies, such as blockchain [33], robotic automation process [39] and automated decision-making [32], artificial intelligence [54], cloud computing [55], big data and machine learning [36], are mentioned as crucial for improving processes and interactions with stakeholders in the digital economy.

Digital technologies have the potential to encourage a paradigm shift to greater transparency, accountability, and citizen-centricity [32], sustainability and innovation ecosystems [56] [57] [51] cost savings, and increased operational efficiency and quality [58], create public value [41] [45] and regional development [37]. Special mention is made of the "potential of digitalization in reshaping governance structures and accountability challenges that accompany digital government transformation" [43]. Tax transparency is another advantage of IT pointed out in the literature. Emphasis is placed on the role of technologies in the design and implementation of a tax policy that stimulates innovation, ensures efficiency, improves the quality of tax services, and prevents tax evasion [42]. These findings align with the broader literature, which underscores the global trend of leveraging digitalisation to revolutionise public services and enhance citizen engagement [13] [5] [9] as well as automating repetitive tasks, enhancing data quality, and promoting corporate governance [19] [20].

Despite the prevalence of an optimistic perspective of digital transformation outcomes, the obstacles, the risks, and failures are also pointed out in literature. Digital Government Units are described as an instrument to overcome the failings of public sector IT [43]. In addition to the existing IT governance organisational units, Digital Service teams are expected to fulfil the digital transformation of government, "bridging the gap between traditional forms of IT governance and modern, agile or networked IT governance forms" [31]. Additionally, the training and education of workers, the creation of a culture for adopting digital tools and changing the bureaucratic structures of the organisation emerge as crucial conditions for the successful utilisation of new technologies [45]. Digital transformation in the public sector also raises concerns regarding interoperability, cybersecurity and data integrity, as pointed out by [18]. Further, results draw attention to how effective digital transformation depends on knowledge management [35] and how it is directly demanding new practices and professional competencies as well as digital leadership and capacity building of human resources to drive digital transformation in Public Administration [21] [44] [59].

In the context of accounting within the public sector, the analysis underscores the need for transparency, credibility, and reliability in financial reporting, which are essential for maintaining accountability and trust in government entities [44]. Moreover, technologies have significantly transformed accounting and public finance by automating processes to reduce errors and optimise time; increasing transparency and accountability through online portals; preventing fraud using analytical tools; improving the planning and monitoring of public budgets; and integrating systems for a more comprehensive and unified view of accounting and finance [44]. These findings are in line with other authors in the field [5] [8] [9] [10].

Cluster 2 focuses on the socio-economic impacts and resource efficiency of digital public services, emphasising the dimensions of human resource development, digital leadership, and performance evaluation. Digital public services are highlighted for their positive impact on the economy, society, and resource efficiency, promoting sustainable development and improving social well-being [46] [47]. Thus, there is a great need to push Public Administration towards a change of mentality, framing in technical and organisational aspects the implementation of digital systems [50]. Administrative modernisation and investment in human resource development are considered essential to drive digital transformation in Public Administration. Along with digital (technical) skills, the demand for soft skills is growing in the context of digitalisation of Public Administration [53]. It reinforces the importance of inclusive leadership in digitisation, promoting an effective work environment and enhancing resource management in Public Administration. At this level, Kristensen and Andersen [48] call for attention to leadership of digitalisation and "their ability to cope with the high speed of digital change and deep shift in organizational culture".

Studies in the two clusters converge with previous studies: digital transformation is highly impacted by several external factors, including the adoption of cutting-edge technology by different stakeholders in public organisations [45], "the sense of urgency, the need for change, the definition of a shared vision, and the creation of a collaborative environment" [38] or, as stated by, [31], there are different internal and external factors that influence digital transformation. Internal factors include the management model and bureaucracy, while external factors include legislation, the administrative, political, social,

economic, technological and environmental components [29]. Literature on impeding and driving factors of digital transformation in public sector suggests that more effort is required to include public managers in the current debate on DGT. So, to be successful, digital transformation requires a multifaceted set of strategic actions falling into three main pillars, namely "culture and skills", "infrastructures and technologies", and ecosystems" [34]. Therefore, the creation of monitoring tools is seen as crucial to guide public policies, assess the impact of changes, support the definition of corrective measures, and maintain the sustainability of public finances, ideas also advocated by [8] and [23].

The results also reinforce and support the idea that digital transformation is revolutionising Public Administration, seeking to improve the performance of public services, increase transparency, and strengthen accountability [44] [3] [2]. The results highlight the importance of streamlining bureaucracy, administrative modernisation, and transparency as key pillars of Public Administration in the current context [23].

Overall, the results indicate that over time, technologies have revolutionized public finance management, offering restructuring processes to minimize errors and improve efficiency; provide greater transparency and accountability through online platforms; facilitate fraud prevention through analytical solutions; offer better planning and oversight of public budgets; and perform systems integration for a comprehensive and unified perspective on accounting and finance.

In conclusion, the synthesis of empirical findings and theoretical insights stresses the multifaceted nature of digital transformation in Public Administration. By leveraging digital technologies, governments can better manage modern governance complexities, improve service delivery, and strengthen accountability mechanisms. However, fully realising the potential of digitalisation requires addressing challenges and investing in human capital and leadership. Integrating digital innovation with sound governance principles promises a new era of responsive, transparent, and accountable Public Administration.

5. Conclusion

The exploration of Financial Management and Accounting in the Public Sector underscores its profound impact on modern governance. Through a synthesis of empirical findings and theoretical insights, this article illuminates the multifaceted nature of this transformation. It becomes evident that digitalisation is not merely a buzzword; it represents an essential paradigm shift reshaping Public Administration.

The clusters identified through Bibliometric analysis reveal crucial dimensions of this transformation. Cluster 1 underscores the transformative potential of technologies in enhancing organisational performance, citizen engagement, and knowledge management. From blockchain to big data analytics, technologies offer unprecedented opportunities while also posing challenges such as cybersecurity risks and interoperability issues. Cluster 2 delves into the socio-economic ramifications of digital public services, highlighting the imperative of human resource development, digital leadership, and performance evaluation. This cluster emphasises the interconnectedness of administrative modernisation, inclusive leadership, and sustainable development, advocating for a holistic approach grounded in governance principles.

This comprehensive analysis underscores the transformative potential of digitalisation in public administration, accentuating the evolving role of public sector accounting and financial management. From evolving accounting standards to the imperative of transparency and accountability, there is a clear call for new practices and competencies to navigate digital complexities. The findings demonstrate that, over time, technologies have significantly transformed accounting and public finance by automating processes to reduce errors and optimise time; increasing transparency and accountability through online portals; preventing fraud with analytical tools; improving the planning and monitoring of public budgets; and integrating systems for a more comprehensive and unified view of accounting and finance. However, realising this vision necessitates concerted efforts to address challenges, invest in human capital, and cultivate digital leadership capabilities. Ultimately, the fusion of digital innovation with robust governance principles holds the promise of ushering in an era of effective, efficient, and citizen-centric Public Administration.

The main limitation of this study was not the methodology since this was chosen carefully and with scientific method; instead, it was the fact that the literature analysed only extended to the WOS and SCO databases, and the study performed an in-depth analysis only of the most cited articles, potentially overlooking valuable insights from other sources and less-cited studies.

Overall, this review and mapping of the literature provide a detailed overview of the existing knowledge on digital transformation in accounting and public finance; highlight the most important and relevant works, identify key research areas and keywords, and offer insights into topics of high academic interest. This study is a valuable resource for researchers and professionals because state-of-the-art knowledge helps to implement digital transformation in Public Administration, reduce process inefficiencies and increase the quality, credibility and timeliness of information to support the decision-making process.

Acknowledgment. This work is financed by Portuguese national funds through FCT - Fundação para a Ciência e Tecnologia, under the project UID/05422/2023: Centre for Organisational and Social Studies of Polytechnic of Porto.

References

1. D. Agostino, E. Bracci, and I. Steccolini, “Accounting and accountability for the digital transformation of public services,” FINANCIAL ACCOUNTABILITY & MANAGEMENT, vol. 38, no. 2, SI. WILEY, 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, pp. 145–151, May 2022. doi: 10.1111/faam.12314.
2. J. I. Criado and J. R. Gil-Garcia, “Creating public value through smart technologies and strategies: From digital services to artificial intelligence and beyond,” International Journal of Public Sector Management, vol. 32, no. 5, pp. 438–450, Jan. 2019, doi: 10.1108/IJPSM-07-2019-0178.
3. A. Cordella and C. M. Bonina, “A public value perspective for ICT enabled public sector reforms: A theoretical reflection,” Government Information Quarterly, vol. 29, no. 4, pp. 512–520, Oct. 2012, doi: 10.1016/j.giq.2012.03.004.
4. M. Fatemi and M. R. Behmanesh, “New public management approach and accountability,” International Journal of Management, Economics and Social Sciences (IJMESS), vol. 1, no. 2, pp. 42–49, 2012.

5. A. F. Lino, R. R. de Azevedo, and G. S. Belote, “The influence of public sector audit digitalisation on local government budget planning: evidence from Brazil,” Journal of Public Budgeting, Accounting and Financial Management, vol. 35, no. 2. Emerald Publishing, pp. 198–218, 2023. doi: 10.1108/JPBAFM-05-2022-0090.
6. S. Osborne, “The New Public Governance?,” Taylor & Francis. Accessed: May 24, 2024. Online.. Available: https://www.tandfonline.com/doi/abs/10.1080/14719030600853022
7. N. Hyndman and M. Liguori, “Public Sector Reforms: Changing Contours on an NPM Landscape,” Financial Accountability & Management, vol. 32, no. 1, pp. 5–32, 2016, doi: 10.1111/faam.12078.
8. L. Ferry, V. S. Radcliffe, and I. Steccolini, “The future of public audit,” FINANCIAL ACCOUNTABILITY & MANAGEMENT, vol. 38, no. 3, SI. WILEY, 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, pp. 325–336, Aug. 2022. doi: 10.1111/faam.12339.
9. C. J. Cordery and D. C. Hay, “Public Sector Audit,” Routledge & CRC Press. Accessed: May 24, 2024. Online.. Available: https://www.routledge.com/Public-Sector-Audit/Cordery-Hay/p/book/9780367650629
10. C. Free, V. S. Radcliffe, C. Spence, and M. J. Stein, “Auditing and the Development of the Modern State,” Contemporary Accounting Research, vol. 37, no. 1, pp. 485–513, 2020, doi: 10.1111/1911-3846.12497.
11. A. Di Vaio, R. Hassan, and C. Alavoine, “Data intelligence and analytics: A bibliometric analysis of human–Artificial intelligence in public sector decision-making effectiveness,” Technological Forecasting and Social Change, vol. 174, no. C, 2022, Accessed: May 24, 2024. Online.. Available: https://ideas.repec.org//a/eee/tefoso/v174y2022ics004016252100634x.html
12. Y. Charalabidis and Z. Lachana, “Towards a science base for digital governance,” ACM International Conference Proceeding Series. Association for Computing Machinery, pp. 383–389, 2020. doi: 10.1145/3396956.3400062.
13. I. Munteanu and K. Newcomer, “Leading and Learning through Dynamic Performance Management in Government,” Public Administration Review, vol. 80, no. 2, pp. 316–325, 2020, doi: 10.1111/puar.13126.
14. Y. Ramírez and Á. Tejada, “Digital transparency and public accountability in Spanish universities in online media,” Journal of Intellectual Capital, vol. 20, no. 5, pp. 701–732, Jan. 2019, doi: 10.1108/JIC-02-2019-0039.
15. S. Royo, A. Yetano, and J. García-Lacalle, “Accountability Styles in State-Owned enterprises: The good, the bad, the ugly ... and the pretty,” Revista De Contabilidad, vol. 22, no. 2, 2019, doi: 10.6018/rcsar.382231.
16. T. Polzer and G. Goncharenko, “The UK COVID-19 app: The failed co-production of a digital public service,” FINANCIAL ACCOUNTABILITY & MANAGEMENT, vol. 38, no. 2, SI. WILEY, 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, pp. 281–298, May 2022. doi: 10.1111/faam.12307.
17. S. Sharma, A. K. Kar, M. P. Gupta, Y. K. Dwivedi, and M. Janssen, “Digital citizen empowerment: A systematic literature review of theories and development models,” Information Technology for Development, vol. 28, no. 4, pp. 660–687, Oct. 2022, doi: 10.1080/02681102.2022.2046533.
18. L. Mura, T. Zsigmond, and R. Machová, “The effects of emotional intelligence and ethics of SME employees on knowledge sharing in Central-European countries,” Oeconomia Copernicana, vol. 12, no. 4, pp. 907–934, 2021.
19. S. Marwah and R. Thapar, “A Study of Factors Affecting Consumers’ Behavioural Intention Towards Online Shopping: An Exploratory Study,” presented at the International Conference on Emerging Trends in Business and Management (ICETBM 2023), Atlantis Press, May 2023, pp. 155–164. doi: 10.2991/978-94-6463-162-3_14.
20. K. C. Laudon and J. P. Laudon, Laudon, Management Information Systems: Managing the Digital Firm, 16th Global Edition. Pearson, 2020.

21. J. Schmitz and G. Leoni, “Accounting and Auditing at the Time of Blockchain Technology: A Research Agenda,” Australian Accounting Review, vol. 29, no. 2, pp. 331–342, 2019, doi: 10.1111/auar.12286.
22. M. J. A. Gonçalves, A. C. F. da Silva, and C. G. Ferreira, “The Future of Accounting: How Will Digital Transformation Impact the Sector?,” Informatics, vol. 9, no. 1, Art. no. 1, Mar. 2022, doi: 10.3390/informatics9010019.
23. I. Steccolini, “Accounting and the post-new public management: Re-considering publicness in accounting research,” Accounting, Auditing & Accountability Journal, vol. 32, no. 1, pp. 255–279, Jan. 2018, doi: 10.1108/AAAJ-03-2018-3423.
24. D. Špaček, M. Navrátil, and D. Špalková, “New development: Covid 19 and changes in public administration—what do we know to date?,” Public Money and Management, vol. 43, no. 8, pp. 862–866, 2023, doi: 10.1080/09540962.2023.2199545.
25. D. Petrakaki, “Re-locating accountability through technology From bureaucratic to electronic ways of governing public sector work,” INTERNATIONAL JOURNAL OF PUBLIC SECTOR MANAGEMENT, vol. 31, no. 1. EMERALD GROUP PUBLISHING LTD, HOWARD HOUSE, WAGON LANE, BINGLEY BD16 1WA, W YORKSHIRE, ENGLAND, pp. 31–45, 2018. doi: 10.1108/IJPSM-02-2017-0043.
26. Y. Xiao and M. Watson, “Guidance on Conducting a Systematic Literature Review.” Accessed: May 24, 2024. Online.. Available: https://journals.sagepub.com/doi/full/10.1177/0739456X17723971
27. M. Aria and C. Cuccurullo, “bibliometrix: An R-tool for comprehensive science mapping analysis,” Journal of Informetrics, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.
28. K. Meng et al., “Wearable Pressure Sensors for Pulse Wave Monitoring,” Advanced Materials, vol. 34, no. 21, p. 2109357, 2022, doi: 10.1002/adma.202109357.
29. A. Scupola and A. Zanfei, “Governance and innovation in public sector services: The case of the digital library,” GOVERNMENT INFORMATION QUARTERLY, vol. 33, no. 2. ELSEVIER INC, 525 B STREET, STE 1900, SAN DIEGO, CA 92101-4495 USA, pp. 237–249, Apr. 2016. doi: 10.1016/j.giq.2016.04.005.
30. L. Li, F. Su, W. Zhang, and J.-Y. Mao, “Digital transformation by SME entrepreneurs: A capability perspective,” INFORMATION SYSTEMS JOURNAL, vol. 28, no. 6, SI. WILEY, 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, pp. 1129–1157, Nov. 2018. doi: 10.1111/isj.12153.
31. I. Mergel, “Digital service teams in government,” GOVERNMENT INFORMATION QUARTERLY, vol. 36, no. 4. ELSEVIER INC, 525 B STREET, STE 1900, SAN DIEGO, CA 92101-4495 USA, Oct. 2019. doi: 10.1016/j.giq.2019.07.001.
32. A. Ranerup and H. Z. Henriksen, “Value positions viewed through the lens of automated decision-making: The case of social services,” GOVERNMENT INFORMATION QUARTERLY, vol. 36, no. 4. ELSEVIER INC, 525 B STREET, STE 1900, SAN DIEGO, CA 92101-4495 USA, Oct. 2019. doi: 10.1016/j.giq.2019.05.004.
33. A. Ahl et al., “Exploring blockchain for the energy transition: Opportunities and challenges based on a case study in Japan,” RENEWABLE & SUSTAINABLE ENERGY REVIEWS, vol. 117. PERGAMON-ELSEVIER SCIENCE LTD, THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, ENGLAND, Jan. 2020. doi: 10.1016/j.rser.2019.109488.
34. F. Brunetti, D. T. Matt, A. Bonfanti, A. De Longhi, G. Pedrini, and G. Orzes, “Digital transformation challenges: strategies emerging from a multi-stakeholder approach,” TQM JOURNAL, vol. 32, no. 4, SI. EMERALD GROUP PUBLISHING LTD, HOWARD HOUSE, WAGON LANE, BINGLEY BD16 1WA, W YORKSHIRE, ENGLAND, pp. 697–724, Jul. 21, 2020. doi: 10.1108/TQM-12-2019-0309.
35. A. Alvarenga, F. Matos, R. Godina, and J. C. O. Matias, “Digital Transformation and Knowledge Management in the Public Sector,” SUSTAINABILITY, vol. 12, no. 14. MDPI,

ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND, Jul. 2020. doi: 10.3390/su12145824.
36. M. Zekic-Susac, S. Mitrovic, and A. Has, "Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities," INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT, vol. 58. ELSEVIER SCI LTD, THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, OXON, ENGLAND, Jun. 2021. doi: 10.1016/j.ijinfomgt.2020.102074.
37. O. Popelo, O. Garafonova, S. Tulchynska, M. Derhaliuk, and D. Berezovskyi, "Functions of public management of the regional development in the conditions of digital transformation of economy," AMAZONIA INVESTIGA, vol. 10, no. 43. UNIV AMAZONIA, SEDE PRINCIPAL CALLE 17 DIAGONAL 17 CON CARRERA 3F-BARRIO PORVENIR, FLORENCE, 00000, COLOMBIA, pp. 49–58, Jul. 2021. doi: 10.34069/AI/2021.43.07.5.
38. L. Tangi, M. Janssen, M. Benedetti, and G. Noci, "Digital government transformation: A structural equation modelling analysis of driving and impeding factors," INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT, vol. 60. ELSEVIER SCI LTD, THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, OXON, ENGLAND, Oct. 2021. doi: 10.1016/j.ijinfomgt.2021.102356.
39. C. Flechsig, F. Anslinger, and R. Lasch, "Robotic Process Automation in purchasing and supply management: A multiple case study on potentials, barriers, and implementation," JOURNAL OF PURCHASING AND SUPPLY MANAGEMENT, vol. 28, no. 1. ELSEVIER SCI LTD, THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, OXON, ENGLAND, Jan. 2022. doi: 10.1016/j.pursup.2021.100718.
40. A. de B. Machado, S. Secinaro, D. Calandra, and F. Lanzalonga, "Knowledge management and digital transformation for Industry 4.0: a structured literature review," KNOWLEDGE MANAGEMENT RESEARCH & PRACTICE, vol. 20, no. 2. TAYLOR & FRANCIS LTD, 2-4 PARK SQUARE, MILTON PARK, ABINGDON OR14 4RN, OXON, ENGLAND, pp. 320–338, Mar. 04, 2022. doi: 10.1080/14778238.2021.2015261.
41. A. Scupola and I. Mergel, "Co-production in digital transformation of public administration and public value creation: The case of Denmark," GOVERNMENT INFORMATION QUARTERLY, vol. 39, no. 1. ELSEVIER INC, 525 B STREET, STE 1900, SAN DIEGO, CA 92101-4495 USA, Jan. 2022. doi: 10.1016/j.giq.2021.101650.
42. O. Pylypenko, H. Matviienko, A. Putintsev, I. Vlasenko, and N. Onyshchuk, "Government Tax Policy in the Digital Economy," CUESTIONES POLITICAS, vol. 40, no. 72. UNIV ZULIA, FAC CIENCIAS JURIDICAS & POLITICAS, AV 4 BELLAVISTA CON CALLE 74, EDIF FUNDALUZ, PISOS 10 & 4, MARACAIBO, 4002, VENEZUELA, pp. 279–296, Jun. 2022. doi: 10.46398/cuestpol.4072.15.
43. A. Clarke, "Digital Government Units: What Are They, and What Do They Mean for Digital Era Public Management Renewal?," INTERNATIONAL PUBLIC MANAGEMENT JOURNAL, vol. 23, no. 3. ROUTLEDGE JOURNALS, TAYLOR & FRANCIS LTD, 2-4 PARK SQUARE, MILTON PARK, ABINGDON OX14 4RN, OXON, ENGLAND, pp. 358–379, May 03, 2020. doi: 10.1080/10967494.2019.1686447.
44. D. Agostino, I. Saliterer, and I. Steccolini, "Digitalization, accounting and accountability: A literature review and reflections on future research in public services," FINANCIAL ACCOUNTABILITY & MANAGEMENT, vol. 38, no. 2, SI. WILEY, 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, pp. 152–176, May 2022. doi: 10.1111/faam.12301.
45. F. Kitsios, M. Kamariotou, and A. Mavromatis, "Drivers and Outcomes of Digital Transformation: The Case of Public Sector Services," INFORMATION, vol. 14, no. 1. MDPI, ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND, Jan. 2023. doi: 10.3390/info14010043.
46. L. T. Ha, "Socioeconomic and resource efficiency impacts of digital public services," ENVIRONMENTAL SCIENCE AND POLLUTION RESEARCH, vol. 29, no. 55.

SPRINGER HEIDELBERG, TIERGARTENSTRASSE 17, D-69121 HEIDELBERG, GERMANY, pp. 83839–83859, Nov. 2022. doi: 10.1007/s11356-022-21408-2.
47. O. Voronov, L. Kurnosenko, I. Bezena, N. Petryshyn, S. Korniievskyi, and B. Ilychok, "Public Administration of Planning for the Sustainable Development of the Region in the Context of Total Digitalization," International Journal of Sustainable Development and Planning, vol. 18, no. 1. International Information and Engineering Technology Association, pp. 61–67, 2023. doi: 10.18280/ijsdp.180106.
48. K. Kristensen and K. N. Andersen, "C-suite Leadership of Digital Government," Digital Government: Research and Practice, vol. 4, no. 1. Association for Computing Machinery, 2023. doi: 10.1145/3580000.
49. A. M. Magaz-González, M. García-Tascón, C. Sahelices-Pinto, A. M. Gallardo, and J. C. Guevara Pérez, "Technology and digital transformation for the structural reform of the sports industry: Building the roadmap," Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology. SAGE Publications Ltd, 2023. doi: 10.1177/17543371231197323.
50. M. Marocco, E. Cacciaguerra, and I. Garofolo, "An operational framework for implementing digital systems in public administrations' processes in the design phase," Architectural Engineering and Design Management. Taylor and Francis Ltd., 2023. doi: 10.1080/17452007.2023.2187752.
51. J. Costa and J. C. O. Matias, "Open Innovation 4.0 as an Enhancer of Sustainable Innovation Ecosystems," SUSTAINABILITY, vol. 12, no. 19. MDPI, ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND, Oct. 2020. doi: 10.3390/su12198112.
52. L. T. Ha, "Socioeconomic and resource efficiency impacts of digital public services," ENVIRONMENTAL SCIENCE AND POLLUTION RESEARCH, vol. 29, no. 55, pp. 83839–83859, Nov. 2022, doi: 10.1007/s11356-022-21408-2.
53. P. Krpálek, K. Berková, A. Kubišová, K. K. Krelová, D. Frendlovská, and D. Spiesová, "Formation of professional competences and soft skills of public administration employees for sustainable professional development," Sustainability (Switzerland), vol. 13, no. 10. MDPI AG, 2021. doi: 10.3390/su13105533.
54. M. A. Wimmer, A. C. Neuroni, and J. T. Frece, "Approaches to Good Data Governance in Support of Public Sector Transformation Through Once-Only," ELECTRONIC GOVERNMENT (EGOV 2020), vol. 12219. in Lecture Notes in Computer Science, vol. 12219. SPRINGER INTERNATIONAL PUBLISHING AG, GEWERBESTRASSE 11, CHAM, CH-6330, SWITZERLAND, pp. 210–222, 2020. doi: 10.1007/978-3-030-57599-1_16.
55. I. Nanos, E. Papaioannou, E. Androutsou, and V. Manthou, "The role of cloud computing and citizens relationship management in digital government transformation," International Journal of Internet Marketing and Advertising, vol. 13, no. 2. Inderscience Publishers, pp. 120–136, 2019. doi: 10.1504/IJIMA.2019.099495.
56. A. Scupola and A. Zanfei, "Governance and innovation in public sector services: The case of the digital library," Government Information Quarterly, vol. 33, no. 2, pp. 237–249, Apr. 2016, doi: 10.1016/j.giq.2016.04.005.
57. A. Ahl et al., "Exploring blockchain for the energy transition: Opportunities and challenges based on a case study in Japan," Renewable and Sustainable Energy Reviews, vol. 117, p. 109488, Jan. 2020, doi: 10.1016/j.rser.2019.109488.
58. C. Flechsig, F. Anslinger, and R. Lasch, "Robotic Process Automation in purchasing and supply management: A multiple case study on potentials, barriers, and implementation," Journal of Purchasing and Supply Management, vol. 28, no. 1, p. 100718, Jan. 2022, doi: 10.1016/j.pursup.2021.100718.
59. D. Špaček, M. Navrátil, and D. Špalková, "New development: Covid 19 and changes in public administration—what do we know to date?," Public Money and Management, vol. 43, no. 8. Routledge, pp. 862–866, 2023. doi: 10.1080/09540962.2023.2199545.

Ambrósio José Silva Teixeira is an Invited Assistant Professor in Management Planning and Control. I am the coordinator of the Finance Reform and Monitoring of Planning and Public Policies Unit of Madeira.

Xavier Martínez-Cobas is titular professor of financial economics and accounting at the Universidade de Vigo. He holds a master's degree in accounting and auditing and a PhD in economics and business. He has been vice-rector of institutional relations at the Universidade de Vigo (2003-2006), CEO of Real Club Celta de Vigo (2006-2007), CEO of Xornal de Galicia (2008-2009) and commissioner of strategic plans at the Universidade de Vigo (2010-2018). He is trustee of the Isla Couto Foundation and the Penzol Foundation, both of public interest in Spain. His main lines of research are community, regional and transboundary development from the intellectual capital theory approach, and the interaction between foreign direct investment and energy transition.

Álvaro Rocha was listed in 2023 and 2024 as World's Top 1% Scientist by Stanford University and Elsevier, World's Top 0.05% Scientist by ScholarGPS, and World's Top 1% Scientist by ResearchGate for the fields of Information Science and Information Systems. He is Professor at ISEG, University of Lisbon, Invited Professor at University of Calabria, Honorary Professor at Amity University, President of ITMA - Information and Technology Management Association, Vice-Chair of IEEE SMC Portugal Chapter, and Book Series Scientific Manager at Springer-Nature. He holds the title of Honorary Professor, and holds Habil. in Information Science, Ph.D. in Information Systems and Technologies, M.Sc. in Information Management, and BCs in Computer Science. He is researcher at the ADVANCE (the ISEG Centre for Advanced Research in Management), and a collaborator researcher at CINTESIS (Center for Research in Health Technologies and Information Systems). His main research interests are maturity models, cybersecurity, management information systems, intelligent systems, e-government, e-health, and information technology in education. He is also Founder and Editor-in-Chief of both following Scopus and/or WoS journals: JISEM (Journal of Information Systems Engineering & Management) and RISTI (Revista Ibérica de Sistemas e Tecnologias de Informação / Iberian Journal of Information Systems and Technologies). Moreover, he has served as Vice-Chair of Experts for the European Commission's Horizon 2020 Program, and as an Expert at the COST - intergovernmental framework for European Cooperation in Science and Technology, at the European Commission's Horizon Europe Program, at the Government of Italy's Ministry of Universities and Research, at the Government of Latvia's Ministry of Finance, at the Government of Mexico's National Council of Science and Technology, at the Government of Polish's National Science Centre, at the Government of Cyprus's Research and Innovation Foundation, and at the Government of Slovak's Research Agency

Maria José Angélico Gonçalves holds a Ph.D. in Software Engineering, with a focus on reusable components and applications in human–machine interfaces, from the University of Vigo. She also earned an MSc in Computer Engineering from the Faculty of Engineering of the University of Porto (FEUP, 1996) and a BSc in Computer Science and Applied Mathematics from Portucalense University (1988). She is a Professor of Information Systems at the Porto Accounting and Business School and a researcher at

CEOS.PP (Center for Organizational and Social Studies of the Polytechnic of Porto). In addition, she collaborates as a researcher with OSEAN (Outermost Regions Sustainable Ecosystem for Entrepreneurship and Innovation). Her main research interests lie in Technologies and Information Systems. Over the years, she has taught a wide range of undergraduate and postgraduate courses and has participated in funded research projects in diverse areas, including Knowledge Management, Digital Transformation, the use of Technology in public and private organizations, and Cybersecurity. Her academic contributions include articles published in scientific journals, books, book chapters, and conference proceedings. She has also supervised and examined master's and doctoral theses, organized and participated in scientific events, and remains an active member of international research networks, including RIBCI. From 2021 to 2024, she served as Chair of the Technical-Scientific Council of the Porto Accounting and Business School.

Amélia Ferreira da Silva is a Coordinating Professor at Porto Accounting and Business School, Polytechnic University of Porto. She has taught Management Accounting since 2000. She holds a PhD in Accounting from the University of Vigo, Spain. Her research interests are in accounting and management control in healthcare, accountability in public organizations, business failure prediction, and digital transformation in accounting.
She supervised several Master´s dissertations and PhD's thesis. She has participated in several projects, namely STAMP (2020-1-UK01-KA203-0 KA203-080299), DIPCAT (2018-1-UK01-KA203-048027), ASSET (2023-1-IT02-KA220-HED-000156791) and GAMSTRA - Gamification Strategies for Management Skills. She has also several scientific publications and collaborations as a peer reviewer in index journals. She was head of CEOS.PP – Centre for Organizational and Social Studies of Porto Polytechnic, by FCT – Fundação para Ciência e Tecnologia. Since 2022, she has been the coordinator of the accounting department at Porto Accounting and Business School.

Received: March 11, 2025; Accepted: June 02, 2025.

Computer Science and Information Systems 22(4):1599–1615 https://doi.org/10.2298/CSIS250601063S

HRSP: A High-Risk Social Personnel Risk Assessment Model Based on Graph Attention Label Propagation Algorithm

Xin Su[1], Heng Zhang[2], Xuchong Liu[1], Chunming Bai[3], Wei Liang[4], and Ning Jiang[5]

[1] Hunan Police Academy
[2] Zhuzhou Public Security Bureau Economic Development Zone Branch
[3] Xiangtan University
[4] Hunan University of Science and Technology & School of Computer Science and Engineering
[5] Jingshan People's Hospital, no.448 xinshi road Jingshan city 431899, Hubei, China
hbjsxrmyy@163.com

Abstract. The current society is complex and changeable, and the post-pandemic era profoundly affects people's work and life. Identifying the potential risks of high-risk individuals in society and carrying out early warning and control work effectively is the focus of current public security work and is also the key to maintaining social stability and people's peace. This work first analyzes and constructs a knowledge graph of high-risk individuals based on their backgrounds, trajectories, and related information. Subsequently, we propose a high-risk personnel risk assessment model based on a graph attention-label propagation algorithm. The model employs a multi-label feature selection method, a basic classifier based on a graph attention network for the label propagation algorithm, and an adversarial data augmentation algorithm to enhance the gradient-based adversary during training. In the experiment, we train the model using a public-security-field personnel dataset, and the accuracy of the proposed method reaches 90.2%, Ablation experiments demonstrate the effectiveness and stability of the proposed method. Constructing a knowledge graph specifically for high-risk individuals based on backgrounds, trajectories, and related data,Proposing a risk assessment model using a graph attention-label propagation algorithm, incorporating multi-label feature selection and adversarial data augmentation, which enhances training effectiveness.

Keywords: Knowledgegraph, Graph Attention Label Propagation, Data Augmentation.

1. Introduction

Vicious incidents caused by individuals with extreme or mental issues have been a severe threat to social security and stability in recent years. The risk assessment of High-Risk Social Personnel (HRSP) still depends on the subjective experience of public security police, making it difficult to determine the risk changes promptly and accurately determine the risk. The limitations make the polices lack real-time and precise risk management for high-risk personnel. The urgent issue that the public security organization needs to solve is how to effectively assess the risks of high-risk individuals and identify potential-risk individuals[1,11].

High-risk personnel risk assessment technology is an essential area of smart policing. Efficient and accurate risk assessment technology can provide more intelligent and efficient support for public security departments to carry out risk prevention and control work and contribute to the response speed and accuracy of policing. The risk assessment of high-risk personnel based on deep learning is of great significance at the theoretical level[20,4,7]. First, it provides a new direction for developing risk assessment techniques for high-risk individuals. Most of the traditional risk assessment methods rely on the subjective experience judgment of police in the actual situation and carry out qualitative analysis of the risk characteristics of high-risk personnel while using deep learning technology to the risk assessment task of high-risk personnel can use algorithm model to conduct a more accurate and objective analysis of the risk characteristics and relational data of high-risk personnel. Second, the deep learning algorithms can identify potential high-risk personnel and risk rules that police cannot find from massive data on high-risk personnel, help police obtain a deeper understanding of the formation and evolution of the risk of high-risk personnel, and provide a new perspective for risk management and control.

In recent research, [17] use ordinal value quantification to calculate the risk factors for terrorist risk assessment. [19] use the data mining method to mine hidden risk factors from big data and assess the recurrence risk of correctional personnel. However, in those studies, two significant challenges need to be addressed. First, in the era of big data, the characteristics of high dimension, large magnitude, and multiple redundant features of human data are directly applied to machine learning, leading to efficiency decline and dimensional disasters. Second, a person has risk factors and numerous complex relationships with others, which also affect their risk coefficients. Therefore, the direct application of machine learning in risk assessment methods is ineffective in identifying potentially at-risk personnel.

In response to the first issue, we propose a Relief-GAs multi-label [10] feature selection method for feature selection to achieve feature dimensionality reduction. To address the second issue, we apply the knowledge graph to the risk assessment model, taking people as entity nodes and relationships between people as node edges [13]. Building a knowledge graph of high-risk individuals expands the scope of personnel data and mines more potential information by leveraging the relationships between nodes, reflecting the coupling risk factors among individuals. To improve the model's accuracy, we propose an enhanced graph attention network model [14]. The model serves as the base classifier for the label propagation algorithm, to predict and classify the risk level of high-risk personnel nodes with multiple relationships for risk assessment. Graph attention network models suffer from overfitting when trained on large-scale datasets, and real-world graph datasets involve many test nodes. We add the FLAG algorithm to iteratively add node features during training, allowing the model to maintain stability in response to input data's small fluctuations, enabling it to generalize to out-of-distribution samples and improve its performance during the test. Finally, comparative testing and ablation experiments on multiple datasets demonstrate the efficiency and effectiveness of our model [12].

In summary, the main contributions of this work are as follows:

1. To propose a new feature selection method, the Relief-GAs multi-label feature selection method first uses Relief to remove irrelevant features and then uses a genetic algorithm to find the optimal feature subset;

2. To construct a knowledge map of high-risk individuals, effectively exploring potential risks among high-risk individuals;
3. To improve the graph attention network and replace the basic predictor of the label propagation algorithm with the improved model, resulting in higher prediction accuracy.

The remainder of this work is organized as follows. The related research is explained in the second and third sections, including performing feature selection and improving graph attention networks for risk level prediction. Section 4 evaluates the effectiveness of the proposed method, and Section 5 is a summary.

2. Related works

In this section, we summarize the risk assessment methods of high-risk individuals and those in other fields.

2.1. High-risk Personnel Risk Assessment

Among the existing research methods, domestic and foreign scholars mainly analyze personnel risks through qualitative methods. There is few research on evaluating personnel risks by quantitative methods. [8] use naive Bayesian networks combined with four risk characteristics (static risk, violation score, sudden risk, psychological risk) manually annotated based on police experience to predict unknown risks of supervised personnel, with an accuracy rate of 84% and a recall rate of 86%. [3] construct a judgment matrix based on drug users' typical characteristics (physiological characteristics, social characteristics, drug exposure characteristics, and data characteristics) using the Analytic Hierarchy Process to predict the risk of social drug users. [18] propose a risk assessment method for key personnel by using random forest screening dataset features, selecting the optimal feature combination as the evaluation indicator, establishing a risk assessment system, using the AHP method to determine indicator weights, and combining the evaluation indicator scoring table.

2.2. Research In Other Fields

Several studies in other fields are related to risk assessment; for example, in food safety, food safety risks are analyzed and evaluated. Currently, research focuses on employing machine learning methods for risk assessment analysis. For example, [16] use BP neural network algorithm to combine features (food category, production province, sampling location) to classify and predict food risks with an accuracy rate of over 95%. Lou et al. [15] propose a risk prediction model based on differential automatic regression moving average and support vector machine (ARIMA SVM). Geng et al. [5] proposed an improved hierarchical clustering radial basis function (AHC-RBF) neural network for food risk prediction and early warning.

3. Methodology

3.1. Construction Of A Knowledge Map For High-risk Individuals

Based on the existing data on individuals and their relationships, we can construct a high-risk personnel knowledge graph, as shown in Figure 1.

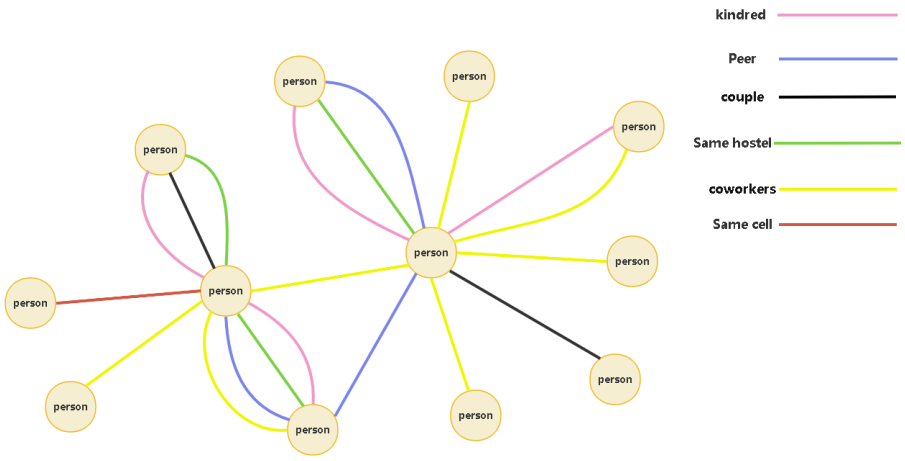


Fig. 1. Schematic diagram of high-risk individual knowledge graph construction

The entity of the knowledge graph is a person, and the attributes of the entity node are the person's risk features (criminal record, recent violation, mental illness, disreputable person, historical disputes, registered personnel, and basic personal information). The various relationships between people serve as node relationships. The entity set of the knowledge graph for high-risk individuals is $X = person(v1)$, with only one type of entity labeled as circles in the graph. The relationship type set $E = \{relative(r1)$, husband and wife$(r2)$, colleague $(r3)$, colleague $(r4)$, same hotel $(r5)$... same ward $(r6)\}$. In the figure, different colors represent different types of relationships. It can be seen that there are multiple edges between two entity nodes. By constructing a knowledge graph of high-risk individuals, we can explore their potential connections.

3.2. Overall Process Of Attention Label Propagation Method

This work proposes a graph attention label propagation method based on the graph attention mechanism consisting of five parts and the algorithm depicted in Figure 2. This method first uses the Relief-GAs multi-label feature selection algorithm to reduce the dimensionality of feature data (section 3.3). The dimensionality-reduced feature data is then used as the basis for a simple prediction classifier using an improved graph attention network. The resulting prediction results are refined using a residual propagation algorithm to correct errors in the prediction, and then the final prediction results are smoothed

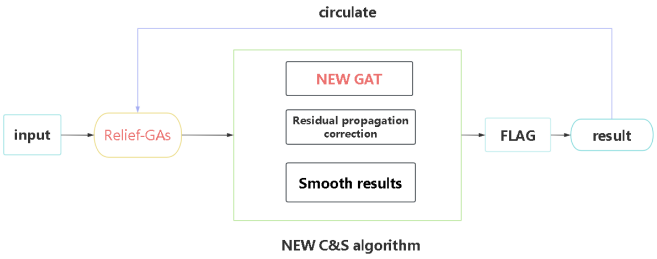


Fig. 2. Describes the five components of the label propagation method based on the graph attention mechanism

using a graph structure (section 3.4). During model training, the FLAG data augmentation technique improves adversarial interference based on gradients and achieves data augmentation (section 3.5).

3.3. Relief GAs Multi-label Feature Selection Method

In the era of big data, personal data has multiple characteristics, such as high dimensionality, large data volume, and numerous redundant features [9]. To avoid dimensionality disasters and improve data value density and model evaluation efficiency, feature selection methods can be used to reduce data dimensionality. Existing feature selection methods can be classified into filter-based, wrapper-based, and embedded-based. Relief is an efficient filter-based feature weight algorithm, but it assigns high weight to all features with high correlation with the class and cannot effectively remove redundant features. We propose a Relief-GAs multi-label feature selection method to address the shortcomings. The algorithm removes irrelevant features using

$$W(A) = W(A) - \sum_{j=1}^{K} \frac{diff(A, R, H_j)}{mK} + \sum_{cclass(R)} \frac{\frac{p(C)}{1-P(class(R))} \sum_{j=1}^{K} diff(A, R, m_j(C))}{mK}. \tag{1}$$

Then uses a genetic algorithm to find the optimal feature set.Among them, $p(C)$ is the proportion of the category, $p(class(R))$ is the proportion of the category of a randomly selected sample, $diff(A, R_1, R_2)$ represents the sample R_1,R_2. The distance on feature A, where m is the number of samples, k is the number of nearest neighbor samples, and $M_j(C)$ represents the $j-th$ nearest neighbor sample in class C.

The algorithm process is as follows:

1. Randomly select a sample R from the training set using the ReliefF algorithm, extract K nearest neighbor samples $H_j(j = 1, 2, \ldots k)$from similar sample sets of R, and then find K nearest neighbor samples $M_j(C)$ from different sample sets, Finally, update the feature weights according to $W(A)$to obtain the average weight of each

feature in a single label dataset and select features with a mean greater than 0 as relevant features to filter the features.
2. Randomly generate feature sets and train the model.
3. Evaluate the fitness of each feature set and remove feature subsets with poor adaptability.
4. Cross-construct a new feature set from the remaining feature sets.
5. Repeat iterations to achieve optimal results.

3.4. Risk Profile

The core of the model is divided into three parts: basic predictor, residual propagation correction, and smoothing prediction results. First, the basic predictor is used to predict the risk of personnel, and the error between the predicted and the truth is corrected by residual propagation. Finally, based on the assumption of the label propagation algorithm, adjacent nodes with similar labels are used to smooth the final prediction results. Each part is introduced as follows.

Basic predictor. The basic predictor is divided into four layers, as shown in Figure 3. Firstly, input the feature data and relational data. Since relational data cannot be directly applied to machine learning, we use the relational quantization layer to quantify relational data as the edges' weights in the graph. The obtained multiple quantitative relationship data and feature data pass through the entity layer, where the feature data is aggregated based on each relationship data, and the numerous aggregated feature matrix is obtained. After the relationship layer, the multiple aggregation matrices are fused to obtain the final feature matrix, which is then used to obtain the classification result in the classification layer. The detailed workflow for each layer is provided below. Since character graph data differs from other graph data during node categories prediction, there are usually multiple relationships between two-character nodes, and node categories are more dependent on the node's relationships. Therefore, this work modifies the graph attention network to serve as the fundamental predictor for label propagation algorithms.

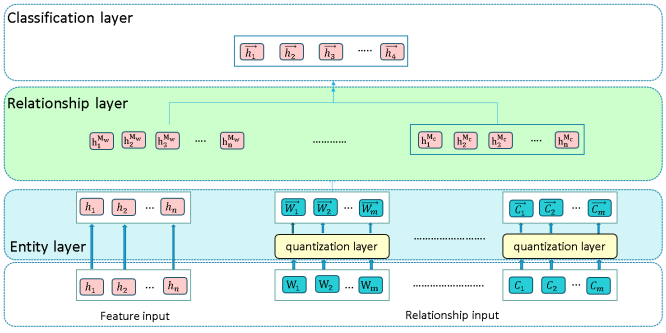


Fig. 3. Detailed structure of the basic classifier used for the label propagation algorithm diagram

Relationship quantification layer,Due to the knowledge graph of high-risk individuals with single entity types and multiple relationship types, different relationship types have different degrees of impact on risk coefficients, and relational data cannot be directly applied in machine learning. Therefore, it is necessary to quantify the relationships between individuals as

$$s_{ij}^{M_K} = a^{M_K} \cdot \frac{h_i \cdot h_j}{\sqrt{\sum_{f=1}^{n} h_{if}^2} \cdot \sqrt{\sum_{f=1}^{n} h_{jf}^2}} \,. \tag{2}$$

In the equation: $s_{ij}^{M_K}$ is the between entity i and entity j quantized value of the relationship under relationship matrix M_K, a^{M_K} is Mapping function under relationship matrix M_K , h_{if}^2 is the value of the $f-th$ dimension of the feature vector of entity i.

Physical attention network layer, The size of a person's risk is related to their risk characteristics and interpersonal relationships. For example, if a person has an intimate relationship with high-risk individuals, their risk level will be increased; that is, by designing an entity attention network layer to achieve the goal. Firstly, learn the attention coefficient between entity a and all connected neighboring entities b in the relationship matrix M_K , where the relationship quantification value is greater than the preset threshold, and calculate it according to equation

$$e_{ij}^{M_K} = a\left(\left[W^{M_K}\ h_i\right| \left|W^{M_K}\ h_j\right]\right). \tag{3}$$

where h_i and h_j are the original feature vectors of entities i and j , and M_K is the weight matrix of the entity attention network layer in the relationship matrix M_K . $||$ is a symbol that represents a connection operation, a is the mapping function , $e_{ij}^{M_K}$ is the attention coefficient in the relationship matrix M_K . Subsequently, normalize the attention coefficient and calculate it in accordance with equation :

$$\partial_{ij}^{M_K} = \frac{\exp\left(LeakyReLU\left(e_{ij}^{M_K}\right)\right)}{\sum_{s \in N_i^{M_K}} exp\left(LeakyReLU\left(e_{is}^{M_K}\right)\right)}. \tag{4}$$

where LeakyReLU is a nonlinear activation function $s \in N_i^{M_K}$, M_K where all relationship quantization values of entity i are greater than the threshold, ∂_{ij} is the normalized attention coefficient.By Using equation $\partial_{ij}^{M_K}$, we can obtain the attention coefficient. By linearly combining this coefficient with adjacent points whose relationship quantization value exceeds the threshold, we aggregate the features after the entity attention network layer under the relationship matrix M_K and get a new feature vector $h_i^{M_K}$. To enhance model training stability, we set the attention mechanism head to $k = 8$, and averaged the feature vectors as equation :

$$h_i^{M_K} = \sigma\left(\frac{1}{K}\sum_{K=1}^{K}\sum_{j \in N_i^{M_K}} \partial_{ij}^{M_K} \cdot W^K \cdot h_j\right). \tag{5}$$

In the above equation, the σ is a nonlinear activation function; the J is the adjacency point j where all relationship quantization values under the relationship matrix M_K are greater than the threshold, $\partial_{ij}^{M_K}$ is the attention coefficient between entities obtained from the

$k-th$ attention mechanism head under the relationship matrix M_K. W^K is the feature transformation matrix corresponding to the $K-th$ attention mechanism head.

Relationship attention network layer,In the knowledge graph of high-risk individuals, each relationship between individuals corresponds to a feature matrix. It is essential to merge new feature vectors from each relationship feature matrix to obtain more valuable entity feature vectors. We design a relationship-level attention network layer. The input for the entity attention network layer serves as the input for the relationship attention network layer, which is calculated using equation:

$$\gamma^{M_K} = \frac{exp\left(\frac{1}{N}\sum_{i\in h} a \cdot sigmoid\left(W \cdot h_i^{M_K} + b\right)\right)}{\sum_{K=1}^{K} exp\left(\frac{1}{N}\sum_{i\in h} a \cdot sigmoid\left(W \cdot h_i^{M_K} + b\right)\right)}. \tag{6}$$

where γ^{M_K} is the attention coefficient, A is the attention mechanism vector, W is a parameterized matrix, b is an offset vector, and sigmoid is a nonlinear activation function. Perform linear combination like the physical network layer, set the attention mechanism header to $p = 8$, and perform averaging, calculated according to equation:

$$\vec{h_i} = \partial\left(\frac{1}{p}\sum_{p=1}^{p}\sum_{k=1}^{k}\gamma^{M_K} \cdot h_i^{M_K}\right). \tag{7}$$

Entity classification laver,The last layer of the model is the entity classification network layer. The core is to aggregate the features of entity I from $\vec{h_i} \in R^F$ to R^C according to equation:

$$\vec{h_i'} = sigmoid(W_c \cdot \vec{h_i}). \tag{8}$$

The set of feature vectors aggregated through these four layers of networks is $\vec{h_i} = \{f_1, f_2, \ldots, f_c\}$, which corresponds to the final classification feature values of the entity. The feature dimension C is used as the risk category to be classified. According to the risk level, the risk is divided into 4 categories (high risk, medium risk, low risk, no risk), and $c = 4$. By normalizing with the softmax function, the probability of the risk level of entity i can be obtained, according to equation :

$$P_i^{f_x} = \frac{\exp\left(f_x\right)}{\sum_{c=1}^{c}\exp\left(f_c\right)} \quad , x \in [1, C]. \tag{9}$$

where $P_i^{f_x}$ is the probability value that entity i belongs to f_x , f_x is an eigenvalue in the entity I eigenvector. After obtaining the basic prediction results, proceed to the next level for error correction.

Error correction in basic prediction by residual propagation. The basis for label propagation is the assumption that adjacent nodes have the same label, which means that the label information of nodes is positively correlated along the graph edges. Therefore, the prediction error of nodes is also positively correlated along the edges, which can improve the accuracy of basic prediction results by combining label information association errors. Firstly, define an error matrix E, where the error of the training set is the residual between its predicted results and the actual label, and the remaining errors are zero:

$$E_T = Z_T - Y_T \; EV = 0 EU = 0. \tag{10}$$

where E_T represents the error of the training set, only when the predicted results are identical to the real labels, the residual of the corresponding training node is zero, and the validation set error E_V and test set error E_U are zero.

Using label propagation technology to smooth errors on the graph and the optimization target is:

$$E = argmintrace\left(W^T\left(1-S\right)W\right) + \mu|\left|W-E\right|\right|_F^2. \tag{11}$$

where W is the final error matrix to be obtained, and S is the normalized adjacency matrix $S = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, The first term of the formula promotes the smoothness of the error on the graph, which is equivalent to $\sum_{j=1}^{c} W_j^T(I-S)W_j$, where W_j represents the jth column of matrix W, The second term of the formula controls the degree of deviation between the final solution and the initial error. The solution can be obtained by iterating the equation $E^{t+1} = (1-\alpha)E + \alpha SE^t$ to rapidly converge to E, where $\alpha = \frac{1}{1+\mu}$, $E^{(0)} = E$. This iteration is the process of error propagation, where the smoothed error is added to the basic prediction and the corrected basic prediction $Z^r = Z + E$ is obtained. This is a post-processing technique that does not participate in the training process of the basic predictor.

Smooth the final prediction using graph structure. The corrected basic prediction result $Z^{(r)}$ is obtained after the Correct process. To obtain the final prediction and fully utilize the structural information of the graph, further smoothing processing is needed on the corrected prediction $Z^{(r)}$. Its motivation comes from the assumption in label propagation algorithms that adjacent nodes are likely to have similar labels. Therefore, another label propagation process is used to promote the smoothness of label distribution on the graph. Start from the best prediction of labels $\widehat{Y} \in R^{n\times c}$, $\widehat{Y}_T = Y_T$, $\widehat{Y}_{V,U} = Z^{(r)}_{V,U}$. Replace label information of the trained node information $\widehat{Y}_T$ with real label Y_T, and replace Label information $\widehat{Y}_{V,U}$ in validation and testing sets with the corresponding label information $Z^{(r)}_{V,U}$ in basic prediction $Z^{(r)}$. Iterate the equation $Y^{(t+1)} = (1-\alpha)\widehat{Y}_T + \alpha s\widehat{Y}^{(t)}$ until convergence to obtain the final prediction result $\widehat{Y}$ and perform row normalization to obtain the final label distribution and the node label prediction. Like the Correct process, the smooth process is also a post-processing process and does not participate in the training process of the basic predictor. Then obtain the final prediction result.

3.5. FLAG Data Augmentation

As a semi-supervised learning task, graph node classification often faces a low proportion of labeled nodes [2]. Whenever there is a large difference in the distribution of labeled and unlabeled nodes, the model is prone to overfitting, resulting in a significant difference between the prediction results of unlabeled nodes and the truth, and insufficient generalization ability of the model. Therefore, a data augmentation algorithm called FLAG based on gradient-based adversarial perturbation is used in the model in this work. During training, the node features are iteratively enhanced to make the model invariant to input data's small fluctuations, allowing the model to generalize to out-of-distribution samples and improve its performance, as depicted in Algorithm 1 the flowchart of the FLAG algorithm. The basic idea of the algorithm is to increase the number of projection gradient

descents during each model training iteration and reduce the number of model training iterations. During each model training process, a perturbation matrix of the same size as the feature matrix is defined and sent to the model together with the feature matrix for training. During each gradient descent step in the training process, the perturbation matrix is updated based on its gradient. The gradients are added together, and finally, the parameters are updated by backpropagation. The algorithm adds perturbations to labeled and unlabeled nodes, and adding this algorithm to the high-risk personnel risk assessment model can improve the model's accuracy. In addition, FLAG can alleviate the model's over-smoothing problem and enable the design of deeper graph neural networks.

Algorithm 1 FLAG: Free Large-scale Adversarial Augmentation on Graph

Require: Graph $G = (V, E)$; input feature matrix X; learning rate τ; ascent step M; ascent step size ω; training epochs N; forward function on graph $f_\theta(\cdot)$ denoted in $H^{(k)} = f_\theta(X; G)$; $L(\cdot)$ as objective function. We omit the READOUT$(\cdot)$ function in $h_G = \text{READOUT}(\{h_v^{(k)} \mid v \in V\})$ for the inductive scenario here.
Initialize θ
for epoch $= 1$ **to** N **do**
  $\delta_0 \leftarrow U(-\omega, \omega)$
  $g_0 \leftarrow 0$
  for $z = 1$ **to** M **do**
    $g_z \leftarrow g_{z-1} + \frac{1}{M} \cdot \nabla_\theta L(f_\theta(X + \delta_{z-1}; G), y)$
    $g_\delta \leftarrow \nabla_\delta L(f_\theta(X + \delta_{z-1}; G), y)$
    $\delta_z \leftarrow \delta_{z-1} + \omega \cdot \frac{g_\delta}{\|g_\delta\|_F}$
  end for
  $\theta \leftarrow \theta - \tau \cdot g_M$
end for

4. Evaluation

4.1. Experimental Environment and Dataset

The experimental computer server is composed of Intel Core i7-9700F CPU, NVIDIA GeForce RTX 3070Ti GPU, 8GB memory, 500GB SSD, running Windows 10 operating system and PyTorch, open-source deep learning framework.

The dataset used for training and evaluation is the public-security-field personnel dataset, which contains three types of personnel data (mental illness, criminal record holders, and drug users). The public-security-field personnel dataset selects three types of personnel information data and personnel relationship data that have been desensitized. The structure of personnel information data and relationship data significantly affects the model's accuracy, and appropriate data preprocessing can make the prediction results more accurate. For personnel information data, based on the Relief-Gas multi-label feature selection method in this paper, the optimal feature combination was obtained, which takes "whether there is a case history, recent violations, mental illness, dishonest individuals, historical disputes, and registered individuals" as a strong correlation factor affecting

human risk. For personnel relationship data selection, "peers, same supervision room, and same hotel" are taken as strong correlation relationships. At the same time, people are divided into four categories based on risk levels, personnel information data, and personnel relationship data. Table 1 provides detailed information corresponding to each dataset.

Table 1. Provides detailed information about the public-security-field personnel dataset, including entities, relationships, features, and categories

Data	entities	Relationship	sides	features	category
Drug	1000	6	7689	12	5
Ex-offenders	2000	6	14302	12	5
Psychopath	2000	6	10394	12	5

4.2. Experimental Content

This work compares the model with mainstream methods based on the neural network model, including selecting Graph Attention Network (GAT), C&S, and GraphSAGE to compare accuracy and recall in the public-security-field personnel dataset. To demonstrate the effectiveness of the model algorithm in different situations, several ablation experiments are conducted to analyze the other modules' impact on performance.

4.3. Results Of Classification Accuracy And Recall

The experimental results of the proposed method are compared with mainstream methods, as depicted in Figure 4. Figure 5 shows that using the proposed model has significant accuracy improvement compared to GAT, C&S, and GraphSAGE. Also, as depicted in Figure 6, it can be observed that the PR curve of this model completely covers those of other models, indicating that the recall and accuracy of the proposed model are superior to others. This is because the proposed model adopts the Relief-Gas multi-label feature selection method, an improved graph attention network as the basic predictor, and incorporates the FLAG algorithm to add gradient-based adversarial perturbations to the input node features to enhance the data. Thus, the model can be generalized to samples outside the distribution and improve the performance. Finally, it is more robust and discriminative. The personnel data validation set consists of 500 nodes, each with a risk level label. The model is used to predict each node's risk, then compare them with the risk level label to calculate the model accuracy. The detailed classification results are shown in Table 2.

Considering that in real life, there are two unfavorable conditions with human data: one is edge missing (missing relationships between people) and the other is node information missing [6] (missing human node features), and this work conducts relevant experiments for these two situations.

Edge missing. In the experiment, preserve 20%, 40%, 60%, 80%, and 100% of the edges in the graph and compare them with other models. In the experiment, 70% of the dataset is taken as the training set and 30% as the testing set.

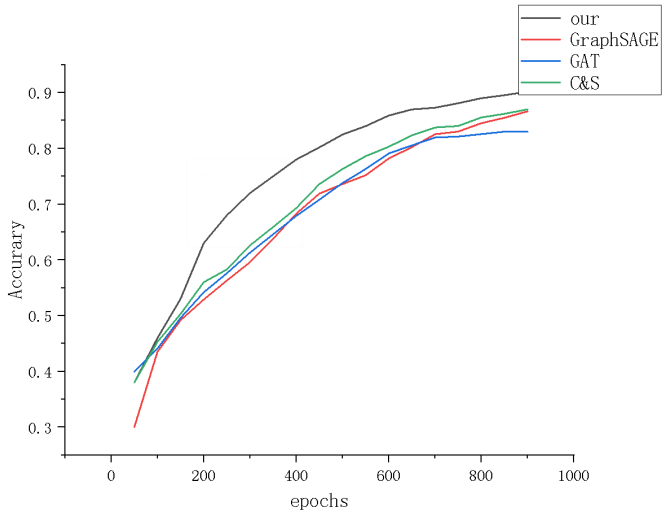


Fig. 4. Training accuracy of different models over epochs. Shows the training accuracy curves for ”our”, GraphSAGE, GAT, and C&S models

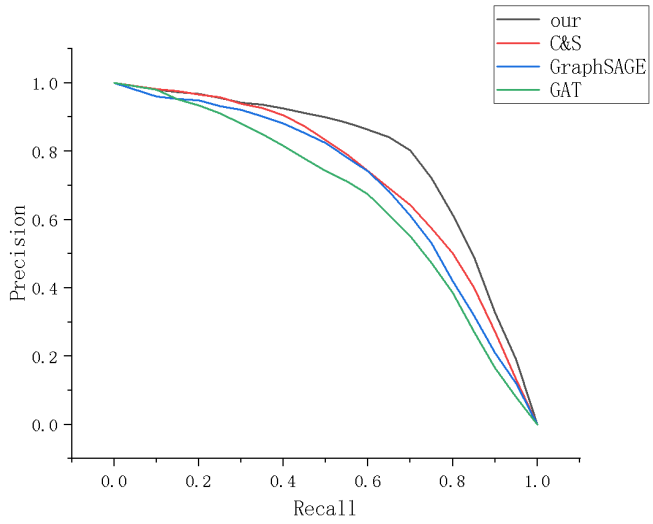


Fig. 5. Compares the PR curves of different models on the public-security-field personnel dataset

The experimental results of different edge missing rates are shown in Figure 6. The graph shows that the more complete the edge information in the graph structure, the higher the accuracy. It verifies that the structural information in the graph can assist in completing node classification tasks. The more complete the structural information, the better the assistance effect. At the same time, it can be found that under different edge missing rates, compared to GAT, C&S models, the classification accuracy of proposed model is higher. This is benefitted from our model's ability to finely mine the interaction information between nodes and the specific interaction information contributing to better classification.

Table 2. Shows the number of correct and incorrect classifications for different risk levels by the model

CATEGORY	CORRECT	INCORRECT
HIGH	31	2
MID	56	7
LOW	46	5
NO	318	35

Missing node information. In the experiment, 40%, 60%, and 80% of non-isolated nodes in the knowledge graph are empty, indicating that they only have structural information. Comparing this model with others, it can be seen from Figure 7 that the classification performance decreases as the amount of data for missing nodes increases. It indicates that node attribute information has an important impact on node classification performance. In the case of varying degrees of information loss, our model performs better compared with GAT and C&S models. This is benefitted from the fact that our model can more finely infer the interaction information between two nodes and thus infer the information of missing nodes.

From the above experiments, it can be concluded that the proposed model utilizes the information and structural information of nodes in the knowledge graph to guide node classification and effectively mines the implicit information between nodes, bringing better classification performance.

4.4. Ablation Studies

From the results of the ablation experiment in Table **??**, the NEW GAT performs feature aggregation on multiple relational nodes, resulting in higher accuracy compared to results of those using traditional GAT as the basic predictor for C&S. At the same time, the FLAG data augmentation algorithm can help the application of graph algorithms and improve model performance. The best feature combination can be found by incorporating the Relief-Gas multi-label feature selection method, and the model's accuracy can be improved. Combining these four methods can significantly improve the accuracy and recall of the model.

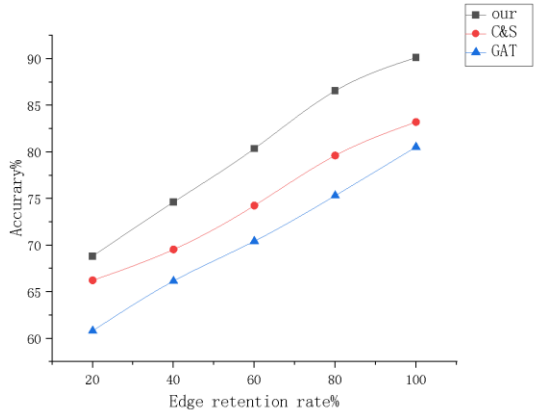


Fig. 6. Shows the impact of different edge retention rates on model classification accuracy

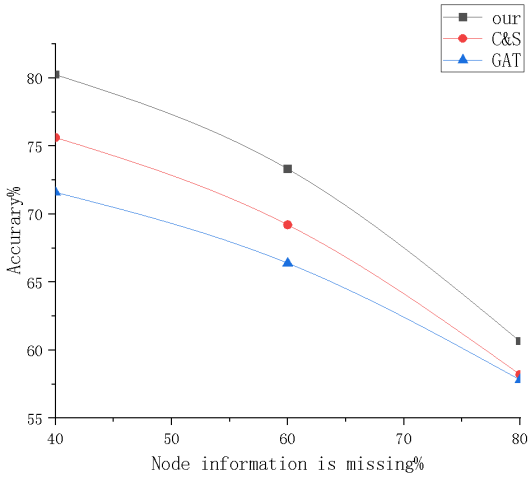


Fig. 7. Shows the impact of different node information missing rates on model classification accuracy

5. Conclusions And Future Work

In the post-pandemic era, how to tap into the potential dangers of high-risk individuals in society and better carry out early warning and control work is the focus of current public security work. With the development of technology and the increase in data volume, traditional risk assessment methods for high-risk personnel can no longer meet the requirements of police officers. This article aims to design a more efficient risk assessment model for high-risk personnel.

A High-Risk Social Personnel Risk Assessment Model Based on Graph Attention Label Propagation Algorithm. By analyzing the main risk factors and relationships that affect the risk coefficient of high-risk individuals, a knowledge graph of high-risk individuals is constructed, and a risk assessment model for high-risk individuals based on graph attention label propagation algorithm is proposed. At the same time, the Relief-Gas label selection algorithm is used to identify the optimal feature set. we train the model using a public-security-field personnel dataset, the accuracy of proposed method reaches 90.2 the recall of proposed method reaches 90.6%. Through comparative experiments with other mainstream graph neural networks, the experimental results show that the proposed model performs better in the graph node classification task.

The experimental results show that the improvements proposed in this paper for graph attention network and label propagation algorithm can improve the performance of the model in the graph data classification task, and the combination of all the improvements in the risk assessment of high-risk personnel can accurately predict the risk level of high-risk groups, improve the work efficiency of public security personnel, and play an important guiding role in social security work.

This model's innovations in the field of high-risk personnel risk assessment lie in its ability to leverage graph attention mechanisms to analyze complex relationships within the constructed knowledge graph, enhancing the accuracy of risk prediction. The practical value of this model is evident in its potential to assist public security agencies in identifying potential threats more efficiently, thereby enhancing public safety. Compared to existing technologies, the superiority of this model is demonstrated through its higher accuracy and recall rates, as well as its robustness in handling real-world data with varying degrees of incompleteness.

In the future, we will continue to optimize the algorithm proposed in this study to improve its performance and generalization ability. Specifically, the following aspects will be considered.

1. Improving the size and quality of the dataset will continue to expand its size and incorporate more scenarios and situations, enhancing the adaptability and generalization ability of the algorithm.
2. The ultimate goal of this paper is to improve the computational efficiency and speed of the algorithm, and successfully apply it to the intelligent policing platform to help public security officers better maintain social order. Therefore, we will continue to optimize the computational efficiency of the algorithm.
3. Explore more network structures and algorithms, continue to focus on relevant algorithms in related fields, and explore more algorithms to improve the effectiveness and performance of the model.

Acknowledgments. This research was supported by the Key Research and Development Program of Hunan Province of China under Grant No. 2022SK2109, Hunan Provincial Department of Education Outstanding Youth Fund under Grant No. 21B0850, Zhejiang Provincial Natural Science Foundation of China under Grant No. LTGG23F020004, Natural Science Foundation of Fujian Province under Grant No. 2023J011460.

References

1. Cai, J., Liang, W., Li, X., Li, K., Gui, Z., Khan, M.K.: Gtxchain: A secure iot smart blockchain architecture based on graph neural network. IEEE Internet of Things Journal 10(24), 21502–21514 (2023)
2. Cai, J., Liang, W., Li, X., Li, K., Gui, Z., Khan, M.K.: Gtxchain: A secure iot smart blockchain architecture based on graph neural network. IEEE Internet of Things Journal (2023)
3. Cai, L.: Research on risk assessment of social drug users. People's Public Security University of China (2019)
4. Diao, C., Zhang, D., Liang, W., Jiang, M., Li, K.: A novel attention-based dynamic multi-graph spatial-temporal graph neural network model for traffic prediction. IEEE Transactions on Emerging Topics in Computational Intelligence 9(2), 1910–1923 (2025)
5. Geng, Z., Liu, F., Shang, D., Han, Y., Shang, Y., Chu, C.: Early warning and control of food safety risk using an improved ahc-rbf neural network integrating ahp-ew. Journal of Food Engineering (2021)
6. Hao, Z., Ke, Y., Li, S., Cai, R., Wen, W., Wang, L.: Node classification method in social network based on graph encoder network. Journal of Computer Applications 40(1), 188–195 (2020)
7. Hu, N., Zhang, D., Xie, K., Liang, W., Li, K., Zomaya, A.: Multi-graph fusion based graph convolutional networks for traffic prediction. Computer Communications 210, 194–204 (2023), https://www.sciencedirect.com/science/article/pii/S0140366423002785
8. Li, S., Li, P., Li, S., Zhao, S.: Research on the application of risk assessment and early warning model for supervised personnel. Police Technology pp. 38–41 (2021)
9. Li, Y., Liang, W., Xie, K., Zhang, D., Xie, S., Li, K.C.: Lightnestle: Quick and accurate neural sequential tensor completion via meta learning. In: IEEE INFOCOM 2023 - IEEE Conference on Computer Communications. pp. 1–10. IEEE (2023)
10. Liang, W., Li, Y., Xie, K., Zhang, D., Li, K.C., Souri, A., Li, K.: Spatial-temporal aware inductive graph neural network for c-its data recovery. IEEE Transactions on Intelligent Transportation Systems 24(8), 8431–8442 (2022)
11. Liang, W., Li, Y., Xie, K., Zhang, D., Li, K.C., Souri, A., Li, K.: Spatial-temporal aware inductive graph neural network for c-its data recovery. IEEE Transactions on Intelligent Transportation Systems 24(8), 8431–8442 (2023)
12. Liang, W., Li, Y., Xu, J., Qin, Z., Zhang, D., Li, K.C.: Qos prediction and adversarial attack protection for distributed services under dlaas. IEEE Transactions on Computers (2021)
13. Liang, W., Xie, S., Cai, J., Xu, J., Xu, Y., Hu, Y.: Deep neural network security collaborative filtering scheme for service recommendation in intelligent cyber-physical systems. IEEE Internet of Things Journal 9(22), 22123–22132 (2021)
14. Liu, Y., Liang, W., Xie, K., Xie, S., Li, K., Meng, W.: Lightpay: A lightweight and secure off-chain multi-path payment scheme based on adapter signatures. IEEE Transactions on Services Computing (2023)
15. Lou, H., Cao, Q., Li, H.: The prediction of food safety risk of china's export to eu based on the arima-svm combination model. Food Industry 41(1), 334–339 (2020)
16. Wang, X., Zuo, M., Xiao, K., Liu, T.: Data mining on food safety sampling inspection data based on bp neural network. Journal of Food Science and Technology 34(6), 85–90 (2016)

17. Wang, Y., Dang, J., Hu, X., Zhang, Y.: Research on the application of matrix method and sequential value method in risk assessment of terrorist suspects. Journal of Zhejiang Police College 2021(01), 83–91 (2021)
18. Wu, R., He, Y., Jia, Y.: Key personnel risk assessment method based on improved ahp. China Safety Science Journal 31(10) (2021)
19. Wu, Z., Fang, Y.: Risk assessment of community correction personnel recidivism based on data mining technology. Guizhou Social Sciences 2016(7), 82–87 (2016)
20. Zhang, S., Ren, F., Liang, W., Li, K., Ling, N.: Gpvo-fl: Grouped privacy-preserving and verification-outsourced federated learning in cloud-edge collaborative environment. IEEE Transactions on Network and Service Management pp. 1–1 (2025)

Xin Su was born in Hunan, China in 1983. He received the Ph.D from Hunan University, Changsha, China, in 2015. He is currently an professor in the Department of Information Technology at Hunan Police Academy, Changsha, China. His research interest is Large Language Model, Machine learning, and Data Analyze.

Heng Zhang was born in Hunan, China in 1980. He received the bacheler degree from People's Public Security University of China, Beijing, China, in 2002. He is currently an political commissar in Zhuzhou Public Security Bureau Economic Development Zone Branch, Zhuzhou, China. His research interest is Intelligence policing.

Xuchong Liu was born in Hunan, China in 1974. He received the Ph.D from Central South University, Changsha, China, in 2010. He is currently an professor in the Department of Information Technology at Hunan Police Academy, Changsha, China. His research interest is Intelligence policing, Big data analyze.

Chuming Bai was born in Shanxi in 2000. He is currently a third-year graduate student at the School of Computer Science at Xiangtan University. His research interests include federated learning, large-scale language models, and deep learning.

Wei Liang was born in Hunan, China in 1978. He received the Ph.D from Hunan University, Changsha, China, in 2013. He is currently an professor in the College of Computer Science and Engineering at Hunan University of Science and Technology, Xiangtan, China. His research interest is Blockchain, Quantum Secure Computing.

Ning Jiang was born in Hubei, China in 1987. He received the ma.eng from wuhan University, wuhan China, in 2012. He is currently as the Director of the Information Department at Jingshan People's Hospital in Hubei Province. His research interests include software engineering, big data, and hospital information management.

Received: June 01, 2025; Accepted: August 22, 2025.

Computer Science and Information Systems 22(4):1617–1636 https://doi.org/10.2298/CSIS240810064A

A study on Multi-scale Attention dense U-Net for image denoising method

MingShou An , XuHang Zhao , Hye-Youn Lim and Dae-Seong Kang

Dong-A University, Dept. of Electronics Engineering, 37 Nakdong-daero 550 beon-gil
Saha-gu, Busan, Korea
anmingshou@xatu.edu.cn
dskang@dau.ac.kr

Abstract. Although many models that have applied learning exhibit good performance, the dataset or image generation and transmission process used for learning may contain noise, which cannot produce the expected results and performance. The representative image denoising technique using deep neural networks generates noisy images by forcibly adding special noise to the original image and learning to make it the same as the original image. However, the performance of deep neural networks depends on depth, and to improve performance, increasing only depth will reach a performance saturation state, which will encounter difficulties. In order to improve these issues, this article applies the Multi-scale Attention model to the representative denoising deep learning model U-Net, to suppress unnecessary information and provide functionality that only emphasizes important information. In a new modular approach, the given input value is divided into two parts based on its internal relationship: the part where the important parts are concentrated and the part where the important parts are concentrated through spatial information. The attention unit based Outburst structure, which combines the two parts after parallel execution, has been implemented, demonstrating better performance than existing models. Moreover, without adding too many parameters, more spatial feature maps than other models are generated by focusing on the effects of components, not only through PSNR and SSIM. The improved performance was also confirmed by removing noisy in images.

Keywords: deep learning, image denoising, Multi-scale Attention, U-Net, outburst structure.

1. Introduction

The existing image processing and analysis involve the entire field of digital image information preprocessing, feature extraction, image restoration and image compression, etc. At present, artificial modeling of human learning and reasoning abilities involves a field of artificial intelligence such as character recognition and image pattern analysis - deep learning. It is worth mentioning that deep learning has emerged in the field of image processing by directly utilizing image spatial information for feature extraction [1]. Representative deep learning techniques for image processing include image classification, object detection, etc. Image classification marks pre determined class information as image data for learning, and the learned model takes the image to be classified as input and outputs class information to distinguish the types of objects. Representative deep

learning technologies in these fields include AlexNet [2], VGGNet [3], GoogLeNet [4], ResNet [5], and many other deep learning models based on CNN [6], which have achieved significant results. The performance of deep learning models may be affected by various noises contained in the input images used during learning. Moreover, in the process of image generation and transmission, there will inevitably be noise involved. Deep learning models that learn through input data also calculate noise during learning, so if the input data is heated by noise, it can lead to performance degradation. So in practical environments, denoising during evaluation is essential [7]. To improve these issues, image denoising techniques are needed. Image denoising technology has been widely applied in multiple fields such as restoring image details and accepting images as inputs. At present, in the field of image denoising, research on the application of deep neural networks is very active in order to improve performance [8].

In order to improve the performance of image denoising, simply increasing the depth of the deep neural network will result in too much computational complexity and may lead to difficulties in performance degradation or saturation [9]. To improve these issues, based on noise pattern prediction, the structure of the neural network was changed to a parallel form instead of increasing the depth of the neural network, thereby expanding the scope of the neural network. By improving the existing Attention units, a new module called multiscale Attention units has been implemented, which can extract more spatial feature information, suppress unnecessary information, and only emphasize important information. Multiscale Attention units are composed of parallel parts that concentrate what is important and where spatial information is important through internal relationships with inputs [10]. In terms of overall structure, noise can be effectively removed by using an Outburst structure [11] that surges feature information in the latter half of the neural network. This can be widely applied to systems that require denoising in image processing.

2. Related Work

2.1. DnCNN

DnCNN(Denoising Convolutional Neural Network) [12] is a deep learning technique that utilizes CNN to implement image denoising. Using Additive White Gaussian Noise (AWGN) as noise, and training the model to remove noise. The existing denoising techniques have long computation time, complex parameter settings, and require a large amount of computation and direct manual interference. The focus of DnCNN is not to directly remove noise from noisy images, but to separate noise from noisy images. The use of CNN ensures the flexibility of image dinoising and achieves performance improvement through residual learning [13] and Batch Normalization [14].

In addition, they also modified the network structure based on VGGNet(Visual Geometry Group) to achieve image denoising. In the technology used here, residual learning is used to improve accuracy, even if the model becomes deeper, it can maintain generalization well, rather than adding parameters, and the calculation is not complicated. And Batch Normalization, through its own generalization process, has low sensitivity and is not affected by parameter size during learning. It can greatly set the learning rate and achieve fast learning. The size of the each convolutional filter of DnCNN is 3x3. The

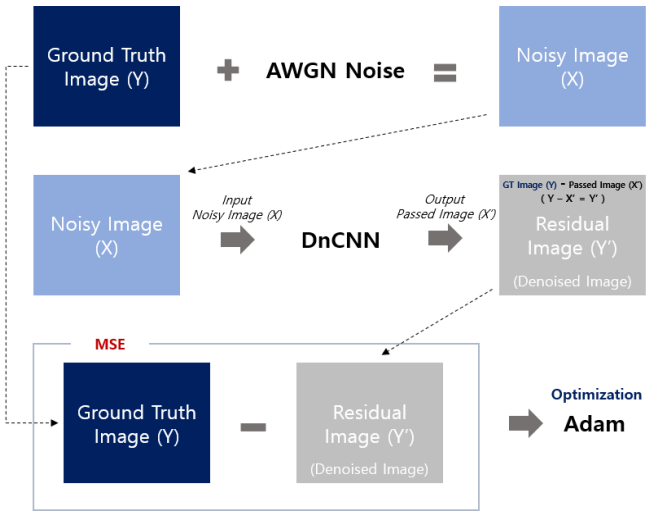


Fig. 1. The structure of DnCNN

model in the Denoising field determines the receptive field size based on the effective patch size. High noise levels typically require a larger effective patch size to capture more detailed features. DnCNN provides a noise level fixed at 25, analyzing the effective patch size of leading noising methods for depth design.The structure of DnCNN is shown in Fig.1, which creates a Noise Image (X) by obtaining the Ground truth image (Y) and adding AWGN noise. Create an image ($X^{'}$) using a CNN network in Noisy Image (X), subtract the resulting image from the original image (Y), and generate a Residual Image ($Y^{'}$). The final calculation of MSE for Y and $Y^{'}$ is reflected in the optimization of the CNN network. Fig.2 shows the average PSNR with and without batch normalization (BN) and residual learning (RL). Residual learning converges faster and more stably than original mapping learning. If residual learning and batch normalization are used simultaneously, it will converge faster than original mapping and exhibit better denoising performance, especially helping SGD and Adam achieve better performance results[15].

2.2. Attention Model

The research on network structure has developed in multiple aspects such as depth and breadth. So far, Attention[16] has focused on research in specific fields, and has not conducted much research in the field of imaging. Recently, in combination with Residual technology, various networks have been studied in the field of image related fields, and Attention as a component of the model is being developed. Attention establishes a complementary relationship and demonstrates meaningful results in improving deep learning performance. Multiple benchmark tests such as ImageNet classification, COCO detection, VOC detection, and multiple models such as ResNet, WideResNet, ResNext, and MobileNet have been validated [17].

Firstly, from the structure of the Attention module, convolutional features are extracted from the generalized Signmode state map and element wise product is performed.

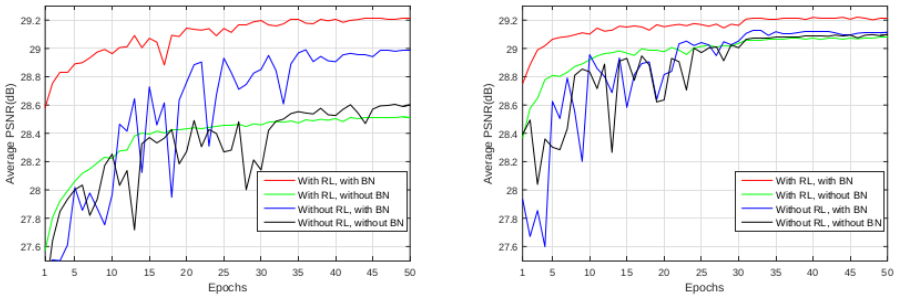


Fig. 2. The average of PSNR with BN and RL

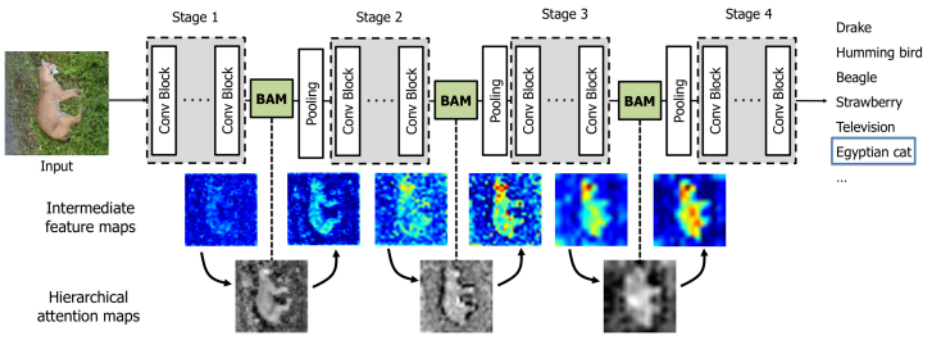


Fig. 3. A general deep learning model with BAM

The performance of Attention plays an important role in achieving greater performance improvement with less computation. The attention map is decomposed into channel wise/spatial wise for calculation. Typical Attention modules include Bottleneck Attention Module (BAM) and Convolutional Block Attention Module (CBAM). These consist of very simple pooling and convolutions. Modularize self tension to easily connect to any CNN. In addition, end-to-end training can be conducted together with existing networks. Fig.3 shows a general deep learning model with added BAM structure. Before the amount of information in the bottleneck interval in the neural network structure is reduced, by adding a BAM structure, can increase the information of important parts while reducing the information of unimportant parts. As a follow-up study, CBAM is a variant of the combination of pooling, spatial, and channel attention, which exhibits higher performance than BAM. The CBAM in Channel attention is different from the previous BAM, as it combines average pool and max pool. The max pool and avg pool values of the 3D feature map are used as meaningful feature maps in global attention. Two pooling features share values with the same meaning, so a shared MLP can be used while reducing the number of parameters. In spatial attention, it is also symmetric, and spatial attention is calculated using a convolution [18].

2.3. U-Net

U-Net [19] combines the concepts of ResNet and Autoencoder, and is a model constructed based on the end-to-end fully convolutional network (FCN) [20] proposed in the Biomedical field for image segmentation. This modifies the FCN structure to provide more accurate segmentation in situations with limited data. As shown in Fig.4, the network used for overall image flow and the network used for precise localization are composed in a symmetrical form.

U-Net has a U-shaped structure, which can be roughly divided into three parts based on the center. The first is to extract the semantic information of image pixels over a large range as the encoding role Contracting Path. The second is to match the semantic information with pixel position information as the decoding role Expansive Path. Finally, there is a conversion interval from the contraction path to the expansion path. This structure eliminates the existing problem of unnecessary repetition of overlapping patches, thereby improving performance. The Contracting Path involves repeating the 3x3 conv layer operation twice, which reduces the size of the feature map. The activation function uses ReLU. In each low sampling process, the size of the feature map will be reduced by half, but the number of channels will increase by twice. At each upsampling, the size of the feature map will double and the number of channels will decrease by half for the Expansive Path. Contrary to Contracting Path, we can use it to expand the size of feature maps. Then, the final layer is processed through a 1x1 conv layer.

3. Proposal Method

Noise may be generated during image generation, transmission, or processing, which can lead to performance degradation during image analysis. In addition, the presence of noise in images can also affect the performance of deep learning models. To improve these issues, this article proposes a Multi-scale Attention U-Net image denoising method. The

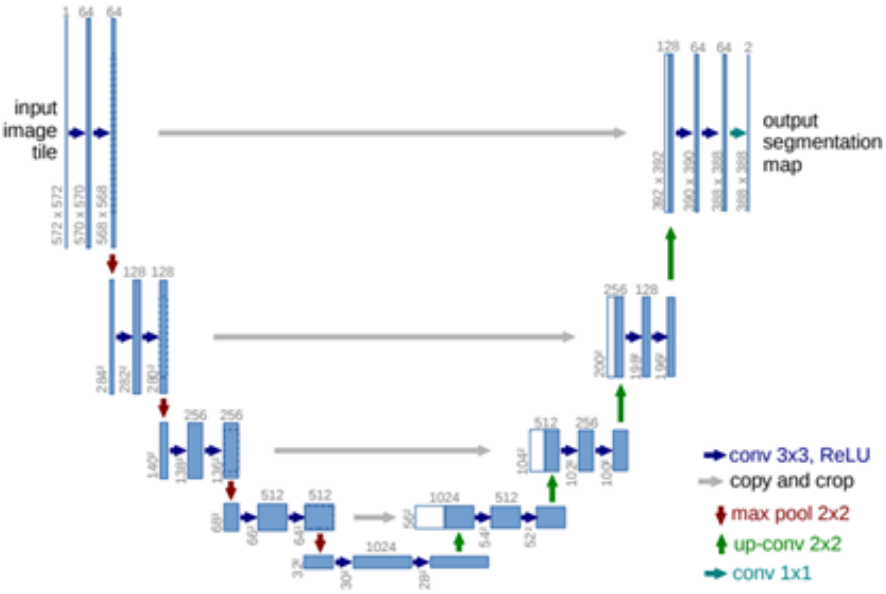


Fig. 4. The structure of U-Net

recommended approach is to follow the process shown in Figure 5 and apply multi-scale attention unit to U-Net based on the typical denoising deep learning model DnCNN, and convert it to an Outburst structure. Using the Attention model to suppress unnecessary information and only emphasize important information. In addition, the second half of the neural network can effectively achieve noise reduction by using the Outburst structure to rapidly increase feature information.

3.1. Improved Attention DnCNN

The performance of deep neural networks depends on depth, but to improve performance, simply increasing depth may reach a performance saturation state. In order to solve this problem, this article did not increase the depth of the neural network, but instead connected two neural networks in parallel, expanded the width of the network, and formed a multi-layer tension module that composed of channel and spatial attention unit. It extracted more scale feature maps and improved the performance saturation state. One of the two networks used for parallel connections is built based on DnCNN. Another approach is to add a Dilated Layer in the front and back half of the middle layer, giving the network scalability and helping to extract more complex feature maps.

The Multi scale Attention unit is configured to execute the Spatial function again after parallel execution of the Channel and Spatial regions and merging their respective outputs. In the field of Channel, what is more important for a given input value is focused on the information about the channel, rather than spatial information. In addition, as shown in Fig. 6, as all nodes are connected using a fully connected layer, the weights will be shared and output more prominent information together with the input value multiple.

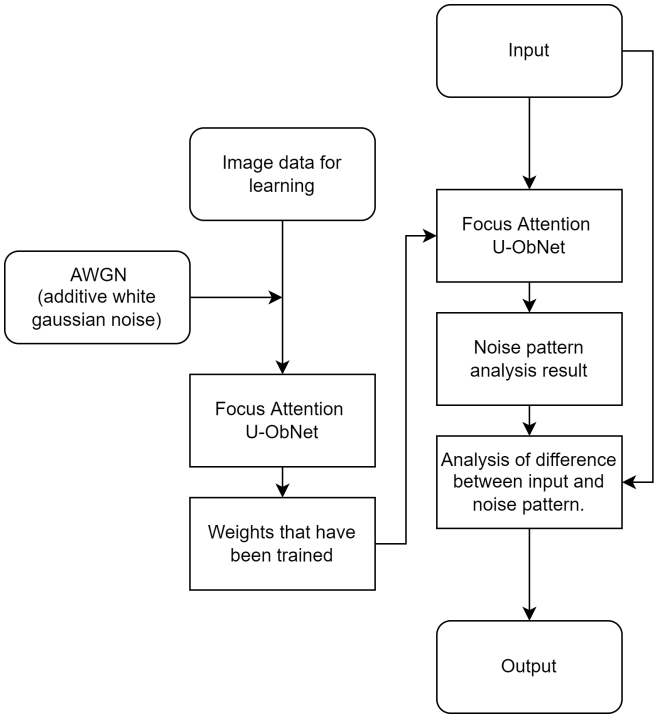


Fig. 5. The flowchart of our proposed method

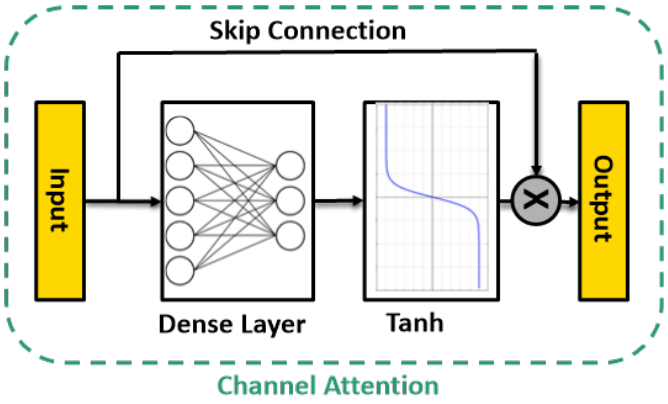


Fig. 6. The structure of Channel Attention

The Spatial region does not consider channel information and only focuses on the spatial importance of input values. As shown in Fig. 7, it is composed by reducing the number of channels and simply continuously arranging convolutional layers. When channels exist, they can cause a sharp increase in parameters during the operation process, which has a significant impact on learning speed. By using a 1x1 conv layer to reduce the computational complexity of initial input values and sharply reduce information about channels, the three-dimensional input is transformed into two-dimensional, thus concentrating only the spatial area. Secondly, after passing through the 1x1 conv layer, only the two-dimensional spatial layer as a channel can continuously apply the 3x3 conv layer, thereby outputting spatial features with only important features as more features.

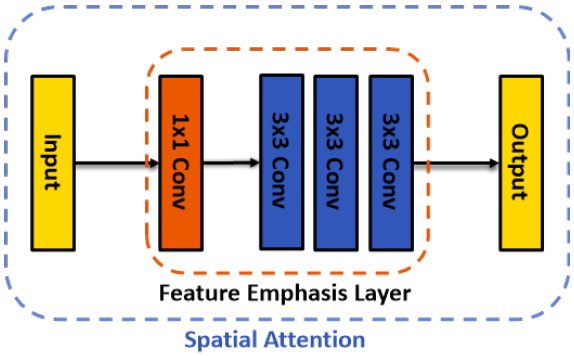


Fig. 7. The structure of Spatial Attention

Finally, merge the output of the Channel region with the output of the Spatial region. By combining channel information and spatial information, important information about channels and spaces can be grasped, and the necessary effective information can be extracted. From a broader perspective, it's like outputting only important and necessary information. Therefore, this output value is used as input in the Spatial module, highlighting spatial and channel elements as a whole, and extracting various important features. The last application of the Spatial module requires the use of both spatial and channel information. Unlike the Spatial module applied earlier, the 1x1conv layer used here serves to delete channel information, therefore the 1x1conv layer is composed in the form of deletion. As shown in Fig. 8, from a larger perspective, the three modules are combined in parallel and serial forms as a multi-scale salient module application.

When two networks(DnCNN and Multi-scale Attention) are connected in parallel, connect the Multi-scale Attention module to the input and output layers. Then, as shown in Fig. 9, the initial model DnCNN proposed in this paper using a Multi-scale Attention module was re-learned through Skip Connection between each network input and output.

In the initial model, it is not simply to increase the depth of the neural network, but to combine two neural network models in parallel, expanding the breadth of the model. In addition, by using ResNet's skip connections, more features can be extracted by referring to the compression function of Autoencoder and previous values. Based on the structure of U-Net and combined with Attention DnCNN, the performance of image denoising meth-

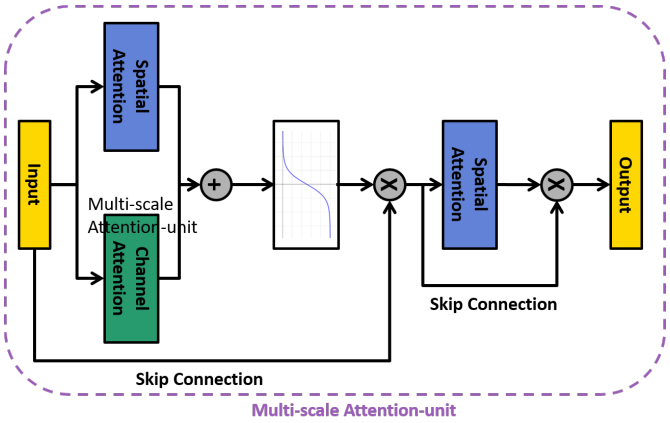


Fig. 8. The structure of Multi-scale Attention

ods can be improved. The existing U-Net adopts the core technologies of Autoencoder and ResNet shortcuts. Moreover, the existing U-Net undergoes four downsampling, but to prevent spatial feature loss at low resolutions, the recommended method only performs two times.

3.2. Outburst Structure

The initial model of this article is constructed in parallel rather than serial form. On the one hand, it adopts the structure of DnCNN, and on the other hand, it uses a dilated layer instead of ordinary convolutional layers to extract feature maps from multiple different levels, while attempting to reduce the number of parameters. Although the performance has been improved, the model also has a flaw.

If layers are added to achieve better performance, the number of parameters will actually increase, leading to a decrease in learning speed and performance. The model developed to compensate for its shortcomings utilizes the structure of U-Net. U-Net can also be seen as a model composed of a combination of the structure of an autoencoder and skip connections. Basically, autoencoders coordinate the spatial size of layers through downsampling and upsampling, reducing the number of parameters required for learning, which is an effective method. Although it is a model of the same length, the size of the kernel decreases and the computational load also decreases. Therefore, the learning speed has also been improved. The decoder has the advantage of having the same output size as the input image, and by extracting features from small-sized images, more feature maps can be obtained, thereby improving learning speed and performance. Every time downsampling is performed, the number of upsampling executions in the decoder region will increase equally. Jumping connections with kernels of the same size can prevent spatial loss caused by downsampling. The improved model is based on the fact that simply passing the original value in skip connections is inefficient. Therefore, for the scalability of the feature map, highlighted information will be passed when connecting the initial model. From the overall structure perspective, this method is also a parallel connection

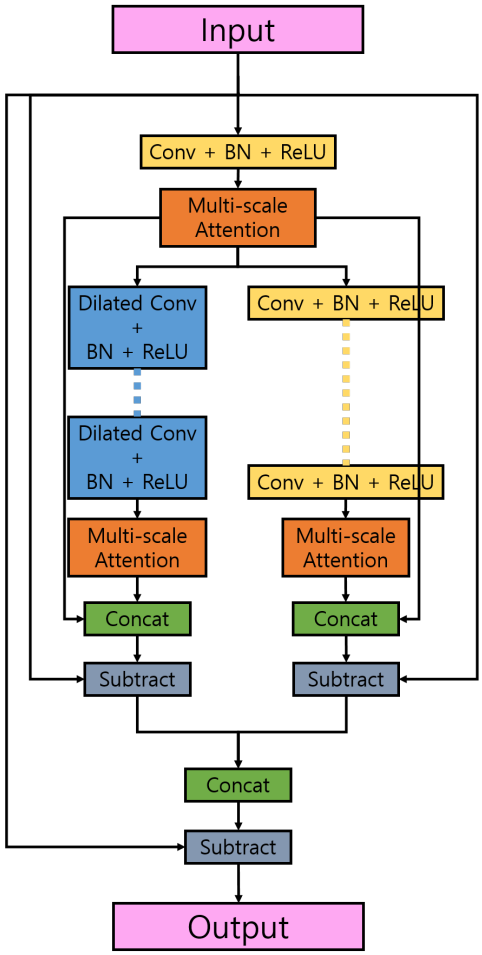


Fig. 9. The structure of Attention DnCNN

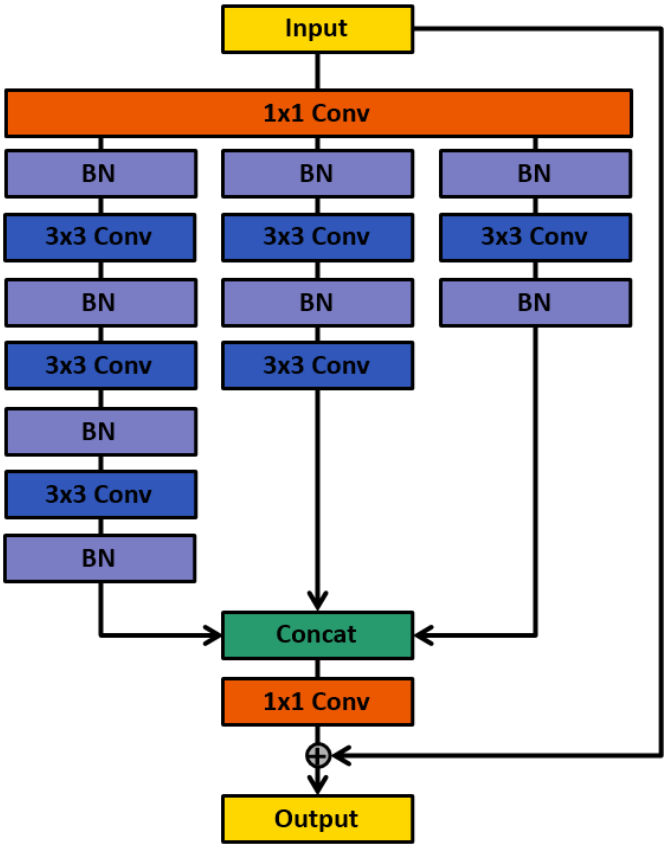


Fig. 10. The outburst structure of our method

rather than a serial connection, thus enhancing the scalability of the network. The performance of pattern analysis in deep learning basically depends on how many different feature maps are extracted. So far, if the model is improved based on parallel connections without increasing the number of parameters, the focus of this article is on how to effectively increase the extracted feature maps. A novel Outburst structure is proposed by explosively adding feature maps using the structure of xception.

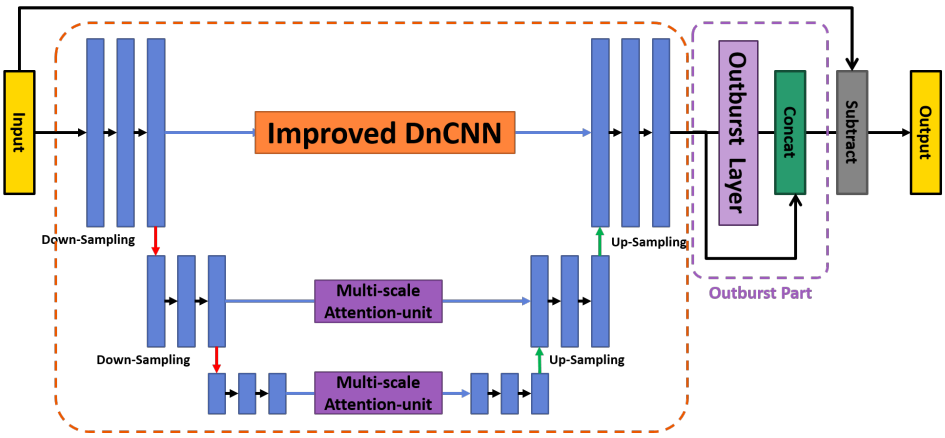


Fig. 11. The structure of our proposed method

The Outburst structure, as the name suggests, refers to an explosion, which refers to a rapid change. The sharp changes also mean the diversity and quantity of feature maps. Outburst can be roughly divided into two parts. If the current model is still part of a calm flow, then a sharp change is made in the latter half by adding the deformation structure of the xception shown in Fig.10. The calm part, also known as the preheating zone, plays a role in consolidating the feature map and extracting many features with sufficient diversity, so it is very important to maintain it well. After the bottleneck structure compression process, it enters the Outburst region and is endowed with diversity by various convolutional layers. Continuous convolutional layers make the features very prominent. The final output end will combine all of these to output. The output value does not remove noise, but rather captures the pattern of noise, indicating its association with the larger framework of DnCNN introduced earlier. Eliminating noise by having a difference in the output of the noise mode through Outburst with the input image containing noise.

3.3. The structure of our proposed method

The suggested methods can be divided into Multi scale Attention unit and Outburst structures, as shown in Fig.11. From the suggested model configuration, the first thing to see is the structural changes. Attention DnCNN uses the calm part as the preheating stage, while the Outburst part is associated with the input of the deformation structure that guides the xception. Effectively extract multiple feature maps from the calm part and endow the feature maps extracted from the Outburst part with scalability. Secondly, the Multi scale

Attention unit is used to assist in extracting various feature maps during the warm-up stage. Compared with traditional Attention modules, the number of layers has increased, allowing for more efficient feature extraction using channel and spatial information.

4. Experimental Results

4.1. Experimental method

In this paper, the learning dataset used 20000 images with dimensions of 180 * 180. The test dataset was tested using grayscale images Set12 and BSD68[21]. The loss function uses Adam, The performance of the epoch remains unchanged after more than 80 attempts, and if executed further, it will deteriorate. Therefore, only 80 attempts were made. For smooth learning, set the learning rate to 1e-3. At the beginning, use a higher learning rate to quickly reduce the value of the loss function before 80 iterations, and then set it to 1e-7 to adjust the safety and details of the loss function.

In order to conduct performance evaluation, experiments were conducted comparing PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index Map) considering human visual image quality differences with other models. PSNR evaluates the quality loss information of generated and compressed videos. The less the loss, the higher the price can be confirmed. If it is a lossless video, its MSE will be 0 and cannot be defined.

$$PSNR = 10log_{10}(\frac{R^2}{MSE}) \tag{1}$$

SSIM (Structural Similarity) is an indicator that measures the similarity between two images. This indicator was first proposed by the Laboratory for Image and Video Engineering at the University of Texas at Austin. Among the two images used by SSIM, one is an uncompressed and undistorted image, and the other is a distorted image.

$$SSIM = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \tag{2}$$

Table 1 shows the results of the comparison between our method and other methods tested on BSD68 dataset. The test results indicate that the PSNR value of our method is better than others.

Tables 2 shows the PSNR and SSIM values tested using our method and other methods on datasets Set12 and BSD68. Our model tested results on the Set 12 dataset are PSNR=30.52/SSIM=0.9315, and on the BSD68 dataset are PSNR=29.43/SSIM=0.9106. The test results indicate that the PSNR and SSIM values tested using our method perform better than other models. From these results, it can be seen that if used for practical noise reduction, universality is feasible.

Table 3 shows the PSNR values of each method tested on 12 types of images in the SET12 dataset. Although it can be seen that the performance of our method is not as good as other models in some class images, it overall shows good performance, with the highest average PSNR value. It can be said that overall, our model showed the better performance than others.

Fig. 12 shows the denoising effect of Monarch, one of the Set12 images, which contains noise. By zooming in on specific parts of the butterfly wing texture, it can be seen that the recommended model has the best level of denoising.

Table 1. The comparison results of our method and other methods

Num	Models	PSNR(dB)
1	BM3D[22]	28.57
2	EPLL[23]	28.68
3	CSF[24]	28.74
4	WNNM[25]	28.83
5	TNRD[26]	28.92
6	IRCNN[27]	29.15
7	FFDNet[28]	29.19
8	ECDNet[29]	29.22
9	DnCNN[12]	29.23
10	ADNet[30]	29.25
11	Ours	29.43

Table 2. The average values of PSNR and SSIM of methods with SET12 and BSD68 datasets

Target	Dataset	BM3D	TNRD	DnCNN	IRCNN	Ours
PSNR	Set12	29.97	30.05	30.44	30.38	30.52
	BSD68	28.57	28.92	29.23	29.15	29.43
SSIM	Set12	0.8505	0.8515	0.8618	0.8601	0.9315
	BSD68	0.8017	0.8148	0.8278	0.8249	0.9106

Table 3. PSNR values of methods with 12 classes of SET12 dataset

Class	BM3D	MLP	TNRD	DnCNN	IRCNN	Ours
C.man	29.45	29.61	29.72	30.18	30.08	30.20
House	32.85	32.56	32.53	33.06	33.06	33.32
Peppers	30.16	30.30	30.57	30.87	30.88	30.96
Starfish	28.56	28.82	29.02	29.41	29.27	29.31
Monarch	29.25	29.61	29.85	30.28	30.09	30.49
Airplane	28.42	28.82	28.88	29.13	29.12	29.15
Parrot	28.93	29.25	29.18	29.43	29.47	29.52
Lena	32.07	32.25	32.00	32.44	32.43	32.54
Barbara	30.71	29.54	29.41	30.00	29.92	30.18
Boat	29.90	29.97	29.91	30.21	30.17	30.23
Man	29.61	29.88	29.87	30.10	30.04	30.13
Couple	29.71	29.73	29.71	30.12	30.08	30.16

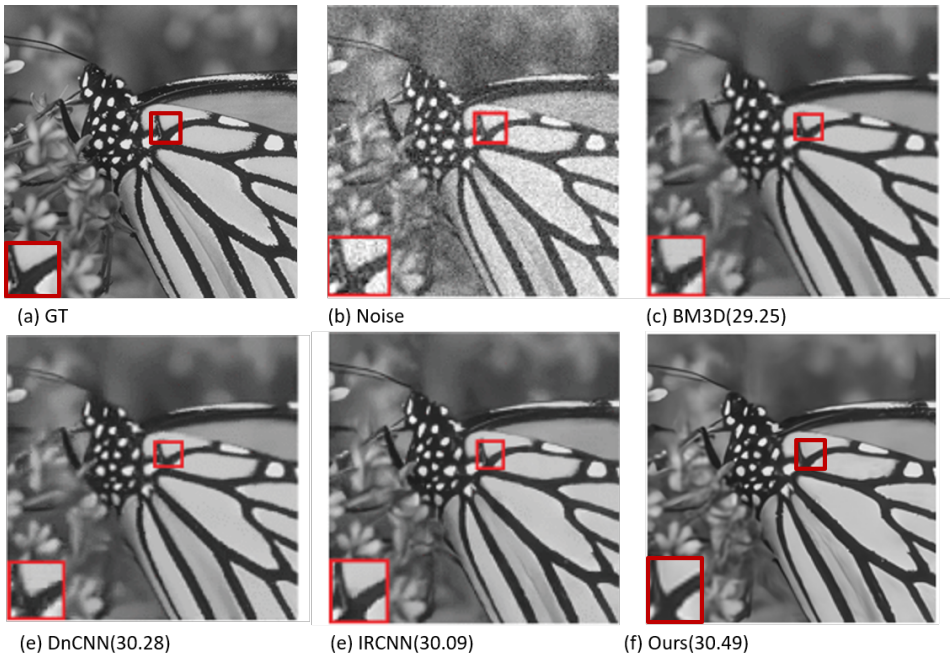


Fig. 12. The results of denoising by models (1)

Fig. 13 and Fig. 14 show the denoising results of one of the images from the BSD68 test dataset by models. After comparing 6 models with the our model, it can be confirmed that our model has the best PSNR values, which are 30.18/38.52, respectively.

5. Conclusions

Noise may be present during image generation, transmission, or processing, which can lead to performance degradation during image analysis. In addition, due to images containing noise, the performance of deep learning models will also decrease during deep learning. To improve these issues, this article proposed a Multi-scale Attention U-Net. The rapproach is to apply Multi-scale Attention unit to denoising network based on DnCNN and convert it into an Outburst structure for image processing. Attention Unit is a new modular approach that can suppress unwanted information and master the function of emphasizing only important information. By analyzing the internal relationships of a given input value, it is divided into two parts: the part that concentrates the important parts and the part that concentrates the important parts through spatial information. The two parts are executed in a parallel structure and then merged. And without adding too many parameters, under the action of Attention unit, more spatial feature maps were generated than other models, not only through PSNR SSIM. The improved performance was also confirmed by removing noisy images. In addition, as the feature part of the overall structure, the latter half of the neural network uses Xception's deformation of the Outburst structure to significantly increase feature information, endowing various feature maps with

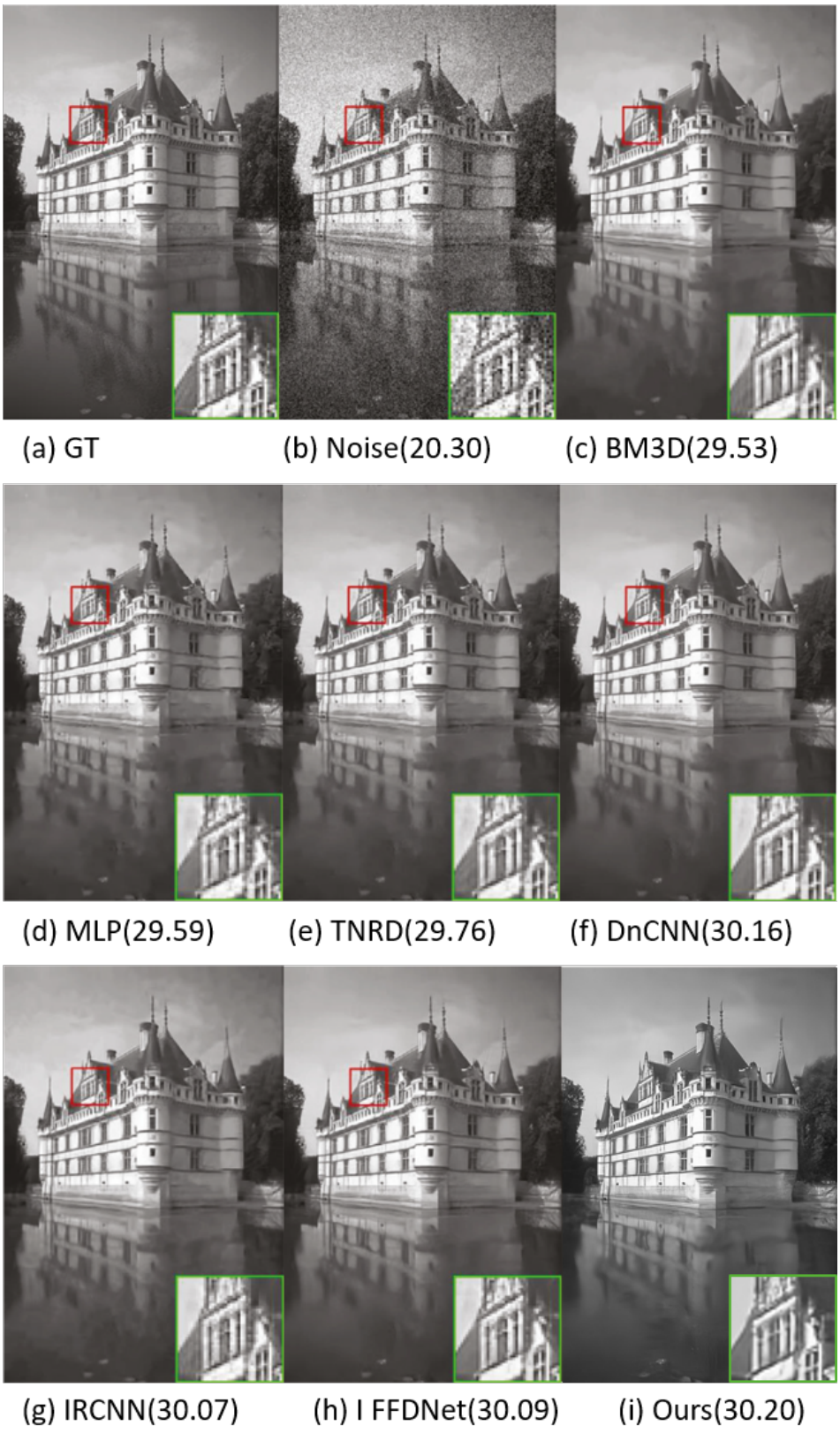


Fig. 13. The results of denoising by models (2)

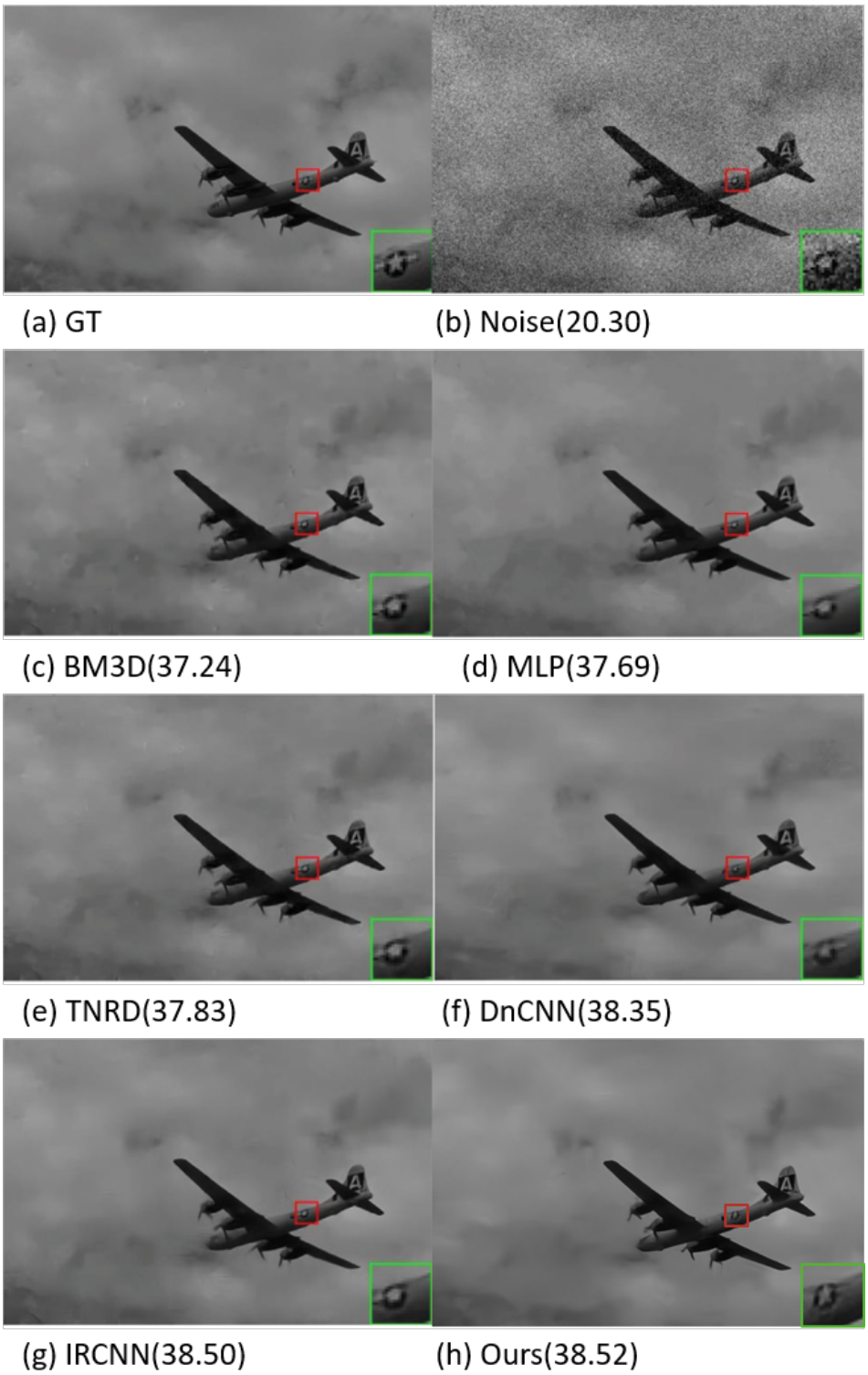


Fig. 14. The results of denoising by models (3)

scalability and demonstrating improvements in noise pattern analysis. The recommended method is to use Set12 and BSD68 test data, compare PSNR and SSIM values with existing models and early versions of the proposed model, and expand specific sections to make their differences more apparent. The experimental results showed that the overall denoising level improved the performance on average compared to other models, especially in terms of numerical results, which confirmed that the model proposed by SSIM for evaluating human visual image quality differences performed the best.

Acknowledgments. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No.RS-2023-00247045)

References

1. F. Emmert-Streib, Z. Yan, H. Feng, S. Tripathi, and M. Dehmer.: An Introductory Review of Deep Learning for Prediction Models With Big Data, Frontiers in Artificial Intelligence, Vol. 3, pp. 4, Feb. 2020.
2. Shanthi, T., and R. S. Sabeenian.: Modified Alexnet architecture for classification of diabetic retinopathy images, Computers and Electrical Engineering, Vol. 76, pp. 56-64, 2019.
3. Majib, Mohammad Shahjahan, et al.: Vgg-scnet: A vgg net-based deep learning framework for brain tumor detection on mri images." IEEE Access 9, pp. 116942-116952, 2021.
4. Tang, Pengjie, Hanli Wang, and Sam Kwong.: G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition, Neurocomputing, Vol. 225, pp. 188-197, 2017.
5. Wightman, Ross, Hugo Touvron, and Hervé Jégou.: Resnet strikes back: An improved training procedure in timm, arXiv preprint arXiv:2110.00476, 2021.
6. Arkin, Ershat, et al.: A survey: object detection methods from CNN to transformer, Multimedia Tools and Applications Vol. 82, No. 14, pp. 21353-21383, 2023.
7. C. Tian, Y. Xu, and W. Zuo.: Image denoising using deep CNN with batch renormalization, ISSN, Vol 121, pp. 461-473, Jan. 2020.
8. J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila.: Noise2Noise: Learning Image Restoration without Clean Data, International Conference on Machine Learning, ICML, Vol. 7, No. 12, pp. 4620-4631, Jan. 2018.
9. Li, Zhuo, Hengyi Li, and Lin Meng.: Model compression for deep neural networks: A survey, Computers, Vol. 12, No. 3, pp. 60, 2023.
10. Yan, Puti, et al.: STDMANet: Spatio-temporal differential multiscale attention network for small moving infrared target detection, IEEE transactions on geoscience and remote sensing, Vol. 61, pp. 1-16.
11. Zhu, Junqi, et al.: Evaluation of deep coal and gas outburst based on RS-GA-BP, Natural Hazards, Vol. 115, No. 3, pp. 2531-2551, 2023.
12. Murali, Vineeth, and P. V. Sudeep.: Image denoising using DnCNN: An exploration study, Advances in Communication Systems and Networks: Select Proceedings of ComNet 2019, pp. 847-859, 2020.
13. Shafiq, Muhammad, and Zhaoquan Gu.: Deep residual learning for image recognition: A survey, Applied Sciences, Vol. 12, No. 18, pp. 8972, 2022.
14. Garbin, Christian, Xingquan Zhu, and Oge Marques.: Dropout vs. batch normalization: an empirical study of their impact to deep learning, Multimedia tools and applications, Vol. 79, No. 19, pp. 12777-12815, 2020.
15. danielseo, "[Computer Vision] DnCNN", gihyun.log, Jun 2021, https://velog.io/@danielseo/Computer-Vision-DnCNN.
16. Niu, Zhaoyang, Guoqiang Zhong, and Hui Yu.: A review on the attention mechanism of deep learning, Neurocomputing, Vol. 452, pp. 48-62, 2021.

17. Ren, Zhongzheng, et al.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.
18. Li, Xiaohui, Haiying Xia, and Lidan Lu.: ECA-CBAM: Classification of Diabetic Retinopathy: Classification of diabetic retinopathy by cross-combined attention mechanism, Proceedings of the 2022 6th international conference on innovation in artificial intelligence, 2022.
19. Siddique, Nahian, et al.: U-net and its variants for medical image segmentation: A review of theory and applications, IEEE Access, Vol. 9, pp. 82031-82057, 2021.
20. Wu, Jizhong, et al.: Fault detection based on fully convolutional networks (FCN), Journal of Marine Science and Engineering, Vol. 9, No. 3, pp. 259, 2021.
21. Yin, Haitao, and Siyuan Ma.: CSformer: Cross-scale features fusion based transformer for image denoising, IEEE Signal Processing Letters, Vol. 29, pp. 1809-1813, 2022.
22. Yahya, Ali Abdullah, et al.: BM3D image denoising algorithm based on an adaptive filtering, Multimedia Tools and Applications, Vol. 79, pp. 20391-20427, 2020.
23. Yu, Qiqiong, et al.: EPLL image denoising with multi-feature dictionaries, Digital Signal Processing, Vol. 137, pp. 104019, 2023.
24. Uetani, Hiroyuki, et al.: A preliminary study of deep learning-based reconstruction specialized for denoising in high-frequency domain: usefulness in high-resolution three-dimensional magnetic resonance cisternography of the cerebellopontine angle, Neuroradiology, Vol.63, pp. 63-71, 2021.
25. Averbuch, Amir, et al.: Cross-boosting of WNNM image denoising method by directional wavelet packets, arXiv preprint arXiv:2206.04431, 2022.
26. Jia, Nan, et al.: Background noise suppression using trainable nonlinear reaction diffusion assisted by robust principal component analysis." Exploration Geophysics, Vol. 51, No. 6, pp. 642-651, 2020.
27. Zou, Xueyan, et al.: Intelligent diagnosis method of bearing fault based on ICEEMDAN and Ghost-IRCNN, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, Vol. 237, No. 13, pp. 3115-3130, 2023.
28. Cao, Like, Jie Ling, and Xiaohui Xiao.: Study on the influence of image noise on monocular feature-based visual slam based on ffdnet, Sensors, Vol. 20, No. 17, pp. 4922, 2020.
29. Qi, Huiqing, Shengli Tan, and Zhichao Li.: Anisotropic weighted total variation feature fusion network for remote sensing image denoising, Remote Sensing, Vol. 14, No. 24, pp. 6300, 2022.
30. Liu, Zhen, et al.: ADNet: Attention-guided deformable convolutional network for high dynamic range imaging." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

Mingshou An received his PhD in Electronics Engineering at Dong-A University. He is a Lecturer at the School of Computer Science and Engineering, Xi'an Technological University. His research interests include issues related to artificial intelligence, pattern recognition, deep learning, and industrial measurement. He is author of a great deal of research studies published at national and international journals, conference proceedings.

XuHang Zhao received his Master's in Computer Science and Engineering at Xi'an Technological University. His research interests include issues related to artificial intelligence, pattern recognition, deep learning, and industrial measurement. He is author of a great deal of research studies published at national and international journals, conference proceedings.

Hye-Youn Lim received her PhD in Electronics Engineering at Dong-A University. She is a Lecturer at the Electronics Engineering, Dong-A University. Her research interests include issues related to artificial intelligence, pattern recognition, and deep learning. She is author of a great deal of research studies published at national and international journals, conference proceedings.

Dae-Seong Kang received his PhD in Electrical Engineering at Texas A&M University. He is a Professor at the Electronics Engineering, Dong-A University. His research interests include issues related to artificial intelligence, pattern recognition, and deep learning. He is author of a great deal of research studies published at national and international journals, conference proceedings.

Received: August 10, 2024; Accepted: July 15, 2025.

Computer Science and Information Systems 22(4):1637–1663 https://doi.org/10.2298/CSIS240929065L

The Intersection of Digital Wellbeing and Collection Exhibition: A Study on the Impact of AR Interactive Display Models on Visitor Experience

Min-Feng Lee[1], Guey-Shya Chen[2], Hui-Chien Chen[2], and Jian-Zhi Chen[3]

[1] Department of Information Management, National Taichung University of Science and Technology, Taiwan
antoniolee@nutc.edu.tw
[2] Institute of Educational Information and Statistics, National Taichung University of Education, Taiwan
grace@mail.ntcu.edu.tw
gigi.baby@yahoo.com.tw
[3] Corporate Synergy Development Center (CSD), Taiwan
t296210@gmail.com

Abstract. This study presents an innovative AR interaction model for small exhibits, integrating physical and virtual display techniques. Combining hand detection sensors and 3D modeling, it allows direct manipulation of virtual objects, enhancing interactivity and immersion. Specifically, the model utilizes Leap Motion for gesture interaction, enabling intuitive and natural user engagement with virtual exhibits. A mixed-method approach assessed its impact on 200 randomly assigned participants. Results show significant improvements in interactivity, immersion, learning effectiveness, and overall satisfaction compared to traditional methods. Quantitative analysis revealed statistically significant differences (p ¡ 0.001) with large effect sizes (Cohen's d ¿ 0.8). Qualitative findings corroborated these results, highlighting increased engagement and deeper understanding. The practical implications of this research extend to museums and cultural heritage display settings, where the AR interactive model offers a scalable and engaging solution for enhancing visitor experiences. The findings further indicate that this AR interactive display model not only enhanced exhibition effectiveness but also positively impacted visitors' digital wellbeing, notably by fostering social interaction and mitigating digital fatigue.

Keywords: Augmented Reality (AR), Interactive Display, Virtual Exhibit Interaction, Visitor Experience, Digital Wellbeing.

1. Introduction

The rapid development of digital technology has made Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR) essential tools for cultural heritage preservation and exhibition. These technologies significantly enhance visitors' immersive experiences and interactivity. Bekele et al. [4] comprehensively reviewed the application of these exhibition technologies in the cultural heritage field, emphasizing their importance in exhibiting and protecting artifacts and cultural heritage. In current exhibition applications, the integration of physical and virtual exhibitions has become a trend in museum exhibitions.

Petrelli et al. [5] pointed out that this combination can provide diverse experiences, meeting the needs of different visitors. Additionally, researchers such as Eghbal-Azar et al. [6] found that digital guide systems have significant advantages in information selection and delivery, further promoting this type of integration.

Immersive experiences and interactivity are at the core of modern exhibition technologies. Younan and Treadaway [7] emphasized the role of 3D digital models in enhancing interactivity and immersion, while Jin et al. [8] confirmed the significant impact of VR and AR technologies on improving visitor experiences. Other related studies have focused on small exhibit exhibitions. Scopigno et al. [9] and Balletti et al. [10] explored the application and integration of digital manufacturing and 3D printing technologies in exhibitions. Recent research [11] has also focused on the potential of gesture recognition technology in cultural heritage exhibitions.

Based on these developments, this study explores a novel interactive exhibition mode that integrates physical and virtual interactive exhibitions within the same showcase. Through hand detection sensors, visitors can directly manipulate virtual exhibits, which is particularly suitable for small artifacts. By utilizing 3D modeling technology to provide rich interactive experiences, this research aims to promote innovation in exhibition technology and provide theoretical and design foundations for future exhibition models. Moreover, this study investigates how AR interactive display technology simultaneously enhances exhibition effectiveness and promotes visitors' digital wellbeing, offering a novel perspective on the role of exhibition spaces in the digital era.

The contributions of this study are threefold:

Development of an Innovative AR Interactive Display Model: This study introduces a novel AR-based framework that seamlessly integrates physical artifacts with virtual interaction, using gesture recognition technology to enhance user engagement, particularly for small-scale exhibits.

Empirical Evidence of AR's Impact on Visitor Experience: Through a mixed-method approach, the study provides comprehensive data demonstrating AR's effectiveness in improving interactivity, immersion, cognitive engagement, and digital wellbeing compared to traditional exhibition methods.

Advancement of Digital Literacy and Inclusive Learning: The research highlights AR's role in promoting digital literacy and fostering inclusive learning environments, emphasizing its potential to support diverse learning styles and enhance accessibility for a broad range of audiences.

1.1. Research Objectives

With the rapid development of digital technology, the application of exhibition technologies in museums and the cultural heritage field has gained increasing attention, especially with the rise of AR, VR, and MR technologies, which have brought significant changes to exhibition methods and visitor experiences [4,5]. The primary objective of this research is to design and develop a novel interactive exhibition mode that integrates physical exhibits with virtual interactive exhibitions within the same showcase. By utilizing hand detection sensors and 3D modeling technology, visitors can directly manipulate virtual exhibits in front of the showcase, such as zooming in, zooming out, and rotating, to achieve a more detailed observation experience.

The study will particularly focus on the following specific objectives:

Design and development of an integrated physical and virtual exhibition model: Based on current trends in exhibition technology, this research will explore how innovative hand detection technology can enhance visitor interactivity and immersion, enabling direct manipulation of virtual exhibits. This design will build upon previous research [6,7], further optimizing and innovating to meet the specific needs of small exhibit displays.

Evaluation of the model's impact on visitor experience: Through experimental design and comparative studies, this research will systematically assess the impact of the exhibition model on visitor interactivity, immersion, and satisfaction. Previous studies have shown that AR and VR technologies can significantly enhance visitor experiences, particularly in terms of immersion and engagement [8,9]. Therefore, this study will employ surveys, interviews, and observations to collect data and conduct detailed analyses to validate the effectiveness of this model.

Technical implementation and challenge analysis: This research will also conduct an in-depth analysis of the technical implementation process of the exhibition model, including the application of projection technology, the practicality of hand detection, and the challenges of system integration [12,13]. The successful application of these technologies will provide valuable insights for future innovation and development in exhibition technology.

The ultimate aim of this research is to design and develop this novel interactive exhibition model, thereby advancing the innovative application of exhibition technology in museums and cultural heritage and providing a theoretical foundation and practical applications for future exhibition models.

2. Related Literature

With the rapid development of digital exhibition technologies, particularly in the fields of museum and cultural heritage exhibitions, the application of Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR) technologies has become increasingly widespread. These technologies have significantly altered the way visitors experience exhibitions and have brought profound changes to the presentation of exhibition content and methods of interaction. For instance, research indicates that the application of AR technology in museum artifact exhibitions can significantly enhance visitor engagement and learning outcomes, while VR technology has improved visitors' depth of perception and sense of involvement in immersive exhibitions [14,15].

In recent years, the development of Augmented Reality (AR) and Virtual Reality (VR) technologies in education has significantly influenced immersive learning experiences. According to the research "Analyzing augmented reality (AR) and virtual reality (VR) recent development in education." [1], AR and VR enhance cognitive engagement, support personalized learning environments, and foster higher motivation and interaction among learners. This comprehensive review highlights the exponential growth of AR and VR applications in educational contexts, emphasizing their potential to bridge the gap between theoretical knowledge and real-world application. These findings align with the objectives of this study, which explores the integration of AR technologies to improve visitor experiences in exhibition environments.

In addition to AR, VR, and MR, recent advancements in Ambient Intelligence have further expanded the potential of interactive exhibition technologies. Ambient Intelligence

refers to environments enriched with embedded sensors, networks, and intelligent systems that proactively support user activities [3]. By integrating AR technologies with Ambient Intelligence, exhibitions can achieve adaptive, context-aware interactions that respond dynamically to visitors' behaviors and environmental conditions. This integration enhances both the interactivity and personalization of exhibitions, providing visitors with immersive and engaging experiences tailored to their preferences.

Furthermore, the integrated application of hand detection technology and 3D modeling technology has also gained widespread attention in exhibition technology. Hand detection technology, by providing intuitive and natural interaction methods, has effectively enhanced visitors' sense of engagement [16]. Concurrently, the development of 3D modeling technology has enabled virtual exhibits to be presented more realistically, improving the authenticity and interactivity of exhibition content [17]. The advancements in these technologies have not only enhanced the interactivity of exhibitions but also significantly increased visitors' sense of immersion. Research shows that interactive exhibition technology can enhance visitors' immersive experience by incorporating elements such as gesture control, while combining AR with hand detection technology can create a more immersive exhibition environment, further improving user satisfaction [18]. Additionally, the integration of Ambient Intelligence concepts supports the creation of intelligent, responsive environments that adapt in real-time to optimize visitor engagement and learning outcomes. These research findings provide important academic background for the design of the exhibition interaction model in this study and lay the technological foundation.

2.1. Current Exhibition Technologies and Interaction Methods

In recent years, exhibition technologies in the fields of museums and cultural heritage have seen continuous development, especially with the advancements in Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR). These technologies have significantly transformed exhibition methods and visitor interaction experiences. They provide powerful tools to enhance the interactivity and immersion of exhibition content, turning exhibitions from static displays into dynamic and immersive experiences.

AR and VR technologies have been widely applied in cultural heritage and museum exhibitions, allowing for the virtual recreation of historical scenes or providing additional layers of information. For instance, Blanco-Pons et al. studied how AR and VR technologies can enhance cultural heritage experiences through virtual reconstructions, finding that these technologies significantly boost visitor engagement and learning outcomes [14]. De Paolis et al. explored the usability of VR in cultural heritage from a user perspective and found that VR offers highly immersive experiences, enhancing visitor satisfaction and depth of understanding [19].

Gesture recognition technology has also emerged as a significant highlight in exhibition technologies in recent years. Zerrouki et al. (2024) demonstrated that gesture recognition technology can significantly improve the interactivity of virtual museums, allowing visitors to interact with exhibits in a natural way, thereby enhancing their overall experience [16]. Moreover, Kyriakou et al. explored the combination of gesture recognition and AR technology, finding that this combination can further enhance user experience in virtual museums, particularly in improving immersion and operational convenience [21].

The application of 3D modeling technology is also becoming increasingly widespread, especially in the construction of virtual museums. Carvajal et al. studied the application

of 3D modeling technology in virtual museums and pointed out that this technology can accurately digitize physical exhibits, providing richer exhibition content and higher interactivity [17]. These technological advancements provide crucial technical support for the innovative design of exhibition models.

In recent years, the importance of digital wellbeing in exhibition spaces has become increasingly prominent. Studies have indicated that appropriately designed digital interactions can enhance visitors' learning experiences while mitigating digital fatigue [22]. A study conducted on 279 participants to assess the psychological impact of AR museum experiences on visitors revealed that such experiences contribute to improved attention restoration levels, stress reduction, and anxiety alleviation [23]. This research further explores the potential of AR technology in this context.

In summary, the development of current exhibition technologies and interaction methods has not only greatly enhanced visitor immersion and engagement but also laid a solid foundation for the innovative application of exhibition technologies in the future. This study will build on these technological trends to further explore how to effectively integrate these technologies to design more interactive and immersive exhibition models.

2.2. Visitor Experience Research

In exhibition technology, research on visitor experience has always been a crucial topic. In recent years, with the popularization of technologies such as Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR), the methods of museum and cultural heritage exhibitions have undergone significant changes, profoundly impacting visitor experiences. Research shows that these immersive technologies not only significantly enhance visitors' sense of immersion but also improve their learning outcomes and interactive engagement.

The research of Blanco-Pons demonstrated that AR and VR technologies can effectively enhance the experience of cultural heritage exhibitions, particularly in strengthening audience engagement and depth of understanding [14]. Furthermore, The research of Hulusic explored the application of VR technology in cultural heritage, finding that visitors can achieve a higher sense of immersion in virtual environments, which not only increased visitor satisfaction but also enhanced their understanding of exhibits [15].

Regarding gesture interaction technology, Huang et al. investigated the application of gesture recognition technology in museum exhibitions. Their research showed that this technology can provide more natural and intuitive interaction methods, significantly improving visitor experiences [16]. The new research about the further explored the combination of gesture recognition and AR technology, discovering that this integration not only enhanced the immersion of virtual museums but also improved user operational convenience [20].

Moreover, the application of 3D modeling technology in virtual museums has also significantly improved visitor experiences. Carvajal et al. studied the application of 3D modeling technology in virtual exhibitions, pointing out that this technology can digitize physical exhibits, providing more realistic and interactive exhibition content, thereby enhancing the overall visitor experience [17].

These studies indicate that through the application of advanced technologies, exhibition methods are no longer limited to static viewing but have become dynamic, highly interactive, and immersive experiences. The application of these technologies provides

important theoretical basis and practical foundation for future innovation in exhibition models.

3. Innovative Exhibition Model Design

This study proposes a novel exhibition model aimed at integrating physical exhibitions with virtual interactive displays, with a particular focus on optimizing the presentation of small exhibits, especially those with details that are difficult to observe with the naked eye. The exhibition model proposed in this research utilizes hand detection sensors and 3D modeling technology, allowing visitors to directly manipulate virtual exhibits, such as zooming in, zooming out, and rotating them, to achieve more detailed observation. The application of this technology not only enhances the interactivity of the exhibition but also improves visitor immersion and satisfaction. The display cabinet designed in this study also incorporates advanced projection technology, ensuring that virtual content is precisely projected onto transparent glass while avoiding issues with glaring light, further enhancing the exhibition's effectiveness. The innovative exhibition model in this study aims to address the shortcomings of existing exhibition technologies, creating a more engaging and educational experience for visitors by combining virtual and physical exhibition methods.

3.1. Visitor Experience Research

The innovative exhibition model proposed in this study aims to break the limitations of traditional exhibition methods by integrating physical displays with virtual interactive technology, providing visitors with a completely new interactive viewing experience. The core concept of this model is to achieve seamless integration of physical and virtual exhibits within the same display case, creating an exhibition environment that is both realistic and highly interactive.

According to research by Dieck et al. [24], mixed reality technology has enormous potential in cultural heritage exhibitions, capable of significantly enhancing visitor engagement and depth of understanding. Based on this concept, the model in this study employs advanced projection technology to accurately present virtual content on the front glass of the display case. This approach not only preserves the authenticity of physical exhibits but also enriches the exhibition content through virtual elements, echoing the idea of enhancing cultural heritage accessibility proposed by Paladini et al. [25].

Another innovation of this model is the introduction of hand detection sensors, allowing visitors to directly manipulate virtual exhibits through gestures. The design of this interaction method was inspired by the research of Trajkova. [26], who found that natural gesture interaction can significantly enhance museum visiting experiences. Through this method, visitors can easily zoom in, zoom out, and rotate virtual exhibits, achieving in-depth observation of small or intricate exhibits, which is difficult to achieve in traditional exhibition methods.

To better illustrate the transparent projection interactive mode of this study, please refer to Figure 1 and Figure 2.

Integration of Transparent Projection Interactive Mode:

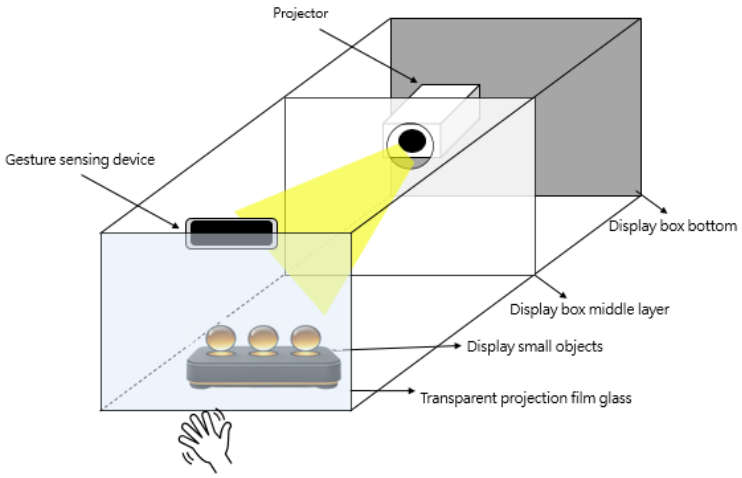


Fig. 1. The Transparent Interactive Projection Framework in this study

(a) Single-hand gesture interface

(b) Dual-hand gesture interface

Fig. 2. Actual operation interfaces of single-hand and dual-hand gesture sensing

To achieve this goal, this study adopts the transparent projection interactive mode as the core technology, as shown in Figure 1. The transparent projection interactive mode is an emerging exhibition technology that cleverly combines physical displays with virtual interactions. By projecting high-resolution images onto transparent screens or glass, visitors can view virtual information and 3D models related to the exhibits through the transparent screen while observing the physical exhibits. This combination of virtual and real exhibition methods not only enriches the exhibition content but also enhances the interactivity and sense of engagement for visitors.

Application of Hand Detection Sensors:

To further enhance interactivity, this study introduces hand detection sensors based on the transparent projection interactive mode, as shown in Figure 2. Visitors can directly interact with virtual content through gestures, such as zooming in, zooming out, rotating virtual exhibits, and even triggering specific animations or information. This intuitive interaction method allows visitors to no longer be passive recipients of information, but to actively explore, creating a more personalized and attractive viewing experience.

Furthermore, the research mode pays particular attention to addressing the glare problem that projection technology may cause. Referring to the research on museum lighting design by Lee. [27], this study employs special projection technology and materials to ensure that virtual content is clearly visible without negatively impacting visitors' visual comfort.

This method of integrating physical and virtual exhibitions is not only suitable for small exhibits but can also be extended to various types of cultural relic exhibitions. As Hauser [28] pointed out, this mixed exhibition method can provide visitors with multi-layered information and a deeper cultural experience. By combining 3D modeling technology with physical exhibits, the mode in this research can present different details of the exhibits and a 360-degree all-around viewing perspective, thereby greatly enriching the educational value of the exhibition.

Overall, this innovative exhibition mode successfully combines physical exhibits with virtual interaction organically by integrating various advanced technologies, opening up new possibilities for exhibitions and cultural heritage displays. It not only enhances the interactivity and attractiveness of the exhibition but also provides visitors with a more in-depth and comprehensive way of learning and exploration. In addition to the aforementioned aspects, this study also investigates the impact of AR interactive display models on visitors' digital wellbeing, encompassing dimensions such as social interaction, learning experiences, and comfort levels with technology use. This research objective aligns with the growing significance of digital wellbeing in exhibition spaces and aims to understand how innovative display technologies can contribute to visitors' overall digital health and experience.

3.2. Technical Implementation Details

The technical implementation of this innovative exhibition model is based on the integration and application of various advanced technologies, with the aim of enhancing the interactivity of the exhibition and the immersive experience of visitors. The following are the detailed implementation specifics for each technology.

Hand detection technology plays a central role in this exhibition model, allowing visitors to manipulate virtual exhibits within the display cabinet through gestures. This technology is based on optical gesture recognition, utilizing a depth camera to capture visitors' hand movements and instantly converting them into control commands. The application of this technology has been proven to significantly enhance interactivity and user experience, particularly in interactive exhibition environments such as museums [29]. Moreover, recent studies have shown that the combination of hand detection technology with AR can further enhance user immersion and operational convenience in virtual museums [16].

In this study, Leap Motion was chosen as the hand detection sensor. Leap Motion is a high-precision, low-latency hand tracking device that can capture subtle movements of fingers and palms, providing highly accurate gesture recognition. Through Leap Motion,

visitors can interact with virtual exhibits in a natural and intuitive way, performing actions such as zooming in, zooming out, and rotating, as if manipulating real objects. The advantage of Leap Motion lies in its high precision and low-latency gesture recognition capabilities, offering a smooth and natural interactive experience. Additionally, Leap Motion is compact and easy to integrate into the display cabinet, without affecting the overall exhibition effect. The application of Leap Motion in this study's exhibition model enables more precise and intuitive gesture interaction, providing visitors with a richer and more immersive interactive experience.

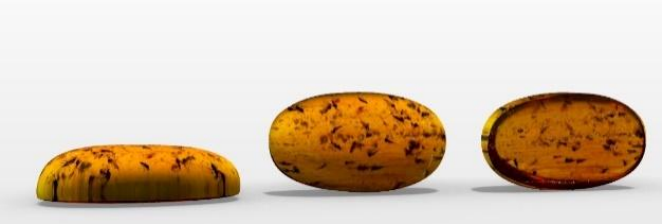


Fig. 3. 3D modeling image of an exhibit in this study

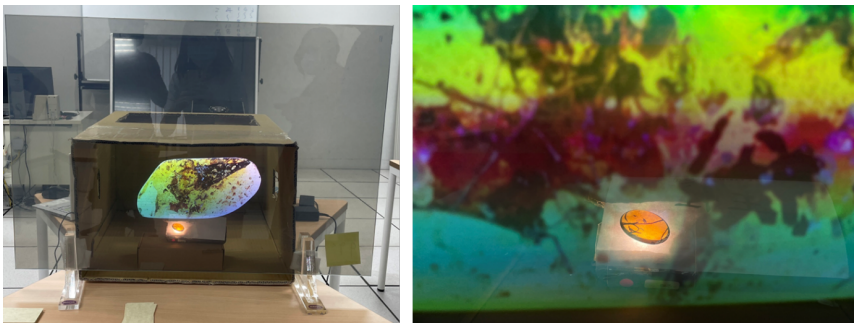


Fig. 4. Testing the transparent projection film (ANSLCF) in the laboratory for image projection and film transparency

This study applies 3D modeling technology to achieve virtual display of small exhibits. Research shows that 3D modeling technology can significantly enhance the realism and interactivity of exhibition content [30], thereby enhancing visitors' immersive experience. This study uses Blender software to model amber exhibits, improving the accuracy and realism of the models through a series of intricate processes. Figure 3 shows a 3D modeling image of one of the exhibits. This process includes high-resolution image acquisition, image processing, photogrammetry, mesh reconstruction, UV unwrapping and texturing, material setup, lighting and rendering, and final optimization. Professional photography equipment is used to capture high-resolution photos, Adobe Photoshop is used for image preprocessing, Blender's photogrammetry plugin is utilized to generate initial

3D point cloud models, and then precise polygon models are created through mesh reconstruction. The UV unwrapping and material setup stages simulate the translucent properties and internal details of amber, while lighting and rendering ensure optimal presentation of the model in various environments. The final optimization step ensures smooth operation of the model in real-time interactive environments. This process successfully creates highly realistic and interactive virtual amber exhibits, not only accurately replicating the appearance of physical amber but also presenting its unique internal structure and optical properties. This allows visitors to explore these small exhibits in unprecedented ways, greatly enhancing the educational value of the exhibition and the experience of detailed observation.

The interactive transparent projection mode in this study employs advanced projection technology to precisely project virtual content onto the transparent glass of the display case. To avoid the glare often associated with traditional projection, the study uses high-contrast projection equipment combined with special light-transmitting materials. This ensures that the projected content is clearly visible in various lighting conditions without compromising visitor comfort. This technology has already been successfully implemented in museum lighting design, demonstrating its potential for enhancing exhibition effectiveness [31].

Specifically, the transparent projection is achieved using the high-transparency projection film (Anisotropic Nano-Structure Light Control Film, ANSLCF) from BENQ MATERIALS CORPORATION. This special projection film combines polymers, liquid crystals, nanoparticles, and optical design. It features high transparency and flexibility, capable of displaying images with up to 8K resolution without being constrained by screen size, allowing for customized designs based on exhibition space requirements. Another significant advantage of the ANSLCF transparent projection film is its 80% light transmission rate, which enables the projection content to display vivid colors without causing glare, while maintaining a clear view through the film. This technology effectively avoids common visual hotspots and light pollution issues seen in traditional projection technologies, offering an ultra-wide viewing angle that provides a consistent viewing experience from different perspectives. Additionally, the film's flexible material makes installation more convenient, allowing for versatile applications in various complex exhibition environments. Figure 4 shows the projection and see-through effects of the high-transparency projection film (ANSLCF) tested in the laboratory. The image demonstrates both the projection on the ANSLCF and the ability to view exhibition objects behind the film through it.

The transparent projection interactive display system designed in this study can dynamically switch based on whether there are visitors in front of the display cabinet, providing an intuitive and interactive exhibition experience. As shown in Figure 5, when no one is present, the display cabinet appears in the same state as a traditional exhibition, directly showcasing the physical exhibits. However, as soon as a visitor approaches the display cabinet, as illustrated in Figure 6, the system immediately activates, displaying a 3D model of the exhibit and related information on the transparent projection screen, thus enriching the exhibition content. The model designed in this study not only captures the attention of visitors but also provides more opportunities for in-depth understanding of the exhibits. Furthermore, as shown in Figure 7, the system supports gesture interaction, allowing visitors to manipulate the virtual 3D exhibits using one or both hands. This in-

Fig. 5. Transparent Projection Interactive Display – Only Showing Regular Exhibits When No Visitors Are Present

teractive method enables visitors to freely zoom in, rotate, or move the 3D amber model for detailed observation, truly allowing them to ”explore” every detail of the exhibit, significantly enhancing the interactivity and depth of the visitor experience.

4. Innovative Exhibition Model Design

This research employed a mixed-method approach, combining questionnaires, interviews, and behavioral observations to comprehensively evaluate visitor experiences. The questionnaire design was based on the Immersive Experience Questionnaire (IEQ) [33], assessing aspects such as immersion, presence, engagement, and enjoyment. Interviews were conducted to explore visitors' subjective feelings and opinions in depth. The experiment was divided into two groups: one experiencing the innovative display mode and the other experiencing the traditional display mode. After data collection, statistical analysis methods were used in the study to compare differences between the two groups and analyze questionnaire and interview data to gain a deeper understanding of visitor experiences.

4.1. Research Design

This study adopted a mixed research methodology, combining quantitative and qualitative analyses to comprehensively assess the impact of the innovative exhibition model on visitor experience. The research design was based on the approach used by Hammady et

Fig. 6. Transparent Projection Interactive Display – Visitor in Front of the Display Cabinet

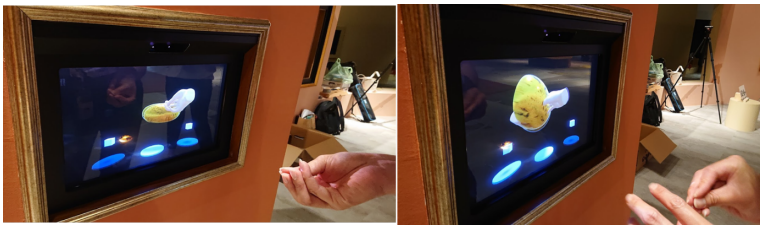


Fig. 7. Visitors can use single-hand or double-hand gestures to operate the 3D amber image in the transparent projection interactive system

al. [33] in evaluating AR applications in cultural heritage and was appropriately adjusted to suit the characteristics of this study.

The study employed a quasi-experimental design, randomly dividing participants into two groups: the experimental group (experiencing the innovative display mode) and the control group (experiencing the traditional display mode). Both groups viewed the same exhibits but presented in different ways. The experimental group used an innovative display case that integrated physical exhibits with virtual interaction, while the control group used standard glass display cases. This design was inspired by the method used by Trunfio et al. [34] in evaluating the impact of mixed reality technology on museum visitor experiences.

This study recruited a total of 200 participants, randomly divided into experimental and control groups, with 100 people in each group. These participants ranged in age from 18 to 45 years old, encompassing students, professionals, and the general public, ensuring diversity and representativeness of the sample. The following description of the data is presented as shown in Figure 8.

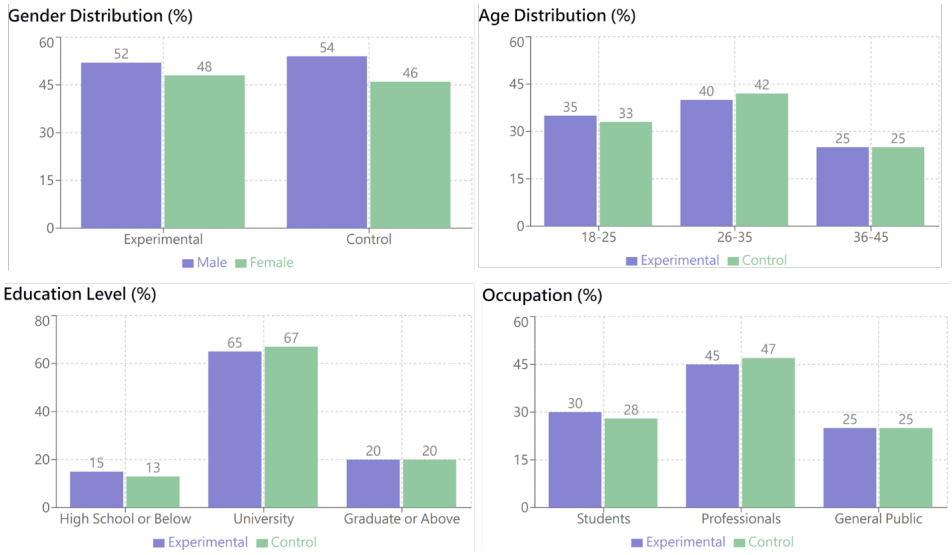


Fig. 8. Demographic information of participants in this study

Regarding participant demographics, the experimental group had a gender ratio of 52% male to 48% female. The age distribution was 35% for 18-25 years old, 40% for 26-35 years old, and 25% for 36-45 years old. In terms of education, 15% had high school education or below, 65% had university degrees, and 20% had graduate degrees or above. Occupation-wise, 30% were students, 45% were professionals, and 25% were from the general public.

The control group had a gender ratio of 54% male to 46% female. The age distribution was 33% for 18-25 years old, 42% for 26-35 years old, and 25% for 36-45 years old. Regarding education, 13% had high school education or below, 67% had university degrees,

and 20% had graduate degrees or above. In terms of occupation, 28% were students, 47% were professionals, and 25% were from the general public.

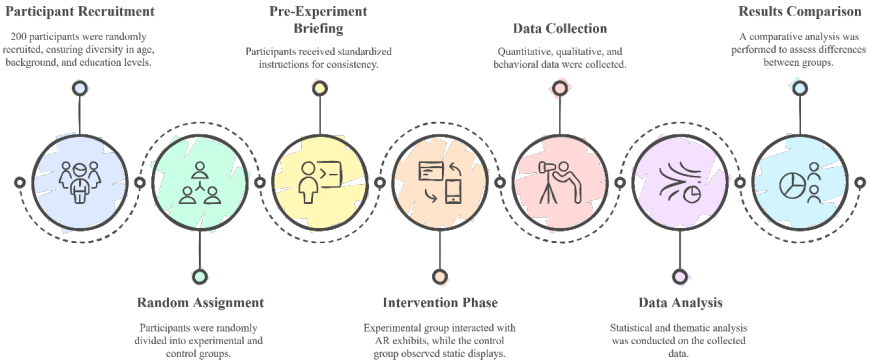


Fig. 9. Experimental design flowchart

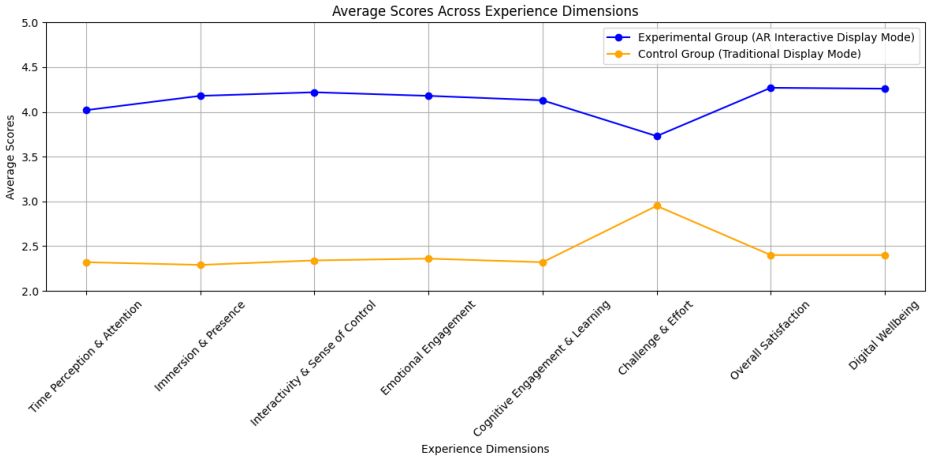


Fig. 10. Mean scores by dimension for the experimental and control groups

By ensuring similarity between the two groups in terms of gender, age, education, and occupational distribution, the research design aimed to reduce potential confounding

variables and enhance the comparability and general applicability of the experimental results.

For the questionnaire survey, the study used a modified version of the Immersive Experience Questionnaire (IEQ) [35], assessing participants across 7 dimensions: 1. Time perception and attention, 2. Immersion and presence, 3. Interactivity and sense of control, 4. Emotional engagement, 5. Cognitive engagement and learning, 6. Challenge and effort, 7. Overall satisfaction and willingness to recommend, and 8. Digital Wellbeing. Overall satisfaction and willingness to recommend. This questionnaire included 35 items, evaluated using a 5-point Likert scale.

Regarding semi-structured interviews, 20 participants from each group were selected for in-depth interviews to explore their subjective experiences and opinions. The design of the interview guide was based on research of Vongkusolkit et al. [26].

In addition to traditional data collection methods, this study integrated a cloud-based data collection framework inspired by the work of Alamri et al. (2013). Real-time behavioral data were captured through sensors embedded in the AR interactive display, including gesture interactions, viewing duration, and navigation paths. This data was securely transmitted to a cloud platform for storage and preliminary analysis, enabling efficient, multi-level behavioral tracking and ensuring data integrity. The cloud-based approach facilitated comprehensive interaction analysis, allowing for the identification of patterns in visitor engagement and behavior that were not easily observable through traditional methods.

This study employed both quantitative and qualitative analyses. For quantitative analysis, the statistical methods used included descriptive statistics, independent samples t-tests, and one-way Analysis of Variance (ANOVA) to compare differences between the two groups across various dimensions. For qualitative analysis, thematic analysis was conducted on the interview data from the subjects, using qualitative analysis software to assist with coding and theme extraction. The analysis methods were based on the research of Clini et al. [36].

To enhance the clarity of the experimental design and improve the transparency of the quasi-experimental methodology, a flowchart has been included to illustrate the participant flow and the key steps involved in the experimental and control groups.

Participant Recruitment:

Random Assignment:

Participants were randomly divided into two groups:

Experimental Group (n=100): Exposed to the AR interactive display model with Leap Motion gesture interaction.

Control Group (n=100): Experienced the traditional static display without AR features.

Pre-Experiment Briefing:

Participants received standardized instructions to ensure consistency in understanding the experimental procedure.

Intervention Phase:

Experimental Group: Interacted with the AR-enhanced exhibits using gesture-based controls.

Control Group: Observed traditional exhibits without interactive features.

Data Collection:

Quantitative Data: Collected through the modified Immersive Experience Questionnaire (IEQ).

Qualitative Data: Gathered via semi-structured interviews and observational notes.

Behavioral Data: Captured in real-time through cloud-based sensors for the experimental group.

Data Analysis:

Statistical analysis (t-tests, ANOVA) and thematic analysis for qualitative feedback.

Results Comparison:

Comparative analysis between the experimental and control groups to assess differences in interactivity, immersion, learning effectiveness, and digital wellbeing.

5. Research Analysis

This study employed a mixed research method, combining surveys, interviews, and behavioral observations to comprehensively assess the impact of the innovative exhibition model on visitor experience. The data analysis results showed that the innovative exhibition model outperformed the traditional model across all experience dimensions, with visitors exhibiting significant improvements in interactivity, immersion, and satisfaction.

The questionnaire survey used a modified version of the Immersive Experience Questionnaire (IEQ) to assess participants' experiences across 7 main dimensions. The following sections will explain the results of the experimental group, the control group, and a comparison between the two groups.

5.1. Experimental Group Results

The questionnaire results for the experimental group showed that the average scores for all dimensions were above 4 points (out of a maximum of 5 points), indicating that participants had a high overall evaluation of the innovative display mode. Among these, the dimensions of "Interactivity and Sense of Control" and "Emotional Engagement" scored the highest (both at 4.32), demonstrating that the innovative model performed exceptionally well in these two aspects. On the other hand, the "Challenge and Effort Level" dimension scored relatively lower (2.92), possibly due to the inclusion of reverse-scored items, or indicating that the content of the display was of moderate difficulty for most participants. The overall internal consistency of the experimental group's questionnaire was good (Cronbach's = 0.91), and the internal consistency of each dimension was also within an acceptable range.

5.2. Control Group Results

The questionnaire results for the control group showed that, except for the "Challenge and Effort Level" dimension, the average scores for all other dimensions were below 3 points (the midpoint of the 5-point scale), indicating that participants generally had a lower evaluation of the traditional display method. The "Interactivity and Sense of Control" dimension scored the highest (2.44 points), but still below the midpoint, suggesting that even in traditional displays, this was a relatively better aspect. The "Challenge and Effort Level" dimension scored relatively high (3.56 points), possibly indicating that the

traditional display method was more challenging for participants. The overall internal consistency of the control group's questionnaire was good (Cronbach's = 0.83), and the internal consistency of each dimension was also within an acceptable range.

Comparison between Experimental and Control Groups

To confirm whether the differences between the experimental group (innovative display mode) and the control group (traditional display mode) were statistically significant, this study conducted one-way Analysis of Variance (ANOVA) and independent samples t-tests on the mean scores of each dimension. The following are the analysis processes and results, which can be referenced in Table 1.

Statistical Analysis of Data from Experimental and Control Groups

Comparison of Average Scores Across Dimensions

Table 1. Comparison of Average Scores Across Dimensions

From the comparison of mean scores, it can be seen that, except for the "Challenge and effort level" dimension, the experimental group scored significantly higher than the control group in all other dimensions.

The results demonstrate that the Experimental Group (AR Interactive Display Mode) consistently outperforms the Control Group (Traditional Display Mode) across all experience dimensions, highlighting the positive impact of AR technology on user engagement, learning outcomes, and digital wellbeing. Significant performance gaps are observed in key areas such as Interactivity & Sense of Control (4.22 vs. 2.34), reflecting the effectiveness of gesture-based interaction using Leap Motion, and in Overall Satisfaction and Digital Wellbeing (4.27 and 4.26 vs. 2.4), showcasing AR's ability to foster meaningful, engaging experiences. Additionally, the AR model enhances Immersion & Presence (4.18 vs. 2.29) and promotes Emotional and Cognitive Engagement (4.18 and 4.13 vs. 2.36 and 2.32), suggesting deeper learning and emotional connections with exhibits. Notably, the Challenge & Effort dimension shows a smaller gap, indicating that AR enhances engagement without imposing excessive cognitive demands. These findings underscore AR's transformative potential in exhibition environments, with broad applicability in museums, educational settings, and cultural heritage displays, where it enriches visitor experiences while supporting cognitive and emotional development.

Independent Samples t-Test

To determine whether these differences are statistically significant, independent samples t-tests were conducted for each dimension. The test results can be referenced in Table 2.

Experimental hypotheses:

H0: There is no significant difference in mean scores between the experimental group and the control group

H1: There is a significant difference in mean scores between the experimental group and the control group

Significance level: = 0.05

Table 2. Independent Samples t-test Results

The experimental data show that the differences in all dimensions are statistically significant (p ¡ 0.001). This means that the experiment can reject the null hypothesis H0 and accept the alternative hypothesis H1, indicating that there are significant differences in mean scores between the experimental group and the control group across all dimensions.

Effect Size Analysis

To further evaluate the practical significance of the differences, Cohen's d effect sizes were also calculated in the experiment. The results can be referenced in Table 3.

Table 3. Cohen's d Effect Size Analysis

According to Cohen's standards:

d = 0.2 indicates a small effect

d = 0.5 indicates a medium effect

d = 0.8 indicates a large effect

The effect sizes for all dimensions exceed 0.8, indicating that there are large practical differences between the experimental group and the control group.

Reliability and Correlation Analysis Comparison

The overall questionnaire reliability and dimension correlations for the experimental and control groups are compared in Table 4. In Table 4, the average correlation coefficient for the experimental group is 0.748, while the average correlation coefficient for the control group is 0.593. The correlations between dimensions in the experimental group are generally higher than those in the control group, indicating that the various aspects of the experience are more closely connected in the experimental group.

Table 4. Overall Questionnaire Reliability and Relationship between Items and Dimensions Data

Explanation of Reliability and Correlation Analysis Comparison:

Overall Questionnaire Reliability: In the correlation analysis comparison, the overall questionnaire reliability indicators are as follows:

Overall Questionnaire Cronbach's : The experimental group has a value of 0.936, while the control group has a value of 0.889, with a difference of 0.08 between the two groups. Cronbach's coefficient is an indicator used to measure the internal consistency of a questionnaire, with values ranging from 0 to 1. The higher the value, the better the internal consistency of the questionnaire. Generally, an coefficient greater than 0.7 is considered acceptable, greater than 0.8 is regarded as good, and greater than 0.9 is considered excellent. Based on these data, the following conclusions can be drawn.

The Cronbach's coefficient for the experimental group is 0.936, which is a very high value, indicating that the questionnaire in the experimental group has excellent internal consistency. This means that the items in the questionnaire effectively measure the same underlying concept, reflecting the reliability of the questionnaire. The Cronbach's coefficient for the control group is 0.889, which, although lower than the experimental group's, is still within a good range. This indicates that the control group's questionnaire also has good internal consistency. The difference between the two groups is 0.047, which, while present, is not particularly large. This may suggest that the innovative exhibition model somewhat increased the consistency of responses among participants, but the impact is not highly significant. Overall, both groups' questionnaires demonstrate good to excellent internal consistency, which enhances the credibility of the study's results. High internal consistency indicates that the questionnaire can reliably measure the constructs of interest in the study, thereby providing a solid foundation for subsequent analysis and conclusions.

Relationship Between Items and Dimensions: The correlation between dimensions in the experimental group (0.748) is generally higher than in the control group (0.593), indicating that the new exhibition method created a more holistic and cohesive experience.

Overall, the statistical analysis strongly supports the effectiveness of the new exhibition method, as it significantly outperformed the traditional method in almost every aspect and created a more integrated and coherent visitor experience.

5.3. Interviews and Qualitative Analysis

The interview results further support the data from the questionnaire survey. Visitors generally felt that the innovative display mode provided a richer interactive experience, allowing for a more in-depth understanding of the exhibit details. At the same time, behavioral observation data showed that participants in the experimental group spent more time in front of the display cases, interacted more frequently, and were more willing to try different operation methods.

In addition to quantitative research methods, this study also employed thematic analysis to conduct a systematic qualitative study of feedback from subjects in both the experimental and control groups. The thematic analysis method referenced Braun and Clarke's six-step approach, including: familiarizing with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report [33]. The study categorized visitor feedback into four main dimensions and conducted a systematic qualitative analysis for each dimension, covering interactivity and sense of control, immersion and presence, learning experience and cognitive engagement, and overall satisfaction and willingness to recommend.

In the dimension of interactivity and sense of control, visitors in the experimental group generally expressed high satisfaction with the innovative display mode. They felt that this mode provided a smooth interactive experience, allowing them to intuitively operate virtual exhibits and feel in complete control of the display process. For example, Subject 3 stated that operating virtual exhibits through hand detection technology felt very natural and interesting. In contrast, the control group showed significant deficiencies, with visitors generally reporting a lack of interactivity in the traditional display method and difficulty in deeply understanding the details of the exhibits. This contrast clearly demonstrates the significant advantages of the innovative display mode in enhancing interactivity and sense of control.

In the analysis of immersion and presence, visitors in the experimental group reported that the innovative display mode provided an unprecedented immersive experience, making them feel completely absorbed in the exhibition environment. Subject 7, when describing this experience, mentioned feeling as if they were placed within the background story of the exhibits, forgetting everything about the outside world. In contrast, visitors in the control group indicated that the traditional display mode lacked sufficient immersion and found it difficult to evoke strong emotional resonance. These feedbacks suggest that the innovative display mode has significant advantages in creating immersive experiences, effectively transporting visitors into the exhibition content.

In the dimension of learning experience and cognitive engagement, visitors in the experimental group reported that the innovative display mode not only increased their understanding of the exhibits but also sparked deeper levels of interest and curiosity. Subject 23 mentioned that this display method helped them better understand the background and details of the exhibits. In contrast, while visitors in the control group were able to understand the display content, they showed notably less interest and engagement with the

exhibits. This reflects that the innovative display mode has advantages over the traditional display mode in promoting learning and cognitive engagement.

Finally, in the dimension of overall satisfaction and willingness to recommend, visitors in the experimental group gave high ratings to the innovative display mode, with many expressing their willingness to recommend this display method to others. Subject 80 explicitly stated that they would recommend friends to experience this innovative display mode. In contrast, visitors in the control group generally gave lukewarm evaluations of the traditional display and lacked a clear willingness to recommend it. This result further confirms the success of the innovative display mode in enhancing visitors' overall satisfaction.

Through qualitative analysis of feedback from the experimental and control groups, this study can draw the following important insights:

Interactivity is key: The innovative display mode greatly enhanced visitors' interactive experience through hand detection and 3D manipulation, which is lacking in traditional display methods. Experimental group participants generally reported higher interactivity and sense of control, such as "feeling adept at manipulating virtual exhibits." In contrast, control group participants expressed interactive limitations, like "finding it somewhat difficult to actually observe physical exhibits." This suggests that future display designs should continue to strengthen interactivity, providing visitors with more intuitive and natural ways of operation.

Immersion creates deep experiences: AR technology and immersive design successfully transported visitors into the world of the exhibits, creating more meaningful and memorable experiences. Experimental group participants often described deeply immersive experiences, such as "feeling as if they were in the natural history of the exhibit." This sense of immersion is particularly important for cultural heritage and science exhibitions, helping visitors better understand and feel the background and significance of the exhibits.

Combining learning and emotion: The innovative model not only promoted cognitive learning but also evoked emotional resonance, a combination crucial for deepening the learning experience. Experimental group participants reported deeper understanding and higher interest, such as "this display method helped them better understand the exhibits." Future display designs should pay attention to the role of emotional factors, enhancing learning effects by stimulating visitors' curiosity and emotional investment.

Balancing technology application: While innovative technology brought significant improvements, attention should be paid to the appropriate use of technology to ensure it doesn't cause operational difficulties for some visitors. The design process should consider the needs of visitors of different age groups and technology proficiency levels, providing appropriate guidance and instructions.

Importance of personalized experiences: Feedback showed high satisfaction among visitors with the ability to control and explore exhibits autonomously. This suggests that future display designs should focus more on providing personalized and self-chosen experiences, allowing visitors to explore exhibits according to their own interests and pace.

Overall satisfaction and willingness to recommend: Participants in the experimental group highly rated the innovative display mode and expressed willingness to recommend it to others. This reflects that the innovative model not only enhanced personal experiences but may also attract more visitors through word-of-mouth effects.

These qualitative analysis results strongly support the effectiveness of the innovative AR interactive display mode. This mode significantly outperforms traditional display methods in enhancing interactivity, immersion, learning experience, and overall satisfaction. These findings not only validate the design concepts proposed in the thesis but also provide valuable empirical foundations and innovative directions for future museum and exhibition designs.

6. Discussion and Future Research Directions

This study explores the comprehensive results of the innovative display mode design and visitor experience research. The integration of physical displays with virtual interactive display technology, particularly for small-scale exhibits, has proven effective in enhancing visitor engagement. Utilizing hand detection sensors and 3D modeling technology, this model offers an unprecedented interactive experience, allowing visitors to freely enlarge, shrink, and rotate virtual exhibits. This highly interactive and immersive display method significantly enhances visitor engagement, cognitive involvement, and overall satisfaction.

The experimental results demonstrate that, compared to traditional display modes, the AR interactive display model offers substantial advantages in terms of interactivity, immersion, learning effectiveness, and digital well-being. Visitors highly praised the ability to manipulate virtual exhibits directly, which provided deeper engagement with exhibit details and enriched learning experiences. This indicates that the innovative display mode is not only technically feasible but also holds significant potential to enhance visitor experiences in practical applications.

Linking these findings to current trends in AR development, this study aligns with the broader digital transformation in museums and cultural heritage fields. AR technologies are increasingly employed not just for content presentation but also for fostering immersive storytelling, real-time interaction, and context-aware experiences. As AR applications evolve, they offer opportunities to create dynamic, personalized exhibitions that adapt to visitors' preferences and learning needs.

Drawing insights from Omonayajo et al. [32], who highlighted the transformative role of AR and immersive technologies in smart education environments, this study further explores how AR interactive display models can positively influence visitors‘ cognitive engagement and digital well-being. Omonayajo et al. emphasized that AR technologies promote active learning, improve knowledge retention, and foster inclusive learning experiences, particularly through immersive and interactive design elements.

In the context of exhibition spaces, the integration of AR technologies not only enhances visitor engagement but also supports the development of digital literacy. By providing interactive experiences that encourage exploration and critical thinking, AR displays contribute to a more meaningful and personalized learning journey for visitors. Additionally, such interactive environments can foster inclusive experiences, accommodating diverse learning preferences and promoting accessibility for broader audiences.

These insights underscore the broader implications of AR interactive technologies beyond traditional educational settings, suggesting their potential as powerful tools for enhancing both educational outcomes and digital well-being in cultural and exhibition contexts. Future research could investigate the long-term impact of AR-enhanced exhibition

experiences on digital literacy development, as well as explore strategies for optimizing these technologies to support inclusive, accessible, and engaging learning environments.

Furthermore, the study highlights AR's role in promoting digital literacy. Interactive AR environments encourage critical thinking, problem-solving, and digital navigation skills, which are essential in today's technology-driven society. Additionally, AR fosters inclusive learning experiences by supporting diverse learning styles and accessibility needs. The multi-sensory, adaptable nature of AR technologies makes exhibitions more engaging and accessible to audiences with varying cognitive and physical abilities.

Future Research Directions: Future studies should investigate the long-term impacts of AR-enhanced exhibitions on digital literacy development. Longitudinal research can explore how sustained exposure to AR environments influences cognitive skills, technological proficiency, and visitor engagement patterns over time. Moreover, examining AR's role in fostering inclusive cultural experiences could provide valuable insights into how immersive technologies bridge accessibility gaps and promote social inclusion. Comparative studies across different cultural contexts may also reveal how AR experiences are perceived globally, informing best practices for museum and exhibition design worldwide.

6.1. Model Advantages

The innovative exhibition model designed in this study demonstrates several advantages over traditional exhibition methods, particularly in enhancing the display of small artifacts and increasing audience interactivity. These advantages stem from the following key factors:

Enhanced Interactivity and Immersion: Through hand detection technology, visitors are no longer passive viewers of the exhibits; they can actively participate by directly manipulating virtual exhibits through gestures. This interactivity not only increases visitor engagement but also allows for a deeper understanding of the exhibits. Additionally, the application of 3D modeling technology enables visitors to observe exhibits from different angles and distances, even zooming in to examine details. This experience closely approximates physically interacting with the exhibits, greatly enhancing immersion.

Application of AR Technology: The integration of AR technology allows for a seamless combination of virtual information with physical exhibits, enriching the content of the exhibition. Visitors can not only see the physical appearance of the exhibits but also access additional information through AR, such as historical context, production processes, and scientific principles, making the viewing experience more educational and engaging.

Addressing the Challenges of Displaying Small Exhibits: Traditional exhibition methods often struggle to provide an ideal viewing experience for small exhibits, especially those rich in detail and requiring close observation. The model proposed in this study, through 3D modeling and magnification features, allows visitors to easily observe every detail of the exhibits, solving the challenges associated with displaying small artifacts.

Improving Overall Visitor Experience: The experimental results show that the innovative exhibition model outperforms the traditional model across all experience dimensions. Visitors exhibited significant improvements in interactivity, immersion, and satisfaction. This indicates that the new model not only enhances visitor engagement and learning outcomes but also leaves them more satisfied with the overall exhibition process.

6.2. Challenges and Future Development

Although the innovative exhibition model proposed in this study shows significant potential in enhancing visitor experiences, it still faces challenges related to technology and cost in practical applications. Technical limitations include potential issues with the accuracy and latency of hand detection and AR technology, as well as the difficulties of applying 3D modeling technology to complex or fragile exhibits. In terms of projection technology, presenting clear virtual content under various lighting conditions is also a major challenge. On the cost side, the economic burden of acquiring, maintaining exhibition equipment, and involving professional personnel cannot be ignored.

To overcome these challenges and further enhance the model‘s application value, future research could focus on the following directions: continuously improving hand detection, AR, and projection technologies to increase accuracy, reduce latency, and enhance stability; exploring more convenient and efficient 3D modeling methods; finding more cost-effective technological solutions; developing more interactive and engaging AR content, such as games and quizzes; offering personalized exhibition content and interaction methods based on visitors‘ interests and needs; and expanding the application scope to include large exhibits and scene displays.

Through ongoing technological innovation and application expansion, this integrated physical and virtual interactive exhibition model is expected to play an increasingly important role in museums, cultural heritage displays, education, and other fields, providing audiences with richer and more inspiring experiences and promoting cultural heritage and knowledge dissemination.

7. Conclusion

This study designed and evaluated an innovative transparent projection interactive display mode, which successfully integrated physical display with virtual interactive technology, particularly suitable for exhibiting small items. The research results show that this new display method significantly enhanced visitors‘ experiences in multiple aspects, including interactivity, immersion, learning effectiveness, and overall satisfaction. Through the application of hand detection sensors and 3D modeling technology, visitors were able to directly manipulate virtual exhibits, achieving an unprecedented interactive experience.

Both quantitative and qualitative analysis results confirmed the superiority of this model compared to traditional display methods, especially in improving visitor engagement and deepening understanding of exhibits. However, the study also identified some technical and cost-related challenges, which point to directions for future research.

Academic Contributions: This study contributes to the academic discourse on AR-enhanced exhibitions by presenting a novel interactive model that integrates physical and virtual display techniques. The mixed-method approach provides robust empirical evidence of the model‘s effectiveness, particularly in enhancing cognitive engagement, emotional involvement, and digital wellbeing. By drawing from contemporary theories of immersive learning and digital literacy, the study bridges gaps in the literature regarding AR’s role in cultural heritage contexts. Furthermore, the integration of real-time behavioral data collection through cloud-based systems introduces a methodological innovation that can be applied in future exhibition studies.

Practical Application Value: The findings have significant practical implications for museums, cultural heritage sites, and educational institutions. The AR interactive display model offers a scalable, cost-effective solution for enhancing visitor experiences without the need for extensive physical space modifications. It supports inclusive learning by accommodating diverse visitors' needs, fostering accessibility, and promoting digital wellbeing through interactive, engaging content. Additionally, the model's flexibility allows for adaptation in various exhibition contexts, including science centers, art galleries, and historical archives, thus broadening its applicability across different educational and cultural environments.

Overall, this study provides new design ideas for museums, science education, and cultural heritage display fields, demonstrating the enormous potential of AR technology in enhancing display effects and visitor experiences. It also lays a foundation for the development of future display technologies. Further research could explore the impact of AR interactive exhibits on visitors‘ digital behavior and wellbeing during their visit, as well as how to optimize such technologies to better serve the educational and social functions across various educational settings. Future studies may investigate the long-term effects of AR-enhanced museum experiences on learning outcomes and visitor engagement. Additionally, researchers could examine the potential of AR technologies to facilitate inclusive and accessible learning environments for diverse audiences in educational institutions.

7.1. Research Contributions

This study has made significant contributions in multiple aspects of display technology. Firstly, it proposed an innovative display model that successfully integrates physical display with virtual interactive technology, addressing the limitations of traditional display methods for small exhibits. This model, utilizing hand detection technology and 3D modeling, significantly enhances the visibility of exhibits and visitor interactivity, providing strong empirical support for the application of advanced technologies in museums, science education, and cultural heritage displays.

Secondly, through experimental comparison, the study systematically evaluated the impact of the innovative display mode on visitor experience, demonstrating its significant advantages in enhancing visitors‘ sense of immersion, interactivity, and overall satisfaction. This enriches the research on the relationship between display technology and visitor experience, providing practical evidence for future display designs.

Furthermore, the study offers detailed technical implementation cases for the integrated application of hand detection, 3D modeling, and advanced projection technologies. This not only provides operational guidelines for display technology practitioners but also lays the foundation for further optimization and innovation of these technologies. This contribution has implications for the cross-disciplinary application of future display technologies, especially in better combining physical and virtual displays.

Overall, this study has made valuable explorations and contributions in innovative design of display technology, visitor experience research, and technical implementation strategies. It has profound implications for the future development of display technology.

7.2. Future Prospects

The innovative display model proposed in this study paints an exciting blueprint for the future development of museums and cultural heritage exhibitions. As technologies such as AR, VR, MR, 3D modeling, and gesture interaction continue to advance, interactive experiences will become more natural and fluid, potentially even allowing interaction with exhibits through voice, eye movements, or brain waves.

Immersive experiences will also deepen further, possibly creating fully virtual environments that make visitors feel as if they are truly present. Moreover, personalized learning experiences will become possible, with display systems potentially providing customized content and interaction methods based on visitors‘ backgrounds and needs.

Looking ahead, the application of this innovative model will not be limited to museums but may expand to fields such as education, healthcare, and retail, bringing new development opportunities to these areas.

Overall, this research not only provides an effective solution for displaying small cultural artifacts but also points the way for the future development of display technology. Through continuous technological innovation and application expansion, it will realize its potential in more fields, creating richer and more inspiring experiences for humanity, promoting cultural inheritance and knowledge dissemination.

References

1. Al-Ansi, A.M., Jaboob, M., Garad, A., Al-Ansi, A.: Analyzing augmented reality (AR) and virtual reality (VR) recent development in education. Social Sciences & Humanities Open, Vol. 8, 100532. (2023)
2. Alamri, A., Hossain, A.M., Hassan, M.M., Hossain, S.M., Alnuem, M., Ahmed, T.D.: A cloud-based pervasive serious game framework to support obesity treatment. Computer Science and Information Systems, Vol. 10, No. 3, 1229-1246. (2013)
3. Augusto, J.C., McCullagh, P.: Ambient intelligence: Concepts and applications. Computer Science and Information Systems, Vol. 4, No. 1, 1-27. (2007)
4. Bekele, M.K., Pierdicca, R., Frontoni, E., Malinverni, E.S., Gain, J.: A survey of augmented, virtual, and mixed reality for cultural heritage. Journal on Computing and Cultural Heritage 1 (JOCCH), Vol. 11, No. 2, 1-36. (2018)
5. Petrelli, D., Ciolfi, L., van Dijk, D., Hornecker, E., Not, E., Schmidt, A.: Integrating material and digital: a new way for cultural heritage. Interactions, Vol. 23, No. 4, 44-49. (2016)
6. Eghbal-Azar, K., Merkt, M., Bahnmueller, J., Schwan, S.: Use of digital guides in museum galleries: Determinants of information selection. Computers in Human Behavior, Vol. 57, 133-142. (2016)
7. Younan, S., Treadaway, C.: Digital 3D models of heritage artefacts: Towards a digital dream space. Digital Applications in Archaeology and Cultural Heritage, Vol. 2, No. 4, 240-247. (2015)
8. Jin, Y., Ma, M., Liu, Z.: Evaluating User Engagement and Preference in Virtual Reality and Augmented Virtuality for Interactive Storytelling. Interacting with Computers. (2024). [Online]. Available: https://academic.oup.com/iwc/advance-article-abstract/doi/10.1093/iwc/iwae027/7717779?redirectedFrom=fulltext
9. Scopigno, R., Cignoni, P., Pietroni, N., Callieri, M., Dellepiane, M.: Digital fabrication techniques for cultural heritage: A survey. Computer Graphics Forum, Vol. 36, No. 1, 6-21. (2017)
10. Balletti, C., Ballarin, M., Guerra, F.: 3D printing: State of the art and future perspectives. Journal of Cultural Heritage, Vol. 40, 195-202. (2019)

11. Tasfia, R., Yusoh, Z.I.M., Habib, A.B., Mohaimen, T.: An overview of hand gesture recognition based on computer vision. International Journal of Electrical and Computer Engineering, Vol. 14, No. 4, 4636-4645. (2024)
12. Wang, N., Niu, J., Liu, X., Yu, D., Zhu, G., Wu, X., Li, Y., Su, H.: BeyondVision: An EMG-driven Micro Hand Gesture Recognition Based on Dynamic Segmentation. Int. J. Electr. Comput. Eng., Vol. 14, No. 4, 4636-4645. (2024)
13. Balletti, C., Ballarin, M., Guerra, F.: 3D printing: State of the art and future perspectives. Journal of Cultural Heritage, Vol. 40, 195-202. (2019) (Duplicate entry)
14. Blanco-Pons, S., Carrión-Ruiz, B., Lerma, J.L., Villaverde, V.: Design and implementation of an augmented reality application for rock art visualization in Cova dels Cavalls (Spain). Journal of Cultural Heritage, Vol. 39, 177-185. (2019)
15. Hulusic, V., Gusia, L., Luci, N., Smith, M.: Tangible user interfaces for enhancing user experience of virtual reality cultural heritage applications for utilization in educational environment. ACM Journal on Computing and Cultural Heritage, Vol. 16, No. 2, 1-24. (2023)
16. Zerrouki, N., et al.: Deep Learning for Hand Gesture Recognition in Virtual Museum Using Wearable Vision Sensors. IEEE Sensors Journal. (2024)
17. Carvajal, D.A.L., Morita, M.M., Bilmes, G.M.: Virtual museums. Captured reality and 3D modeling. Journal of Cultural Heritage, Vol. 45, 234-239. (2020)
18. Pallud, J.: Impact of interactive technologies on stimulating learning experiences in a museum. Information & Management, Vol. 54, No. 4, 465-478. (2017)
19. De Paolis, L.T., Gatto, C., Corchia, L., De Luca, V.: Usability, user experience and mental workload in a mobile Augmented Reality application for digital storytelling in cultural heritage. Virtual Reality, Vol. 27, No. 2, 1117-1143. (2023)
20. Giariskanis, F., Kritikos, Y., Protopapadaki, E., Papanastasiou, A., Papadopoulou, E., Mania, K.: The augmented museum: A multimodal, game-based, augmented reality narrative for cultural heritage. In Proceedings of the 2022 ACM International Conference on Interactive Media Experiences, 281-286. (2022)
21. Kyriakou, P., Hermon, S.: Can I touch this? Using natural interaction in a museum augmented reality system. Digital Applications in Archaeology and Cultural Heritage, Vol. 12, e00088. (2019)
22. Al-Mansoori, R.S., Al-Thani, D., Ali, R.: Designing for Digital Wellbeing: From Theory to Practice a Scoping Review. Human Behavior and Emerging Technologies, 2023(1), 9924029. (2023)
23. Shen, J., et al.: Dwells in museum: The restorative potential of augmented reality. Telematics and Informatics Reports, Vol. 14, 100136. (2024)
24. tom Dieck, M.C., Jung, T.H., tom Dieck, D.: Enhancing art gallery visitors' learning experience using wearable augmented reality: generic learning outcomes perspective. Current Issues in Tourism, Vol. 21, No. 17, 2014-2034. (2018)
25. Paladini, A., et al.: Impact of virtual reality experience on accessibility of cultural heritage. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 42, No. W11, 929-936. (2019)
26. Trajkova, M., Alhakamy, A.a., Cafaro, F., Mallappa, R., Kankara, S.R.: Move your body: Engaging museum visitors with human-data interaction. In Proceedings of the 2020 chi conference on human factors in computing systems, 1-13. (2020)
27. Lee, Y.Y., Lee, J.H., Ahmed, B., Son, M.G., Lee, K.H.: A new projection-based exhibition system for a museum. Journal on Computing and Cultural Heritage (JOCCH), Vol. 12, No. 2, 1-17. (2019)
28. Hauser, W., Noschka-Roos, A., Reussner, E., Zahn, C.: Design-based research on digital media in a museum environment. Visitor Studies, Vol. 12, No. 2, 182-198. (2009)
29. Baraldi, L., Paci, F., Serra, G., Benini, L., Cucchiara, R.: Gesture recognition using wearable vision sensors to enhance visitors' museum experiences. IEEE Sensors Journal, Vol. 15, No. 5, 2705-2714. (2015)

30. Barszcz, M., Dziedzic, K., Skublewska-Paszkowska, M., Powroznik, P.: 3D scanning digital models for virtual museums. Computer Animation and Virtual Worlds, Vol. 34, No. 3-4, e2154. (2023)
31. Zhang, Z., Geng, Z., Li, T., Pei, R., Liu, Y., Zhang, X.: Integration of real-time 3D capture, reconstruction, and light-field display. In Stereoscopic Displays and Applications XXVI, Vol. 9391: SPIE, 131-145. (2015)
32. Omonayajo, B., Al-Turjman, F., Cavus, N.: Interactive and innovative technologies for smart education. Computer science and information systems, Vol. 19, No. 3, 1549-1564. (2022)
33. Hammady, R., Ma, M., Powell, A.: User experience of markerless augmented reality applications in cultural heritage museums: 'MuseumEye' as a case study. Virtual Reality, Vol. 24, No. 2, 303-324. (2020)
34. Trunfio, M., Campana, S., Magnelli, A.: Measuring the impact of functional and experiential mixed reality elements on a museum visit. Current Issues in Tourism, Vol. 23, No. 16, 1990-2008. (2020)
35. Jennett, C., et al.: Measuring and defining the experience of immersion in games. International Journal of Human-Computer Studies, Vol. 66, No. 9, 641-661. (2008)
36. Clini, P., Quattrini, R., Frontoni, E., Pierdicca, R., Nespeca, R.: Real/not real: Pseudo-holography and augmented reality applications for cultural heritage. Journal of Computing and Cultural Heritage, Vol. 13, No. 2, 1-21. (2020)
37. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology, Vol. 3, No. 2, 77-101. (2006)

Ming-Feng Lee received the Ph.D. degree from the Institute of Educational Information and Measurement at National Taichung University of Education, Taiwan. He is currently an Assistant Professor at the Department of Information Management, National Taichung University of Science and Technology, Taichung, Taiwan. His current research interests include artificial intelligence, innovative display interaction design, eye-tracking analysis, affective computing, physiological information sensing, mobile computing, and heterogeneous system integration and development. Contact him at antoniolee@nutc.edu.tw

Guey-Shya Chen received the Ph.D. degree from the Information Technology and Electrical Engineering at University of Queensland, Queensland, Australia. Contact her at grace@mail.ntcu.edu.tw

Hui-Chien Chen received the Master. degree from the Institute of Educational Information and Measurement at National Taichung University of Education, Taiwan. She is currently pursuing a doctoral degree at the same institute. Contact her at gigi.baby@yahoo.com.tw

Jian-Zhi Chen received the Master. degree from the Department of Bio-Industry Technology at Dayeh University, Taiwan. He is currently working at Corporate Synergy Development Center (CSD), Taiwan. Contact her at t296210@gmail.com

Received: September 29, 2024; Accepted: June 12, 2025.

Computer Science and Information Systems 22(4):1665–1685 https://doi.org/10.2298/CSIS241001066Y

Application of the Inception-ResNet-V2 algorithm to the analysis of embryo microscope images for the prediction model of assisted reproduction

Yu-Yu Yen[1], Weng Shao-Ping[2], Su Li-Jen[3], Kao Jui-Hung[4,⋆], and Chu Woei-Chyn[1,⋆]

[1] Department of Biomedical Engineering, National Yang Ming Chiao Tung University, Taipei, Taiwan
sheepkelly19.be11@nycu.edu.tw
wchu@nycu.edu.tw
[2] IHMED Reproductive Center, Taipei 106028, Taiwan
drweng@ihmed.com.tw
[3] Department of Biomedical Science and Engineering, National Central University, Taoyuan 320317, Taiwan
sulijen@gmail.com
[4] Department of Information Management, Shih Hsin University, Taipei, Taiwan
kjhtw@mail.shu.edu.tw

Abstract. The World Health Organization (WHO) estimates that approximately14 80 million men and women worldwide, with childbearing potential, need medical assistance due to fertility difficulties, which represents approximately 15% of the population. Similarly, about 15% of couples of maternal ages in Taiwan experience infertility problems. In clinical practice, in vitro fertilization (IVF) is the primary method of artificial reproduction. Using deep learning technology and an Inception-ResNetV2 model, we can create a reliable embryo classification and prediction system, which improves the selection of high-quality embryos and enhances pregnancy success rates. The classification and prediction model achieved 80% precision, AUC= 0.88, sensitivity 73% and 88% specificity. This exceeds the statistics of the Taiwanese National Health Service, where the average pregnancy rate for IVF in 2023 was 27.8 %. The results indicate that our model efficiently classifies embryos for successful implantation at a higher rate than the national statistics in Taiwan.

Keywords: Convolution neural networks, Deep learning; embryo, In vitro fertilization.

1. Introduction

The World Health Organization (WHO) estimates that approximately 80 million individuals worldwide experience infertility and require medical assistance. This figureaccounts for about 15% of the global reproductive population, more than three times the 1 population of Taiwan. Similarly, in Taiwan, around 15% of couples of childbearing age face fertility challenges. Based on this estimate, approximately 300,000 couples in Taiwan are experiencing varying degrees of fertility difficulties [1, 2]. Many factors can cause

⋆ Corresponding authors

infertility or low pregnancy ability. According to the Health Promotion Administration of the Ministry of Health and Welfare of Taiwan, the average pregnancy rate for IVF in Taiwan in 2023 was 27.8%, with a live birth rate of approximately 20.4%. Among these live births, 25.8% were twins, and 0.2% were triplets [3]. Clinical treatments for infertility often involve interventions at various stages, such as ovulation timing, intrauterine insemination (IUI), and more advanced techniques like in vitro fertilization (IVF).

Couples attempting to conceive without using contraception but who are unsuccessful in achieving pregnancy or carrying it to term are considered infertile. The likelihood of pregnancy generally declines with age, particularly after age 30. Conception and live birth rates decrease significantly as women age [4, 5]. The woman's age considerably influences the success of her conception and the quality of the oocyte. A woman's age significantly impacts her ability to conceive and the quality of her oocytes. As women age, the probability of conception decreases yearly while the risk of miscarriage increases. A woman produces approximately one million oocytes during her lifetime,although the number of viable oocytes decreases with age. By the late thirties, the natural conception rate drops below 10%. Among infertility cases, 40% are attributed to female factors, 40% to male factors, and the remaining 20% to combined factors or unexplained causes [6].

With recent advancements in reproductive technology, treatment methods have significantly progressed. In 1978, through the collaborative efforts of British physiologist Dr. Robert Edwards and obstetrician and gynecologist Patrick Steptoe, the first test tube baby was born [7]. This woman had a natural pregnancy and gave birth to a healthy baby. IVF is the breakthrough treatment for infertility. In recent years, reproductive technology has been changing with each passing day, such as the birth of new treatments in ovulation (Gonadotropin, recombinant follicular stimulating hormone (rFSH), etc.), improvement of ovulation stimulation (Gonadotropin-Releasing Hormone (GnRH) enhancer and antagonist treatment), the use of vaginal ultrasound to retrieve the oocyte, major advances in vitro fertilization (Intracytoplasmic sperm injection (ICSI), improvements of vitro culture technology(extended from three days of culture to five days), and changes in embryo implantation methods(replacement uterine implant from implantation of the fallopian tube), various technologies have become more mature, making test tube baby reproductive technology a very important auxiliary method for all infertile couples to seek a child.

IVF treatment is not a panacea; it is simply an artificial assisted reproductive technology that can provide infertile couples with the fastest goal. Many previously helpless problems, such as bilateral fallopian tubes, severe male infertility, etc., are currently only possible to achieve pregnancy through this artificial reproductive technology. According to statistics from the Taiwan Ministry of Health and Welfare, the average IVF pregnancy rate in Taiwan in 2017 was 27.1%, but the live birth rate was only approximately 20.4%, of which 25.8% were twins and 0.2% were triplets [3]. Similarly, data from the European Society of Human Reproduction and Embryology (ESHRE) in 2016 showed that the clinical pregnancy rate per embryo implantation in European countries was only 27.1% [8].

In the early stages of IVF treatment, the low efficiency required the transfer ofmultiple embryos to increase the probability of pregnancy. However, with the accumulation of knowledge and advancements in preimplantation embryonic development, the efficiency of IVF treatments has significantly improved. The transfer of multiple embryos, while increasing the likelihood of achieving pregnancy, also significantly raises the

risk of multiple pregnancies, which can exacerbate maternal and fetal complications during and after pregnancy. These complications include fetal death, developmental arrest, preeclampsia, eclampsia, placental abnormalities, and primary postpartum hemorrhage, all of which become more prevalent with multiple pregnancies [9-11]. Therefore, modern IVF treatment gradually emphasizes the safety of mother infant. The improvement in embryo selection technology is to reduce the number of embryos required for transplantation. Due to the improvement in embryo culture,technology, and culture medium, the efficiency of development from the fertilized egg to the blastocyst stage has increased. Advancements in embryo culture technology and media have significantly improved the development of embryos from fertilization to the blastocyst stage. By extending the in vitro culture period, the developmental potential of each embryo can be better assessed, and advances in cryopreservation techniques further enhance the ability to preserve embryos with high developmental capacity [12-14]. Related studies have compared embryo implantation and pregnancy potential on the basis of their morphology. The goal of selecting a single, optimal embryo for transfer is to achieve a singleton birth [15]. Although the positive impact of single embryo transfer (SET) on mother-infant safety is a well-known fact, it is still not possible to use single embryo transfer as the only method in most centers. The main reason is that current technology is not stable and efficient in selecting embryos with the best implantability, and to maintain an acceptable pregnancy rate, multiple embryo transfers are still necessary to have no choice [16]. Therefore, to increase the clinical results of single embryo transfer, how to further develop new technologies or markers to improve embryo screening remains one of the main topics today [17].

2. Materials and Methods

After ovulation stimulation and in vitro fertilization (IVF) during assisted reproductive technology cycles, embryos are cultured in vitro for two to five days before being implanted in the uterus. The fertilization process and subsequent development of each embryo vary greatly, with some embryos exhibiting robust growth while others may divide more slowly, stop dividing altogether, or develop cytoplasmic fragments. Given the limitations in pregnancy success rates associated with IVF treatment, many assisted reproduction facilities opt to implant a relatively large number of embryos to increase the likelihood of achieving pregnancy. This practice is often driven by the patient‘s strong desire for a successful pregnancy, which consequently increases their tolerance for the risks associated with multiple pregnancies. However, this approach significantly ncreases the risk of high-risk pregnancies and preterm births, placing a substantial burden on medical resources [18]. On this basis, it is considered necessary to establish a set of screening criteria for the treatment cycle of blastocyst-stage embryo implantation and to combine embryo culture values and embryo images using convolutional neural networks (CNN) algorithms. It is used to predict the potential of embryos for pregnancy and provide the clinician with appropriate counseling for the infertile couple after the embryos have been implanted following this standard of selection criteria.

2.1. Importance of embryo quality in IVF treatment

Many studies have pointed out that embryo types from the first day to the sixth day after fertilization can be used as a basis for embryo selection. Therefore, many scoring systems

have been proposed to improve the pregnancy success rate of IVF treatment [19-21]. The clinical guidelines provided by the American Society for Reproductive Medicine (ASRM) have evolved significantly over time, increasingly advocating for personalized approaches tailored to different age groups and varying embryo types, either on the third or fifth day of development. In other words, in clinical practice, the emphasis is on adapting to the patient's age to reduce the number of embryo implantations as much as possible, thus achieving the goal of reducing multiple pregnancies while maintaining a high success rate. At present, three days (division period) or five days (blastocyst period) represent the vast majority of in vitro culture worldwide, and most artificial reproduction centers still mainly have three days of culture. The selective single embryo transfer method during the cleavage stage on the third day is ineffective [22, 23]. This suggests that currently, it is not possible to directly select the embryo with the highest implantation potential for single embryo transfer to achieve the optimal pregnancy rate. Therefore, there are two follow-up solutions for artificial reproduction institutions. One is to continue cultivating embryos in the blastocyst stage before implantation, and the other is establishing screening criteria for embryos in the division stage.

The majority of assisted reproduction institutions focus on using embryos on the third day of the cleavage stage for implantation. Several factors drive this practice: First, extending the in vitro culture period to five days significantly increases the demand for manpower and incubator space by more than 40%. Additionally, the requirements for culture medium differ significantly. Pyruvate is the primary energy source during the cleavage stage, while glucose becomes essential during the morula and blastocyst stages. Consequently, the cost of maintaining cultures for five days increases substantially. Second, although culturing embryos for five days can help select more viable embryos, this process primarily enhances the implantation rate without necessarily improving pregnancy or live birth rates. Is it worth risking embryos due to prolonged culture time due to the risk of degradation? The insistence on extending the time of in vitro culture remains an issue that should be carefully considered by various artificial reproduction institutions. If it is relatively difficult to culture for five days to the blastocyst stage, a set of screening criteria is provided for embryos in the division stage. Especially in the United States and Taiwan, medical insurance does not cover IVF treatment [24, 25].

Patients are very passionate about pregnancy success, which relatively increases patients ardent expectation and degree of tolerance for a twin pregnancy, resulting in many artificial reproduction institutions with implants that include a larger number of embryos to increase their success rate of pregnancy. In view of this, the American Society of Reproductive Medicine provides guidelines for the number of embryo implants for its members, mainly based on age [26]. According to the clinical guidelines of the American Society of Reproductive Medicine on the number of embryos implanted in 2004, 2008, and 2009, patients continue to be divided into four age groups, namely under 35 years, 35 to 37 years, 38 to 40 years, and older than 40. The corresponding number of embryo implantations is suggested according to different age groups. The aim is to reduce the pregnancies of multiple births above triplets in the young population. Even in European countries where single embryo implantation is emphasized, single embryo implantation is restricted to young groups of people under the age of 38 years. The clinical use of the embryo scoring system on the third or fifth day of the type of embryo is relatively limited to young groups of people [27].

2.2. Selection and transfer in IVF treatment

During the fertilized period of fertilized oocytes, mitochondria are distributed mainly around the prokaryotic, especially where the two prokaryotics face each other. However, its distribution is not necessarily uniform. It is separated from the interface of the first split. Sometimes, there is more on one side and less on the other [28]. In prokaryotic stage embryos (fertilized oocytes), an uneven distribution of mitochondria will likely persist through subsequent developmental stages. During the two-cell stage,

mitochondria initially concentrate at the distal ends of the two nuclei before eventually surrounding them. Similarly, at the four-cell stage, mitochondria tend to concentrate at the distal ends of the nuclei in most cells. However, some cells may receive a greater number of mitochondria, while others may receive fewer, leading to variability in mitochondrial distribution across cells [29]. If the number of mitochondria in the cell is less, the ability to synthesize adenosine triphosphate is reduced. These embryonic cells will often stop dividing or even become embryonic fragments and die [28].

The oocyte must undergo several cell cycles (mitosis) to form a blastocyst and subsequent fetal tissue. In addition to precise gene regulation, this process also requires the supply of energy. In mammals, cell energy comes mainly from adenosinetriphosphate. The synthesis of adenosine triphosphate can be divided into two pathways: mitochondrial and non-mitochondrial. Before the stage of compaction or embryonic mulberry, oxidative phosphorylation is used as the pathway of the adenosine triphosphate production pathway, and the main energy source is pyruvate [30, 31]. After the embryonic stage of the mulberry, glucose is preferred as an energy source, but glycolysis is used as the production pathway of adenosine triphosphate [30]. Clinically, in vitro culture is typically divided into two stages. The first stage is enriched with pyruvate, which optimally supports embryo development during the first three days. The second stage is glucose-based, supporting development during the fourth and fifth days.

Embryos that advance to the eight-cell stage within the first three days will likely accumulate substantial amounts of free radicals and oxidative byproducts due to oxidative phosphorylation, making them particularly vulnerable to oxidative stress. Therefore, preserving the mitochondrial membrane potential is essential for supporting embryonic development during this critical early period. At this stage, pyruvate is preferred as the energy source over glucose. It is inferred that if the concentration of free radicals and oxidetives is too high in the first three days, the inner mitochondrial membrane potential will be affected and cannot be maintained, further affecting embryonic development. Suppose a reducing agent that removes the toxicity of free radicals can be provided. In that case, it may help embryo development even more, increase the rate of blastocyst development, and also help increase the pregnancy rate of IVF treatment [31]. The incidence of multiple pregnancies has been steadily increasing each year. In response, the American Society for Reproductive Medicine (ASRM) has developed comprehensive clinical guidelines for its members. These guidelines provide recommendations on the optimal number of embryos to implant, criteria for embryo selection based on maternal age, and assessments of embryo quality. By adhering to these protocols, clinicians aim to enhance reproductive outcomes while minimizing the risks associated with multiple gestations [26].

In recent years, due to the technological advancement of the embryo in vitro culture at National Taiwan University and even in artificial reproduction centers throughout Taiwan, in addition to culturing embryos until the third day for embryo implantation, some

institutions even cultivate blastocysts until the fifth day. Recent clinical guidelines of the American Society for Reproductive Medicine recommend the number of embryos implanted in the division stage on day 3 and the blastocyst stage on day 5, respectively [26]. All artificial reproduction agencies also expect to be able to establish their own guidelines for the pregnancy success rate of IVF treatments for embryo implantation on days 3 and 5. Finding suitable clinical guidelines based on age, embryo quality, number of implants, and other variables by improving embryo selection techniques to reduce the number of embryos required for transplantation. Due to the technological improvement of the embryo in vitro culture and culture medium, the efficient development of fertilized oocytes in the blastocyst stage has been improved. The developmental potential of individual embryos can be better recognized by prolonging in vitro culture time, and the advancement of vitrification and freezing technology can also maximize the preservation of fertile embryos. The embryos with the highest developmental capacity are chemically preserve [12-14].

Additionally, related research compares embryo implantation and pregnancy capacity with embryo morphology. Select the best single embryo transplant for single birth [15, 32]. Although the positive impact of single embryo transfer (SET) on maternal and infant safety is a well-known fact, it is still not possible to use single embryo transfer as the only method in most centers. The main reason is that current technology is not stable and does not efficiently screen embryos with optimal implantability. To maintain an acceptable pregnancy rate, multiple embryo transfer is sometimes not an option but is necessary. Therefore, to increase the clinical results of single embryo transfer (SET), how to further develop new technologies or markers to improve embryo screening remains one of the current important issues.

2.3. Related work

The reproductive technology for test tube babies basically consists of five procedures: ovulation stimulation, egg retrieval, in vitro fertilization, embryo culture, and embryo implantation into the mother's body. The capacity of each embryo to grow during fertilization varies, with some embryos growing well, others slowly dividing or even stopping, and others appearing as cytoplasmic fragments. The highest quality of the first three embryos can be applied to guide the physician and the spouse in coordinating the number of embryos to be implanted. Embryo quality has a significant impact on the success rate of an IVF pregnancy, and some embryos that develop in vitro stop dividing, which is considered to undergo cellular aging. The need to establish a set of selectioncriteria to divide embryos is essential when the 5-day growth period to the blastocyst stage is relatively challenging. The development of new techniques or markers to improve embryo screening is still one of the most important issues today [16], and the following is a description of this research.

Loewke et al. employed an AI model to predict pregnancy rates, achieving an area under the curve (AUC) between 0.6 and 0.7, outperforming traditional manual morphological classification at each stage. A bootstrap analysis predicted that implementing AI could enhance pregnancy rates by 5% to 12% per site, compared to manual classification using an inverted microscope. However, sites utilizing low magnification stereo zoom microscopes did not exhibit the anticipated improvements with AI implementation. Visualization techniques and attribution algorithms indicated significant overlap between the

features identified by the AI model and those used in the manual scoring system. Two sources of bias were identified-associated with the type of microscope and the embryo retention micropipette apparatus-and were subsequently mitigated. The analysis further revealed that a 0.1 (10%) increase in AI scores correlates with a corresponding increase in pregnancy rates [33].

Sujata et al. evaluated embryo quality by visual morphology during in vitro fertilization (IVF) to transfer potential embryos. However, the success rate of in vitro fertilization remains low due to differences in the selection process. The main objective is to improve the rate of implantation by predicting the quality of the embryos that are transferred from day 2 to day 3 [34].

Brás de Guimarães et al. developed an artificial neural network (ANN) supported by a decision tree to predict the probability of live birth after in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI) treatments prior to the first embryo transfer. The study analyzed 26 demographic and clinical variables across 1,193 IVF/ICSI treatment cycles conducted at the Centro de Infertilidade e Reprodução Medicamente Assistida between 2012 and 2019. The ANN demonstrated an accuracy of 75.0%, with the area under the receiver operating characteristic (AUROC) curve measuring 75.2% (95% confidence interval: 72.5-77.5%) [35].

To enhance the precision of convolutional neural networks (CNNs) in image classification, C. Peng, Y. Liu, X. Yuan, and Q. Chen have undertaken a comparative analysis of distinct classification model structures. They have proposed an enhanced Inception-ResNet-v2 model, which is based on CNN. A multiscale depth-separable convolution has supplanted the original convolutional structure. This modification has reduced the number of parameters necessary to capture disparate sensory field features. The model has also been endowed with a channel filtering module to filter and merge channels, thereby enhancing the efficiency and accuracy of feature extraction. Data enhancement techniques and other methods optimize the model's performance. Experiments demonstrate that the proposed model surpasses most existing models in multiple datasets, achieving a maximum classification accuracy of 94.8% [36].

Barnett-Itzhaki et al. employed machine learning algorithms, specifically Support Vector Machines (SVM) and Neural Networks (NN), to predict outcomes such as the number of eggs retrieved, mature oocytes, fertilized oocytes, high-quality embryos, positive β-hCG results, clinical pregnancies, and live births. Using age, BMI, and clinical characteristics, these models outperformed traditional logistic regression models in predictive accuracy. The precision of the NN and SVM models ranged from 0.69 to 0.9 and 0.45 to 0.77, respectively, while the logistic regression model exhibited a lower precision range of 0.34 to 0.74 [37].

3. Results

3.1. IOTA Decentralized Ledger Technology

Inception-ResNet-v2 [38] is an enhanced variation of the earlier Inception V3 model, incorporating advancements inspired by Microsoft's ResNet architecture. This improved network architecture reduces computational complexity by fusing feature maps at different scales. Specifically, it replaces the 5x5 and 7x7 convolutions with multiple 3x3 convolutions, thus decreasing the computational effort required for processing.

3.2. Inception-ResNet-v1

The architecture consists of five Inception-ResNet-A modules, ten Inception-ResNet-B modules, five Inception-ResNet-C modules, and the Reduction-A and Reduction-B modules. These are sequentially processed after the stem, which connects the input image to the Inception-ResNet modules. Feature extraction is achieved by applying stride, padding, and max-aggregation in convolutional layers, enabling the capture of feature vectors from the facial images. Subsequently, these vectors are processed through average pooling, dropout, a fully connected layer, and normalization of L2 (as depicted in Figure 1).

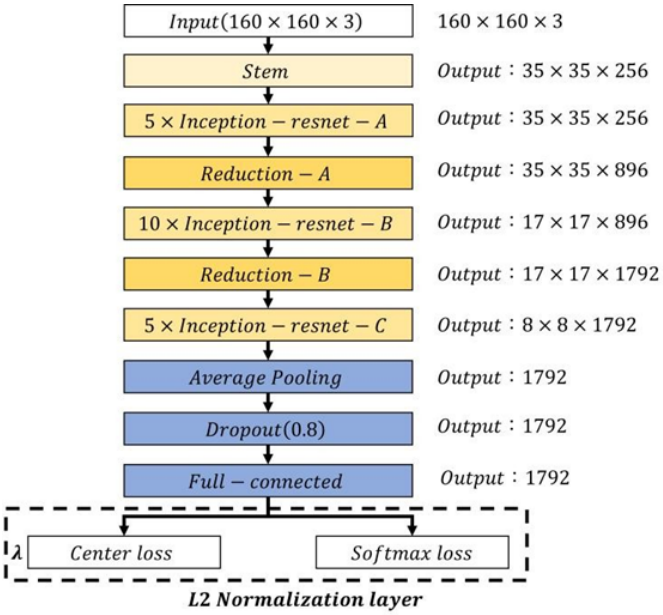


Fig. 1. Inception-ResNet-v1 architecture

3.3. Inception-ResNet-v2

The critical difference between Inception-ResNet-v1 and Inception-ResNet-v2 lies in their output dimensions, with Inception-ResNet-v2 having a more significant number of output dimensions, primarily attributed to variations in their stem structures. Additionally, Inception-ResNet-v2 contains a higher number of parameters within each module. The extraction of feature vectors from face images is achieved by applying stride, padding, and max pooling within the convolutional layers. This extraction process culminates in average pooling, dropout (to mitigate overfitting), a fully connected layer, and an L2 normalization layer. Inception-ResNet blocks, or residual inception blocks, are integral components of the Residual-Inception network. These blocks are computationally more efficient than the original Inception blocks. The Inception-ResNet architecture incorporates residual connections from ResNet into the Inception framework, allowing the output

of each Inception-ResNet layer to add its input value, thereby increasing the network's depth.

The Inception-ResNet module is a meticulously engineered convolutional block designed to produce distinct features while simultaneously reducing the number of parameters within the network. At the end of each Inception-ResNet layer, a 1x1 convolutional kernel is used for dimensionality enhancement, a feature absent in the Inception layer. This is significant since we used 1x1 convolutional cores for the purpose of reducing Inception's computation. We did not resize the image to 299×299. This does not have any change in the number of channels, but only in the size of the feature map generated during the process. After the convolution layer and the Inception module, the feature map size is 5×5 with 1792 dimensional vectors (number of channels). Kaiming He et al. [39], proposed Incep-tionResNet-v2. Architecturally, Inception-ResNet-v1 is quite similar to Incep-tion-ResNet-v2, but the difference lies in a deeper and more complex hierarchy, with more parameters corresponding to higher accuracy.

The main difference between the two is in the preprocessing part, the latter adopts a more complicated stem structure, and the 384-dimensional vector of the stem output dimension of Inception-ResNetv2 is larger than the 256-dimensional vector of Inception-ResNet-v1. The complicated stem structure caused a slightly slower training speed than Inception-ResNet-v1, but produced better performance. The following are the differences between Incep-tion-ResNet-v1 and Inception-ResNet-v2 in terms of stem, Inception-ResNet-A module, Inception-ResNet-B module, Inception-ResNet-C module, and differences between Reduction-A module and Reduction-B module. Although the structure of Reduction-A is the same, the discrepancies lie in the number of parameters. The number of parameters k, l, m, and n of Inception-ResNet-v1 is 192,192, 256, and 384, while the number of parameters k, l, m, and n of Inception-ResNet-v2 is 256, 256, 384, and 384 (as shown in Figure 2˜Figure 7).

Our findings indicate that scaling the residuals in a residual network can lead to instability during the early stages of training when the number of filters exceeds 1000. To address this issue, the learning rate is gradually adjusted to a stable level. At the same time, the residual scaling factor is maintained between 0.1 and 0.3 to ensure consistent and regular training (as illustrated in Figure 8).

The advancement of ResNet's residual learning propagation demonstrated that feed-forward and feedback signals could be transmitted directly through the network. As a result, the non-linear activation function (e.g., ReLU) in the shortcut connections was replaced by Identity Mappings. Furthermore, Inception-ResNet-V2 employs Batch Normalization in every layer, which, following normalization, simplifies the training process and improves the model's adaptability to uncertain data, surpassing the performance of earlier methods.

3.4. Research Methodology and Implementation Steps

In this research, Inception-ResNet-v2 pre-trained models were adopted to identify the success and failure of fertilization rate by employing the convolutional neural network in deep learning with migration learning. Each pre-trained model was verified by stratifying K-fold, and the assessment was performed by accuracy (ACC), area under the curve (AUC), sensitivity, and specificity. Consequently, the image modeling training and evaluation process are as follows (as shown in Figure 9).

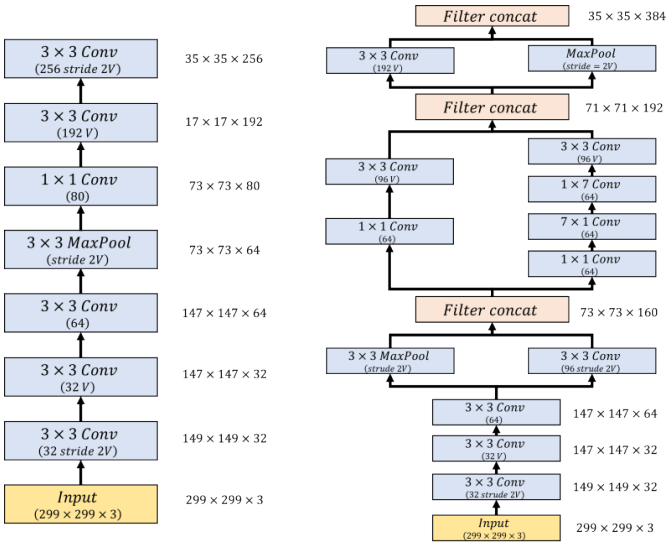


Fig. 2. Inception-ResNet stem differences

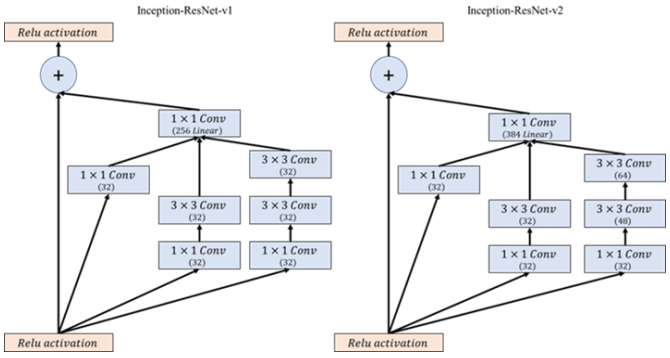


Fig. 3. Differences between the Inception-ResNet-A module

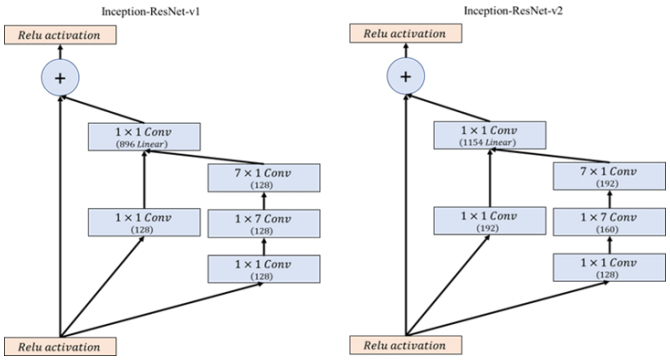


Fig. 4. Differences between the Inception-ResNet-B module

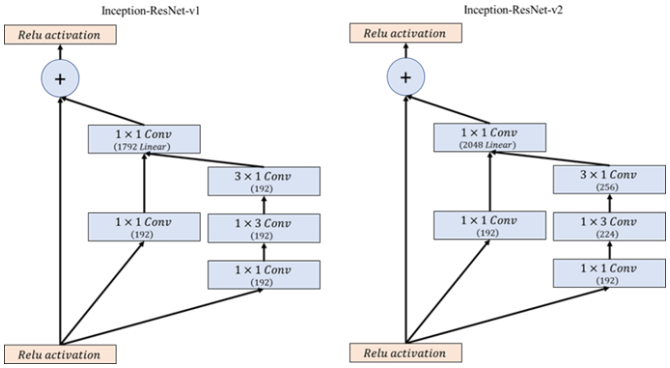


Fig. 5. Differences between the Inception-ResNet-B module

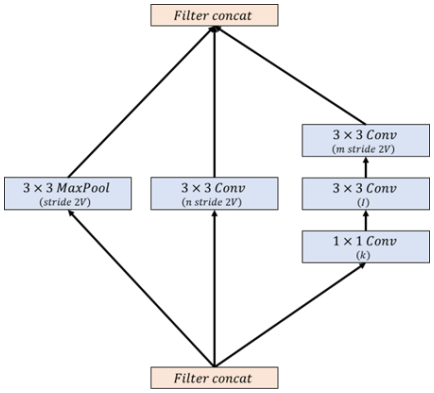


Fig. 6. Reduction-A Module Differences

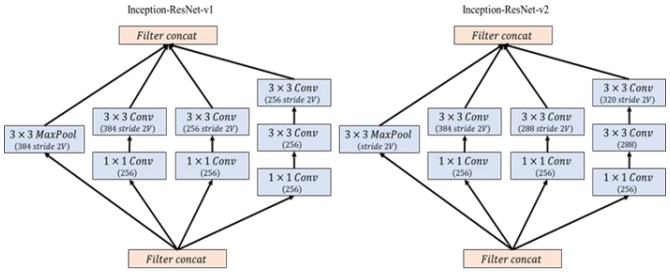


Fig. 7. Reduction-B Module Differences

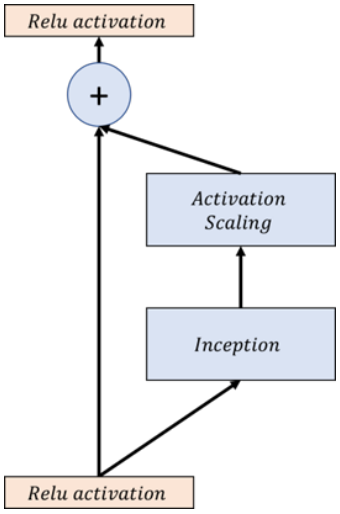


Fig. 8. Scaling of the Residuals

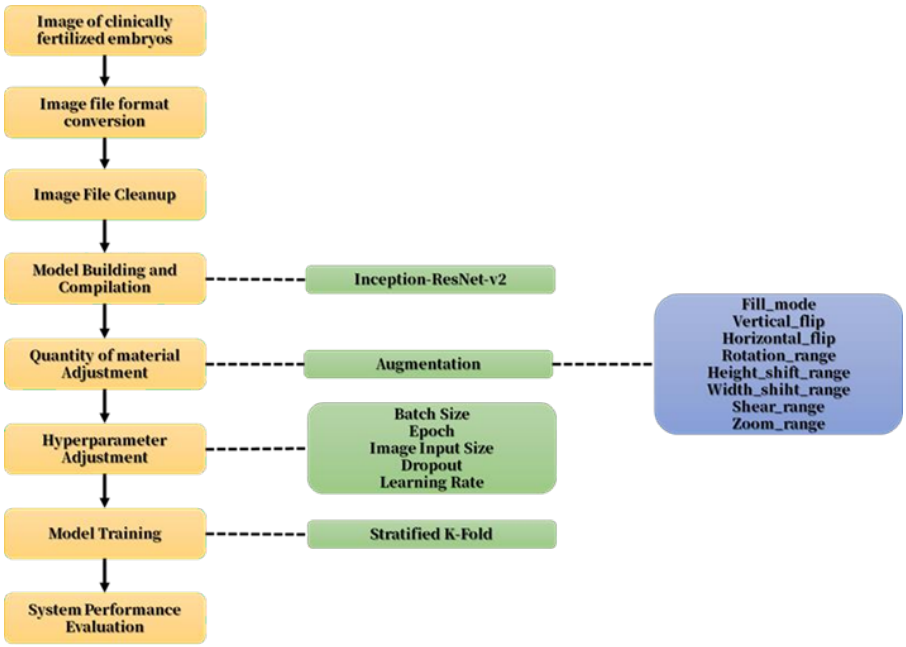


Fig. 9. Image-Modeling and Training Flywheel

3.5. Source and population of data

The data set for this research was acquired from Integrative Holistic Medicine, with a collection of 460 microscopic images of clinically conceived embryos. Of the 460 microscopic images, 150 were in PNG format (Portable Network Graphics), and 310 were in JPEG format (Joint Photographic Experts Group). Subsequently, the Python image library (PIL) was applied to convert the files to JPEG in a standardized manner. Among them, 230 embryos that did not conceive and 230 embryos that sustained a successful conception were also distinguished. The following are examples of embryo image collection (as shown in Figure 10).

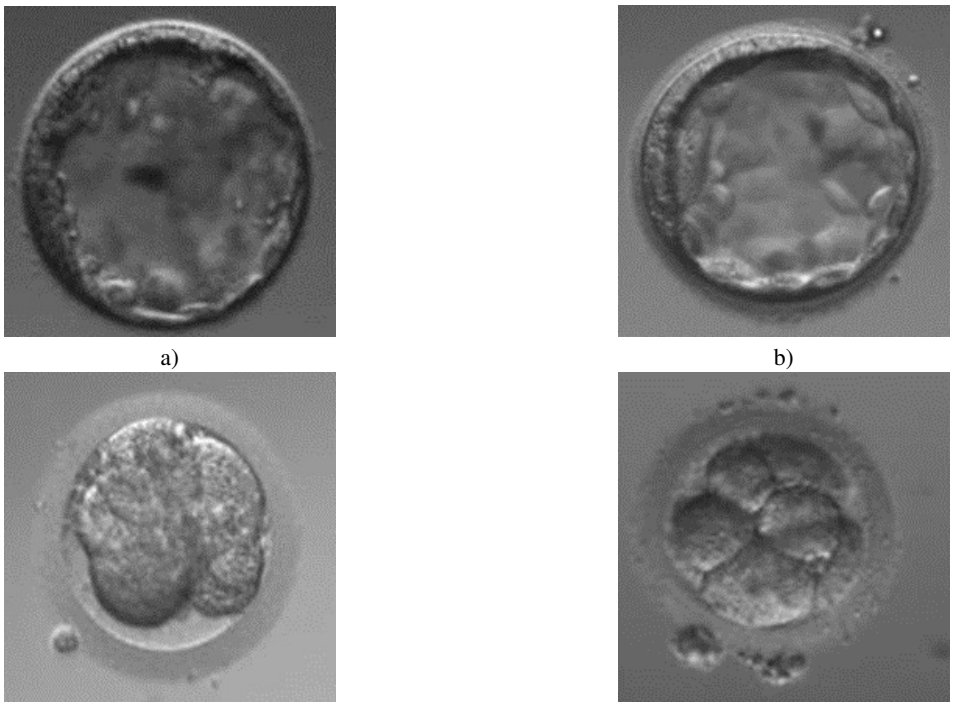


Fig. 10. Example of microscopic images of fertilized embryos selected from the database. (a) and (b) are fertilized embryos with failed implantation; (c) and (d) are fertilized embryos with successful implantation

4. Discussion

4.1. Experimental setup

This research adopts a transfer learning approach when constructing the model. At first, the convolutional substrate neural network weights are frozen, which prohibits updating

the weights while training is in progress. Meanwhile, the original classifier of the dense layer of the pre-trained model is abandoned, while the new classifier of the dense layer is established to compose a new model for retraining. The new model adopts relu and softmax for the activation function and adds batch normalization and dropout. When compiling the new model, the loss function of the new model adopted the categorical cross-entropy (CCE). The Optimizer is configured using Adam with a learning rate=0.001. Batch size of 16 randomly selected samples from the training set. To avoid overfitting, it is mandatory to stop training and keep the best model if there is no improvement after 10 epochs.

In total, 460 images were recorded, 230 without successful conception and 230 with successful conception, all in JPEG format (Joint Photographic Experts Group). The 230 data sets were randomly selected as 80% of the training set and 20% of the test set, while the original embryo image data sets had different sizes of image sources. Considering that the input image size is scaled, the calculation of feature point acquisition will change, which means that the model training results may be affected. Cut the image size to 512 pixels long * 512 pixels wide to facilitate subsequent data training. To deal with the problem of scarcity of data, we made use of the data enhancement method, which employed random horizontal_flip and vertical_flip, shear_range, zoom_range, rotation_range, width_shift_range, height_shift_range, and fill_mode. When we perform migration learning, we discard the original classifier of the dense layer and create a new classifier of the dense layer; therefore, we experiment with adjusting the input image size.

4.2. Accuracy analysis

In the field of machine learning, classification is a common task in supervised learning, for which binary classification is the most commonly applied approach. The actual output of the binary classification algorithm is a prediction score, which indicates the degree of certainty with which the system determines the class to which a given observation belongs to the positive category. For the person who uses this score, if the observation should belong to a positive or negative category, then he/she is required to select a classification threshold by comparing the scores with that value to interpret the scores. Any observation with a score above the threshold was predicted to be in a positive category. In contrast, those with a score below the threshold were predicted to be in a negative category.

As we know, accuracy is one of the most common indicators to evaluate classification models. In layman's terms, this refers to the proportion of results for which the model redicts the correct outcome, i.e., TP + TN divided by the number of all datasets. The alue is equal to the number of correctly predicted samples divided by the total number f the samples in the range [0,1].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Nevertheless, when we encounter the problem of Imbalance Class Classification, if we only focus on accuracy as the main measurement of model performance, we have the so-called Accuracy Paradox, which makes the accuracy metric meaningless. Therefore, other statistics are available to help us more objectively evaluate whether the classification model is good or bad with unbalanced data sets.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

To make the predicted results of the classifier more objective and accurate, precision and recall rates are employed (as one of the evaluation indicators, two metrics are extensively used in the field of information retrieval and statistical classification). The recall represents the percentage of data that are correctly predicted to be a positive category among all data that are in positive categories. It is described as follows:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Generally, we would like to get the precision and recall of a model that is not too poor, so we use the F1 indicator as a composite measure of the imbalance classification problem. The F1 indicator is the harmonic mean of precision and recall. It is described as follows:

$$F_\beta = (1+\beta^2)\frac{Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{4}$$

Where β is the weight of the precision and recall control, and takes the range of values [0,2]. As we can see from the formula, the larger the value, the greater the emphasis on recall, while the smaller the value, the greater the emphasis on precision. The general classification task is usually both precision and recall, that is, the β-value is taken as 1, which means the F1-measure.

4.3. Confusion matrix

A confusion matrix is a visualization tool that is particularly suitable for supervised learning, which is the most widespread and fundamental way of evaluating classification models. Each column of the matrix represents a category prediction of the data, while each row indicates the actual category of the data. With this mechanism, it is easier to determine whether the model is confusing between the two distinct categories. Once the confusion matrix is obtained, it can be utilized to calculate the accuracy, precision, and recall of the corresponding categories in the model. In addition, we observe the performance of the model in each category. With the visualization tool, it is straightforward to observe which categories are less easy to distinguish, such as how many categories A have been assigned to category B. In the matrix, all correct predictions are on the diagonal, whereas incorrect predictions are presented outside the diagonal. In this regard, we can formulate target-oriented improvement strategies to make the model more distinguished for each category (as shown in Table 1).

Table 1. Confusion matrix

Confusion matrix	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

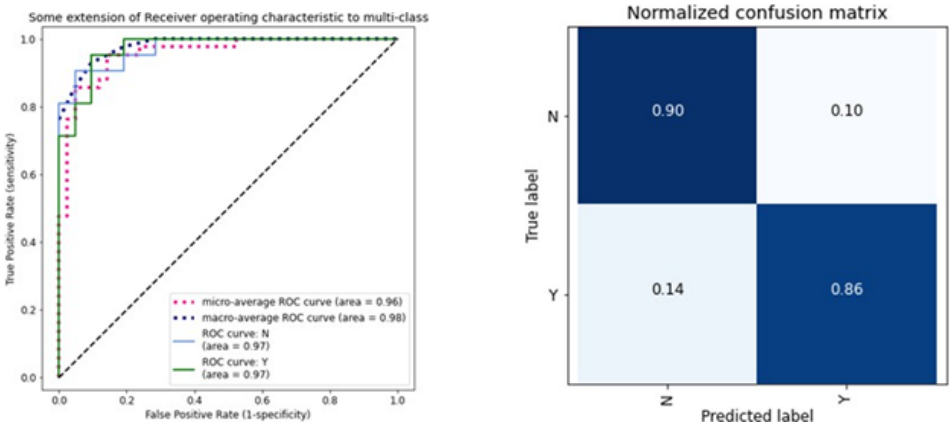
4.4. Analysis of results

Neural networks offer a different approach to pattern recognition and have been used in wide-ranging fields and have proven to be an effective diagnostic tool for many diseases or as a supplement for predicting treatment outcomes. In recent years, due to the popularity of deep learning research, various sophisticated neural network architectures have emerged; inception-resnet-v2 was proposed and applied in image recognition tasks, and the performance of these architectures that are highly task-dependent. In our research, we use InceptionResNetV2 to train a model based on embryo image datasets to evaluate the algorithm, classify, model, and train the embryos according to their morphological quality. For the dataset, it was classified into good and poor, using Group K-Fold to separate the training set and the test set. Furthermore, the test set section is split into validation and testing, with the files being recorded and then the values documented during backtesting. The following table shows the values of recall, precision, and F1-scores of the validation data, which indicate that precision N is 0.86, precision Y is 0.90, recall N is 0.90, recall N is 0.86, and both result in an F1-score of 0.88 (as shown in Table 2).

Table 2. Accuracy, sensitivity, and specificity of the validation set

Precision	recall	f1-score	support
N 0.90	0.90	0.88	21
Y 0.86	0.86	0.88	21
accuracy		0.88	42
macro avg 0.88	0.88	0.88	42
weighted avg 0.88	0.88	0.88	42

To present the values through visualizations, the following figure illustrates the chaotic matrix and the ROC curve of the validation data (as shown in Figure 11).



a) b)

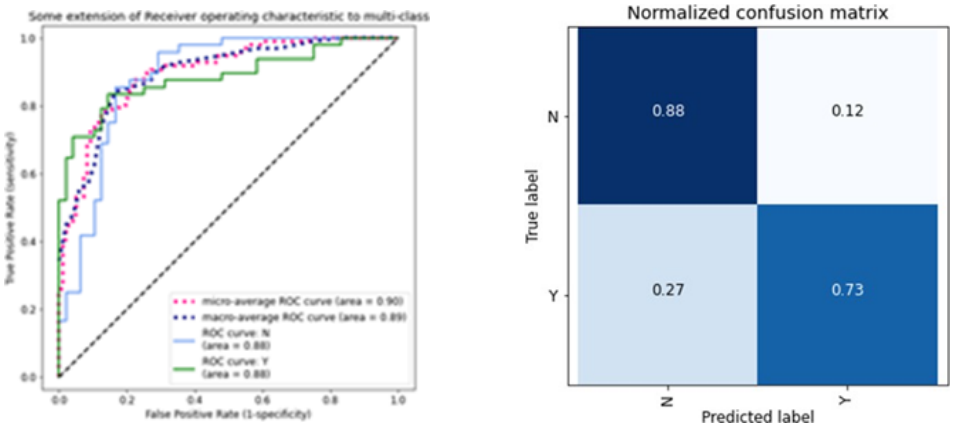
Fig. 11. a) Confusion matrix of the test set. b)Receiver operating characteristic curve plot

While the test data without the training can be used to realize the real situation of the trained model. The following table shows the recall, precision, and F1-score values of the test data. As can be seen, the precision N is 0.76, the precision Y is 0.85, the recall N is 0.88, the recall Y is 0.73 in the category F1-score of N is 0.82, and in the category, the F1-score of Y is 0.79 (as shown in Table 3).

Table 3. Accuracy, sensitivity, and specificity of the test set

Precision	recall	f1-score	support
N 0.76	0.88	0.82	48
Y 0.85	0.73	0.79	48
accuracy		0.80	96
macro avg 0.81	0.80	0.80	96
weighted avg 0.81	0.80	0.80	96

The following figure illustrates the chaotic matrix and the ROC curve for the verification test data (as shown in Figure 12).



a) b)

Fig. 12. a)Confusion matrix of the validation set. b) The curve of the receiver operating characteristic

With advances in machine learning, our future work will focus on many worthwhile applications and targets. For example, to predict the outcome of embryos at early time points of development, the utilization of research, neural networks help with daily clinical tasks and predict the outcome of networked embryos.

5. Conclusions

This study analyzed microscopic images of fertilized embryos using the Inception-ResNet-V2 algorithm. Inception-ResNet is a hybrid model that merges the Inception network with the Residual network. Currently, Inception-ResNet-V2 represents the most

advanced network architecture in the ImageNet dataset, incorporating transfer learning to build classification models. The model's accuracy was validated through K-Fold cross-validation and hierarchical analysis. Combining image enhancement techniques with hyperparameter adjustments, such as modifying image size, dropout layers, and learning rates, we developed a classification model that achieved a maximum accuracy of 80%, exceeding the typical success rate of manual selection.

Despite the promising results, few studies have applied deep learning techniques to the imaging of fertilized embryos. The classification model developed in this research provides a valuable reference for clinical applications, demonstrating improved accuracy in recognizing fertilized embryos. If further systematized, this classification model could serve as a highly effective tool to assist with the manual selection of embryos during the IVF process, thereby reducing workload and minimizing errors associated with manual selection. Ultimately, this would enhance the standardization and efficiency of the embryo selection process, benefiting infertility patients by increasing their chances of successful conception.

When deep learning is performed, the amount of data required is considerable because of the requirement for the machine to learn it within the model. Although the amount of data required for the adjustment of parameters is very large, it may not be possible to adjust the parameters effectively when the amount of data is too small. The number of datasets in this research is 460, which is still not enough for deep learning technology, yet the image quality is not consistent because the datasets are gathered in different years and the machine version has changed over time. Also, when capturing image data, some of them were manually cut, while others were programmed. While the time difference between the intercepted images may cause ambiguity or shift, the different sizes also affect the model training results. In the future, if we can acquire a higher quality and quantity of embryo images or obtain maternal biochemical information for supplemental analysis, it will help improve the precision of embryo selection.

References

1. Zegers-Hochschild, F., et al., *The International Committee for Monitoring Assisted Reproductive Technology.* Human reproduction, **24**(11): p. 2683–2687.(2009)
2. Kissin, D.M., et al., *Assisted reproductive technology surveillance* : Cambridge University Press.(2019)
3. Health Promotion Administration, M.o.H. and Welfare, *The Assisted Reproductive Technology Summary 2021 National Report of Taiwan*, Taiwanese Government Taipei.(2023)
4. Gnoth, C., et al., *Definition and prevalence of subfertility and infertility.* Human reproduction, **20**(5): p. 1144–1147.(2005)
5. Gurunath, S., et al., *Defining infertility|a systematic review of prevalence studies.* Human reproduction update, **17**(5): p. 575–588.(2011)
6. Yu Ng, E.H., et al., *High serum oestradiol concentrations in fresh IVF cycles do not impair implantation and pregnancy rates in subsequent frozen–thawed embryo transfer cycles.* Human Reproduction, **15**(2): p. 250–255.(2000)

7. Warnock, M., *Report of the committee of inquiry into human fertilisation and embryology*. Vol. 9314: HM Stationery Office.(1984)
8. Geyter, C.D. *European pregnancy rates from IVF and ICSI 'appear to have reached a peak'*.[cited 2025 9/4]; Available from: https://www.eshre.eu/Annual-Meeting/Vienna-2019/Media/2019-Press-releases/EIM. (2009)
9. Wennerholm, U.-B., et al., *Incidence of congenital malformations in children born after ICSI.* Human reproduction, **15**(4): p. 944–948.(2000)
10. Sullivan, E.A., et al., *Single embryo transfer reduces the risk of perinatal mortality, a population study.* Human Reproduction. **27**(12): p. 3609–3615.(2012)
11. Lieberman, B., *An embryo too many?* Human reproduction (Oxford, England), **13**(10): p. 2664–2666.(1998)
12. Gardner, D.K., et al., *Culture and transfer of human blastocysts increases implantation rates and reduces the need for multiple embryo transfers.* Fertility and sterility, **69**(1): p. 84–88.(1998)
13. Rienzi, L., et al., *embryo and blastocyst cryopreservation in ART: Systematic review and meta-analysis comparing slow-freezing versus vitrification to produce evidence for the development of global guidance.,23.* DOI: https://doi. org/10.1093/humupd/dmw038: p. 139–155.(2017)
14. Gardner, D.K., *Blastocyst culture: toward single embryo transfers.* Human Fertility, **3**(4): p. 229–237.(2000)
15. Ahlström, A., et al., *Trophectoderm morphology: an important parameter for predicting live birth after single blastocyst transfer.* Human reproduction, **26**(12): p. 3289–3296.(2011)
16. Gardner, D.K., et al., *Diagnosis of human preimplantation embryo viability.* Human reproduction update, **21**(6): p. 727–747.(2015)
17. Gardner, D.K. and B. Balaban, *Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': is looking good still important?* MHR: Basic science of reproductive medicine, **22**(10): p. 704–718.(2016)
18. Lee, T.-H., *The effect of microenvironmental free radicals on embryo development for assisted reproduction technology cycles*. (2012)
19. Henderson, I., et al., *Predicting the outcomes of assisted reproductive technology treatments: a systematic review and quality assessment of prediction models.* F&S Reviews, **2**(1): p. 1–10.(2021)
20. Vestergaard, E.T., et al., *The follicle-stimulating hormone (FSH) and luteinizing hormone (LH) response to a gonadotropin-releasing hormone analogue test in healthy prepubertal girls aged 10 months to 6 years.* European Journal of Endocrinology, **176**(6): p. 747–753.(2017)
21. Visser, D. and F.l.R. Fourie, *The applicability of the cumulative embryo score system for embryo selection and quality control in an in-vitro fertilization/embryo transfer programme.* Human Reproduction (Oxford, England), **8**(10): p. 1719–1722.(1993)
22. Papanikolaou, E.G., et al., *In vitro fertilization with single blastocyst-stage versus single cleavage-stage embryos.* New England Journal of Medicine, **354**(11): p. 1139–1146.(2006)
23. Majumdar, G., et al., *Relationship between morphology, euploidy and implantation potential of cleavage and blastocyst stage embryos.* Journal of human reproductive sciences, **10**(1): p. 49–57.(2017)
24. Jain, T., B.L. Harlow, and M.D. Hornstein, *Insurance coverage and outcomes of in vitro fertilization.* New England Journal of Medicine, **347**(9): p. 661–666.(2002)
25. Howell, E.P., et al., *Preconception evaluation before in vitro fertilization.* Obstetrical & gynecological survey, **75**(6): p. 359–368.(2020)
26. Medicine, P.C.o.t.A.S.f.R. and P.C.o.t.S.f.A.R. Technology, *Guidelines on number of embryos transferred.* Fertility and Sterility, **92**(5): p. 1518–1519.(2009)
27. Penzias, A., et al., *Guidance on the limits to the number of embryos to transfer: a committee opinion.* Fertility and sterility, **107**(4): p. 901–903.(2017)

28. Van Blerkom, J., P. Davis, and S. Alexander, *Differential mitochondrial distribution in human pronuclear embryos leads to disproportionate inheritance between blastomeres: relationship to microtubular organization, ATP content and competence.* Human reproduction, **15**(12): p. 2621–2633.(2000)
29. Cozzolino, M., D. Marin, and G. Sisti, *New Frontiers in IVF: mtDNA and autologous germline mitochondrial energy transfer.* Reproductive Biology and Endocrinology, **17**(1): p. 55.(2019)
30. Colavita, M. and G. Tanzer, *A cryptanalysis of IOTA′s curl hash function.* White paper, p. 1–13.(2018)
31. Wilding, M., et al., *Energy substrates, mitochondrial membrane potential and human preimplantation embryo division.* Reproductive Biomedicine Online, **5**(1): p. 39–42.(2002)
32. Lou, H., et al., *Does the sex ratio of singleton births after frozen single blastocyst transfer differ in relation to blastocyst development?* Reproductive Biology and Endocrinology, **18**(1): p. 72.(2020)
33. Loewke, K., et al., *Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos.* Fertility and sterility, **117**(3): p. 528–535.(2022)
34. Sujata, P.N., S. Madiwalar, and V. Aparanji. *Machine learning techniques to improve the success rate in in-vitro fertilization (IVF) procedure*. in *IOP Conference Series: Materials Science and Engineering*. IOP Publishing.(2020)
35. Brás de Guimarães, B., et al., *Application of artificial intelligence algorithms to estimate the success rate in medically assisted procreation.* Reproductive Medicine, **1**(3): p. 181–194.(2020)
36. Peng, C., et al., *Research of image recognition method based on enhanced inception-ResNet-V2.* Multimedia Tools and Applications, **81**(24): p. 34345–34365.(2022)
37. Barnett-Itzhaki, Z., et al., *Machine learning vs. classic statistics for the prediction of IVF outcomes.* Journal of assisted reproduction and genetics, **37**(10): p. 2405–2412.(2020)
38. Szegedy, C., et al. *Rethinking the inception architecture for computer vision*. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016)
39. He, K., et al. *Identity mappings in deep residual networks*. in *European conference on computer vision*. Springer.(2016)

Yu-Yu Yen, she has been working as a lecturer in the Center for General Education at Shih Hsin University since 2022. She is also currently enrolled in a PhD program in the Department of Bio- medical Engineering, National Yang Ming Chiao Tung University.

Shao-Ping Weng, is a specialist in reproductive medicine at IHMED International Healthcare. His clinical and research expertise includes recurrent in vitro fertilization (IVF) failure, advanced maternal age infertility, immunological infertility, and premature ovarian insufficiency.

Li-Jen Su, is the Director of the Service Learning Development Center at National Central University (NCU). He also serves as the Director of the Education and Research Center for Technology-Assisted Substance Abuse Prevention and Management at NCU, as well as the Core Facilities for High Throughput Experimental Analysis in the Department of Biomedical Sciences and Engineering at NCU.

Jui-Hung Kao, is an associate professor at Shih Hsin University since 2023. The empirical research topics focus on spatial data analysis, medical management research, and long-term medical policy.

Woei-Chyn Chu, earned his Ph.D. in electrical and computer engineering from the University of California at Irvine in 1991, he holds the position of distinguished professor in the Department of Biomedical Engineering at National Yang Ming Chiao Tung University in Taipei, Taiwan, ROC.Dr. Chu has made significant contributions through his patented work, particularly in integrating RFID technologies for clinical applications and real-time tracking.

Received: October 01, 2024; Accepted: May 18, 2025.

Computer Science and Information Systems 22(4):1687–1705 https://doi.org/10.2298/CSIS241002067L

A Study of Real-Time Operations by Converting Human Skeleton Coordinates to Digital Avatars

Fei-lung Lin[1], Jui-Hung Kao[2], Yu-Yu Yen[3,4,⋆], Kuan-Wen Liao[5], and Pu Huang[6,⋆]

[1] Institute of Technical and Vocational Education, National Taipei University of Technology, Taipei, Taiwan
t110499005@ntut.edu.tw
[2] Department of Information Management, Shih Hsin University, Taipei, Taiwan
kjhtw@mail.shu.edu.tw
[3] Center of General Education, Shih Hsin University, Taipei, Taiwan
melyen@mail.shu.edu.tw
[4] Department of Biomedical Engineering, National Yang Ming Chiao Tung University, Taipei, Taiwan
sheepkelly19.be11@nycu.edu.tw
[5] Department of Information Management, Shih Hsin University, Taipei, Taiwan
m111660002@mail.shu.edu.tw
[6] School of political science and law, Shaoguan University, Shaoguan, China
20201047@sgu.edu.cn

Abstract. This study aims to develop a real-time motion recognition system that translates skeletal human movements into a virtual environment. This will be achieved through the use of advanced tech-niques for the accurate capture of human skeletons and coordinate conversion. This paper investi-gates the acquisition and processing of motion data for virtual characters using depth cameras to obtain depth information. This study identifies six specific actions: left kick, right kick, left punch, right punch, squatting, and sitting. The experimental process successfully integrated RGB+D cameras, Media Pipe, and OpenCV into Unreal Engine models to capture and display human skeletal and joint positions in real-time. The experimental results show that the system achieved a precision of 100% for all motion detections, with an accuracy of more than 94%. How-ever, the recall rate for specific actions was lower, reaching 88%.

Keywords: Mixed Reality, Confusion Matrix, Motion Recognition.

1. Introduction

The application of image recognition technology has rapidly become a critical aspect of modern tech-nology, with the capacity to simulate and even surpass the capabilities of the human visual system. There is potential for advancement in several fields, including virtual reality (VR), augmented reality (AR), autonomous driving, and intel gent surveillance. Technology has the potential to enhance efficiency across a range of industries and

⋆ Corresponging authors

expand the scope of applications, thereby contributing to the advancment of these sectors. Nevertheless, accurately capturing and converting between physical and virtual environments remain a significant challenge, particularly in integrating digital avatars. While image recognition technology has demonstrated considerable success across various domains, technical challenges still exist to achieve higher precision and a more comprehensive application.

The increasing prevalence of mobile devices, such as smartphones and tablets, has led to a surge in demand for rapidly identifying captured subjects. This can be achieved by using cameras to scan (QR) codes, which can be used to obtain URLs or product information. Alternatively, document capture can facilitate text recognition and scanning functions. The aforementioned functionalities are made possible by image recognition technology, which also plays a pivotal role in the advancement of fields such as virtual reality (VR) and augmented reality (AR). VR technology enables the simulation of an entirely virtual environment, facilitating immersive user interactions. In contrast, AR technology overlaps virtual information with the real world, thereby achieving a fusion of reality and virtual elements.

The fundamental aspect of both technologies is the perception and recognition of the surrounding environment. This is paramount for determining the position and orientation of virtual objects, enabling interaction with the real world. This study aims to develop novel methodologies and techniques for precisely and accurately capturing the human skeleton. This process involves a thorough investigation of different types of sensors and imaging devices, instilling confidence in the research process. The selection and optimization of hardware is also a key part of this study, ensuring stable operation in a variety of scenarios. The proposed enhanced algorithms will improve the precision and practicality of the capture process. Furthermore, the study examines methods to incorporate additional human characteristic values to improve the reliability of skeletal capture. Another significant challenge is the efficient conversion of coordinate information into virtual space. Virtual space modeling techniques facilitate the real-time translation of physical human skeletal coordinates into actions within a virtual environment. This includes the calibration and transformation of coordinate systems to achieve consistency and accuracy. In addition, the data processing workflow was optimized to facilitate the efficient transformation of coordinates, thereby enabling real-time synchronization.

The system developed through this research is expected to open new avenues for virtual reality and gaming applications and extend to various fields such as social networks, remote conferencing, and virtual performances. The system enables users to interact in real-time within virtual spaces, participating in diverse activities through digital avatars. Technology has the potential to offer not only a novel entertainment experience but also commercial value and social impact. This could be achieved through technology to communicate with friends, attend virtual meetings, or perform online.

2. Materials and Methods

2.1. Human Motion Recognition

The field of computer vision has long been concerned with the recognition of human motion. Traditional methods have employed contour detection techniques to track the

human body and infer movements by calculating the torso's range of motion [1]. However, these conventional approaches have certain limitations, particularly their susceptibility to variations in camera angles and environmental backgrounds, which can lead to reduced stability in practical applications.

Several neural network models have been developed to recognize human motion. Among these models, the 2D Convolutional Neural Network (2DCNN)[2], which encompasses Convolutional Neural Networks (CNN) and Two-Stream Networks, has been extensively utilized in image recognition. However, these models are limited in processing image sequences with temporal information, which prevents them from in erring the temporal order of actions. To address this issue, some researchers have proposed recurrent neural network models capable of simultaneously learning temporal and spatial features. These include recurrent neural networks (RNN) [3], long-short term memory networks (LSTM) [4], and gated recurrent units (GRU) [5]. Furthermore, there are 3D Convolutional Neural Network (3DCNN) models [6] , including Inflated 3D ConvNet (I3D) [7] , Pseudo-3D Convolutional Neural Network (P3D) [8], and Separable 3D Convolutional Neural Network (S3D) [9], as well as models that incorporate attention mechanisms. However, these models typically have many parameters and high computational costs, which presents a significant challenge for their practical application in real-world scenarios.To achieve an optimal balance between accuracy and computational cost, researchers have proposed the implementation of new model architectures. The trade-off above models includes the first convolutional network (FstCN)[10], the temporal relation network (TRN) [11], the efficient channelized video model (ECO) [12], the multi-mode fusion network (MFNet), and the R (2 + 1) D convolutional neural network (R (2 + 1) D). These models have been designed to achieve high accuracy while maintaining computational efficiency, making them more suitable for practical human motion recognition applications. One motion recognition method, the Temporal Shift Module (TSM), is depicted in the accompanying illustration. This model can achieve the performance of 3DCNN models while only requiring the computational complexity of 2DCNN model.

The fundamental concept is to extract spatial features through a 2DCNN architecture, subsequently shifting the feature maps of selected channels by one frame in either the forward or backward direction of the temporal dimension. This approach integrates temporal and spatial information without needing temporal convolution operations. This method enables simultaneous learning of temporal features while preserving the integrity of spatial features, and it does so with a relatively low computational cost.

A motion recognition system solely focused on spatial and short-term temporal features may be susceptible to overlooking the significance of long-term temporal information for accuracy. Consequently, a neural network framework called Temporal Segment Network (TSN) [13] was devised to capture long-term motion information. The input video sequence is divided into K segments, with a randomly selected snippet extracted from each. Subsequently, each snippet is processed through a convolutional neural network, generating K classification scores. The above scores are subsequently integrated to create the final recognizable human action.

In video data processing, most CNN models are designed with the primary objective of two-dimensional image analysis. A straightforward approach treats the video as a sequence of static images, applying a 2DCNN for frame-by-frame recognition. However, this approach cannot capture motion information in the temporal domain. To effectively

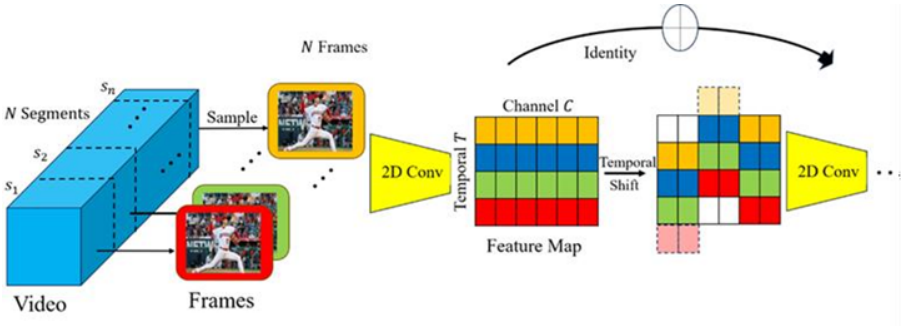


Fig. 1. TSM Architecture Diagram

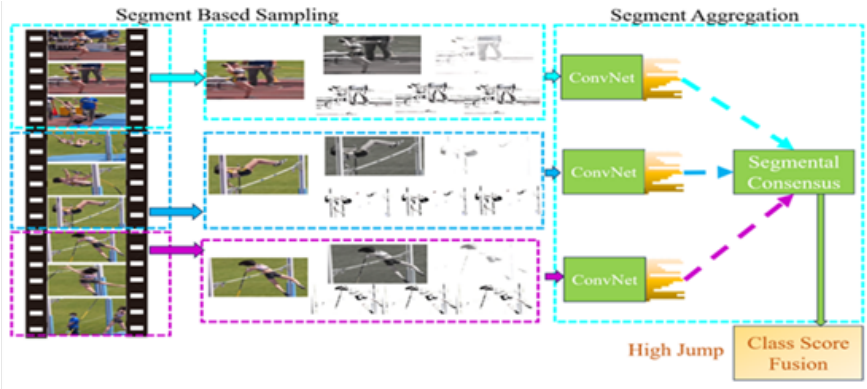


Fig. 2. TSN Architecture Diagram

incorporate temporal details, this study explores the use of 3DCNN to recognize human actions. The operation of a 3DCNN on the input video simultaneously learns spatiotemporal features, providing a more accurate representation of the video.

In the context of 3DCNN, small segments are extracted from consecutive input video frames for processing. This is exemplified in the following diagram. The convolution operation is performed by sliding a three-dimensional kernel over the input, simultaneously modeling temporal and spatial features. The data used by 3DCNN is sequential, typically comprising multiple frames of a video or a series of integrated segmented images. The input data is represented as a 3D cube, and the convolutional kernel is also a cube. The convolutional kernel performs sliding-window operations over the input data's spatial dimensions (length, width, and depth) to compute inner products, resulting in a single value in the output data. In the convolutional layer, each feature map is connected to multiple adjacent consecutive frames from the previous layer. This enables the capture of motion information. The formula for a 3D CNN is as follows:

$$\nu_{ijk}^{L} = \tanh\,[2061?](b^{1} + \sum_{m} \sum_{p=0}^{p-1} \sum_{q=0}^{Q-1} \sum_{r=0}^{R-1} \varpi_{pqr}^{m} \bullet v_{(i+p)(j+q)(k+r)}^{l-1}) \tag{1}$$

In this formula, ν_{ijk}^{L} represents the feature map value at position ([2148?],j,k) in the I-th layer, bI denote the bias term of the I-th layer, [D835?][DF14?][D835?][DC5A?] signifies the 3D convolution kernel weight of the m-th feature map in the (I−1)-th layer, and $\nu_{(i+p)(j+q)(k+r)}^{\iota-1}$ represents the feature map value at position ([2148?]+p , j+q , k+r) in the (I−1)-th layer. The tanh function is used as the activation function.

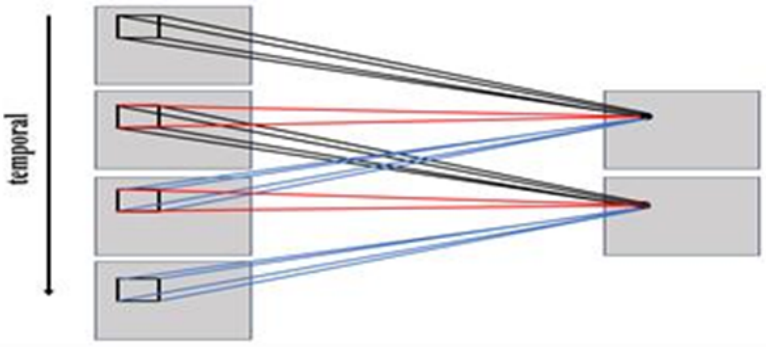


Fig. 3. 3D Convolution

2.2. Discussion on Target Detection and Human Skeletal Tracking Techniques

Target detection is a technology that identifies and locates specific objects or targets within images or videos. The successful application of deep learning, particularly convolutional neural networks (CNNs), has led to significant breakthroughs. The contem-porary methods of target detection are based on deep learning models, including the You Only Look Once (YOLO) [14] and Single Shot MultiBox Detector (SSD) algo-Rithms [15]. These

methods facilitate the rapid and precise recognition of targets. Themfield of human skeletal detection is concerned with the technology used to identify and track critical skeletal points of the human body in images or videos. There have been notable advances in this field in recent years due to the advent of deep learning- base methodologies. Several skeletal models, including OpenPose [16] and AlphaPose [17],have been used in various applications.

In the field of motion capture, various technologies such as OpenPose,DeepLabCut [18], and Kinect SDK [19] are widely used, but they have limitations in terms of computational resource requirements, applicability, and deployment flexibility. OpenPose offers high accuracy and multi-person pose estimation capabilities; however, its high computational cost makes real-time applications challenging. DeepLabCut is designed for biomedical applications and can improve recognition accuracy for specific subjects through transfer learning, yet it relies heavily on high-performance GPUs for training and inference, making it unsuitable for multi-object detection. Kinect SDK, on the other hand, depends on specialized hardware, limiting its adaptability in diverse environments. In contrast, MediaPipe [20] and OpenCV [21] provide superior efficiency and cross-platform compatibility, making them ideal for real-time motion capture applications. MediaPipe features a built-in Pose Landmarker, capable of running on both CPUs and GPUs, offering lightweight and efficient pose esti mation. OpenCV provides robust image processing capabilities for data preprocessing and feature extraction. Compared to other technologies, MediaPipe and OpenCV are more suitable for low-resource environments while ensuring stable real-time analysis, making them the preferred choice for this study.

This study selected OpenCV and MediaPipe as the primary tools for implementing the detection of the target above and the human skeletal detection methods. OpenCV provides a comprehensive image processing library, while MediaPipe offers efficient human skeletal detection capabilities. OpenCV (Open-Source Computer Vision Library) is a software library that assists developers in processing and analyzing various images and videos to perform multiple computer vision tasks. OpenCV has evolved significantly, becoming a cross-platform tool that supports many programming languages, including C++, Python, and Java. RGB image processing represents a fundamental function within the OpenCV framework. RGB stands for the red, green, and blue channels, and mixing these three colors at different intensities can represent a wide range of colors. In OpenCV, images are stored in matrix form, with the color information of each pixel represented by three matrices corresponding to the red, green, and blue channels. OpenCV splits an RGB image into three separate color channel matrices when reading it. These matrices contain the intensity values for each pixel in each channel. Once the image has been read, the intensity distribution can be observed by displaying the image for each channel. Separating the RGB channels of the image produces three separate greyscale photos, each representing the intensity of a one-color channel. In contrast, these three separate channels can also be combined into a single RGB image.

This study also used MediaPipe, an open-source visual computing framework developed by Google, used primarily for various vision-related tasks such as pose estimation,hand tracking, and face recognition. It uses CNNs to perform a detailed analysis of human images, accurately identifying key skeletal points. These models are trained on large datasets of images annotated with human skeletal points. With the optimizations and enhancements MediaPipe provides, they can efficiently and reliably perform full-body

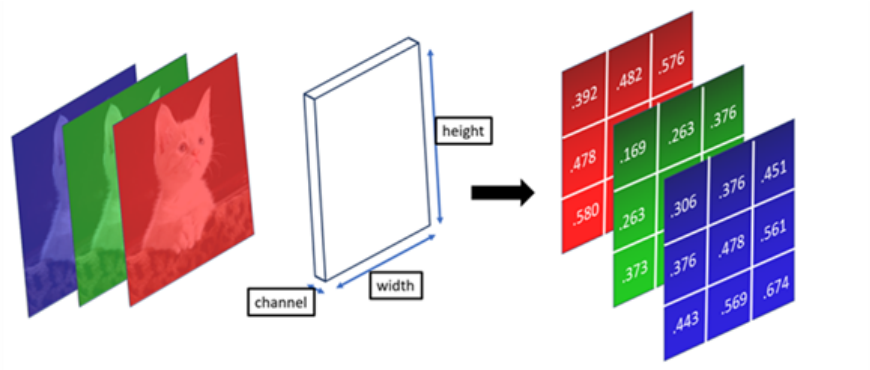


Fig. 4. OpenCV Processing Diagram

skeletal recognition in real-time video streams. The core concept of these techniques is to detect critical points on the human body (such as the head, shoulders, elbows, wrists, knees, and ankles) in images, thus accurately reconstructing human posture.

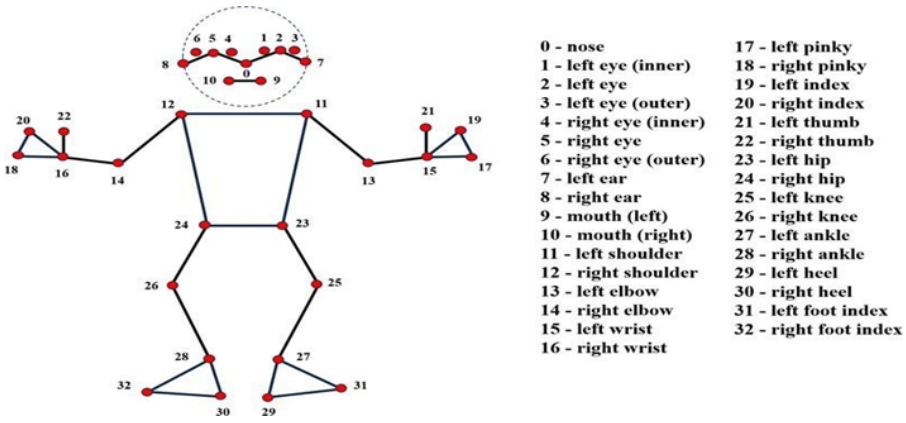


Fig. 5. MediaPipe Full-Body Skeletal Keypoints

3. Research Methodology

3.1. Research Process

Human figure recognition and the use of 2D skeletal information for feature correction and fitting are first discussed. It then explores how depth cameras can complement 3D human skeleton estimation and coordinate system transformation. The system's use of depth image information to obtain the critical points of the skeleton's third axis and to perform a

coordinate system transformation that facilitates the conversion of skeletal rotation values is explained. Next, we discuss the method for calculating the skeletal orientation. This involves using the 3D skeletal data to construct vectors and convert the skeletal rotation values. This step is critical for accurately describing the direction of human motion. Then, data transfer and numerical conversion are examined. The data transfer methods and numerical conversion techniques for skeletal data used in this study are described in detail, including how the calculated skeletal data is transferred to the Unreal Engine and how consistency and accuracy of the data are achieved. Finally, following the overall research process and framework, each step will be detailed, from data acquisition, processing, and skeletal estimation to the final presentation of the results. Each stage's functionality and synergistic interactions will be described to show this system's complete architecture and operational process.

3.2. 2D Human Skeleton Estimation and Feature Adjustment

The human location, skeletal positioning, and feature-matching process using technologies such as OpenCV and MediaPipe are detailed. The literature mentions using OpenCV's object recognition models to accurately identify individuals' position and appearance features and using MediaPipe's human skeletal models to locate human joint points accurately. Additionally, this study highlights three user-defined vital points added during the skeletal positioning process to enhance the convenience of subsequent computations and applications. In practice, the accuracy of depth information requires appropriate positional adjustments and calibrations. This includes maintaining

the optimal distance between the camera (Intel RealSense D435i) and the subject (approximately 2 meters), and making the necessary corrections and adjustments based on the actual environment.

During the measurement, the position and angle of the camera should be adjusted to minimize potential measurement errors. These adjustments help improve the depth of the information's accuracy and increase the data's reliability in application scenarios. To further improve the accuracy of depth information, it is recommended that multiple tests and calibrations are performed in different application environments to determine the optimum camera settings and operating methods. This approach effectively reduces errors caused by environmental variations, thus improving the overall stability and accuracy of the system. This process involves three main steps: object detection, skeleton positioning, and feature adjustment.

(1) Use OpenCV's object recognition models for human localization and identification.

(2) MediaPipe's human skeleton model positions the skeleton on the human image identified by OpenCV. Based on deep learning, MediaPipe's pose estimation model can accurately locate human joints from pictures and obtain the XY coordinates of each joint. The model adapts and adds three new vital points: the center of the shoulders, the center of the hips, and the center of the torso.

(3) Feature adjustment and correction for misidentified individuals are performed based on skeletal and joint position information. This includes eliminating or correcting potential errors in joint positioning, such as misidentifying parts of the scene or nonhuman targets as humans, which could lead to misapplication of the skeletal model and abnormal

values. These adjustments ensure that the information and feature values obtained more accurately reflect the proper posture and structure of the target individual.

3.3. Transformation of skeleton rotation value

In virtual character motion simulation, accurate and smooth rotation calculations are crucial for user experience. Euler angles, often used for rotation, can suffer from gimbal lock, which limits the system's ability to control rotation freely. To avoid this, quaternions are commonly used in fields like computer graphics, robotics, and motion capure, as they can represent rotations without gimbal lock and provide smoother interpolation [22]. A quaternion consists of a real part ω and three imaginary parts x, y, z, and can represent a 3D rotation. The quaternion is defined as:

$$q = \omega + xi + yj + zk \tag{2}$$

Where ω is the cosine of half the rotation angle and (x, y, z) are the components of the rotation axis multiplied by the sine of half the angle. To calculate the rotation of a skeleton, we first identify two points on a bone, say A and B, with initial coordinates. An array is used to record the three-dimensional coordinates of all skeletal points in this initial frame. The second identified frame is designated as the latest, and the 3D coordinates of all skeletal points are recorded in another array. The 3D coordinates identified in the second frame will continuously update this array until the program terminates. Given that the skeleton comprises numerous segments, each formed by two points, the direction of a bone segment following rotation can be determined by calculating the vectors formed by the two endpoints of the segment between the initial frame and the most recent frame. This process yields a series of vectors that represent the skeleton's direction of rotation. These vectors can then be used to calculate the skeleton's rotation, with the rotation values being represented as quaternions.

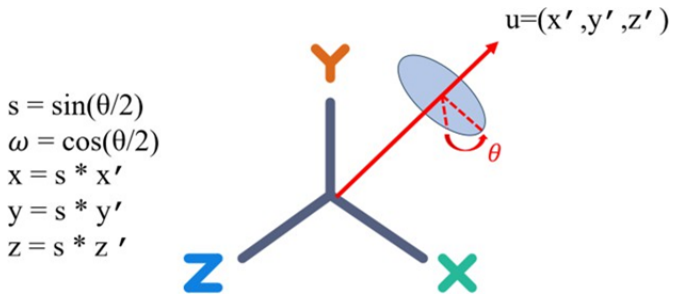


Fig. 6. Quaternion Illustration

The quaternion for this rotation is derived as:

$$q = (\cos\tfrac{\theta}{2}, \sin\tfrac{\theta}{2}\widehat{r}) \tag{3}$$

where [D835?][DC5F?]ˆ is the unit vector of the rotation axis, and θ is the angle between the vectors. If the norm of the cross product of the vectors is zero, it implies the vectors are parallel, and no rotation is needed. In such a case, the quaternion is set to [1, 0, 0, 0]. However, if the vectors are not parallel, the rotation axis is calculated as the unit vector of the cross product, and the rotation angle is the angle between the vectors. This information is used to generate the corresponding rotation quaternion. Through this process, we can accurately describe and apply the rotation from point A to point B in the motion reproduction of virtual characters, ensuring smooth and realistic motion without gimbal lock, and enhancing the overall animation quality [23].

3.4. Data Transmission and Numerical Conversion

It is of the utmost importance that this process undergoes rigorous data accuracy and transmission stability verification. It is of the utmost importance that this process be carried out to guarantee the correct application of external data within the Unreal Engine. This process enables the accurate tracking and presentation of the skeletal direction and coordinates. The precision and fluidity with which virtual characters can perform movements depend on the data's accuracy and the reliability of their transmission. This ensures that users will experience enhanced interaction with the virtual environment. Once a successful socket connection has been established, the Received Message event will initiate the reception of the overall skeletal rotation values transmitted from Python. Rotation values are associated with nine distinct segments of the skeletal system. The data related to each segment comprises seven values: the world coordinates (X, Y, Z) and the rotation quaternion (X, Y, Z, W). The primary rationale for selecting

the quaternion method is that it offers a straightforward and accurate approach to handling rotational data. Although traditional Euler angles are intuitively appealing, they are susceptible to gimbal lock issues during calculations, which increases the complexity of data processing. On the contrary, quaternions circumvent this issue, offering a stable and accurate description of rotations, dealing with rotational data. Although traditional Euler angles are intuitively appealing, they are susceptible to gimbal lock issues during calculations, which increases the complexity of data processing. On the contrary, quaternions circumvent this issue, offering a stable and accurate description of rotations.

The diagram below illustrates how the first three Read float nodes are employed to receive the world coordinates of the skeletal points. The coordinate data represent the positions of the skeletal rotation points in three-dimensional world space. Subsequently, the data from the aforementioned skeletal rotation points is transmitted to the Make Vector node, which is incorporated into the location array. This approach effectively manages and stores all the skeletal rotation points' position coordinates. The final four Read float nodes receive the quaternions of the skeletal rotation points. The quaternions thus describe the rotational direction of each skeletal rotation point, representing its rotational state in the world coordinates. Subsequently, the quaternions of these skeletal rotation points are transmitted to the Make Quat node and incorporated into the quaternion array, ensuring accurate storage and utilization of the quaternion data. Subsequently, the quaternions are converted into the rotation type. This design enhances the efficiency and accuracy of data processing and reduces the need for data conversion in subsequent applications. Finally, the Received Message event is linked to nine quaternion methods, thus completing the

data processing. This process ensures that all data points are correctly received and processed, guaranteeing the system's stable operation and high performance.

To align the spatial coordinates identified by the camera with those of the Unreal Engine, this study devised a rudimentary human skeleton within the scene (level). The skeletal information from the TCP socket blueprint is called the Level Blueprint, with the location and rotation information being passed sequentially to each skeleton joint [24]. This ensures the precise synchronization of the data. Once the initial configuration has been completed, the skeletal data will be imported into the relevant pins. This process enables the data captured by the camera to be transmitted to the corresponding skeleton within Unreal Engine [25]. Subsequently, the skeletal data are transmitted to the scene, where they are converted to the skeleton, as illustrated in the schematic diagram. Despite the use of world coordinates by both the camera and the scene in Unreal, an offset exists due to differing origins. Consequently, direct use of the coordinates captured by the camera will result in positional errors. To guarantee the synchronization of coordinates, it is necessary to incorporate an offset into the location data before its transmission to the skeleton. This step aims to correct the origin offset, ensuring data consistency and accuracy. Consequently, the discrepancies above are eliminated, thus eliminating errors caused by differences in the origins of the coordinate system.

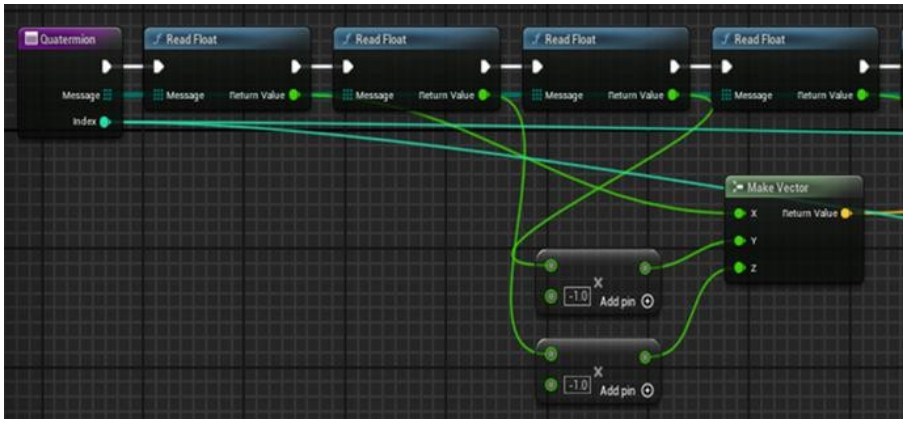


Fig. 7. Socket Connection Event Blueprint 1

4. Research Results and Discussion

The approach proposed by Lao et al. [26], which centers around virtual characters in theatrical performances, aims to enhance learners' language abilities and stimulate their interest in learning. The value of this method lies in its ability to help learners better understand the language through realistic performances and interactions with virtual characters, while also improving learning outcomes via immediate feedback and personalized interaction. The effectiveness of virtual characters in language learning has been demonstrated

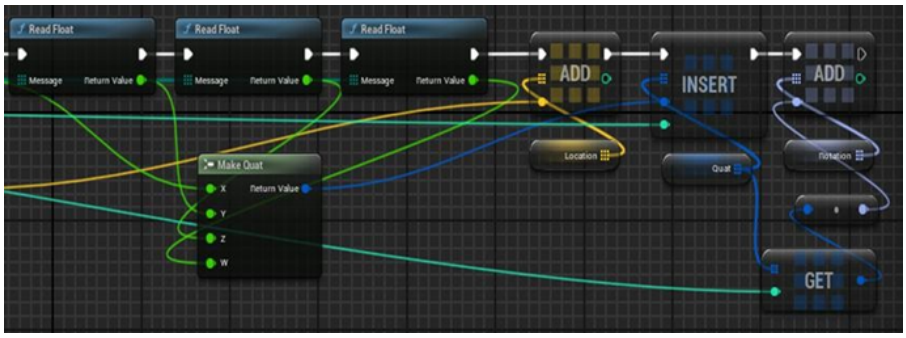


Fig. 8. Socket Connection Event Blueprint 2

in numerous studies. This approach not only simulates real-life communication scenarios, allowing learners to practice in a safe environment, but also creates an interactive learning atmosphere, further increasing learners' engagement and motivation.

In this study, six typical movements of virtual characters (such as kicking left, punching right, squatting, etc.) were selected as research subjects. These movements represent fundamental human motions and are frequently used in virtual character performances, making them highly relevant and practical. These movements were chosen because they effectively test the technology of converting human skeleton coordinates into digital avatars and ensuring that the virtual character's motions accurately reflect the user's actual movements HTML [27]. The experiment used a self-developed system to record the user's motion data in real-time and convert this data into the movements of a digital avatar in Unreal. The goal was to verify the accuracy of the technology that converts human skeleton coordinates and synchronizes movements in real-time. The system design is based on previous research, such as open-source skeleton tracking and real-time animation synchronization technologies, which have been shown to effectively capture and reproduce motions. Recording videos during the experiment helps with subsequent analysis and comparison, providing rich data support to ensure the accuracy of the technology and facilitate continuous improvements.

The analysis of the videos showed that the system could accurately reproduce the user's movements with high precision and low latency, confirming the effectiveness and feasibility of the proposed method [28]. This not only achieves ideal results in the performance of the virtual character's movements but also enhances the interactive experience for the learner. In conclusion, virtual characters have great potential in language learning and other application scenarios. These technologies are expected to be further applied in various fields, bringing innovative changes to education, entertainment, and social industries. To verify this study's primary objective, the system's accuracy in recognizing six specific actions, left kick, right kick, left punch, right punch, squatting, and sitting, will be evaluated using a confusion matrix as the primary analysis tool and multiple tests. The objective is to obtain accurate data to improve the reliability of the test results.

4.1. Experimental Environment

The experimental environment for the system in this study begins with the configuration and activation of the server side to ensure that the server runs correctly. Subsequently, the user establishes a connection by inputting the server's fixed IP address, ensuring stable data transmission from the server to the user side. The server is primarily responsible for processing and managing the data in this process. The server receives data from input devices, performs the necessary data processing and analysis, and then transmits the processed data to the user side. It is paramount that the server is capable of efficient processing and stability, as these are the two key factors that ensure that the system functions correctly.

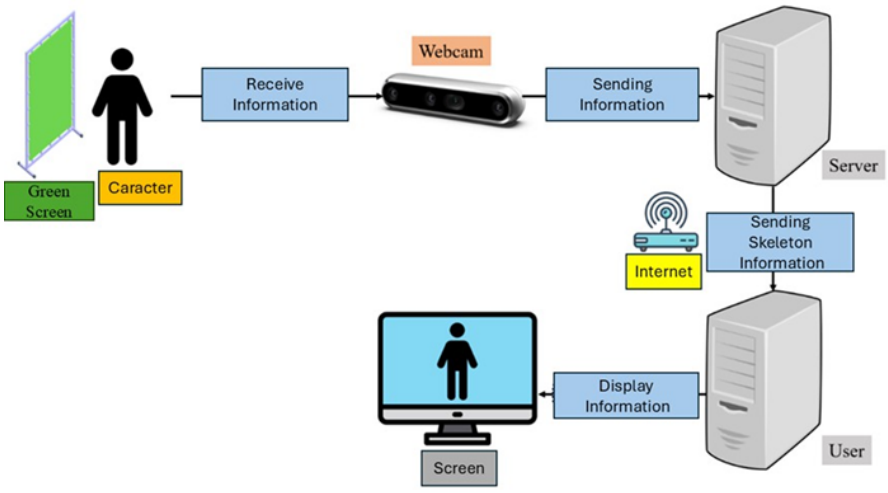


Fig. 9. Schematic Diagram of the Experimental Environment

The user interface connects to the server, which transmits the required data. Subsequently, the data is employed to simulate and regulate the virtual character's movements. It is of the utmost importance that the user interface can receive data from the server in real-time, as this data is used to simulate the virtual character's movements to ensure that the character's actions are accurately and smoothly reproduced. It is of the utmost importance that the connection between the server and the user side is stable and reliable, as this is a prerequisite for the system to function optimally. Any instability in the connection could result in interruptions or delays in data transmission, consequently affecting the performance of the virtual character's movements. It can be reasonably concluded that the server and user environments must exhibit high stability and speed to ensure optimal performance.

4.2. Experimental Evaluation Metrics

This study employs the confusion matrix as the principal metric for experimental evaluation to quantify the system's performance in capturing and synchronizing human skeletons in virtual space. The confusion matrix is a widely employed methodology for assessing the efficacy of classification models, offering a range of metrics, including precision and recall, to comprehensively evaluate the system's performance. The confusion matrix is a specific tabular structure used to describe the performance of a classification model in classification tasks. The model's classification results are presented by comparing the actual and predicted classifications.

True Positive (TP): The number of instances where the actual action is performed, and the Unreal model correctly displays the action.

True Negative (TN): The number of instances where the actual action is not performed, and the Unreal model correctly displays that the action is not performed.

False Positive (FP): The number of instances where the actual action is not performed, but the Unreal model incorrectly displays the action as being performed.

False Negative (FN): The number of instances where the actual action is performed, but the Unreal model incorrectly displays that the action is not performed. The confusion matrix enables the determination of the system's performance, which is evaluated using the following metrics.

Accuracy: This represents the proportion of correct predictions made by the model and is an essential indicator of overall performance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

Precision: represents the proportion of samples predicted to be a specific action that is an action, reflecting the reliability of the model's predictions for that action.

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

Recall: represents the proportion of samples with a specific action correctly predicted as that action, reflecting the model's sensitivity.

$$Recall = \frac{TP}{TP+FP} \tag{6}$$

Next, the system's accuracy in recognizing six specific actions described in the literature is analyzed. These actions include kicking left, kicking right, punching left, punching right, crouching, and sitting. The analysis will determine whether these actions are accurately rendered in Unreal Engine.

4.3. Experimental Testing and Analysis

This study focuses on the system's accuracy in recognizing six types of action out of 100 instances. We will calculate and analyze various metrics in detail, including accuracy, precision, and recall, to evaluate the system's recognition performance. These metrics will demonstrate the system's accuracy in recognizing these six actions and reveal any

potential recognition errors, providing a basis for subsequent system improvements. The analysis starts with 100 instances of video recording and analysis of

27 the right-kicking action (followed by the other five actions). Figure 9 shows the correct identification of the right kick action, while another figure shows the incorrect identification of the same action.

Fig. 10. Schematic Diagram of the Experimental Environment

In this analysis of action recognition, we found that the system made some errors in identifying the action kick right. Out of 100 instances, the system correctly identified kick right 45 times but made five incorrect identifications. Additionally, 50 actions were not kick right, and the system correctly identified all 50 actions as not kick right. Therefore, the number of non-kick right actions misidentified as kick right is zero. Below is the confusion matrix for this action analysis and the calculation results for the confusion matrix of the kick right action.

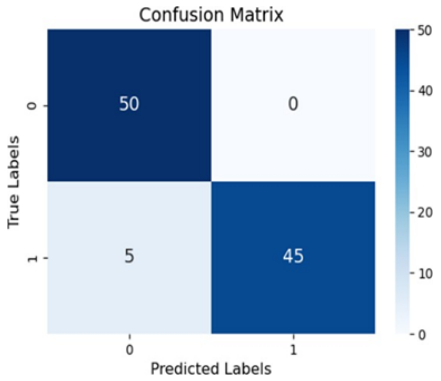


Fig. 11. Kick Right Confusion Matrix and Calculation Results

4.4. Experimental Testing and Analysis

To evaluate the system's accuracy in recognizing six actions, we analyzed experimental data from 100 instances. We calculate each action's accuracy, precision, and recall to evaluate the system's recognition performance. As shown in Table 1, the experimental results indicate that although the system performs well in recognizing most actions there is still room for improvement.

Table 1. Results of the experimental analysis of the confusion matrix

	Precision (%)	Accuracy(%)	Recall(%)
Kick Left	96	100	92
Kick Right	95	100	90
Punch Left	97	100	94
Punch Right	95	100	90
Squat	94	100	88
Sit	98	100	96
Overall Average	95.84	100	91.67

We can draw the following conclusions based on a comprehensive analysis of the above results.

High accuracy: The system achieved a 100% accuracy rate for all actions, which was almost always correct once the system identified an action. This reflects the system's strong ability to avoid false positives.

Insufficient recall: Despite the high accuracy, the recall rate was relatively low, especially for certain actions such as 'crouching' and 'kicking right,' with recall rates of 88% and 90%, respectively. This indicates the need for further adjustments to improve the system's recall, demonstrating our commitment to its ongoing development.

Overall high accuracy: The average accuracy rate for all actions was more than 94%,

demonstrating the overall stable recognition performance of the system. However, the accuracy of individual actions must be improved to achieve a more comprehensive recognition capability.

5. Conclusions

The system developed in this study demonstrated a precision rate of 100% for all action recognitions. This indicates that it was almost always correct once the system identified an action. This reflects the system's robust capacity to avoid false positives, particularly in recognizing everyday actions such as walking, running, and waving. The system demonstrated the ability to identify these actions with a low incidence of misclassification accurately.

Despite the high precision, the recall rate was relatively low, particularly for specific actions such as crouching and kicking right, with 88% and 90% recall rates, respectively. This indicates that the system occasionally fails to identify these actions correctly. Further optimization may be required for these actions' feature extraction and recognition models.

For instance, the system continues encountering difficulties recognizing complex and fast actions. This requires the acquisition of additional training data and the implementation of optimized algorithms to enhance recall rates. The system demonstrated an accuracy rate of more than 94% for all actions, indicating stable and reliable recognition performance. This suggests that the system performed consistently and reliably in recognizing most actions, whether simple or complex sequential movements, maintaining a high accuracy rate. However, further enhancements are necessary to achieve a more comprehensive recognition performance, particularly with respect to the accuracy of specific individual actions.

This study's experimental results demonstrate the system's potential in action recogition while identifying specific areas for improvement. Collective data indicate that the system achieved a precision rate of 100%, an accuracy rate of 95.84%, and a percent recall rate of 91.67%. These figures demonstrate that the system has a high recognition capability, yet there is room for improvement. Further research can be conducted to build upon these findings to optimize the system's accuracy and stability in recognizing various actions, thereby enhancing its applicability in real-world scenarios. In future experimental procedures, individual differences in participants' physical characteristics will be progressively considered, and environmental variables during the experiments—such as temperature, clothing effects, and body shape—will be controlled to maintain the system's accuracy at a consistent level. Additionally, in terms of future research directions, building on the system's current foundation, we aim to use the recognition of these six basic movements as a basis to develop more refined motion representations in the next phase. As the number of recognized basic movements increases, it may become feasible to apply the system to recognize sequences of continuous actions—such as Tai Chi or wellness exercises—thereby adapting the system to a wider range of application scenarios.

References

1. Sathe, P.S., *Tracking, Recognizing and Analyzing Human Exercise Activity*. University of Akron.(2019)
2. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, **25**.(2012)
3. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning representations by back-propagating errors.* nature, **323**(6088): p. 533–536.(1986)
4. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation. **9**(8): p. 1735–1780.(1997)
5. Cho, K., et al., *Learning phrase representations using RNN encoder-decoder for statistical machine translation.* arXiv preprint arXiv:1406.1078.(2014)
6. Tran, D., et al. *Learning spatiotemporal features with 3d convolutional networks*. in *Proceedings of the IEEE international conference on computer vision*. (2015)
7. Carreira, J. and A. Zisserman. *Quo vadis, action recognition? a new model and the kinetics dataset*. in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017)
8. Qiu, Z., T. Yao, and T. Mei. *Learning spatio-temporal representation with pseudo-3d residual networks*. in *proceedings of the IEEE International Conference on Computer Vision*. (2017)
9. Xie, S., et al. *Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification*. in *Proceedings of the European conference on computer vision (ECCV)* (2018)

10. Long, J., E. Shelhamer, and T. Darrell. *Fully convolutional networks for semantic segmentation*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2015)
11. Zhou, B., et al. *Temporal relational reasoning in videos*. in *Proceedings of the European conference on computer vision (ECCV)*.(2018)
12. Zolfaghari, M., K. Singh, and T. Brox. *Eco: Efficient convolutional network for online video understanding*. in *Proceedings of the European conference on computer vision (ECCV)*.(2018)
13. Wang, L., et al. *Temporal segment networks: Towards good practices for deep action recognition*. in *European conference on computer vision*. Springer.(2016)
14. Bochkovskiy, A., C.-Y. Wang, and H.-Y.M. Liao, *Yolov4: Optimal speed and accuracy of object detection.* arXiv preprint arXiv:2004.10934.(2020)
15. Liu, W., et al. *Ssd: Single shot multibox detector*. in *European conference on computer vision*. Springer.(2016)
16. Cao, Z., et al. *Realtime multi-person 2d pose estimation using part affinity fields*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2017)
17. Fang, H.-S., et al., *Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time.* IEEE transactions on pattern analysis and machine intelligence, **45**(6): p. 7157–7173.(2022)
18. Mathis, A., et al., *DeepLabCut: markerless pose estimation of user-defined body parts with deep learning.* Nature neuroscience, **21**(9): p. 1281–1289.(2018)
19. Izadi, S., et al. *Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera*. in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. (2011)
20. Lugaresi, C., et al., *Mediapipe: A framework for building perception pipelines.* arXiv preprint arXiv:1906.08172, (2019)
21. Bradski, G. and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc.(2008)
22. Shoemake, K. *Animating rotation with quaternion curves*. in *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*.(1985)
23. Alvarado, E., D. Rohmer, and M.P. Cani. *Generating Upper-Body Motion for Real-Time Characters Making their Way through Dynamic Environments*. in *Computer Graphics Forum*. Wiley Online Library.(2022)
24. Edeline, K., et al., *Using UDP for internet transport evolution.* arXiv preprint arXiv:1612.07816. (2016)
25. Qiu, W., et al. *Unrealcv: Virtual worlds for computer vision*. in *Proceedings of the 25th ACM international conference on Multimedia*. (2017)
26. Huang, X., et al., *A systematic review of AR and VR enhanced language learning.* Sustainability, **13**(9): p. 4639.(2021)
27. Younes, M., *Learning and simulation of sport strategies (boxing) for virtual reality training*, Université de Rennes.(2024)
28. Yan, Z. and J. Yi, *Dissecting Latency in 360 Video Camera Sensing Systems.* Sensors, **22**(16): p. 6001.(2022)

Fei-lung Lin, he was born and raised in Taiwan. Currently studying at the Institute of Technological and Vocational Education, National Taipei University of Technology, Taiwan. A strong personal interest in science, technology, and educational environments drives my pursuit of academic research.

Jui-Hung Kao, is an associate professor at Shih Hsin University since 2023. During his tenure as project manager at the Research Center for Humanities and Social Sciences in

2014, he was responsible for the administrative business of research and program execution, which combined statistical methods with spatial information visualization and is good at writing programs and data analysis. The empirical research topics focus on three parts: spatial data analysis, medical management research, and long-term medical poicy.

Yu-Yu Yen, she has been working as a lecturer in the Center for General Education at Shih Hsih University since 2022.She is also currently enrolled in a PhD program in the Department of Biomedical Engineering, National Yang Ming Chiao Tung University.

Kuan-Wen Liao, was born and raised in Taiwan; a fervent interest in the intersection of technology and management has driven my academic and professional pursuits. In July 2024, I was honored to receive my Master's degree in Information Management from the distinguished Shih Hsin University.

Pu Huang, has been a faculty member at Shaoguan University since September 2020. His research fields such as administrative management, social governance, and public policy analysis.

Received: October 02, 2024; Accepted: June 27, 2025.

Computer Science and Information Systems 22(4):1707–1756 https://doi.org/10.2298/CSIS241130068L

Implementing Persona in the Business Sector by A Universal Explainable AI Framework Based on Byte-Pair Encoding

Zhenyao Liu[1,⋆], Yu-Lun Liu[2], Wei-Chang Yeh[2], and Chia-Ling Huang[3]

[1] School of Economics and Management, Taizhou University
Taizhou 225300, Jiangsu Province, China
zyliu@tzu.edu.cn
[2] Integration and Collaboration Laboratory, Department of Industrial Engineering and Engineering Management, National Tsing Hua University
Hsinchu 300044, Taiwan
yulun.lawrence@gmail.com
yeh@ieee.org
[3] Department of International Logistics and Transportation Management, Kainan University
Taoyuan 33857, Taiwan
clhuang@mail.knu.edu.tw

Abstract. In the commercial realm, particularly for businesses targeting consumers (B2C), the challenge of acquiring and retaining valuable potential customers is paramount. As chip technology continues to advance at breakneck speed, in line with Moore's Law, various innovative AI technologies have emerged, yet this also highlights the infamous "black-box" issue. Naturally, this has paved the way for the rise of Explainable AI (XAI) and machine learning. In response, this study proposes a universal explainability framework to tackle both the black-box conundrum and the limitation of customer list sizes. The framework leverages the fundamental Byte-Pair Encoding (BPE) algorithm from large language models to tokenize natural language data, integrating the results into customer data as feature columns, thereby constructing comprehensive Persona. Crucially, domain experts are involved in the model-building process, selecting and recommending features. These experts utilize depth-first search to identify additional, similar feature columns, which are then used as target categories for machine learning models. The final step involves classification tasks and prediction evaluations. The proposed framework demonstrates its effectiveness and generalizability through validation on public datasets, increasing the number of potential customers by 7.5 times compared to traditional modeling approaches. In case studies, the framework outperforms customer lists generated by experts based on past experience, yielding 2.4 times more customers, 3.8 times higher response rates, and 9 times more total respondents. More importantly, both the model-building process and predictive outcomes are interpretable through domain knowledge, enabling businesses to transfer experience and expertise, thus laying a solid foundation for large language models within the industry.

Keywords: Natural Language Processing, Byte-Pair Encoding, Persona, Explainable Machine Learning, Business Sector.

⋆ Corresponding author

1. Introduction

In the aftermath of the COVID-19 era, technological advancements have accelerated various industries. However, the adoption of technologies such as machine learning has progressed slowly in certain sectors [1, 2] . Many scholars argue that despite the improvements in tools and computational capabilities over the years, the impact of these technologies has not reached the expected levels and continues to shape our daily lives [3, 4]. Although artificial intelligence has become an indispensable part of modern life, the use of opaque "black-box" models in highly sensitive domains—such as healthcare, biomedicine, public policy, human life, judicial rights, finance, consumer products, and any area related to personal privacy and life—remains particularly problematic [5].

Consequently, literature on Explainable AI (XAI) began to experience exponential growth from 2020 onward. Research in this area spans various fields, including digital health, law, public transportation, finance, and defense, reflecting the increasing recognition of the importance of transparency and interpretability [5]. On the other hand, major technology companies released numerous large language models (LLMs) in 2023, which gained considerable popularity. These include OpenAI's ChatGPT [6] and Meta AI's Llama [7, 8], with ChatGPT alone amassing over 180 million users [6].

In fact, large language models (LLMs) have not only introduced technological transformations in domains directly related to natural language, such as customer support [9], search engines [10, 11], and text translation [12, 13], but have also been broadly applied across other interdisciplinary areas, including medicine, code assistance, education, and finance [14–17]. This signifies that LLMs possess adaptability and potential for language-related tasks across various industries and environments.

With the pulse of technological advancement, both the economy and society have undergone significant changes, profoundly altering consumer shopping and retail capabilities [18]. For businesses targeting the consumer (B2C) market, competition has become increasingly fierce, making the retention of valuable and indispensable potential customers crucial [18]. In customer relationship management, whether in automotive, aviation, retail, or e-commerce sectors, unique methods of customer segmentation are employed. Various techniques are utilized to predict future demand peaks and adjust pricing and marketing strategies, all with the aim of gaining a deeper understanding of consumer behavior and habits [19].

Beyond e-commerce and retail, the financial industry is also a prominent example of B2C commerce. Marketing campaigns are one of the primary methods for achieving corporate objectives and are crucial for banks in attracting and retaining customers. Moreover, if a company's marketing activities or strategies are not executed effectively, it can face significant challenges in meeting annual targets [20], which in turn can impact overall business performance [21] and corporate profitability [22]. In any industry focused on sales strategies, it is essential to gain a deeper understanding of each consumer, including their purchasing habits and preferences [23], and to develop appropriate marketing strategies based on their buying patterns and attributes. Marketing strategies aim to deliver greater value to both customers and the company at a lower cost. In the business realm, failing to carefully consider the process through which potential consumers purchase or receive products can lead to wasted resources [24]. Consequently, calculating the return on investment (ROI) of marketing expenditures across activities and strategies such

as physical advertising, promotional campaigns, and digital advertising is a complex yet crucial issue for decision-makers [18].

Given the aforementioned context, decision-makers or domain experts responsible for marketing strategies often reject the use of potential customer lists generated by models, primarily due to concerns about cost and career development. They tend to distrust "black-box" models and are apprehensive about the possibility of these technologies replacing their roles [25]. Furthermore, as many companies lack additional funds for validating the lists produced by these models, skepticism regarding the validity of model-generated lists persists. Consequently, experts prefer to rely on their own experience to plan marketing activities and identify potential customers [26], choosing to preserve their job security while potentially causing the company to lag behind technological advancements.

This study aims to address the issue of domain experts' reluctance to accept marketing lists generated by models by proposing an explainability framework. In addition to leveraging the fundamental principles of large language models (LLMs) to provide interpretability to data through natural language, previous literature has also employed RFM models to enhance users' understanding of data [27]. Furthermore, techniques such as LIME and SHAP can be used to supplement the interpretability of model results [28], and the use of graph-based co-occurrence descriptions can elucidate the weights and relationships between features [29], thereby improving the efficiency of information retrieval.

Moreover, involving domain experts in the construction of models to help them understand the significance of the predictive results can not only increase the number of potential customers but also enable decision-makers and experts to connect marketing activities with corporate value. This fosters trust in the model-generated results and alleviates the tension between machine learning and domain expertise [25], while also preventing manipulation of marketing variables [18]. Meeting these conditions will facilitate the integration of decision-makers' and experts' domain knowledge and experience into the models [30, 31], thereby enhancing the company's value and position in the era of large language models.

Therefore, this study develops a universal explainability framework to address issues related to domain experts' inability to accept model-generated lists and the limitations on the number of items in these lists. The framework will incorporate the following functionalities and conditions:

- The framework proposed in this study is designed to be applicable to any B2C business within the commercial sector. It will enable companies to obtain potential customer lists that are both understandable and interpretable, and that exceed the number of potential customers typically identified through conventional experience and models.
- This framework must possess both reproducibility and generalizability, allowing any industry dealing with natural language data to apply it in order to provide additional interpretability to their data and models.
- This approach leverages the fundamental principles of large language models (LLMs), specifically, tokenization algorithms, to provide additional and effective feature columns to natural language data. This enhancement makes customer data more descriptive, thereby improving readability for users and facilitating a clearer understanding of customer purchasing behavior and habits.

- In the steps of the explainability framework, involving domain experts in understanding the operation of the framework and the model-building process not only facilitates the transfer of their knowledge and experience into the model but also reduces the tension between their professional status and technological advancements.
- The predictive results of this method should surpass those generated based on past experience in terms of list size, response rate, and overall number of respondents. Additionally, the method should be applied in practical cases to achieve more effective, diverse, and precise customer relationship management and marketing strategies.

2. Preliminaries

2.1. XAI and XML

As the applications of AI and ML become increasingly widespread, the methods have also grown in complexity. Consequently, business stakeholders have become more concerned about the potential drawbacks of these models, including data-specific biases [32]. To address these concerns, Lundberg et al. introduced SHapley Additive exPlanations (SHAP) as an industry standard for interpreting machine learning models [33]. However, such interpretability often falls short of satisfying most users, leading to the consideration of post-hoc explanation methods, such as textual explanations, visual explanations, and example-based explanations [34, 35].

Due to the "black box" issue inherent in artificial intelligence and machine learning, three key elements have emerged to define XAI and XML.

- **Transparency**: A ML method is considered transparent if the model itself is easy to understand and the extraction process is transparent. This encompasses model transparency, design transparency, and algorithmic transparency [36].
- **Interpretability**: Users must be able to understand the basis on which algorithmic decisions are made within the model. They should also be capable of explaining the algorithmic criteria and hyperparameter variables within the model in comprehensible terms [36].
- **Explainability**: The definition of this element varies [36], but it is commonly understood as the user's ability to explain why the model made a specific decision, understand the rationale behind a particular prediction, and even integrate domain knowledge with the prediction to provide contextual explanations. This deeper understanding is essential for achieving true explainability [36].

With the rapid advancement of XAI and XML, their applications have become increasingly widespread across various fields [37].
In the medical field: Soares et al. utilized computed tomography (CT) scans to identify COVID-19 [38]. Morais et al., in collaboration with domain experts, examined the performance of XAI in cancer diagnosis from the perspective of experts, offering explanations that extend beyond the experts' viewpoint [39].
In the field of public policy and the judicial system: Dressel and Farid highlighted the widespread use of criminal risk assessment systems. They emphasized the necessity of providing explanations in key decisions within these systems to maintain fairness and avoid racial bias [40].

Applications based on natural language processing are also being explored in the research domain. Several authors have improved user trust in applications through the use of XAI techniques for anomaly and fraud detection [41]. Additionally, Mathews proposed an interpretable tweet classification method based on LIME, which enhances the explainability of application results, thereby increasing user engagement and trust [42].

A significant portion of applications is found in autonomous driving systems. In a fully automated system, the driving system is expected to operate in any unknown environment [43], which impacts trust and transparency compared to black-box systems. Therefore, from the perspectives of public perception and trust, as well as regulatory and legal considerations, XAI is critically important. Transparency, interpretability, and explainability are essential for developing more reliable, safe, and regulation-compliant autonomous driving systems [43].

In other domains, Murindanyi et al. utilized four tree-based machine learning methods and four standard machine learning methods to predict customer churn at Czech banks. By incorporating post-hoc explanation techniques such as LIME and SHAP, they achieved satisfactory predictive results [44]. Clement et al. proposed the XAIR process for the development of XAI, which mirrors the five steps of software development: requirements analysis, design, development, evaluation, and deployment. This process is presented as a comprehensive framework for other scholars to reference [45].

From the literature review presented in this section, it is evident that both XAI and XML share many similar elements and principles. XAI or XML fundamentally relies on three key elements: transparency, interpretability, and explainability. The application of XAI or XML in commercial domains is relatively limited, as these fields place greater emphasis on domain experts' experience. While techniques such as LIME and SHAP have demonstrated effective explanatory capabilities in literature, they may still be deemed insufficient by experts lacking data-related knowledge, leading to a lack of persuasion and practical application in industry. Additionally, due to the cost sensitivity in commercial sectors, extensive experimental costs and expenditures for model validation are often unacceptable. However, if interpretability frameworks can improve the reliance on experience in commercial fields, they are likely to contribute significantly and offer future advancements in these domains.

The issue of marketing lists being rejected by domain experts, as examined in this study, will be addressed through an interpretability framework that meets the three key elements: transparency, interpretability, and explainability. This framework will be designed to extract natural language data from customers, transforming the extracted natural language results into customer personas. Additionally, it will incorporate recommendations from domain experts to form a comprehensive solution.

2.2. Persona

Many studies have indicated that constructing persona aids in better understanding user needs, thereby facilitating personalized and precise information services [46].
Given that the early development of persona was driven by the needs of designers, scholar Travis, who specializes in user experience research, provided the following definition of the persona extraction process [47]:

- **Primary Research**: Whether the persona is determined based on real customer data or contextual interviews.

- **Empathy**: Whether the persona evokes user understanding and empathy by incorporating elements such as names, photographs, or product-related descriptions.
- **Realism**: Whether the persona appears authentic to experts in the field or frontline personnel who interact directly with customers.
- **Singularity**: Whether each persona is unique in its composition and distinct from other characters.personnel who interact directly with customers.
- **Objectives**: Whether the persona includes product-related goals and provides key descriptions that articulate these objectives.
- **Quantity**: Whether the number of personas meets the team's requirements, is sufficient for the team to remember their characteristics, and designates one persona as the primary character.
- **Applicability**: Whether the team can practically apply the persona in decision-making processes.

The seven elements outlined above play a crucial role in the effective implementation of persona techniques. They offer sufficient flexibility and adaptability, enabling practitioners to creatively explore and develop various applications of personas in practice [48].

To fulfill various objectives, personas are widely applied in software design [49], advertising [50], and technology products [46]. However, the ultimate aim remains that these personas should effectively inform and guide planning and decision-making processes [51].

Although the definition of personas is relatively broad, a review of the literature reveals that most scholars agree that personas are inherently goal-oriented. Practitioners must have a clear understanding of the purpose behind persona extraction and whether the resulting personas fulfill the initially defined objectives. Moreover, the process of developing personas should ensure that the seven essential elements are met, thereby ensuring that the personas align with expectations.

In practical applications, beyond obtaining personas through pre-classified data combined with statistical and regression methods, many scholars also employ clustering and supervised learning techniques for persona extraction. Some even derive personas from predictive outcomes. However, these approaches often fall short of achieving the initially set objectives [49, 50].

The explainability framework proposed in this study differs from traditional methods of persona extraction in the literature. It utilizes BPE to extract customers' natural language data, directly generating personas that enable experts and decision-makers to describe their behaviors and characteristics. These personas will meet the criteria of the seven key elements, ensuring not only realism and uniqueness but also alignment with the intended objectives. Additionally, BPE enhances the model's transparency, interpretability, and explainability.

2.3. The BPE

When discussing Byte Pair Encoding (BPE), it is common to reference the increasingly popular large language models (LLMs) in recent years. These models, characterized by an extensive number of parameters, are designed to understand and process natural language by modeling the semantics and probabilities of text sequences within vast datasets.

Through pre-training tasks, such as Masked Language Modeling or autoregressive prediction, large language models learn to comprehend and generate natural language effectively [52].

A well-designed pretrained transformer language model requires the implementation of various subword tokenization methods [53], among which the most renowned is BPE [54].BPE, proposed in 1994, is a straightforward data compression technique that employs a single unused byte to iteratively replace the most frequent pair of bytes in a sequence [54].

The following are the steps involved in BPE:
Step 1: we initialize the symbol vocabulary using the character vocabulary and represent each word as a sequence of characters, appending a special end-of-word symbol to the end of each word. This symbol aids in restoring the original state after translation.
Step 2: we begin iteratively calculating the frequency of all symbol pairs, where a symbol pair is a combination of each character in the vocabulary. The most frequent symbol pair is then replaced with a new symbol. For instance, if the most frequent pair is A and B, it will be replaced by a new symbol AB. In the subsequent iteration, A and B are ignored, and the frequency is calculated using AB in combination with other characters.
Step 3: each occurrence of a new symbol represents a merging operation. In other words, each merging operation generates a new symbol, which also signifies an n-gram of characters.

From the above steps and explanation, it is evident that an increase in the number of merging operations results in a larger symbol vocabulary and a corresponding increase in the granularity of the characters [55].
The BPE method merges the most frequent pairs of symbols in the entire text. Although it may appear as though BPE is performing a form of word concatenation, this is actually due to the high frequency of certain pairs. These high-frequency pairs persist and thus appear as concatenated sequences. Consequently, the most frequent pairs in the text will become prevalent in the final vocabulary. This characteristic of BPE is also why it can be effectively applied to various languages.

After understanding the operation and fundamental principles of BPE, one might consider why, for English text tokenization, spaces are not used for segmentation. From a human perspective, using spaces for tokenization seems to be the most intuitive approach. However, employing spaces or punctuation marks for segmentation results in an excessively large vocabulary. Any variation of a word would be included in the vocabulary, and if a word has multiple forms, the vocabulary size can grow exponentially. Such a large vocabulary necessitates an enormous matrix for input and output layers, increasing both memory and computational complexity [56].
Consequently, various tokenization algorithms avoid using spaces or punctuation for segmentation. This is why the BPE algorithm includes an end-of-word symbol in its implementation, a practice that also contributed to GPT-2 achieving optimal performance during its initial training [57].

In summary, BPE is a tokenization method, also referred to as a segmentation algorithm, and serves as a preprocessing technique for natural language data. It can also be applied to address the Out-Of-Vocabulary (OOV) problem [55] in natural language processing. Tokenization involves the mechanism of segmenting or dividing sentences and words into their smallest possible units [58].

The application domains of BPE include various fields. In the realm of language translation, BPE is characterized by its adaptability to different languages [59]. Additionally, numerous practical use cases of BPE have been documented: in the field of network diagnosis and detection [55], the medical and healthcare sector [60], experiments involving symbolic music for music generation and composer classification [61], and addressing the linguistic complexity on social media platforms [62].
This study proposes an interpretability framework for the business domain. The framework employs an improved version of BPE to enhance data feature dimensions, transforming tokenization results into feature fields to form Personas. By incorporating domain experts' recommendations on target fields, the framework utilizes Depth-First Search (DFS) to expand feature fields.

2.4. The DFS

DFS is a technique that has been extensively applied as a solution method for problems in combinatorial theory and AI [63]. The search process of DFS is closely related to graph theory, necessitating the introduction of certain graph-related definitions. These definitions are derived from Harary's research [64].

Let G be a graph, such that $G = (V, E)$, where V is the set of nodes and E is the set of edges. The set E consists of unordered pairs of nodes, each representing an edge. When manually drawing a graph, nodes are typically represented by circles, and the connections between these circles correspond to the edges.

Suppose we aim to search through the graph G. Initially, none of the nodes in G have been explored. We begin at an arbitrary vertex and select an edge to follow, traversing it to reach a new node, and continue this process. At each step, we choose an unexplored edge leading from the current node and traverse it. Once an edge has been traversed, it will not be explored again. This process continues until all edges in G have been traversed exactly once. This procedure constitutes the search [65].

The detailed steps of the DFS algorithm are presented below. Please refer to Figure 1 for illustration.

In summary, DFS is a graph traversal method that begins at an arbitrary vertex and explores as far as possible along each path before backtracking to visit any unvisited vertices when no further progression is possible [66].

From the perspective of text classification and Information Retrieval (IR), the concept of weighting is also applied. Blanco and Lioma proposed a graph-theoretic approach applied within the IR field, where text is modeled as a graph with edges representing the relationships between words. These relationships are then assigned corresponding weights. This method has been shown to perform on par with standard techniques in IR [29]. To address Word Sense Disambiguation (WSD), Rahmani et al. developed an unsupervised co-occurrence graph based on a corpus, which does not rely on the inherent structure and properties of the language. In other words, ambiguous words are assigned additional weights, altering the contextual structure [67].

DFS is also applied in practical cases. Du et al. proposed an algorithm that combines deep convolutional neural networks with DFS to address the problem of identifying power outage locations. In their approach, convolutional networks are used as a safety assessment tool, followed by DFS to find suitable interruption path locations. This method not only improves accuracy but also performs thousands of times faster than traditional

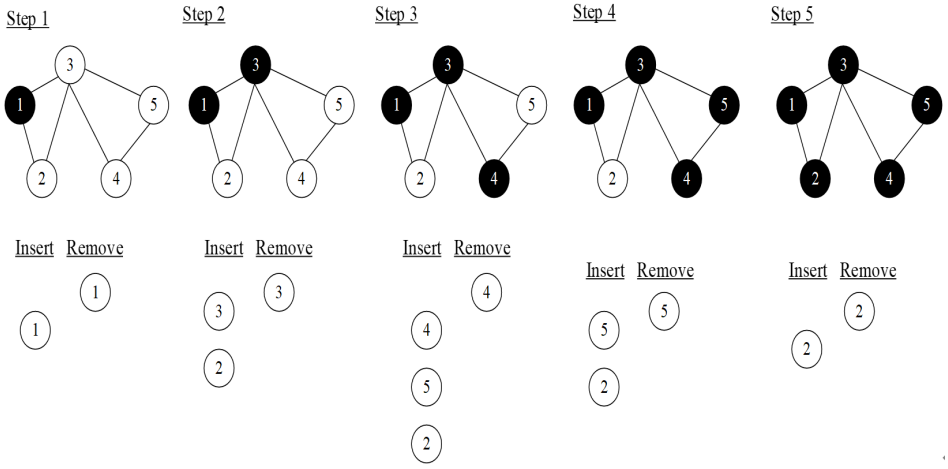


Fig. 1. The steps of the DFS algorithm

methods [68]. Mei and Gül introduced a new approach for detecting crack patterns in foundational design. After enhancing the CNN model, DFS is used for post-processing to remove isolated pixels and improve accuracy [69].

In summary, DFS is a graph-based search method that involves exploring data from one vertex to another, delving deeper into subsequent vertices, and backtracking to unvisited vertices to traverse the entire graph. Therefore, in the field of IR, many scholars use graphs to represent statements by altering edge weights. This approach not only provides models with enhanced features but also improves context and addresses issues with specialized vocabulary. In various practical cases, the characteristics of DFS combined with models enable the identification of diverse root causes and assist expert systems in effectively proposing solutions.

This study proposes a generalizable interpretability framework for application in the commercial domain. The framework enhances feature dimensions using BPE and, in conjunction with features provided by domain experts, employs DFS to identify similar feature fields. Features are represented in a graph, with edge weights based on co-occurrence to illustrate the relationships between features. The results of the DFS, combined with expert recommendations, are treated as categorical answers for a Light Gradient Boosting Machine (LightGBM). Finally, the framework performs a classification task to predict a list of potential customers highly similar to the expert recommendations.

2.5. Light Gradient Boosting Machine, LightGBM

LightGBM is an algorithmic framework based on Gradient Boosting Decision Tree (GBDT) [70]. This algorithm employs a leaf-wise tree growth strategy, designed for maximum efficiency, offering faster training speeds and minimal memory usage when handling large datasets [71].

During the model training process, decision trees are employed to generate base classifiers, and the weight parameters for each classifier are calculated iteratively. The final

model is then constructed by integrating all the base classifiers and their corresponding weights. This can be expressed by the following equation, where $f_m(X)$ represents the base classifiers and ∂_m denotes the weight parameters for each classifier, as shown in Equation (1) [72].

$$f_m(X) = \partial_1 f_1(X) + \partial_2 f_2(X) + \cdots + \partial_m f_m(X) \tag{1}$$

LightGBM offers superior predictive performance and memory efficiency compared to other classification algorithms [73]. LightGBM is highly effective in handling class imbalance issues and demonstrates strong performance in such scenarios [74]. It significantly enhances the predictive accuracy of Intrusion Detection Systems (IDS) and is notably efficient in flow classification tasks [75]. Moreover, when addressing class imbalance issues, methods such as Synthetic Minority Oversampling Technique (SMOTE) can be employed to adjust the sample distribution, yielding excellent results [76]. In the business domain, numerous studies have demonstrated that LightGBM outperforms other algorithms in terms of precision and F1 scores [77]. Additionally, applications of LightGBM often incorporate the RFM model to include customer purchasing behavior as additional features or utilize RFM combined with clustering algorithms to categorize customers before making predictions [23]. Regardless of the specific application, LightGBM consistently delivers outstanding classification performance. Although the study employs the SMOTE technique to address class imbalance, its application to extremely large-scale datasets—such as the case study involving 185 million transaction records—may hinder model generalization due to computational inefficiency and the potential introduction of noisy synthetic samples.

The interpretability framework proposed in this study involves several key components: enhancing feature dimensions through Byte Pair Encoding (BPE), incorporating expert recommendations, and utilizing Deep Feature Synthesis (DFS) to derive similar features. The final classification task is performed using LightGBM, with precision, recall, and F1 score serving as the evaluation metrics for users. Given that class imbalance is a common issue in business applications, the SMOTE technique is employed to adjust the sample distribution. Additionally, to enhance feature representation and interpretability in persona, the RFM model's monetary value is incorporated into the feature set, providing new interpretative dimensions. There are also some alternative approaches:

- **Stratified Sampling and Cost-Sensitive Learning**: During data preprocessing, oversampling the minority class or incorporating class weights during model training (e.g., using the `scale_pos_weight` parameter) can help mitigate imbalance more efficiently.
- **Ensemble Methods**: Combining undersampling techniques (e.g., RandomUnderSampler) with boosting algorithms (e.g., RUSBoost) can reduce redundancy in the majority class while preserving performance.
- **Application of Focal Loss**: Introducing dynamic weights into the loss function can down-weight the contribution of well-classified (majority class) samples and emphasize learning from hard (minority class) examples.

3. Research Methodology

To propose an interpretable framework to address the issues of customer lists being unacceptable and limited in number in the business domain, the framework will utilize the characteristics of natural language to extract customer labels. These labels will not only possess industry knowledge and interpretability but also serve as personas. By augmenting the data features with these labels and incorporating them into model training, we aim to obtain predictive results for classification tasks.

The method proposed in this study consists of three main steps:
The first step is to identify the objectives and obtain relevant raw data, which represents customer-related feature data.
The second step involves the framework proposed in this study, which first preprocesses the data using Byte-Pair-Encoding (BPE). This process extracts fact tags (F-tags) from the raw data and adds them as feature fields. Experts then define target tags (T-tags) based on the objectives, domain knowledge, and fact tags. Subsequently, Depth-First Search (DFS) is employed to identify tag combinations based on the target tags, and experts determine derivative tags (D-tags) from these results. Finally, the derivative tags are used as the basis for actual class labels, making them the target variables in the model.
The third step involves model prediction and value evaluation. By assessing the model's accuracy, recall, and F1 score, the next steps are determined. If the metrics do not meet the standards, model parameters, DFS parameters, or tags are adjusted, and the model is retrained. If the standards are met, special customers are excluded, and the resulting list is evaluated against the objectives identified in the first step. If the list does not meet the objectives, the process returns to the third step and repeats until the objectives are satisfied.

This section sequentially introduces the implementation details of the proposed framework, named Tag-Framework: Section 3.1 discusses the definition and acquisition of labels and experts. Section 3.2 explains the adjustments made to Byte-Pair Encoding (BPE) to find more root results and analyzes its complexity. Section 3.3 provides detailed explanations and examples of the adjustments made to Depth-First Search (DFS) to find tag combinations similar to the target tags. Section 3.4 analyzes the improvements to DFS in terms of time and space complexity. Section 3.5 describes the evaluation metrics for the model and the process of value evaluation.

3.1. Definition and Acquisition of Labels and Experts

Labels will vary depending on their source: first, apart from being cleaned and preprocessed according to the characteristics of the data, all data must undergo BPE preprocessing. Moreover, the generation of labels relies on the involvement of experts, specifically domain experts. According to the research and definition by Wong et al., domain experts typically lack training in data analysis, visualization, and statistics [78]. Such experts may include sociologists who analyze social phenomena in their work, sales professionals familiar with certain types of products or marketing strategies, or individuals who have deliberately practiced in areas like chess, music, healthcare, or education [79]. These experts possess advanced knowledge, business rules, and processes within their respective fields, serving as the primary source of information for the team [80], but they usually have limited awareness of technical aspects such as visualization or technology [78].

The domain experts in this study were selected based on a case study approach, involving three credit card marketing project managers (PMs) and three managers from the investment products department, each with over eight years of experience in the financial industry. These experts participated in the experiment, model development, and label selection.

Fact Tag (F-tag): Derived from the original data through BPE processing, unsuitable tags and function words are excluded. The results are then matched with extracted fields using regular expressions (regex). If a match is found, the corresponding subword is retained as a fact tag. The fact tag will serve as a feature field in the original data, with the field's value determined by the data characteristics or as a binary $0/1$ value.

Target Tag (T-tag): Determined by experts from the F-tags based on domain knowledge or target characteristics, the T-tag can consist of one or more fact tags. Experts select the T-tag from precise feature fields; otherwise, the selected value may not correspond to any existing feature, and the features are determined through experience from the data to meet the definition of an interpretable model.

Derivative Tag (D-tag): Using the T-tag as the root node, DFS is employed to identify tag combinations that are similar to the T-tag. DFS includes two parameters: depth ($pair_{len}$) and similarity ratio ($pair_{proportion}$). Differences in parameters and simple examples are detailed in Section 3.3. Upon discussion with experts, the tag combination can then be finalized as the derivative tag. Assuming DFS as the Approx function, the tag set and corresponding formulas are represented as shown in Equations (2) and (3).

$$T_{\text{tag}} \subseteq F_{\text{tag}} \tag{2}$$

$$D_{\text{tag}} = \text{Approx}(T_{\text{tag}}, F_{\text{tag}}) \tag{3}$$

According to the formulas, the T_{tag} is a subset of the F_{tag}, while the D_{tag} is derived from tag combinations identified by DFS that are close to the T_{tag}, with the final decision made by experts. In simple terms, the F_{tag} is generated by BPE, the T_{tag} is selected from the F_{tag} and determined by experts, and the D_{tag} is derived from various tag combinations found by DFS, with the final decision also made by experts, as illustrated in Figure 2.

3.2. Adjustment and Complexity Analysis of BPE

One of the objectives of this framework is to enable preprocessing of any natural language data to ensure its generalizability. Therefore, based on the conclusions drawn from the literature review, the BPE tokenization algorithm was selected for adjustment. Using alternative methods, such as splitting by punctuation or whitespace, would reduce generalizability, limiting applicability to other languages and potentially exceeding hardware constraints. Since BPE is not the most frequent tokenization method, even terms that appear only once in the dataset would be transformed into feature columns in subsequent steps, which would significantly impact memory usage [57]. Furthermore, when a term that occurs only once in the dataset becomes a feature column, it leads to a situation where only one record holds a value while all others are 0.

Another reason for adopting BPE in this study's framework is its application in the commercial sector. In addition to general consumer electronics and household products, many financial product names lack spaces and punctuation rules, and most are phrases consisting of single sentences, such as: "Green Power Global ESG Green Power ETF

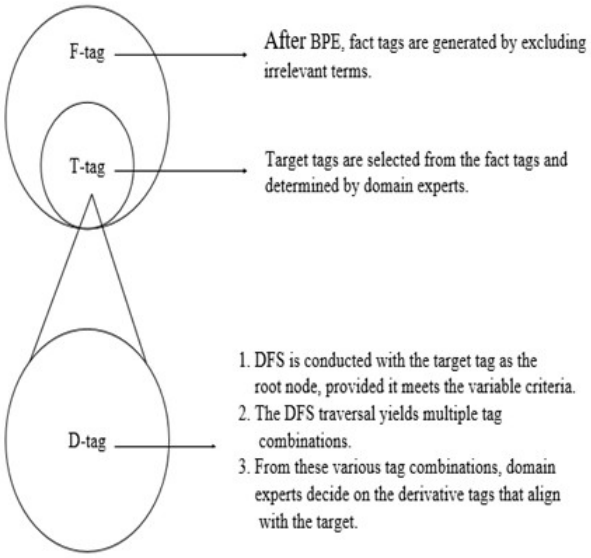


Fig. 2. Label Relationship Diagram

Fund",“US Treasury 20-Year U.S. Government Bond 20+ Year Fund",“Japan Leveraged 2x Tokyo Stock Exchange Daily Fund (Currency Hedged)",“Super Enjoy Life Variable Annuity Insurance",“Triple True Medical Hospitalization Insurance (Outward Type)",“BNP Paribas 12-Month Non-Principal-Protected Structured Note",“Franklin Templeton AI New Technology Fund N (Accumulation) (USD) (Back-End Load)", and “BlackRock Emerging Markets Bond Fund (Stable Distribution) (Monthly Distribution) (AUD Hedged)". If traditional tokenization methods were used, it would either require a customized dictionary or fail to generate reasonable morphemes, which would hinder their conversion into customer labels or feature columns. However, the characteristics of BPE can effectively resolve this issue.

The original BPE algorithm employs 2-gram characters to obtain the most frequent words. To ensure generalizability across languages such as Chinese and English, as well as specialized terminology in various industries, the byte size was modified to iterate over the data in forms of 2, 4, 6, and 8 grams. After extracting the most frequent words using 2-gram characters, the process continues with 4-gram, 6-gram, and 8-gram iterations to capture a wide range of morphemes. This approach ensures that the proposed framework can successfully extract morphemes from datasets in any natural language. The pseudocode is illustrated in Figure 3.

This code is divided into three phases, which will be explained in detail below, along with an analysis of their time and space complexities. Data Processing Phase: This phase includes converting full-width characters to half-width, removing non-essential symbols, and converting all text to lowercase to ensure data consistency. Since each character must be processed individually, the time complexity of this process is proportional to the length of the input data, denoted as $O(n)$.

```
class ExtractFtag:
    Method __init__(eng_series, eng_regex=r'[()|\[\]]'):
        Normalize eng_series to lowercase, remove specified symbols, and convert full-width characters to half-width.
        Split normalized text into words.
        Flatten list of words.

    Method strQ2B(ustring):
        Convert full-width characters in ustring to half-width.
        Return converted string.

    Method get_bpe_vocab_count(str_in_list):
        Calculate and return frequency count of sequences in str_in_list, appending '</w>'.

    Method get_stats(vocab):
        Compute and return frequency of adjacent character pairs in vocab.

    Method merge_vocab(pair, v_in):
        Merge specified character pair in v_in, update vocabulary.
        Return updated vocabulary.

    Method run_bpe(iter_num, length_thresholds=[2, 4, 6, 8]):
        For each length threshold, refine vocabulary using BPE:
            Initialize vocabulary.
            For each iteration:
                Merge highest frequency character pair.
                Record character pair if it meets current length criterion and is not previously recorded.
```

Fig. 3. Extraction of Fact Labels and BPE Pseudocode

Word Frequency Construction Phase: In this phase, the algorithm traverses the entire text to create a table mapping each unique word to its corresponding frequency. If the text contains m unique words, the time complexity of this process is $O(m)$. At the end of this phase, additional space is required to store both the processed text and the word frequency table, resulting in a space complexity of $O(m + n)$.

Iteration and Merging Phase: The core of BPE lies in repeatedly iterating and merging the most frequently occurring pairs of characters until either the iteration limit k is reached or no more character pairs can be merged. During each iteration, the algorithm calculates the frequency of all possible character pairs and selects the most frequent pair for merging. In the worst-case scenario, each iteration involves a comprehensive search through all the words, resulting in a time complexity of approximately $O(m^2)$ per iteration. Therefore, the total time complexity is $O(k \cdot m^2)$. During this phase, the frequency of each character pair is recorded. Assuming the maximum number of character pairs is p, the space complexity for this phase is $O(p)$.

In summary, the overall time complexity of the BPE algorithm designed in this study can be expressed as $O(n + m + k \cdot m^2)$, while the space complexity can be expressed as $O(n + m + p)$. However, in practical applications, adjustments and optimizations will be made based on the characteristics and structure of the data, so the actual runtime and space usage are expected to be lower than the worst-case scenario.

3.3. Adjustments to DFS and Example Explanation

One of the objectives of this framework is to address the issue of a limited number of potential customers. If targets are provided by experts and predictions are based on these targets, the number of individuals on these lists cannot be further increased. Furthermore, due to a lack of interpretability, the list may be rejected by experts or even result in a number lower than what experts would propose based on their experience. Therefore, this study employs DFS to expand the features selected by experts, allowing for the acquisition of features similar to the targets and thereby increasing the number of individuals provided by the model's predictions. DFS was originally designed to traverse an entire graph or tree until all discovered nodes are visited, as detailed in Section 2.4. However, the purpose of employing DFS in this framework is to identify label combinations that approximate T-tag in order to expand target features. Therefore, it is necessary to determine an appropriate approximation ratio through industry knowledge or expert consultations. If the parameters do not meet the specified conditions, the search will not continue further.

Therefore, this study designs two parameters for DFS: depth ($pair_{len}$) and similarity ratio ($pair_{proportion}$), to flexibly identify labels that meet the requirements. Based on Equation (3), the formula (4) representing this design is as follows:

$$D_{\text{tag}} = \underset{T_{\text{tag}}}{\text{DFS}}(F_{\text{tag}}, \text{pair}_{\text{len}}, \text{pair}_{\text{proportion}}) \tag{4}$$

The DFS function will use the initially set T-tag as the root node and search for fact labels that meet the criteria based on the variables $pair_{len}$ and $pair_{proportion}$. Here, a larger value for $pair_{len}$ indicates a greater depth extending downward from the root node, resulting in more nodes. Conversely, a larger value for $pair_{proportion}$ signifies a higher degree of association between the root node and subsequent nodes, leading to a higher proportion of co-occurrence. If $pair_{len}$ is set to 3 and $pair_{proportion}$ is set to 0.7, then the weight of the edge between the root node and the next node must be greater than 0.7, and the number of nodes below the root node must be at least 3 for it to be considered a candidate for further search. Finally, all candidate nodes are traversed, and only after this traversal are they added to the label combination list. This process continues until there are no more candidate nodes. A schematic representation is shown in Figure 4.

As shown in Figure 4, we first assess whether the weight score between the root node and the next node exceeds $pair_{proportion}$. Subsequently, we evaluate whether the total number of nodes in the graph, including the root node, is at least $pair_{len}$. Only if the root node meets both conditions will it be included in the list of candidate nodes. Therefore, the edge weight is 0.2, which does not satisfy condition 1, so this node is ignored in scenario 1; the edge weight is 0.8, satisfying condition 1, and the total number of nodes is 3, which meets condition 2, so this node is added to the list of candidate nodes in scenario 2; the edge weight is 0.8, satisfying condition 1, and the total number of nodes is 4, which meets condition 2, so this node is added to the list of candidate nodes in scenario 3.
Furthermore, if the value of $pair_{proportion}$ is set closer to 1, the association between the root node and the next node will be higher. Similarly, a larger value for $pair_{len}$ indicates a greater number of nodes. Therefore, when both DFS parameters are set to larger values, the conditions for satisfying nodes become more stringent, resulting in fewer label combinations.
Conversely, if both parameters are set to smaller values, the number of label combinations

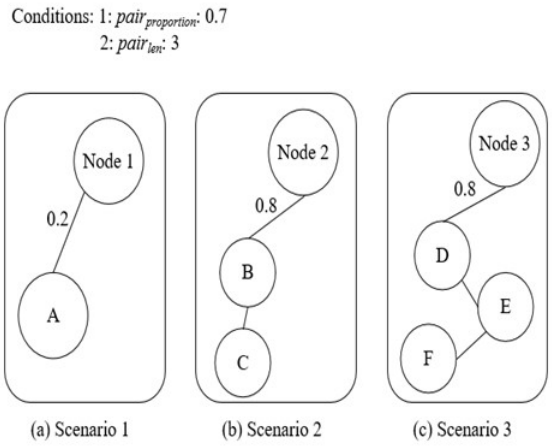


Fig. 4. DFS Search Schematic

will be significantly higher. The actual settings should be determined based on the desired objectives and discussions with experts regarding these variables.

The association between nodes can be confirmed through the weight scores (edge scores) between nodes and the set $pair_{proportion}$. This weight score is calculated based on the values of fact labels and co-occurrence (the frequency with which two features appear together). The calculation is explained as follows.

This score is derived from summing the value of a label with the values of other columns where it appears simultaneously. This approach reflects the correlation between the occurrence of labels together [71]. In other words, if one P-tag frequently co-occurs with another P-tag across multiple records, the score on the edge between these two P-tags will be higher.

After all scores are calculated, normalization will be performed. Since we are primarily interested in the relationship between the highest score and other scores to confirm the association between labels, each label's score is divided by the highest score. This process ensures that all edge weight scores fall within the range of 0 to 1. A score closer to 1 indicates a higher degree of correlation between the two labels.

In summary, after performing a descending order sort, the edge weights can reveal the co-occurrence between each pair of labels. Additionally, the top 10 most common P-tags and labels with 0 co-occurrence, as well as records with a root node score of 0, will be excluded. An example is provided in Table 1.

Using Table 1 as an example, we calculate the co-occurrence weight between the fact label A (node A, root node) and other labels (nodes). We then proceed to traverse the data.

- In the first row, the score for A is 0.5. The corresponding values for B, C, and D are 0.2, 0, and 0.1, respectively. Since the label C has a corresponding value of 0, its co-occurrence with A is also 0, and thus it does not contribute to the weight. The scores for labels B and D will be updated accordingly, with B's score increasing to 0.7 (0.5 + 0.2) and D's score increasing to 0.6 (0.5 + 0.1).

Table 1. Example Table for Weight Calculation

	F-tag			
No.	**A**	**B**	**C**	**D**
1	0.5	0.2	0.0	0.1
2	0.0	0.3	0.4	0.0
3	0.6	0.1	0.2	0.3

- In the second row, the score for A is 0 and the root node is 0, so this record can be ignored.
- In the third row, the score for A is 0.6. The corresponding values for B, C, and D are 0.1, 0.2, and 0.3, respectively. Therefore, the scores for labels B, C, and D will all increase: B's score will be 0.7 (0.6 + 0.1), C's score will be 0.8 (0.6 + 0.2), and D's score will be 0.9 (0.6 + 0.3).
- Following the calculation of the final scores, the sum of the scores obtained by all labels results in the final scores for label A with respect to the other labels: B: 1.4, C: 0.8, and D: 1.5.
- After normalization and descending order sorting, the sequence of labels D, B, and C is determined, with the highest score being 1.5. By dividing all label scores by 1.5, the final co-occurrence scores of labels D, B, and C with label A are obtained as 1, 0.93, and 0.53, respectively. The results indicate that label D has the highest co-occurrence with A, followed by B, while label C has the lowest co-occurrence.
- Therefore, we can create a graph with A as the root node, connected to nodes D, B, and C. The corresponding edge scores are 1, 0.93, and 0.53, respectively, as illustrated in Figure 5.

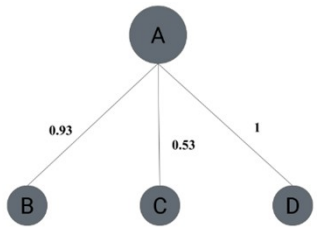


Fig. 5. Graph Generated with A as the Root Node

3.4. Complexity Analysis of DFS

The D-tag code is divided into two phases. The following sections will explain the code and analyze its time and space complexities. Please refer to Figure 6 for details.

```
Class Dtag
    Method __init__(dataFrame)
        self.dataFrame = dataFrame
        self.columns = list of dataFrame's columns
        self.linkNodeReport = empty dictionary

    Method findMajorityColumns(number=10)
        count non-zero values for each column in dataFrame
        sort columns by count in descending order, take top number
        return set of top column names

    Method linkNode(rootName, majoritySet=empty set)
        get scores for rootName
        create matrix excluding rootName column
        initialize scores array with zeros, length = number of columns in matrix

        for each row in matrix
            if root score is non-zero
                find indices of non-zero values in row
                update scores array with non-zero scores + root score

        sort columns in matrix by scores in descending order
        update scores array to match sorted order
        record in linkNodeReport: sorted columns and scores, excluding majoritySet and zeros
```

Fig. 6. Pseudocode for Deriving Labels

Phase One: Column Identification
In this phase, the entire dataset is traversed, examining each row and column to compute the count of non-zero values and perform sorting for each column. If n represents the number of rows (samples) and m denotes the number of columns (features), the time complexity for this method is $O(n * m)$. The time complexity for the sorting operation is $O(n*log_m)$. Since $m*n$ is significantly larger than $m*log_m$, the overall time complexity for the first phase will be dominated by $O(n * m)$. Phase Two: Node Connection

In this phase, calculating the weight scores for each node requires traversing the entire dataset, resulting in a time complexity of $O(n * m)$. Additionally, sorting the nodes has a time complexity of $O(m * log_m)$. Therefore, the total time complexity for the second phase is $O(n * m + m * log_m)$.

In both phases described above, the space complexity is determined by the number of columns in the data and the scores and sorted nodes for each column, resulting in a space complexity of $O(m)$. Overall, the time complexity for processing data with the D-tag class is primarily determined by the traversal of the data. As the number of samples and features in the dataset increases, the computational load will also increase, leading to a time complexity of $O(n * m)$.

Refer to Figure 7 for the analysis of the time and space complexities of label searching and DFS pseudocode. The time complexity of DFS is typically expressed as $O(V + E)$, where V represents the number of nodes and E represents the number of edges. In this study, the DFS code is adapted based on the D-tag, so in the worst-case scenario, if each node is connected to every other node, the number of edges approaches V^2. Consequently, the time complexity is close to $O(V^2)$. Regarding space complexity, the primary considerations are the storage of the node set during traversal and temporary nodes. Therefore, in the worst-case scenario, where all nodes' visit states and paths need to be stored, the space complexity is $O(V)$.

```
Class DtagSearch inherits Dtag
    Method __init__(dataFrame, majoritySet, ptagRoot)
        super().__init__(dataFrame)
        self.majoritySet = majoritySet
        self.ptagRoot = ptagRoot
        initialize adjacencyDict and adjacencyDictRaw as empty dictionaries
        for each root in ptagRoot, link node with root and majoritySet
        prepare adjacencyDict and adjacencyDictRaw for DFS

    Method runDFS(pairLength, pairProportion)
        Define inner method DFS(node, adjacencyDict, visited, tempPair)
            if node not in adjacencyDict
                if tempPair length >= pairLength, record tempPair
                return
            for each connectedNode in adjacencyDict[node]
                if connection strength > pairProportion
                    if connectedNode not in tempPair, check length and uniqueness, then record
                    if connectedNode not visited, mark as visited and recurse with DFS

        for each root in ptagRoot
            initialize pairFeature as empty list
            call DFS with root, adjacencyDict, set with root, and list with root
            record DFS results for root in dfsDtagResult
```

Fig. 7. Pseudocode for Deriving Labels and DFS Search

3.5. Model Metrics and Value Evaluation

When evaluating a model, various metrics are used to compare performance, and specific evaluation criteria are applied to datasets with class imbalance issues, as relying solely on accuracy can be misleading [81]. Typically, a confusion matrix is employed to provide statistical data on true and false results [82], as illustrated by the relationships in the following table.

Table 2. Confusion Matrix for Binary Classification

Predicted Class	Actual Class	
	True (1)	False (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Saito et al. have designed various model evaluation metrics [82]. Commonly used metrics include:

- **Precision**: Focuses on evaluating the predicted positive results.

$$\frac{TP}{TP + FP} \tag{5}$$

- **Recall, TP Rate, Sensitivity**: Concerns the results when the actual class is 1.

$$\frac{TP}{TP + FN} \tag{6}$$

- **F1 Score**: Considers all aspects of the confusion matrix simultaneously.

$$\frac{2 \times TP}{2 \times TP + FP + FN} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

The results of the three metrics are considered better the closer they are to 1 and worse the closer they are to 0. Vujović's research indicates that scores greater than 0.93 are classified as excellent, scores above 0.8 as good, and scores above 0.6 as satisfactory [81]. Given that this study places a higher emphasis on the overall performance of the model, the F1 score is used as the primary evaluation metric. Only models with an F1 score greater than 0.8 are considered for value assessment.

In the value assessment, to simulate the customer lists generated by domain experts based on past experience, the target labels are used to create a benchmark model. The predictions from this model are treated as the customer lists produced by experts, and are compared with those generated by the explainable framework to evaluate the differences.

4. Experimental Results and Analysis

This study aims to propose an explainable framework to address the marketing list issue in the business domain.

4.1. Identify Objectives and Acquire Relevant Data

To ensure that the test data closely reflects the marketing list issue in case studies, the selected dataset must meet the following criteria: the dataset should include fields with natural language, with enough detail to allow domain experts to interpret the characteristics of the data. For example, customer purchase records should contain descriptive statements such as product descriptions, product names, or transaction-related information, and the data should be relevant to the customers. Under these conditions, the dataset aligns with the definition of model explainability [83–86]. Consequently, the Amazon Sales public dataset is chosen for the experiment, which consists of 16 feature fields per record.

Table 3. Description of the Public Dataset(1)

Dataset Name	Feature Count	Number of Records	Source
Amazon Sales	16	1465	[87]

– `product_id`: **String** - Product ID, each product has a unique identifier
– `product_name`: **String** - Product Name, including detailed content
– `category`: **String** - Product Category
– `discounted_price`: **Numeric** - Price after Discount
– `actual_price`: **Numeric** - Actual Price
– `discount_percentage`: **Numeric** - Discount Percentage
– `rating`: **Numeric** - Product Rating Score

- `rating_count`: **Numeric** - Number of User Ratings
- `about_product`: **String** - Product Description
- `user_id`: **String** - User ID of the Reviewer
- `user_name`: **String** - Name of the Reviewer
- `review_id`: **String** - User Review ID
- `review_title`: **String** - Review Title
- `review_content`: **String** - Review Content
- `img_link`: **String** - Product Image URL
- `product_link`: **String** - Product Official Website URL

4.2. Explainable Framework

F-tag The public dataset serves as the raw data for this experiment. After data cleaning and conversion between full-width and half-width characters, the `product_name` and `about_product` fields were selected as the target columns for BPE. This process resulted in the extraction of 723 and 293 subwords, respectively. Due to space constraints, only a portion of the subwords is displayed; see Tables 4 and 5 for detailed information.

Table 4. Subwords for the `product_name` Field

Column Name	Number of Subwords	Sample Subwords	Method of Generation
`product_name`	723	accessor accessories adapter apple/dell apple/dell/lenovo black-heart black/char black/chartre black2v9 blackxcd- capicity cappuccin carecase certified …	BPE

After obtaining the subwords and conducting data cleaning to exclude unnecessary and duplicate subwords, the comparison between the `product_name` and `about_product` fields using regular expressions yielded a total of 177 F-tags, as shown in Table 6. The fact labels do not necessarily represent complete words, as the final labels are determined by the frequency of characters in the vocabulary.

The F-tags are treated as feature columns and added to each data entry, resulting in the original dataset having 177 additional columns, making a total of 193 columns after including the fact labels. The F-tag values are expressed as 0/1, where 1 indicates that the record contains the feature associated with the tag, and 0 indicates its absence. This means

Table 5. Subwords for the `about_product` Field

Column Name	Number of Subwords	Sample Subwords	Method of Generation
about_product	293	anti-rust anti-wrink anti-wrinkle- appropri assistant attachment authenti availability backlight bluetooth bn59-013 borosilic breaking cameramem centimet …	BPE

that a record with a value of 1 indicates that the customer's purchase includes a product description or name that possesses the corresponding tag characteristics. See Figure 8 for illustration.

Table 6. Table for Public Dataset of F-tab

Label Category	Number of Labels	Partial Results of F-tag	Generation Method
F-tag	177	['technolog', 'experienc', 'manufacture', 'temperat', 'addition', 'connectiv', 'material', 'features', 'transmis', 'recharge', … , 'notebook', 'straight', 'thorough', 'attachment', 'guidelin', 'instruction', 'upholster', 'sandwich', 'resistance', 'component']	Based on the roots generated from Tables 4 and 5, data cleaning and regular expression matching were conducted to obtain.

T-tag and label combinations In experiments conducted on public datasets, this study incorporated the recommendations of domain experts—credit card marketing product managers (PM). Three experts collaboratively discussed and selected a target label set from 177 factual labels, resulting in the following labels: sensitivity, lightweight, and durability. The experts expressed the desire to identify label combinations from the public datasets that are similar to sensitivity, lightweight, and durability.

Based on the design of the Depth-First Search (DFS) algorithm, we treat the T-tag as the root node to locate the corresponding D-tags. Through trial and error and expert discus-

Original Columns: total 16 columns				**Additional F-tag:** total 177 columns			
Product_id	…	…	Product_link	notebook	straight	…	component
B07JW9H4JI	…	…	…	0	1	…	0
…	…	…	…	…	…	…	…

Fig. 8. The public dataset with the addition of F-tag columns

sions, the parameters for the depth-first search, namely $pair_{len}$ and $pair_{proportion}$, were set to 3 and 0.4, respectively. The DFS results identified multiple label combinations for the three target labels, specifically (74 sets, 74 sets, and 35 sets). For detailed results, refer to Table 7.
The label combinations represent factual labels that co-occur with the target labels, allowing experts to identify which labels are related to their experience-based target labels. Through these label combinations, experts can better understand the relationships between the target labels and other relevant labels.

D-tag Based on the DFS results from the previous step, multiple label combinations were generated. These results need to be discussed with experts, who will determine two derived label sets based on their experience, the characteristics of the data, and the scope of interpretability. The two sets identified are: (convenient, warranty) and (function, protection).
The selection of derived labels must effectively convey the meaning of the target labels. Therefore, the derived labels (convenient, warranty) and (function, protection) were chosen to correspond to the target labels (sensitivity, lightweight, durability).
The aforementioned D-tags are treated as the actual categories for the model and are assigned to each data point. If the data contains the specified combinations, it is labeled as 1; otherwise, it is labeled as 0, as shown in Figure 9. After labeling the data, it is possible to determine which customers purchased products featuring (convenient, warranty) or (function, protection), or whether the product descriptions include items with (convenient, warranty) or (function, protection).
After labeling, separate models were developed for each derived label combination. As a result, with the two derived label sets, two models were generated: one to predict potential customers for (convenient, warranty) and another for (function, protection).

Total 193 columns							
Product_id	…	Function	Protection	…	convenient	warranty	label
B07JW9H4JI	…	1	1	…	0	0	1
…	…	…	…	…	0	0	0
B072NCN9M4	…	…	…	…	1	1	1

Fig. 9. The public dataset with the addition of F-tag columns

Table 7. Partial results table of DFS label combinations

T-tag	Number of combinations	Partial results of label combinations	Generation method
sensitivity	74	[['sensitiv', 'experienc', 'experience'], ['sensitiv', 'features', 'warranty'], . . . , ['sensitiv', 'features', 'warranty', 'devices.', 'charging', 'compatibl', 'experience'],. . . , ['sensitiv', 'function', 'devices.'], ['sensitiv', 'function', 'protection'], ['sensitiv', 'function', 'features'], ['sensitiv', 'function', 'convenient'], ['sensitiv', 'function', 'compatibl'], ['sensitiv', 'function', 'capacity']]	DFS
lightweight	74	[['lightweight', 'compatibl', 'devices.'], ['lightweight', 'compatibl', 'devices.', 'charging'],['lightweight', 'compatibl', 'devices.', 'charging', 'transfer'], ['lightweight', 'compatibl', 'devices.', 'charging', 'warranty'], ['lightweight', 'compatibl', 'devices.', 'charging', 'warranty', 'manufactur'],['lightweight', 'compatibl', 'devices.', 'charging', 'warranty', 'manufacture'], . . . , ['lightweight', 'function', 'devices.'], ['lightweight', 'function', 'protection'], ['lightweight', 'function', 'features'], ['lightweight', 'function', 'convenient'], ['lightweight', 'function', 'compatibl'], ['lightweight', 'function', 'capacity']]	DFS
durability	35	[['durability', 'charging', 'devices.'], ['durability', 'charging', 'devices.', 'compatibl'], ['durability', 'charging', 'devices.', 'compatibl', 'smartphon'], ['durability', 'charging', 'devices.', 'compatibl', 'warranty'],. . . , ['durability', 'charging', 'devices.', 'transmission'], ['durability', 'charging', 'devices.', 'transmis'] ,['durability', 'charging', 'warranty'], ['durability', 'charging', 'connector'],['durability', 'charging', 'smartphon']]	DFS

4.3. Model Prediction and Value Assessment

The experiments conducted on the public dataset were implemented using Python 3.10.5 in Visual Studio Code, on a MacBook Pro 2023 with an M2 chip and 32GB of RAM. Missing values in the dataset were handled by imputing the mean. For BPE processing, the fields (`product_name`, `about_product`) were converted to half-width characters and lowercase English letters, and full-width spaces were replaced with half-width spaces. Following Sections 4.1 and 4.2, the objective of this section is to predict potential customers for products that possess two sets of derived labels. The machine learning model utilized is LightGBM, as described by Aditya et al. in the literature review [20]. This model not only offers excellent data adaptability but also provides accurate and standardized hyperparameter settings. Hyperparameter adjustments were made based on the parameters and data characteristics discussed by Gupta et al. [88]. Table 8 details the hyperparameter settings for the model.

Table 8. Model Hyperparameter Settings Table

Model Type	Parameter Category	Setting Value
LightGBM	`num_leaves`	60
	`max_depth`	8
	`extra_trees`	True
	`random_state`	42
	`sampling_strategy`	0.8
	`train, val`	0.75, 0.25

Based on the settings in the above table, the model's prediction results and scores are shown in Table 9. For the first set of derived labels (convenient, warranty), the precision is greater than 0.8; for the second set (function, protection), the precision reaches 0.99. Both sets of D-tags meet the model standards outlined in Step 3 (precision, recall, and F1 score all exceeding 0.8). Therefore, the list of predicted customers can be subjected to a value assessment, excluding clients who are refused or blacklisted by the company. The final step is to verify whether the list aligns with the domain experts' objectives, thereby producing a final, interpretable list of potential customers.

Table 9. Model Prediction Scores Table

D-tag	Evaluation Category	Score
(convenient, warranty)	precision	0.833
	recall	0.833
	F1 score	0.833
(function, protection)	precision	0.999
	recall	0.833
	F1 score	0.909

In the value assessment, to simulate the customer list proposed by experts based on past experience, modeling based on the target labels was used as a benchmark for comparison. Since there were no individuals meeting all three criteria (sensitivity, lightweight, durability), three combinations were used for data labeling: (sensitivity, lightweight), (sensitivity, durability), and (lightweight, durability). These combinations served as the actual categories for the model. Predictions were made using the same parameter settings, and the results are presented in the table below.

Table 10. Model Prediction Scores Based on Past Experience Table

Actual Category	Evaluation Category	Score
(sensitiv, lightweight, durability)	precision	0.999
	recall	0.666
	F1 score	0.800

Next, we compared the number of individuals on the lists, as detailed in Table 11. At an F1 score threshold greater than 0.8, the model predictions based on the explainability framework identified 17 potential customers, while the directly modeled numbers based on past experience identified only 2. This result indicates that the explainability framework predicts 8.5 times more potential buyers than the past experience model. It also highlights that the targets selected based on past experience may result in no customers meeting the criteria, which contributes to modeling challenges and low explainability, exemplifying the so-called cold start problem.

Table 11. Value Assessment Table

F1 Score Threshold > 0.8	Explainability Framework	Past Experience
Predicted Number of Buyers	11 (Number of Individuals)	2 (Number of Individuals)
Difference	5.5x	1x

In summary, the results of the small-sample experiments on the public dataset validate the effectiveness of the framework. It is evident that the parameters of DFS significantly impact the number of model predictions. Additionally, the choice of label combinations can be affected by whether the DFS parameters are set too low or too high. Furthermore, the derived labels determined by domain experts based on these label combinations can substantially influence the prediction results, potentially leading to class imbalance. Therefore, it is advisable to apply SMOTE to augment and adjust the sample as needed.

4.4. Interpretability and persona

In the interpretability section, the modeling results for derived labels are analyzed using LIME for local explanations and SHAP for global explanations. First, the explanation plot

for (convenient, warranty) is presented in Figure 10, followed by the explanation plot for (function, protection) as shown in Figure 11.

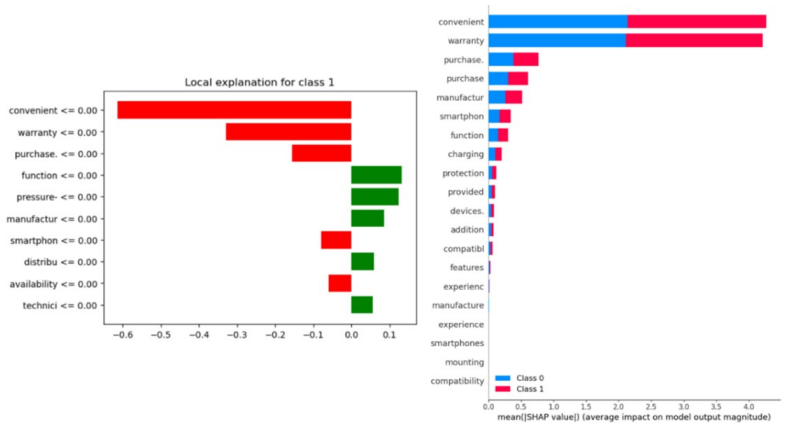


Fig. 10. LIME and SHAP Explanation(1)

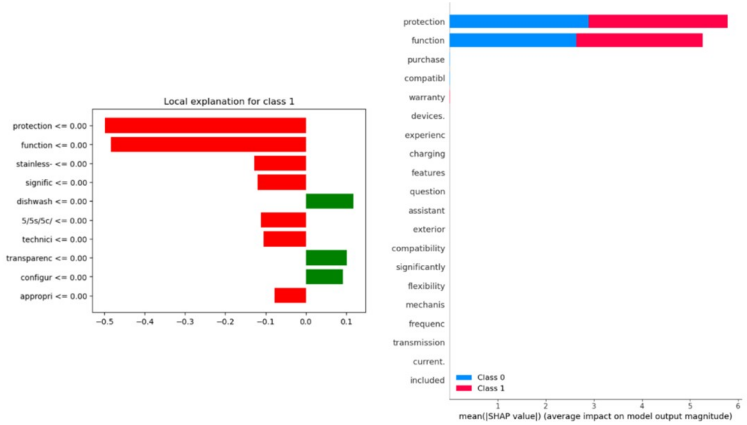


Fig. 11. LIME and SHAP Explanation(2)

Figure 10 shows the LIME explanation on the left. In the LIME plot, the Y-axis is sorted by the features with the greatest impact on the prediction, while the X-axis represents each feature's contribution to the prediction. A negative direction (in red) indicates a lower likelihood of predicting class 1, while a positive direction (in green) suggests a higher likelihood of predicting class 1. Therefore, when the values of 'convenient,' 'warranty,' and 'purchase' are less than or equal to 0, they exhibit a strong negative contribu-

tion (-0.6), influencing the model to not predict class 1. Conversely, if these values are greater than 0, they show a strong positive contribution. 'Function,' 'pressure,' and 'manufactur' have positive contributions, which slightly push the prediction toward class 1.
The right side of Figure 10 presents the SHAP explanation. In the SHAP plot, the Y-axis ranks the features by importance, with the most important ones listed at the top. Blue corresponds to class 0, and red corresponds to class 1. If a feature significantly affects both classes, the corresponding row will display both colors. Therefore, 'convenient' and 'warranty' naturally have a significant impact on the model's output, while 'purchase' also exhibits a certain level of influence. Additionally, 'manufactur' and 'smarphon' show moderate SHAP values, indicating a smaller but still notable effect on the model's predictions.
On the other hand, the label co-occurrence provided by this framework allows for the calculation of co-occurrence scores in the prediction list based on (convenient, warranty), as shown in Table 12.

Table 12. Co-occurrence Scores for D-tags(1)

D-tag	F-tag	Co-occurrence Score
(convenient, warranty)	warranty	1.000
	manufactur	0.372
	convenient	0.344
	manufacture	0.313
	provided	0.288
	features	0.242
	devices.	0.226
	charging	0.217
	compatibl	0.170
	protection	0.161
	experienc	0.155
	experience	0.155
	function	0.145
	addition	0.136
	installation	0.133

The table above lists the top 15 ground truth labels based on co-occurrence ranking. From the table, it can be observed that, in addition to the labels 'convenient' and 'warranty' themselves, the label most closely related to them is 'manufactur,' which has the highest co-occurrence score. This is followed by 'provided' and 'features,' among others. These results help reveal the relationships between labels and can be used to validate the reasonableness of the model's feature explanations in Figure 10. Similarly, in the LIME explanation on the left side of Figure 11, it can be seen that when the values of 'function' and 'protection' are less than 0, they provide a sufficiently negative contribution, leading the model to predict against class 1. In the right-side plot, 'function' and 'protection' exhibit the highest average SHAP value (5.8), meaning the model can determine whether a customer is a potential target based solely on these derived labels. Furthermore, Table

13 shows that the co-occurrence of 'function' and 'protection' is quite high, while the co-occurrence of other labels drops below 0.34, which fully explains the feature ranking results provided by SHAP.

Table 13. Co-occurrence Scores for D-tags(2)

D-tag	F-tag	Co-occurrence Score
(function, protection)	function	1.000
	protection	0.872
	warranty	0.345
	experienc	0.303
	experience	0.303
	charging	0.297
	devices.	0.250
	compatibl	0.228
	features	0.228
	technology	0.218
	technolog	0.218
	functional	0.207
	smartphon	0.202
	performanc	0.202
	convenient	0.180

In the persona analysis, this study randomly selected one instance from the model prediction results for each of the two sets of derived labels and listed the non-zero ground truth labels, as shown in the table below. These results were then analyzed by domain experts.

Table 14. Persona and F-tags(1)

D-tag	F-tag	Value
(convenient, warranty)	experienc	1
	purchase	1
	complete	1
	resistant	1
	comfortabl	1
	lightweight	1
	playback.	1
	experience	1
	convenient	1
	warranty	1

Table 14 indicates that the purchaser values experience in product descriptions and actively makes purchases. They typically prefer products with attributes such as resis-

tance, lightweight, playback capability, convenience, and warranty. Experts suggest that in the prediction list of derived labels associated with (convenient, warranty), the persona corresponds to a tech-savvy professional. Such individuals usually prioritize unique experiences and lightweight products, and they value high-performance gadgets, which aligns with a high relevance to playback and experience. Alternatively, they might be active lifestyle enthusiasts who emphasize comfort, convenience, and durability, indicating that they may engage in outdoor activities or travel and require items that are both durable and portable.

Table 15. Persona and F-tags(2)

D-tag	F-tag	Value
	connector	1
	function	1
	computer	1
	transfer	1
(function, protection)	interfer	1
	compatibl	1
	protection.	1
	convenient	1
	conductor	1

Table 15 reveals that the purchaser values connectors and functionality in product descriptions and desires features such as computers, transfer capabilities, and protection. Experts suggest that in the prediction list of derived labels associated with (function, protection), the persona corresponds to IT professionals, including engineers, remote workers, or technical specialists. Given their emphasis on computer-related labels and their concern for convenience and transfer functions, these characteristics align well with this type of persona. In summary, to ensure that the model's prediction results are sufficiently interpretable, it is necessary to adopt various expert recommendations based on the characteristics and domain knowledge of different datasets. Integrating the RFM model to assign values to ground truth labels can provide experts with additional analytical space in persona. Additionally, using DFS with optimal parameters to explore label combinations and expand the feature scope is essential for the model to identify a larger and more reasonable number of potential customers. In summary, the above steps can be categorized into three major phases: identifying objectives and obtaining relevant data, implementing the interpretability framework, and conducting model prediction and value assessment. The interpretability framework itself comprises three labeling processes: ground truth labels, target labels, and derived labels. The following discussion will explain the general validation process based on these steps.

4.5. Generalizability Validation

Due to the rarity of test sets that fit specific case scenarios, publicly available datasets such as Google Play Store Apps [89] and Amazon Products Sales Dataset 2023 [90] are

not suitable. These datasets either lack user IDs or have product names or descriptions that have already become class labels, making it impossible to enhance interpretability through BPE. However, the effectiveness and feasibility of the framework have been validated through small sample experiments in the preceding sections. To assess the framework's generalizability and reproducibility, this study has selected other similar datasets for experimentation. **Define Objectives and Acquire Relevant Data** In summary, for the generalizability validation, the UCI Online Retail dataset [91], hereafter referred to as Public Dataset 2, was used. Each record in this dataset contains 6 feature columns, with a total of 541,909 online transaction records, as listed below:

Table 16. Description of the Public Dataset(2)

Dataset Name	Feature Count	Number of Records	Source
UCI Online Retail	6	541909	[91]

- `Description`: **String** - Product Name
- `Quantity`: **Value** - quantity purchased for the record
- `InvoiceDate`: **String** - purchase date for the record
- `unit_price`: **Value** - unit price of the product
- `CustomerID`: **Value** - user ID
- `Country`: **String** -user's country of residence

Interpretability Framework

F-tag The selected dataset utilizes the Description column for BPE processing, resulting in 442 word stems. After excluding duplicates and non-applicable stems, a comparison between the stems and the values in the Description column using regular expressions produced 194 ground truth labels. Due to space constraints, only a subset of these stems is displayed in the table below.

After obtaining the ground truth labels, as listed in Table 18, they were incorporated into the dataset as feature columns. Consequently, 194 new columns were added to the dataset, resulting in a total of 200 feature columns.

To represent the purchasing characteristics of customers, the M (Monetary Value) component from the RFM model was used as the value for the ground truth labels. The calculation for the monetary value is as follows: (Total purchase amount by the customer for the products corresponding to the label) / (Number of purchases of that product) / 2. This represents the average spending per product by the customer over a two-year period. Subsequently, the data was merged based on CustomerID, retaining only the CustomerID and ground truth label columns, resulting in a total of 195 columns.

In summary, for consumer A, if there is only one record in the data matching the label "popcor," the total amount for that record is divided by 2 to determine the value of "popcor" for customer A. If consumer B has two purchase records in the data that match the label, the amounts for these two records are summed, divided by the number of records,

Table 17. Partial Word Stems from the Description Column

Field Name	Number of Stems	Partial Stem Results	Generation Method
		artific	
		butterfi	
		butterfli	
		butterfly/	
		campho	
Description	442	...	BPE
		windmil	
		wirele	
		yellow	
		yellow/	
		yuleti	

Table 18. Public Dataset of F-tag Table 2: Label

Lable Categories	Number of Labels	Partial F-tag Results	Generation Method
F-tag	194	['childre', 'strawb', 'butter', 'scandina', 'babush', 'victori', 'dinosa', 'garden', 'sketch', 'popcor', ... , 'revolu', 'toilet', 'square', 'artific', 'glass', 'cabinet', 'candle']	Based on the word stems generated from Table 12, cleaning and regular expression matching were performed to obtain

and then divided by 2 to determine the value of the label "glass" for customer B. The consolidated results will form a new dataset.
Based on the combination of the RFM model and the ground truth labels, we can determine the average spending amount of each customer on products corresponding to the label over a two-year period. This information helps to understand the average purchasing power of each customer in online shopping.

T-tags and combinations of labels The T-tags were determined by domain experts—a credit card marketing project manager—based on the objectives of the task, using the ground truth labels. After discussion among three experts, the target labels were decided to be ('childre', 'candle', 'decorati', 'chocolate'). The experts expressed the aim of identifying label combinations in the public dataset that are similar to those related to children, candles, decorations, and chocolate.
Next, the parameters $pair_{len}$ and $pair_{proportion}$ for the DFS were set. To verify the reproducibility of the framework, these parameters were set to 3 and 0.4, respectively. This configuration means that the label combinations should include at least three nodes and that the co-occurrence between the root node and the subsequent node should be above 0.4. The DFS search results for the target labels were (147 combinations, 147 combinations, 136 combinations, 147 combinations), as detailed in Table 19.

D-tag Based on the experts' experience, data characteristics, and the scope of interpretability, two sets of derived labels were selected from the label combinations: (christ, colour) and (garden, flower). The selection of D-tags must be able to convey the meaning of the target labels. Therefore, the derived labels (Christmas, colorful) and (garden, flowers) were chosen to correspond to the target labels (children, candles, chocolate, decorations).

The D-tags mentioned above will be treated as the actual categories for the model and assigned to each record in the dataset. If a record contains one of the specified combinations, it will be marked as 1; otherwise, it will be marked as 0. After labeling the data, it will be possible to identify which customers purchased products related to either (Christmas, colorful) or (garden, flowers).

Model Prediction and Value Assessment The experimental environment for Public Dataset 2 was the same as that used for Public Dataset 1. Missing values in the dataset were imputed with the mean values. The BPE processing for the column (Description) was standardized to lowercase and half-width characters, with full-width spaces converted to half-width spaces and extra spaces removed.

After completing steps one and two, the task objective is to predict potential customers for products that possess either of the two sets of derived labels. To ensure the stability of the experiment, the machine learning model used is LightGBM, with hyperparameters configured as outlined in Table 8. Using the settings from Table 8, the model prediction results and scores are obtained as shown in Table 20. For the first set of derived labels (Christmas, colorful), the precision of the predictions reaches 0.98; for the second set of derived labels (garden, flowers), the precision is 0.97. Both sets of derived tags meet the model standards from step three (with precision, recall, and F1 score all exceeding 0.8). Consequently, the customer lists predicted by the model can be subjected to value assessment.

Table 19. DFS Label Combination Results(2)

T-tag	Number of Combinations	Partial Label Combination Results	Generation Method
childre	147	[. . . , [’childre’, ’colour’, ’glass.1’, ’christ’, ’decorati’, ’candle.1’, ’candle’, ’black’], [’childre’, ’colour’, ’glass.1’, ’christ’, ’decorati’, ’candle.1’, ’candle’, ’black’, ’cakestand’], . . . , [’childre’, ’colour’, ’glass.1’, ’christ’, ’decorati’, ’drawer’], [’childre’, ’colour’, ’glass.1’, ’garden’]. . . ,]	DFS
candle	147	[. . . , [’candle’, ’candle.1’, ’colour’, ’glass.1’, ’butter’], [’candle’, ’candle.1’, ’colour’, ’glass.1’, ’drawer’], [’candle’, ’candle.1’, ’colour’, ’christ’],. . . .]]	DFS
decorati	136	[[’decorati’, ’christ’, ’colour’], [’decorati’, ’christ’, ’colour’, ’glass.1’], [’decorati’, ’christ’, ’colour’, ’glass.1’, ’silver’], . . . , [’decorati’, ’christ’, ’drawer’], [’decorati’, ’christ’, ’butterf’], [’decorati’, ’christ’, ’lanter’], [’decorati’, ’christ’, ’butterfly’]]	DFS
chocolate	147	[[’chocolate’, ’colour’, ’glass.1’, ’christ’, ’decorati’, ’candle.1’, ’candle’, ’black’, ’garden’], [’chocolate’, ’colour’, ’glass.1’, ’christ’, ’decorati’, ’candle.1’, ’candle’, ’black’, ’garden’, ’cakestand’], . . . , [’chocolate’, ’colour’, ’glass.1’, ’christ’, ’drawer’], [’chocolate’, ’colour’, ’glass.1’, ’christ’, ’butterf’],. . .]	DFS

Table 20. Model Prediction Scores Table 2

D-tag	Evaluation Categories	Score
(christ, colour)	precision	0.983
	recall	0.8093
	F1 score	0.887
(garden, flower)	precision	0.978
	recall	0.866
	F1 score	0.919

In the value assessment, the T-tags (‘childre’, ‘candle’, ‘decorati’, ‘chocolate’) are used as a benchmark based on the experts’ past experience for comparison. The T-tags are treated as the actual categories, and the same parameter settings are employed for modeling. The prediction scores are listed in the table below.

Table 21. Prediction Scores Table 2 Based on Historical Experience

Actual categories	Evaluation Categories	Score
(‘childre’, ‘candle’, ‘decorati’‘, ‘chocolate’)	precision	0.955
	recall	0.773
	F1 score	0.85

At an F1 score level greater than 0.8, a comparison of the number of potential customers was conducted. The interpretability framework identified 512 potential buyers, while the number predicted based on historical experience modeling was 68. The results indicate that the number of potential customers predicted by the interpretability framework is 7.5 times that of the historical experience modeling, as detailed in the table below.

Table 22. Value Assessment(2)

F1 Score >0.8	Interpretability Framework	Historical Experience
Predicted Number of Buyers	512 (Number of Individuals)	68 (Number of Individuals)
Difference	7.5x	1x

Interpretability and Persona In the interpretability section, the modeling results for the derived labels were analyzed using LIME for local explanations and SHAP for global explanations. First, the explanation diagram for (Christmas, colorful) is shown in Figure 12. Following that, the explanation diagram for (garden, flowers) is presented in Figure 13.

In Figure 12, the left side shows the LIME explanation. In the LIME diagram, when the value of ”colour” is less than or equal to 0, it has a strong negative contribution (-0.4), influencing the model not to predict class 1. Conversely, if the value is greater than 0, it has a strong positive contribution. Additionally, when ”christ” exceeds 7.07 and ”scandina” exceeds 0.49, they provide positive contributions, pushing the prediction towards class 1. Even ”silver” values between 0 and 4.1 have a positive contribution towards class 1. On the right side of Figure 4-6, the SHAP explanation is presented. The SHAP diagram indicates that ”colour” and ”christ” have significant impacts on the model’s output, while ”scandina,” ”silver,” and even ”garden” and ”decorati” also exhibit certain degrees of influence. On the other hand, the co-occurrence scores in the prediction list were calculated based on the label pairs (Christmas, colorful) as provided by this framework, as

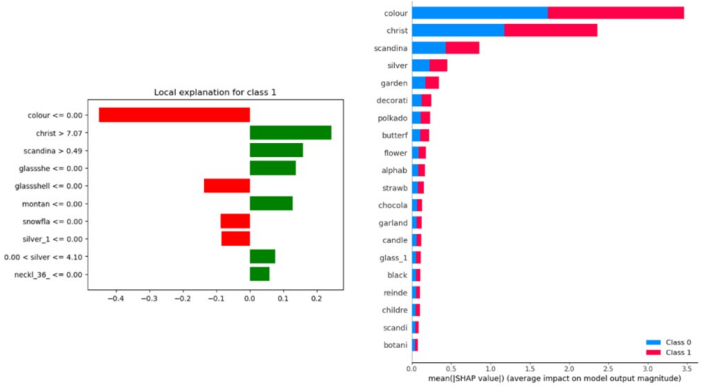


Fig. 12. Public Dataset 2: LIME and SHAP Explanation Diagrams 1

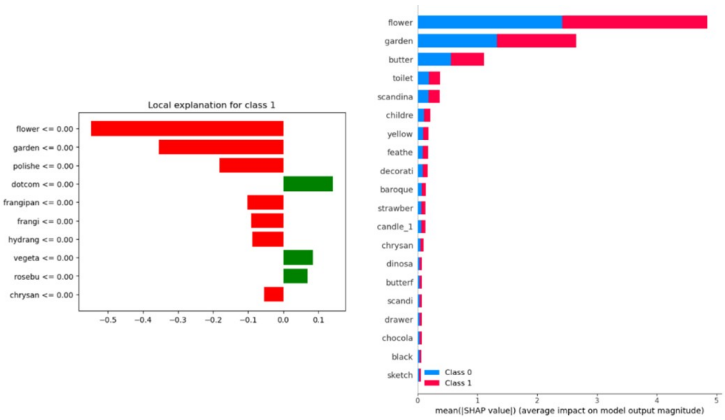


Fig. 13. Public Dataset 2: LIME and SHAP Explanation Diagrams 2

shown in Table 23. The table indicates that, aside from the labels themselves, the label most strongly associated with (Christmas, colorful) is ”flower,” followed by ”cakestand,” which also ranks high in co-occurrence. Conversely, ”garden” and ”silver” exhibit lower levels of co-occurrence, which slightly diverges from the explanation results in Figure 12.

In Figure 13, the LIME explanation on the left indicates that when the values of ”flower,” ”garden,” and ”polishe” are less than or equal to 0, they have a sufficient negative contribution, causing the model to not predict class 1. Conversely, when ”dotcom” is less than or equal to 0, it has a positive likelihood of predicting class 1. On the right side of the figure, the SHAP explanation shows that ”flower” and ”garden” have the highest average SHAP values (5), while ”butter” and ”toilet” also have a notable impact on the model in this classification. Additionally, Table 24 shows that, aside from ”flower” itself, ”christ” and the derived labels appear together quite frequently, followed by the ”garden” label.

Table 23. Based on D-tags, Co-occurrence Value(3)

D-tag	F-tag	Co-occurrence Scores
(christ, colour)	christ	1.000
	colour	0.672
	flower	0.374
	cakestand	0.274
	black	0.221
	glass_1	0.179
	garden	0.164
	drawer	0.158
	silver	0.123
	cabinet	0.112
	candle	0.1096
	candle_1	0.1096
	botani	0.1092
	decorati	0.1035
	alphab	0.1035

The co-occurrence scores for ”black,” ”colour,” ”darwer,” and ”cakestand” with the derived labels are all greater than 0.4, indicating a high level of co-occurrence. This finding slightly diverges from the feature explanation results in Figure 13, but it still provides experts with greater interpretability and comprehensibility of the model's prediction results.

In the persona section, this study randomly selected one instance from the model prediction results for each of the two sets of derived labels and listed the factual labels with RFM values greater than 0, as shown in the table below. These were then analyzed by domain experts.

From Table 25, it can be seen that the purchaser places significant importance on the descriptive features of the product, such as ”decorati” and ”silver,” and even prefers descriptions that include ”feathe.” The first three features have high consumption amounts and frequencies over the past two years, with scores exceeding 5.9. Additionally, the purchaser values specific themes (”charlot”) and stripe elements. Experts suggest that, for the derived labels corresponding to (Christmas, colorful), the persona is likely that of a home decoration enthusiast. This individual enjoys beautifying and decorating items and may be a homemaker, interior designer, or a working professional who values ritualistic elements, potentially even an admirer of Scandinavian style (scandina, scandinavi, scandi).

From Table 26, it is evident that the purchaser has a strong preference for products described with ”doughn” in the product descriptions. The consumption frequency and amount over the past two years for this label are significantly higher, reaching a ratio of 12.5, far surpassing the second most frequent label, ”flower.” Other related labels include ”breakf” (breakfast), ”citron” (lemongrass), and ”thermo” (temperature control) products. Experts suggest that, for the derived labels corresponding to (flower, garden), the persona is likely that of a home baking and breakfast enthusiast. The labels indicate that the customer enjoys baking and preparing breakfast, and could also be a baker selling homemade

Table 24. Based on D-tags, Co-occurrence Value(4)

D-tag	F-tag	Co-occurrence Scores
(flower, garden)	flower	1.000
	christ	0.840
	garden	0.791
	black	0.587
	colour	0.579
	drawer	0.461
	cakestand	0.445
	botani	0.355
	cabinet	0.329
	glass_1	0.273
	candle	0.251
	candle_1	0.251
	butter	0.246
	butterf	0.239
	stripe	0.223

Table 25. Persona and F-tags(3)

D-tag	F-tag	RFM value
(christ, colour)	decorati	7.5
	silver	6.36
	feathe	5.9
	charlot	4.25
	stripe	4.12
	colour	3.75
	christ	3.26
	scandina	1.25
	scandinavi	1.25
	scandi	1.25

Table 26. Persona and F-tags(4)

D-tag	F-tag	RFM value
(flower, garden)	doughn	12.50
	flower	4.20
	garden	4.19
	breakf	4.08
	citron	3.97
	citronel	3.97
	thermo	2.97
	candle	2.86
	candle_1	2.86
	black	1.77

goods. Additionally, the customer might be a nature lover or a retiree, having ample time for floral arrangements and baking activities.

In summary, the experimental results with the UCI Online Retail dataset (Public Dataset 2) demonstrate the generalizability and reproducibility of the explainability framework. It is also observed that when the content of the target labels is more relevant, the label combinations become highly correlated, resulting in numerous similar label combinations within the nodes.

When the target labels are ('childre', 'candle', 'decorati', 'chocolate'), domain experts interpret the relevance from a natural language perspective, noting that candles and decorations are related, chocolate and children are related, and chocolate color is also related to decorations. This aligns with the setting of target labels being interrelated. Consequently, the derived labels that were determined—(Christmas, colorful) or (garden, flowers)—are also relevant. Decorations are associated with candles and Christmas, as well as being colorful and even with flowers. Children are related to both Christmas and gardens. Many garden decorations fall within the category of decorations, indicating that despite different label combinations, they remain highly related. This is evident from Tables 22 and 23, where the co-occurrence rankings of the two sets of derived labels both reflect the presence of the other set of derived labels.

Additionally, the experimental results indicate that directly modeling using the expert-provided target labels can lead to issues such as a small number of customers and a lack of explanatory power, potentially even causing cold start problems. However, by employing the described method, which utilizes DFS to identify co-occurrence and reasonably extend the most relevant features, it is possible to broaden the scope of similar targets and increase the number of potential customers in the model. Moreover, assigning values to factual labels through the RFM model and computing the co-occurrence of labels based on derived labels can effectively enhance the interpretability of the model's potential customer list for experts. This approach increases trust in the model's results and facilitates the analysis of personas.

In summary, by integrating domain knowledge with natural language understanding, experts can provide additional explanations for the labels in the prediction results and reassess whether their target settings are aligned with the outcomes. After verifying the generalizability and reproducibility of the framework, Section 4.6 applies the same steps for case validation.

4.6. Case validation

After validating the effectiveness and generalizability of the interpretability framework using publicly available datasets, this study further applies the interpretability framework to cases in the financial industry where marketing lists have been rejected.

Identify objectives and obtain relevant data For this study's case, marketing lists, transaction data, and customer datasets provided by one of the top three financial holding companies in Taiwan were used, as detailed in Table 27. After data integration and consultation with domain experts, a total of 185,199,048 transaction records were obtained, with each record containing four feature columns, as listed below.

- `customer_id`: **String** - Customer Code for Each Transaction
- `monetary`: **Numeric** - Transaction Amount for Each Entry

Table 27. Dataset Description Table

Dataset Name	Number of Features	Number of Records	Data Period	Purpose
Transaction Data	4	185,199,048	2021/8-2023/8	Obtain F-tags
Customer Information	4	7,314,643	2023/1-2023/8	Training and Prediction
Marketing Lists	N/A	N/A	N/A	Define Target Labels

– `date`: **String** - Year and Month of Each Transaction
– `remark`: **String** - Transaction Notes

Interpretability Framework

F-tag After processing the transaction notes field using Byte Pair Encoding (BPE), a total of 1,465 word stems were generated. Following the removal of duplicates, non-descriptive, and irrelevant stems, regular expression matching was conducted, resulting in 319 ground truth labels. Due to restrictions imposed by financial regulations and personal data protection laws [92, 93], only a portion of the ground truth labels can be provided, as detailed in Table 28.

Table 28. Partial F-tags Table for the Case Dataset

Label Category	Number of Labels	Partial Results of F-tags	Generation Method
F-tag	319	['Stocks', 'Allowance', 'Taipei City', 'Steak', 'Streaming Media Platform', 'Dividend', 'Holiday Cash Flow', ... , 'Policy Loans', 'Funds', 'Credit Card', ...]	Based on the word stems generated by BPE, data cleaning and regular expression matching were performed to obtain

In the customer information dataset, after data integration, a total of 7,314,643 records were obtained. Thus, as of the end of August 2023, there were 7,314,643 customers. This dataset contains four feature columns, consistent with those in the transaction data. The next step involves incorporating the 319 ground truth labels generated by Byte Pair Encoding (BPE) as additional feature columns into the customer dataset, resulting in a total of 323 feature columns.

To represent each customer's spending level, the F-tag values in the case study are expressed using the RFM model's expenditure amount, calculated as (Total Transaction Amount) / (Total Number of Transactions). The total number of transactions is calculated as the sum of the occurrences of transactions corresponding to the label within the customer's two-year transaction history, indicating the total number of transactions involving that label over the two years. The total transaction amount is computed using the 'monetary' field from the transaction data, reflecting the total amount of transactions associated with the label over the two-year period, expressed in ten thousand units.

According to the RFM description provided above. When calculating the fund label field for Customer A, the relevant transaction records associated with the fund label in the transaction data are used. The total transaction amount, expressed in ten thousand units, is then divided by the total number of transactions, and the resulting value is entered into Customer A's fund label field. Similarly, when calculating the credit card label field for Customer B, the relevant transaction records associated with the credit card label in the transaction data are used. The total transaction amount, expressed in ten thousand units, is divided by the total number of transactions, and the resulting value is entered into Customer B's credit card label field.
Through the application of the RFM model as described above, the F-tag values provide insights into the relationship between the total transaction amount and the number of transactions associated with each label for the customer up to the end of August 2023.

T-tag and combinations of labels The marketing list dataset was pre-screened by domain experts, specifically the investment products department manager, to define the target T-tags. Based on the F-tags derived from the transaction data, and following discussions among the three experts, a total of eight T-tags were determined, including but not limited to: (Dividend, Year-End Bonus, Holiday, Retirement Pension, etc.). The remaining four target labels are related to the descriptions of investment products and internal product sensitivity, and therefore, not all T-tags can be disclosed.
With the target labels established, this case study sets the DFS parameters $pair_{len}$ and $pair_{proportion}$ to 3 and 0.6, respectively, to identify similar label combinations with the target label as the root node. The aim was to find combinations where there are at least 3 nodes and the co-occurrence between labels exceeds 0.6. The DFS search results, sorted by the number of combinations, are as follows: (119 combinations, 16 combinations, 108 combinations, 3 combinations, 133 combinations, 59 combinations, 60 combinations, 43 combinations). To avoid violations of financial regulations and personal data protection laws [93], and to prevent the illegal misuse of personal information, the specific label combinations cannot be disclosed.

D-tag Following a discussion among three experts, including the investment products department manager, it was decided that there are 11 derived label combinations. Since the D-tags involve aspects such as the characteristics of investment products, customer response rates, consumer habits, subscription outcomes, and company sensitivity, disclosing these could potentially lead to the inadvertent exposure of specific customer characteristics and legal issues. To prevent misunderstandings related to the use of personal data and any illegal intentions [93], the D-tags cannot be disclosed.

Model Prediction and Value Assessment The experiments for this case study were implemented using Python 3.6 and PySpark 1.4 in the Cloudera Data Science Workbench. The hardware used consisted of an Intel Xeon 64-bit processor with 16 cores and 128GB of RAM.
As described in Sections 4.3 and 4.5, the LightGBM model was used with D-tags treated as the actual class for training. The prediction task involved identifying investment product purchasers between September 1, 2023, and October 31, 2023. The results are detailed in the table below, with an average precision of 0.941, a recall of 0.899, and an F1 score of 0.938 for the 11 D-tags.

Table 29. Model Prediction Scores

D-tag	Evaluation categories	Score
11 D-tag sets	Average precision	0.941
	Average recall	0.899
	Average F1 score	0.938

Since all three evaluation metrics exceed the threshold of 0.8 set within the interpretability framework, the process moves on to the next step of value assessment. This study uses A/B testing to validate actual effectiveness. A list of potential customers for the same investment products identified by domain experts based on past experience is used as the benchmark (Group A), consisting of 91,018 individuals. The proposed framework predicts 226,998 potential buyers (Group B), which is 2.493 times higher than the expert-provided list.

Digital advertisements for investment products were targeted to the customer lists provided by each group (A and B) through an advertising deployment system. Under identical advertising copy, this study tracked and compiled data only for customers within one month after the advertisement was deployed. Due to the challenges in objectively defining transaction tracking and effectiveness of investment products, data on customer repurchase rates or cost-effectiveness could not be obtained for comparison. Therefore, only customer response rates and the number of respondents were compared.

The test results indicate that the customer response rate is 3.8 times higher than that of the list provided by industry experts, and the total number of responses is 9 times greater, as detailed in Table 30.

Table 30. Value Assessment Table-3

Under an F1 Score Greater Than 0.8	Interpretability Framework	Previous Experience
Predicted Number of Purchasers	226,998 (Number of individuals)	91,018 (Number of individuals)
Difference	2.493x	1x
Customer Response Rate Within One Month of Advertisement Deployment	3.8x	1x
Number of Responses Within One Month of Advertisement Deployment	9x	1x

Interpretability and Persona Due to regulatory constraints, it is not possible to display LIME and SHAP model visualizations. However, by utilizing an explainability framework and anonymizing the data, some co-occurrence labels can still be presented. The order of the labels does not reflect their actual ranking, and identifiable labels have been omitted, as shown in the table below.

Table 31. Table-5 of Co-occurrence Scores for D-tags

D-tag	F-tag	Co-occurrence Score
(ETF A, ETF B, ETF C)	technology industry	0.719
	regular installment plans	0.644
	group dining	0.355
	Shin Kong Mitsukoshi	0.181
	afternoon tea	0.140

The table presents a randomly selected de-identified label from 11 derived label sets for co-occurrence analysis. It reveals that among the labels associated with the three ETFs, aside from the strong correlation with their own labels, the ”technology industry” fact label exhibits a particularly high level of co-occurrence. Additionally, terms such as ”regular installment plans” frequently appear in the descriptions of these products, attracting a diverse range of customers. In the persona analysis, Table 32 indicates that the purchaser

Table 32. Persona and F-tags (5)

D-tag	F-tag	RFM value
(ETF A, ETF B, ETF C)	rent	9.50
	afternoon tea	8.27
	group dining	6.83
	technology industry	4.29
	steak	2.21

frequently has the label ”rent” noted in their account records, and they tend to spend a significant amount on afternoon tea and group dining. This individual is also likely employed in the technology industry. Experts suggest that the profile aligns closely with that of an engineer working in a science park. Alternatively, it could describe a financially savvy individual, possibly a landlord, who specializes in managing rental properties. This person predominantly invests in technology-related portfolios, which may explain the high frequency and expenditure on afternoon tea and group dining.

The validation results not only significantly increased customer response rates, response volume, and the number of targeted individuals but also enabled domain experts to interpret the results through personas derived from the data. This enhances data transparency and model interpretability, allowing experts to better understand the predictive outcomes and reducing skepticism toward the model. These findings further demonstrate the practical feasibility of the explainability framework and its potential for increased profitability.

5. Conclusions and Discussion

This study addresses cases where the customer lists generated by the model were deemed unacceptable. Specifically, it focuses on issues such as the inability to explain the lists, the features of the lists failing to persuade domain experts and decision-makers, and the limitations on the number of individuals in the lists based on past experience. To address these challenges, this research proposes an explainability framework as a solution.

The framework integrates BPE and a three-tier labeling system to enhance the interpretability of the model's results. Fact labels expand the feature dimensions of customer data, enabling users to perceive the functional dimensions of the data through natural language [94]. Industry experts can select target labels from the fact labels and determine derived labels for the model's actual categories based on label combinations identified by DFS. This allows domain experts to fully understand that the customer lists predicted by the model are generated from natural language features, and further explanations can be provided through personas. This approach fully satisfies the transparency, comprehensibility, and interpretability requirements of XML [36]. Moreover, the relationship between derived and target labels can be further elaborated to provide contextual explanations, aligning with the definition of XAI [37].

This study develops a general explainability framework to address challenges in the business domain, where industry experts or decision-makers may reject model-generated potential customer lists, and where the number of marketing list recipients is limited by past experience. The experimental results show that:

1. By labeling data with natural language, this framework enhances data interpretability for any user and produces comprehensible potential customer lists. It effectively increases response rates and the number of recipients on the lists, offering a higher chance of generating greater corporate profits.
2. Although designed for the business domain, the framework is repeatable and generalizable, applicable to any dataset involving natural language. It can be adopted to enhance both the feature dimensions and readability of data, helping users better understand its behavior and characteristics.
3. Grounded in the experience of domain experts and decision-makers, this framework successfully transfers prior knowledge and domain expertise into the model. In the future, experts can confidently leverage technological advancements, and managers can more easily monitor changes in customer consumption patterns and habits.

The proposed framework can be further extended to improve its adaptability across industries and alignment with cutting-edge technologies through the following directions:

- Multimodal Data Fusion: In retail scenarios, integrating product images (e.g., clothing design sketches) with textual reviews via vision-language models such as CLIP can generate cross-modal tags, thereby enriching user profiling.
- Federated Learning: In privacy-sensitive domains such as finance and healthcare, distributed model training enables collaborative modeling (e.g., credit risk assessment across banks) while preserving user data privacy by avoiding raw data exchange.
- Replacing DFS with Graph Neural Networks (GNNs): Instead of heuristic DFS-based search, the label co-occurrence structure can be directly modeled using GNNs. Graph Attention Networks (GAT), in particular, can capture complex inter-label relationships (as discussed in Section 4.6.4), offering a more expressive alternative.

- Improving Interpretability with Large Language Models (LLMs): Prompt engineering techniques can leverage models like GPT-4 to automatically generate semantic explanations for tags (e.g., defining the business meaning of "durability"), thereby reducing reliance on domain experts.

While the proposed framework demonstrates strong performance, several limitations should be acknowledged:

- Dependence on Data Quality: The framework's effectiveness relies heavily on the completeness and accuracy of natural language fields (e.g., product descriptions). High levels of noise—such as spelling errors or ambiguous expressions—may lead to suboptimal tag generation by BPE. For example, as shown in Table 4, subwords like "cappuccin" require manual correction to align with intended semantics.
- Expert Involvement Overhead: The DFS-generated tag combinations require manual filtering by domain experts. As illustrated in Section 4.2.3, only 2 out of 74 candidate D-tag combinations were selected for downstream use, which limits the framework's automation in knowledge-scarce scenarios.
- Computational Bottlenecks: Both BPE and DFS may incur substantial memory and time costs when applied to large-scale datasets, such as the 185 million transaction records used in the case study. Distributed computing frameworks (e.g., Apache Spark) or approximate algorithms may be necessary to improve scalability.
- Limited Adaptability to Dynamic Data: The current framework does not account for data distribution shifts over time (e.g., evolving consumer preferences). Future work should explore online learning mechanisms to periodically update the tag taxonomy and maintain robustness under dynamic conditions.

Acknowledgments. This research was supported in part by the National Science and Technology Council, R.O.C. under grant MOST 110-2221-E-007-107-MY3, NSTC 112-2221-E-007-086 and NSTC 113-2221-E-007-117-MY3.

References

1. A. Rakipi, O. Shurdi, and J. Imami, "Utilization of data mining and machine learning in digital and electronic payments in banks," *Corporate and Business Strategy Review*, vol. 4, no. 4, pp. 243–251, 2023.
2. W. Yeh, M. Chuang, and W. Lee, "Uniform parallel machine scheduling with resource consumption constraint," *Applied Mathematical Modelling*, vol. 39, no. 8, pp. 2131–2138, 2015.
3. W. Yeh and S. Wei, "Economic-based resource allocation for reliable grid-computing service based on grid bank," *Future Generation Computer Systems*, vol. 28, no. 7, pp. 989–1002, 2012.
4. K. Pousttchi and M. Dehnert, "Exploring the digitalization impact on consumer decision-making in retail banking," *Electronic Markets*, vol. 28, no. 3, pp. 265–286, 2018.
5. P. Angelov, E. Soares, R. Jiang, N. Arnold, and P. Atkinson, "Explainable artificial intelligence: an analytical review," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. 2021, 2021.
6. J. Achiam and et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
7. H. Touvron and et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

8. "Introducing llama: A foundational, 65-billion-parameter language model." https://ai.meta.com/blog/large-language-model-llama-meta-ai/. Accessed: 2024/4/2.
9. S. Spatharioti, D. Rothschild, D. Goldstein, and J. Hofman, "Comparing traditional and llm-based search for consumer choice: A randomized experiment," *arXiv preprint arXiv:2307.03744*, 2023.
10. H. Corley, J. Rosenberger, W. Yeh, and T. Sung, "The cosine simplex algorithm," *The International Journal of Advanced Manufacturing Technology*, vol. 27, pp. 1047–1050, 2006.
11. B. Arcila, "Is it a platform? is it a search engine? it's chatgpt! the european liability regime for large language models," *Journal of Free Speech Law*, vol. 3, p. 455, 2023.
12. W. Yeh, "Novel binary-addition tree algorithm (bat) for binary-state network reliability problem," *Reliability Engineering and System Safety*, vol. 208, p. 107448, 2021.
13. M. Karpinska and M. Iyyer, "Large language models effectively leverage document-level context for literary translation, but critical errors persist," in *Proceedings of the Eighth Conference on Machine Translation*, 2023.
14. W. Yeh, "A new branch-and-bound approach for the n/2/flowshop/f+ cmax flowshop scheduling problem," *Computers & Operations Research*, vol. 26, no. 13, pp. 1293–1310, 1999.
15. A. Thirunavukarasu, D. Ting, K. Elangovan, L. Gutierrez, T. Tan, and D. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.
16. C. Luo, B. Sun, K. Yang, T. Lu, and W. Yeh, "Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme," *Infrared Physics & Technology*, vol. 99, pp. 265–276, 2019.
17. A. Mbakwe, I. Lourentzou, L. Celi, O. Mechanic, and A. Dagan, "Chatgpt passing usmle shines a spotlight on the flaws of medical education," *PLOS Digital Health*, vol. 2, no. 2, p. e0000205, 2023.
18. N. Chiliya, G. Herbst, and M. Roberts-Lombard, "The impact of marketing strategies on profitability of small grocery shops in south african townships," *African Journal of Business Management*, vol. 3, no. 3, p. 70, 2009.
19. T. Damrongsakmethee and V.-E. Neagoe, "Data mining and machine learning for financial analysis," *Indian Journal of Science and Technology*, vol. 10, no. 39, pp. 1–7, 2017.
20. R. Aditya and D. Satria, "Optimizing bank marketing strategies through analysis using lightgbm," *CoreID Journal*, vol. 1, no. 2, pp. 58–65, 2023.
21. S. Shim, M. Eastlick, and S. Lotz, "Search-purchase (s-p) strategies of multi-channel consumers," *Journal of Marketing Channels*, vol. 11, no. 2-3, pp. 33–54, 2004.
22. A. Faria and W. Wellington, "Validating business gaming: Business game conformity with pims findings," *Simulation & Gaming*, vol. 36, no. 2, pp. 259–273, 2005.
23. P. Chate, *Behavioral Modelling of Customer Marketing Patterns and Review Prediction Using Machine Learning Techniques*. PhD thesis, National College of Ireland, Dublin, 2022.
24. M. Muslim, Y. Dasril, A. Alamsyah, and T. Mustaqim, "Bank predictions for prospective long-term deposit investors using machine learning lightgbm and smote," *Journal of Physics: Conference Series*, vol. 1918, no. 4, p. 042143, 2021.
25. E. Broek, A. Sergeeva, and M. Huysman, "When the machine meets the expert: An ethnography of developing ai for hiring," *MIS Quarterly*, vol. 45, no. 3, 2021.
26. T. Jovanov and M. Stojanovski, "Marketing knowledge and strategy for smes: Can they live without it?," in *Thematic Collection of papers of international significance: Reengineering and entrepreneurship under the contemporary conditions of enterprise business*, pp. 131–143, 2012.
27. Y. Huang, M. Zhang, and Y. He, "Research on improved rfm customer segmentation model based on k-means algorithm," in *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*, 2020.

28. E. Soares, P. Angelov, B. Costa, and M. Castro, "Actively semi-supervised deep rule-based classifier applied to adverse driving scenarios," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019.
29. R. Blanco and C. Lioma, "Graph-based term weighting for information retrieval," *Information Retrieval*, vol. 15, no. 1, pp. 54–92, 2011.
30. X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical Image Analysis*, vol. 69, p. 101985, 2021.
31. C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu, "Integrating machine learning with human knowledge," *iScience*, vol. 23, no. 11, p. 101656, 2020.
32. V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers in Big Data*, vol. 4, p. 688969, 2021.
33. S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
34. A. Arrieta and et al., "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.
35. Z. Liu, W. Yeh, K. Lin, C. Lin, and C. Chang, "Machine learning based approach for exploring online shopping behavior and preferences with eye tracking," *Computer Science and Information Systems*, vol. 21, no. 2, pp. 593–623, 2024.
36. R. Roscher, B. Bohn, M. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020.
37. H. Chia, "The emergence and need for explainable ai," *Advances in Engineering Innovation*, vol. 3, no. 1, pp. 1–4, 2023.
38. E. Soares, P. Angelov, S. Biaso, M. Froes, and D. Abe, "Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification," *MedRxiv*, 2020.
39. F. Morais, A. Garcia, P. Santos, and L. Ribeiro, "Do explainable ai techniques effectively explain their rationale? a case study from the domain expert's perspective," in *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2023.
40. J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, no. 1, p. eaao5580, 2018.
41. A. Smith-Renner, R. Rua, and M. Colony, "Towards an explainable threat detection tool," in *IUI Workshops*, 2019.
42. S. Mathews, "Explainable artificial intelligence applications in nlp, biomedical, and malware classification: A literature review," in *Advances in Intelligent Systems and Computing*, pp. 1269–1292, Springer International Publishing, 2019.
43. A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.
44. S. Murindanyi, B. Mugalu, J. Nakatumba-Nabende, and G. Marvin, "Interpretable machine learning for predicting customer churn in retail banking," in *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2023.
45. T. Clement, N. Kemmerzell, M. Abdelaal, and M. Amberg, "Xair: A systematic metareview of explainable ai (xai) aligned to the software development process," *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 78–108, 2023.
46. Y. Han, "Research on precise service of academic journals based on user profile," *Acta Editologica*, vol. 2, pp. 142–146, 2021.
47. D. Travis, "How to create personas your design team will believe in." `https://www.userfocus.co.uk/articles/personas.html`. Accessed: 2024/4/2.
48. Y. Chang, Y. Lim, and E. Stolterman, "Personas: from theory to practices," in *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, pp. 439–442, 2008.

49. L. W., O. K., L. C.G., and C. H.J., "User profile extraction from twitter for personalized news recommendation," in *16th International conference on advanced communication technology*, pp. 779–783, IEEE, 2014.
50. M. Raghuram, K. Akshay, and K. Chandrasekaran, "Efficient user profiling in twitter social network using traditional classifiers," in *Advances in Intelligent Systems and Computing*, pp. 399–411, Springer International Publishing, 2015.
51. R. Bonnie, "The power of the persona." `https://www.pragmaticinstitute.com/resources/articles/product/the-power-of-the-persona/`. Accessed: 2024/4/2.
52. Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, vol. 4, no. 2, p. 100211, 2024.
53. K. Bostrom and G. Durrett, "Byte pair encoding is suboptimal for language model pretraining," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.
54. P. Gage, "A new algorithm for data compression," *The C Users Journal*, vol. 12, no. 2, pp. 23–38, 1994.
55. J. Zhan and et al., "An effective feature representation of web log data by leveraging byte pair encoding and tf-idf," in *Proceedings of the ACM Turing Celebration Conference-China*, pp. 1–6, 2019.
56. "Summary of the tokenizers." `https://huggingface.co/docs/transformers/tokenizer_summary#summary-of-the-tokenizers`. Accessed: 2024/4/2.
57. Thomwolf, "Bpe tokenizers and spaces before words." `https://discuss.huggingface.co/t/bpe-tokenizers-and-spaces-before-words/475`. Accessed: 2024/4/10.
58. R. A. and S. Borah, "Study of various methods for tokenization," in *Applications of Internet of Things*, pp. 193–200, Springer Singapore, 2020.
59. X. Gutierrez-Vasques, C. Bentz, and T. Samardžić, "Languages through the looking glass of bpe compression," *Computational Linguistics*, vol. 49, no. 4, pp. 943–1001, 2023.
60. N. Tavabi and K. Lerman, "Pattern discovery in physiological data with byte pair encoding," in *Multimodal AI in Healthcare*, pp. 227–243, Springer International Publishing, 2022.
61. N. Fradet, N. Gutowski, F. Chhel, and J. Briot, "Byte pair encoding for symbolic music," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
62. H. Liu, "Byte-pair and n-gram convolutional methods of analysing automatically disseminated content on social platforms," *MDPI AG*, 2020.
63. N. Nilsson, *Principles of Artificial Intelligence*. Springer Berlin Heidelberg, 1982.
64. F. Harary, "The explosive growth of graph theory," *Annals of the New York Academy of Sciences*, vol. 328, no. 1, pp. 5–11, 1979.
65. R. Tarjan, "Depth-first search and linear graph algorithms," *SIAM Journal on Computing*, vol. 1, no. 2, pp. 146–160, 1972.
66. C. Photphanloet and R. Lipikorn, "Pm10 concentration forecast using modified depth-first search and supervised learning neural network," *Science of The Total Environment*, vol. 727, p. 138507, 2020.
67. S. Rahmani, S. Fakhrahmad, and M. Sadreddini, "Co-occurrence graph-based context adaptation: a new unsupervised approach to word sense disambiguation," *Digital Scholarship in the Humanities*, vol. 36, no. 2, pp. 449–471, 2020.
68. Y. Du, F. Li, T. Zheng, and J. Li, "Fast cascading outage screening based on deep convolutional neural network and depth-first search," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2704–2715, 2020.
69. Q. Mei and M. Gül, "Multi-level feature fusion in densely connected deep-learning architecture and depth-first search for crack segmentation on images collected with smartphones," *Structural Health Monitoring*, vol. 19, no. 6, pp. 1726–1744, 2020.

70. A. Syah, F. Helmiah, N. Irawati, and N. Hasibuan, "Depth first search algorithm in the expert system for diagnosis of palm oil growth obstacles," in *4TH INTERNATIONAL CONFERENCE ON CURRENT TRENDS IN MATERIALS SCIENCE AND ENGINEERING 2022*, 2024.
71. G. Logeswari, S. Bose, and T. Anitha, "An intrusion detection system for sdn using machine learning," *Intelligent Automation & Soft Computing*, vol. 35, no. 1, pp. 867–880, 2023.
72. W. Cai, R. Wei, L. Xu, and X. Ding, "A method for modelling greenhouse temperature using gradient boost decision tree," *Information Processing in Agriculture*, vol. 9, no. 3, pp. 343–354, 2022.
73. G. Ke and et al., "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, vol. 30, 2017.
74. B. Wardani, S. Sa'adah, and D. Nurjanah, "Measuring and mitigating bias in bank customers data with xgboost, lightgbm, and random forest algorithm," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 1, pp. 142–155, 2023.
75. Y. Hua, "An efficient traffic classification scheme using embedded feature selection and lightgbm," in *2020 Information Communication Technologies Conference (ICTC)*, 2020.
76. N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
77. J. Ponsam, S. Gracia, G. Geetha, S. Karpaselvi, and K. Nimala, "Credit risk analysis using lightgbm and a comparative study of popular algorithms," in *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, 2021.
78. Y. Wong, K. Madhavan, and N. Elmqvist, "Towards characterizing domain experts as a user group," in *2018 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, pp. 1–10, 2018.
79. P. Fadde and P. Sullivan, "Developing expertise and expert performance," in *Handbook of Research in Educational Communications and Technology: Learning Design*, pp. 53–72, 2020.
80. K. Chandrasekaran, *Domain-Driven Design with Java - A Practitioner's Guide: Create simple, elegant, and valuable software solutions for complex business problems*. Packt Publishing, 2021. https://ddd-practitioners.com/home/glossary/domain-expert/.
81. Vujović, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.
82. T. Saito and M. Rehmsmeier, "Basic evaluation measures from the confusion matrix." https://classeval.wordpress.com/introduction/basic-evaluation-measures/, 2017.
83. P. Le, M. Nauta, V. Nguyen, S. Pathak, J. Schlötterer, and C. Seifert, "Benchmarking explainable ai - a survey on available toolkits and open challenges," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.
84. A. Holzinger, "From machine learning to explainable ai," in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 2018.
85. Z. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
86. C. Molnar, *Interpretable machine learning*. Lulu.com, 2020.
87. J. Karkavelraja, "Amazon sales dataset." https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset. Accessed: 2024/4/2.
88. A. Gupta, A. Raghav, and S. Srivastava, "Comparative study of machine learning algorithms for portuguese bank data," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021.
89. Lavanya, "Google play store apps." https://www.kaggle.com/datasets/lava18/google-play-store-apps. Accessed: 2024/4/10.
90. P. Lokesh, "Amazon products sales dataset 2023." https://www.kaggle.com/datasets/lokeshparab/amazon-products-dataset. Accessed: 2024/4/2.
91. D. Chen, "Online retail." UCI Machine Learning Repository, 2015.
92. "Personal data protection act." https://law.moj.gov.tw/LawClass/LawAll.aspx?PCODE=G0380233. Accessed: 2024/4/2.

93. "Banking act." https://law.fsc.gov.tw/LawContent.aspx?id=GL000624. Accessed: 2024/4/2.
94. A. Caramazza and J. Shelton, "Domain-specific knowledge systems in the brain: The animate-inanimate distinction," *Journal of Cognitive Neuroscience*, vol. 10, no. 1, pp. 1–34, 1998.

Zhenyao Liu is currently an Assistant Professor of the School of Economics and Management, Taizhou University in Jiangsu Province, China. He received Ph.D. degree from the Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan. His research areas are soft computing and machine learning.

Yu-Lun Liu received M.S. degree from the Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan.

Wei-Chang Yeh received the M.S. and Ph.D. degrees from the Department of Industrial Engineering, University of Texas at Arlington. He is currently a Chair Professor of the Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan. Most of his research is focused on algorithms, including exact solution methods and soft computing. He has published more than 300 research articles in highly ranked journals and conference papers.

Chia-Ling Huang is currently a Professor of the Department of International Logistics and Transportation Management, Kainan University, Taiwan.

Received: November 30, 2024; Accepted: June 25, 2025.

Computer Science and Information Systems 22(4):1757–1775 https://doi.org/10.2298/CSIS241204069L

Formative Interviews for a User-Centered Design Study on Developing an Effective Gateway for Health Research Data Search – Towards a Sustainable Wellbeing Environment

Hsiu-An Lee[1,2], Tung Lin[3], Hsin-I Chen[1], Wei-Chen Liu[1], Yen-Ju Shen[1], Wen-Chang Tseng[1], and Chien-Yeh Hsu[2,4] and Yi-Hsin Yang[1,⋆]

[1] National Institute of Cancer Research, National Health Research Institutes
No.367, Sheng-Li Rd., North District, Tainan, 70456 Taiwan
100510@nhri.edu.tw
denise9306@nhri.edu.tw
q09855213@nhri.edu.tw
a0979251512@gmail.com
gdi89009@nhri.edu.tw
yhyang@nhri.edu.tw
[2] Standards and Interoperability Lab, Smart Healthcare Center of Excellence
No.365, Mingde Rd., Peitou Dist., Taipei City 112303, Taiwan
[3] Island Design Lab, F7, No.27, Ln. 66, Sec. 4, Heping E. Rd., Wenshan Dist., Taipei City, Taiwan
tunglin.sy@gmail.com
[4] Department of Information Management, National Taipei University of Nursing and Health Sciences, No.365, Mingde Rd., Peitou Dist., Taipei City 112303, Taiwan
cyhsu@ntunhs.edu.tw

Abstract. Despite the abundance of biomedical databases in Taiwan, there is currently no unified portal that effectively facilitates health research data searches to drive scientific discovery and promote a sustainable wellbeing environment. This study aims to design a user-centered gateway for health research data search, focusing on usability and ensuring that the platform supports the retrieval of fit-for-purpose datasets while maintaining data privacy, accessibility, and transparency.

A user-centered design approach was employed, involving personal interviews with domain experts. An initial set of questions, derived from literature reviews and expert consultations, explored various dimensions of health data usability. The interview results identified key criteria for assessing the effectiveness of health research data searches in supporting sustainable health outcomes.

Seven critical factors were identified for quick confirmation of search requirements: follow-up, publisher, purpose, source, time lag, data custodian, and specific requirements. The interviews also highlighted a lack of familiarity with dataset retrieval tools, emphasizing the need for cultivating user knowledge and habits to promote wider adoption and effective use of the gateway.

As dataset retrieval needs in Taiwan remain a relatively new area, understanding the characteristics of datasets and tailoring search patterns to meet user requirements are essential. This framework provides a foundation for improving health data accessibility. Future research should explore advanced methodologies for addressing

⋆ Corresponding author

diverse user needs, including intelligent recommendation systems to support a sustainable wellbeing environment.

Keywords: Metadata, Big Data, Real-world Data, User-Centered Design

1. Introduction

Digital health, as outlined in the World Health Organization's strategic plan, has the potential to revolutionize global healthcare [2]. The Covid-19 pandemic underscored the critical role of data and artificial intelligence in devising effective strategies to combat the virus. These technologies have significantly contributed to disease trend modeling, precise diagnosis, symptom categorization, result interpretation, vaccine development, therapy advancements, drug innovations, and forecasting medical demand hotspots [5].

As the volume and diversity of data generated and presented in various formats continue to grow, retrieving essential information efficiently has become increasingly challenging. Many large-scale research databases are constantly evolving, yet their content, usage guidelines, and application scopes are fragmented across platforms, hindering quick and accurate information retrieval for users. Addressing this challenge requires the development of a health data gateway adhering to FAIR principles (findable, accessible, interoperable, reusable) [16].

Several countries have taken proactive measures by establishing integrated platforms to facilitate data sharing, thereby enhancing accessibility and fostering collaboration. However, establishing a consistent data presentation framework remains a significant challenge due to diverse user backgrounds and needs. Notable examples such as BBMRI-ERIC in Europe, housing 100 million samples and delineating quality standards for European biobanks, and UK Health Data Research (UK HDR), a collaborative initiative enabling access and utilization of health-related data for research purposes, showcase distinct architectural designs and insights relevant to our proposed platform [3,13].

Taiwan with many large scale biomedical databases currently lacks a comparable integrated platform, motivating our research endeavor. This study seeks to contribute to the scientific community by addressing these challenges and conceptualizing a comprehensive platform that facilitates seamless data integration, promotes collaborative research, and nurtures a more accessible and impactful data ecosystem. Central to this approach is a user-centered design, offering filtering conditions for efficient data retrieval and assessing disparities between datasets. By constructing appropriate data gateways, data can be effectively utilized and an environment for sustainable development of medical technology can be created.

The primary objective of this study is to identify key factors in dataset screening frameworks and data availability criteria. A user-centered gateway plays a pivotal role in determining critical selection factors for data users, shaping data screening pathways, and defining dataset metadata essential for effective data navigation. Our study aims to achieve three main goals: (1) provide metadata for dataset definitions applicable to health data analysis research; (2) propose measures and tools for inclusion in future portals and metadata for research retrieval; and (3) identify areas requiring further research attention. Our researchers have balanced diverse stakeholder needs to design a prototype database search portal (dataset portal). While many national biomedical databases exist in Taiwan, there is a notable gap in developing user-centric platforms that provide intuitive access

and metadata-driven dataset screening mechanisms. This study directly addresses this gap through a structured user-centered design process, aligning with the special issue's focus on computational technologies for sustainable wellbeing environments.

Scientific Contributions of this Study are:

- 1. We introduce a structured metadata filtering framework derived from empirical insights of domain experts.
- 2. We operationalize user-centered design in the context of health dataset discovery, identifying seven metadata-driven screening criteria.
- 3. We develop a prototype interface ("Easy Search") informed by user needs, which bridges qualitative understanding with quantifiable utility indicators.
- 4. We enrich qualitative findings through operational design artifacts and metadata codification, laying groundwork for intelligent data gateway development in sustainable health environments.

2. Materials and Methods

This study employed an in-depth interview-based consensus approach to define critical criteria for dataset filtering. These criteria provide essential information about different dataset metadata and establish operational guidelines for future dataset selection. The researchers reviewed relevant literature to inform this study. Additionally, this research convened a panel of experts comprising individuals from diverse fields, including health information services, medical material testing research and development, drug development, auxiliary medical services, and academic research. These experts collectively discussed the current state of clinical dataset search.

The research focused on gathering insights from this diverse group to achieve its stated objectives. An interview panel was convened specifically to discuss the current state of clinical dataset search. This panel consisted of stakeholders representing various perspectives related to health data, data integration, research, and platform development. The in-depth interviews focused on the following critical points:

- Identifying Key Factors: The interviews aimed to identify essential factors for dataset screening and data availability identification. This involved understanding the criteria that researchers and data users consider important when selecting datasets.
- Designing an Effective Gateway: The panel's input helped in designing a gateway that aligns with identified selection factors, ensuring that the platform effectively meets the needs of data demanders.
- Developing Metadata: Collaboratively defining dataset metadata meaningful for health data analysis research, enabling users to better understand available datasets and their attributes.
- Identifying Research Needs: Recognizing gaps or areas requiring further research, such as understanding specific dataset requirements or addressing challenges related to data availability.
- Balancing Stakeholder Needs: The panel's insights contributed to balancing the needs and expectations of different stakeholders, ensuring that the database search portal prototype addresses various perspectives effectively.

The interview panel comprised experts from diverse backgrounds who collaboratively addressed the study's goals and objectives, ultimately contributing to the development of a comprehensive and impactful health data integration platform.

2.1. Designing Interview Interactions:

(1.) Interview Process and Practical Operation The research process(Fig. 1 shows the Interview and Practical Operation Process.) begins with Interview Design and Interviewee Selection, where the primary focus is on crafting appropriate questions that align with the study's objectives. At this stage, careful consideration is given to selecting the right interviewees whose experiences and insights can provide valuable contributions to the research.

Once the design is finalized, the next step is Interview Preparation. This phase involves refining the interview questions and ensuring that all necessary tools and materials for data collection are prepared. In this study, the interview was conducted online, and prior to the interview, an interview outline along with virtual cards (via a website link) were provided to the interviewees. This allowed them to better understand the purpose of the interview and prepare accordingly.

Following the preparation, the research enters the In-depth Interview phase. This is a key component of the study, where the interviewer engages with the selected individuals to explore their thoughts, experiences, and perspectives in great detail. During the interview, the researchers strictly followed the outline, ensuring the conversation stayed focused. If the interviewee needed to provide additional input or demonstrate practical operations, control of the screen was passed to them. In the third part of the interview, interviewees were asked to operate and explain processes based on their experience. If they lacked recent practical experience, they were encouraged to share their thoughts and needs regarding the database retrieval process.

Finally, the process concludes with the Interactive Interview phase. Unlike traditional one-sided interviews, this stage emphasizes a two-way exchange between the interviewer and the interviewee. The interaction allows for a more dynamic conversation, where both parties contribute to the dialogue, leading to the discovery of deeper insights and a fuller understanding of the subject matter. This interactive format was particularly useful in exploring practical demonstrations and conceptual understanding, further enriching the research findings.

(2.) Interviewee Selection: The researchers of this study employed a purposeful sampling methodology to ensure representation and diversity in the participant group. Experts from various fields were invited to participate in in-depth interviews to assess requirements. The participants were carefully selected to encompass diversity across several working fields, including health information service, medical material testing research and development, drug development, auxiliary medical services, and academic research.

An expert is defined as an individual who meets the following criteria: an active participant in health data analysis and research who can provide insights into the types of data needed, the challenges faced in accessing and using data, and requirements for a user-friendly platform. Also, an individual with expertise in the healthcare and medical fields who can provide insights into the practical applications of health data, the relevance

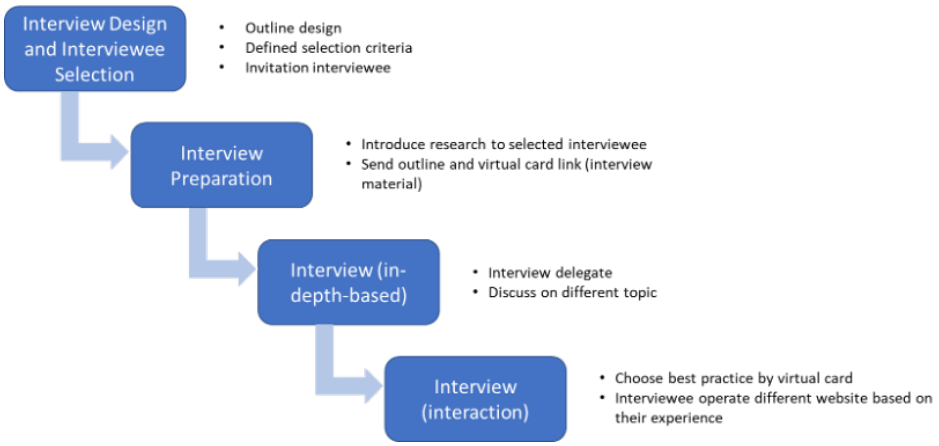


Fig. 1. Interview and Practical Operation Process

of data to medical practice, and the potential benefits of integrated platforms. Additionally, data analysts and data scientists, professionals with experience in data analysis, can provide perspective on technical aspects such as data integration, metadata creation, and the tools needed for effective data analysis.

A total of six participants were invited, and each of them was informed about the purpose of the study, with complete anonymity among experts. This study is primarily intended for the preliminary assessment of data gathering requirements and gateway design elements.

(3.) Key Factor (Factor Card Sorting) for Dataset Screening (Searching): This study summarized some important factors in designing virtual cards based on the initial survey results and referred to the screening field of UK HDR [8], and data explanations provided by UK Biobank [12]. During the interview, the interviewees referred to virtual cards, effectively highlighting the key factors and extending their explanation. The Fig. 2 shows virtual card we used during the interview.

1. Publisher
2. Phenotype
3. Coverage – spatial and follow-up
4. Provenance – purpose, source, collection situation, and time lag
5. Access – delivery lead time, jurisdiction
6. Format and standards – vocabulary, conforms, and language

(4.) Recording of interviews and the coding of questions

2.2. Interview Outline Design

This interview is divided into four sections: the interviewee's background and work experience, database search experience, gateway's functional requirements and expectations,

Fig. 2. Key factor virtual card during the interview

and the interviewee's perspective of the main values and principles of gateway. Due to the varying needs of experts (interviewees) in different fields, some questions were further asked after the survey answers were obtained. The interview guideline is as follows:

Background Survey – 10 mins Respondents introduced the interviewees' work background and current job content.

Dataset Screening Criteria – 40 mins The interviewees were asked about their experience with search databases. In addition to understanding data requirements, it is crucial to grasp the business logic and workflow to the fullest extent possible.

Additional point: If there are concerns related to censorship, inquire about the context of these needs, such as compliance with GDPR norms.

(a) Please share an example of a high-quality or particularly useful dataset, and explain its typical use.

(b) What aspects of data formats or standard models (e.g., the OMOP or the requirements for providing data via FHIR API) are critical to your work?

(c) The "number of items included in the database" is generally indicated for the coverage (rate) of the dataset. What are your needs (e.g., number of observations (value), observation points, etc.) in terms of quickly understanding the coverage (rate)?

(d) Lastly, what features of the dataset are you looking forward to?

Operational Requirements and Expectations of Database Search – 30 mins This section focused on asking the interviewees to share their experience in database searching and, if possible, open the website they use (any website or tool) and show the re-searchers how to use it. Additional point: If the interviewees use more than two portals (websites), please ask them to share the differences, advantages, and disadvantages of their user experience.

(a) Which database search portals (websites) have you used in the last three months?

(b) Please briefly introduce these portals (websites) and explain their significance in your work.

- How can you tell if a database site is useful for your work?
- What information can be used to determine whether a database site is useful? (e.g., other users' opinions, comments, and five-star reviews on the database)
- Have you ever used any portals (websites) that were difficult to understand, counter-intuitive, or otherwise required technical assistance?

(c) Please show an ideal database search portal (website) and explain why it is ideal. If none, what is lacking from the current portal (website) you use?

- What is your favorite feature or function of this site?
- How much time did you spend learning how to use this portal (website)? What learning strategy did you use?
- What was the most confusing or difficult aspect of using this portal (website)?
- Is there anything in the database's search portal (website) architecture that you of-ten find questionable or requires explanation (it cannot be self-evident)?

(d) As a user, what service functions (e.g., structure, metrics, scoring, etc.) do you think the database search portal (website) should include helping you make better use of the dataset?

- Please describe under what circumstances these features positively impact the utilization of datasets.

Gateway Values and Design Principles – 30 mins This section is designed to determine the value and design principles involved in developing a portal for database search.

(a) What principles are critical when developing and designing a database search portal to make it easier to find high-quality data? Additional point: This is an open exploration and is not limited to the content and reference factors of the interviewees' direct answers.

- Which filters are the most/least important for you to use when finding datasets?
- Additional point: This is a semi-structured exploration. If the answer aspect is related to the card, you can follow the opponent's context and ask additional questions. If not, provide a factor card that leads to the category discussion.
- (Provide the factor card.) Based on experience with the database search portal architecture, which factors are particularly useful/useless concerning the card? Why?
- Please try to find a dataset that you find useful on HDR UK.
- How would you rate the importance of the key factors you mentioned in understanding data utility based on your own data needs and experience?
- Are there any other filters you find useful?

(b) Please refer to the table for information about data utility. Which ones do you think are important or helpful to you? Please choose three to five options.

3. Results

Complex topics are deeply explored and analyzed through a series of in-depth one-on-one interviews with experts, allowing a detailed examination of various aspects relevant to the research objectives. Insights gained from these interviews provide valuable perspectives from experts in various fields, contributing to a comprehensive understanding of the subject. Interviews help gather rich qualitative data, allowing the identification of nuanced patterns, perspectives, and underlying themes. A total of 7 result categories were summarized based on the interview, including:

1. How Taiwanese users describe high-quality datasets
2. Experience sharing
3. User needs
4. Website design strategy
5. Metadata filter preference insights
6. Platform prototype development
7. Current data application process and difficulties

The research output incorporates multiple experts, ensuring a holistic view that encompasses different perspectives, disciplines and areas of expertise, thereby enhancing the robustness of the conclusions drawn. The in-depth interview and interactive process allows for detailed exploration of complex concepts, leading to a deeper understanding of complex interrelationships and factors.

1. Preliminary Screening: Initially, we conducted a preliminary screening of experts from various fields to ensure they possess relevant professional knowledge and experience. This may include reviewing their research background, work history, and relevant professional certifications.
2. Criteria Setting: Next, we set criteria for participants to ensure they meet specific conditions required for the study, such as actively engaging in clinical data analysis and research, having experience in gathering usable databases, familiarity with or usage of Taiwan databases, and working in the field of healthcare information and clinical research.
3. Invitation Selection: Based on the preliminary screening and set criteria, we invited experts from different fields to ensure diversity and representation among the interviewees.
4. Confirmation of Participation: We confirmed the willingness and availability of the invited participants to ensure they have sufficient time and resources to participate in the interview process.
5. Interview Conduct: We conducted in-depth interviews to gather insights and opinions from the interviewees to achieve the research objectives and goals.

Through the above selection process, we successfully invited a total of six interviewees who have diverse expertise and experience in different fields. These interviewees have experience in clinical data analysis, gathering usable databases, familiarity with or usage of Taiwan databases, and work in the field of healthcare information and clinical research, providing a range of perspectives and in-depth insights.

In-depth interviews yielded substantive and multifaceted insights. These insights provide a rich qualitative data set that comprehensively explores all dimensions of research

objectives, facilitating nuanced and informed analysis of topics. The background, experience and field of work of the experts are described in the Table. 1. A total of six experienced interviewees from different fields e participated in the in-depth interviews.

Table 1. Background statement of interviewees

No.	Work Field	Research Field	Affiliation	Data analysis experience (years)
1	Drug Developers	Genomics	Researcher	5-10
2	Health Information Service	Epidemic	Researcher	5-10
3	Academic Research	Pharmaceutical Management	Professor	More than 10
4	Medical Material Testing Research and Development	Biomarker	Senior Executive	More than 10
5	Medical Auxiliary Services	Clinical Trials	Researcher	5-10
6	Medical Auxiliary Services	Clinical Trials	Business Manager	More than 10

3.1. How Taiwanese users describe high-quality datasets

(1.) The data has high integrity Respondents expect the data in the target database to be complete and continuous. The target data is considered complete if it contains all the required data fields (such as the clinical data), and it is continuous if the target data has been valid for a period of time. Taiwanese researchers asserted that completeness and continuity ensured quality research. Respondent feedback:

"Whether the data is complete enough will also be affected by continuity. Continuity refers to whether the same question is asked every year in succession. Taiwan's National Health Interview Survey (NHIS) lacks continuity. Some questions were asked in the previous year and will not be asked the following year, causing an interruption. Meanwhile, NHIS in the United States is very continuous. If the continuity is not good, there will be no way to see the difference for several years." - 3

"One of the key factors in determining whether the database is easy to use is completeness. We check to see if the data has been collected at different times. In past experiences, there is a unit that provides a database of about 10,000 patients, but less than half of the people have complete information (such as kidney function and various clinical tests), so there is a gap in completeness." - 4

(2.) Conducive to multi-party cooperation and value-added application The definition of data fields is unified, and there are rules to follow, such as a unified format or filling method to facilitate data cleaning and effective serial file analysis, which is also beneficial for users in terms of collaborating with several stakeholders simultaneously. Respondent feedback:

"I would expect the data connection between different database systems to be easy, and be discussed not only with IT professionals or statisticians, but with clinicians who can participate in the discussion and explore the database. For example, I researched pre-end-stage renal disease (pre-ESRD). We obtained patient-related diagnoses, medication, treatment, examination, medical treatment history, and other relevant information from the case management data sheet (registry data). During analysis, we needed to combine these databases. For example, when analyzing medication, we had to do further analysis from the sorted drug file, which is integrated into many different forms." - 2

"We pay great attention to whether it can be compared to data from different cohorts. Once the data is in our hands, we need to perform data cleaning. I often encounter a

situation wherein the same target is measured, but the same field is stated to be different. Uniformity, maybe capitalization, whether there are spaces or brackets, and so on, or the units are not uniform. It would be best if this part could be unified." - 4

(3.) Good accessibility Users expect the barriers to accessing data to be as low as possible and in line with their research schedules. Good data accessibility involves a friendly application process, short waiting time, and expected frequency of data collection. Users mentioned that hospital-side databases, or databases organized by academic societies, have good accessibility and are more beneficial in achieving research output. Respondent feedback:

"Currently, the health insurance database requires entering the value-added center to access the data. If there is an issue following the analysis, it will be discovered a month later. If another issue is discovered later, there may be a constant need to present it in a monthly update." - 6

(4.) The data is mature enough This characteristic refers to whether the data collection period is sufficient to verify the research hypothesis, meet the research needs of the Taiwan region, and provide users with confidence in their ability to produce a certain level of analysis results. Simultaneously, the data covering a long-time span can avoid the inconvenience caused by time delay, suggesting there is no need to update and analyze the data repeatedly. In other words, the less affected by time, the more mature the material is. Respondent feedback:

"Maturity is a time-dependent outcome. Can I analyze, at least, something like median survival before my study is closed?" - 2

"At present, all hospitals assume that their data are all from patients who will go to them for a long time. If there is no long-term data, there will be no way to know whether these people are representative. Another possibility is that multiple rounds of analysis will be required." - 6

3.2. Experience-sharing

(1.) NHIS – easily accessible, with complete information Respondent feedback:

"The data collected by NHIS are sorted out every year and can be used directly after downloading. There are guidelines on how to obtain and merge the data. However, there are guidelines on SAS coding and file conversion. Therefore, data security and accuracy are very high. In NHIS, a person is an identification code, so data processing is fairly simple. Instead of merging additional files, one can just be pulled. Files are also very easy to obtain without any hindrance." - 5

(2.) Government open information platform - can be browsed quickly, grasps the information overview before applying The Taiwan government's open data platform (https://data.gov.tw/) contains several interpretations of data sets. Some even provide online viewing of demonstration data, which can obtain much information before data selection to ensure that the obtained data is suitable for analysis. Respondent feedback:

”On the open data platform, just press the sample data button to immediately see what the data looks like. The number of people and the amount in the sample data are good points to consider. Whether it is from a research or business perspective, knowing how much is enough can help professionals quickly determine whether to use it.” - 5

3.3. User needs

(1.) Interface design - Gradually develops the habit of using search engines Users in relevant fields in Taiwan have not yet used search engines to find data-bases. They usually identify and recognize usable databases based on referrals and by reading papers and materials.

”I know what topic I am looking for. Using too many keywords can be intrusive. But, if you want to explore how the database can be applied, keywords or filters are a good choice.” - 6

(2.) Information Design - Search results should display cross-domain key information Users expect that the search data set can briefly describe the data contained in the database, such as a list of data fields, data volume, complete transaction numbers, or missing data rate (missing rate). On the database introduction page, users can disclose complete information. At this time, the information can cover the different needs of various fields as completely as possible. Respondent feedback:

”Most databases should not have complete information. However, some research fields may only require some variables depending on the research field. In contrast, some research may require all of them. What the website platform can provide is the approximate missing rate of each database or each field, making it easier for researchers to evaluate the database effectiveness.” - 5

3.4. Website Design Strategy

(1.) Interface interaction - Open for self-downloading, simplifies the application process The interface presents a simple and clear call-to-action design on the search result page, lowers the threshold for data acquisition, and increases the data utilization rate. For example, in-depth cooperation users plan a concise application process for heavyweight or special data sets after logging in, design services for users’ common contact points (such as web pages, phone calls, or emails), and provide users with a painless application process. Respondent feedback:

”When I found the data I wanted to use, I noticed that I couldn’t download it. At this time, I went to find out how to apply. This part of the US TCGA function is deeply hidden. I can’t quickly identify which data should be logged in and cannot be obtained.” - 1

(2.) Quickly browses and grasps the information overview before applying Users can inspect the metadata of the data set for the search results, as well as data samples, to understand the cross-section of the data, give other users concise information in evaluating whether it is a desired high-quality file, and provide the users with the database application rate. Respondent feedback:

”I would like to know the experimental design method and materials involved in the data collection. If you see that the analysis method of this data is consistent with the

analysis method of my previous genetic data, you can save the original unorganized data and directly download the sorted vcf." - 1

(3.) Clear instruction documents to enhance the freedom of data To improve data usability, an explanatory document should be provided on the introduction page of the search database (collection), which should include the definition of data fields and the data connection method to ensure that users can freely utilize the data after downloading.

3.5. Metadata Filter Preference Insights

For user-centered design, this study intended to let users experience the actual operation process and leverage the "operation process" in enhancing the survey's effect. This study included interactions at the end of the interview.

Referring to their virtual cards during the interview, the interviewees effectively pointed out key factors and extended their explanation. Data quality management is the most important item according to statistics on each interviewee's preference insight, and data quality can assist demanders in ensuring that the data they found can be used. The second most important item is interpreting data integrity and compliance, which allows demanders to effectively evaluate whether the data is aligned with their intended use and rules. Meanwhile, the last item points out to data dictionary and semantic library, which can improve data usability and accuracy.

Data utility allowed the interviewees to choose virtual cards during the interviews. Data utility is primarily applicable to data set interpretation. It guides users in determining which indicators help understand whether a data set has good data utility (e.g., facilitating discussions with colleagues, supervisors, or other stakeholders or facilitating research business advancement).

The virtual cards, where each card represented a specific aspect and included possible options. The Utility items included: Documentation Completeness, Availability of documentation and support, Data Model, Data Dictionary, Provenance, Data Quality Management Process, DAMA Quality Dimensions, Pathway coverage, Length of follow-up, Allowable uses, Time Lag, Timeliness, Linkages, and Data Enrichment. During the question and answer process, users selected indicators from the virtual cards that were relevant to their usage scenario, and these selections were then compiled to identify the most frequently mentioned Data Utility items.

The Fig. 3 shows the statistical results of data utility selection. Over half of the interviewees identified four important items: documentation completeness, data quality management process, DAMA quality dimensions, and allowable uses.

(1.)Documentation completeness Documentation completeness refers to the availability of comprehensive documentation for clinical research datasets. This includes detailed information about data sources, data collection methods, variables, data formats, and any transformations or pre-processing steps used. Complete documentation is crucial for researchers to under-stand the datasets, replicate analyses, and interpret the results accurately. It contributes to clinical research transparency and reproducibility.

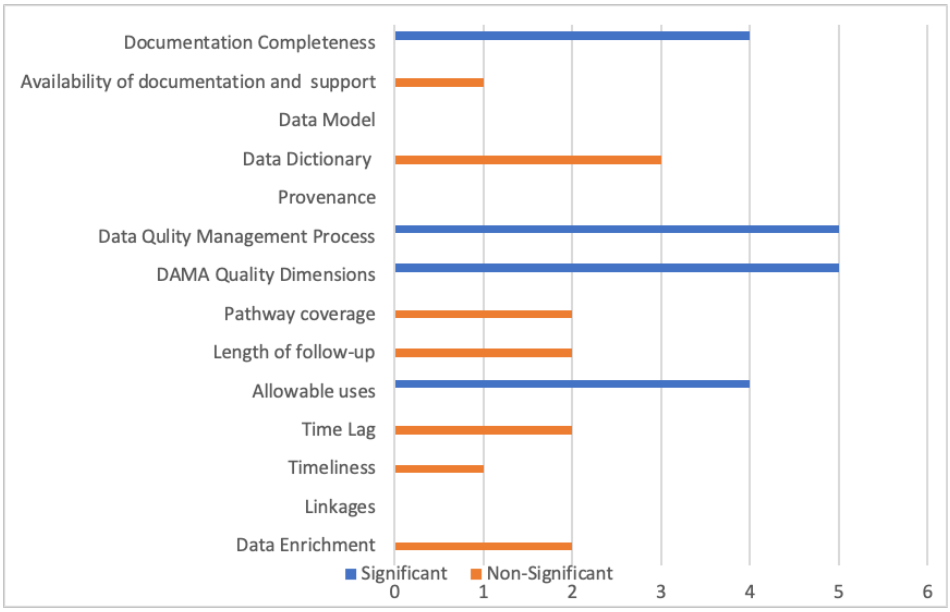


Fig. 3. Data utility selection results

(2.)Data quality management process Data quality management is critical in clinical research dataset searches. It entails the implementation of processes and procedures to monitor, assess, and improve data quality. This includes identifying and addressing data errors, inconsistencies, missing values, and outliers. A robust data quality management process assists researchers in ensuring the reliability and validity of the datasets they use, leading to more accurate and meaningful research outcomes.

(3.)Data quality dimensions Data quality dimensions provide a framework for evaluating and assessing data quality. These dimensions include several factors, such as accuracy, completeness, consistency, timeliness, uniqueness, and relevance. Evaluating these dimensions is crucial in clinical research dataset search as it assists researchers in understanding the strengths and limitations of the datasets they are working with. Addressing these dimensions helps researchers ensure that data fits the intended research purpose.

(4.)Allowable uses Due to privacy regulations, data-sharing agreements, and ethical considerations, clinical research datasets may have specific restrictions on their allowable uses. Therefore, understanding and adhering to these allowable uses is critical in clinical research dataset searches. Researchers must be aware of any limitations or constraints on dataset use to ensure compliance with legal and ethical requirements. This ensures the responsible and ethical use of the data while protecting patient privacy and maintaining data security.

Documentation completeness, data quality management processes, data quality dimensions, and adherence to allowable uses are all critical aspects of clinical research dataset

searches. These four issues contribute to clinical research data reliability, transparency, and ethical use, resulting in more robust and meaningful research outcomes. Based on the interview results and statistical summary, seven explicit factors that metadata could explain were defined. These factors include publisher, purpose, source (documentation completeness), data custodian (data quality management process), follow-up, time lag (data quality dimensions), and requirements (allowable uses).

3.6. Platform Prototype Development

The metadata framework of the dataset was designed based on the interview results, and seven important indicators were selected as the key processes involved in user-friendly use. During the dataset retrieval process, the dataset that meets the users' needs can be located using the "Easy Search," which selects the seven key factors. The prototype of Easy Search based on seven factors as shown in Fig. 4 4. The UI of Easy Search as shown in Fig. 5. The Data screening example as shown in Fig. 6.

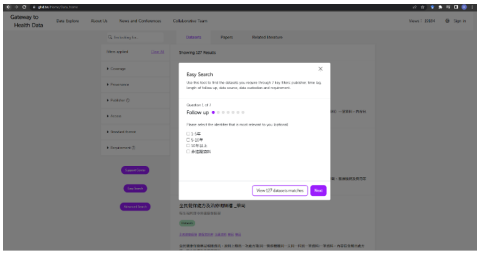


Fig. 4. Prototype of Easy Search based on seven factors

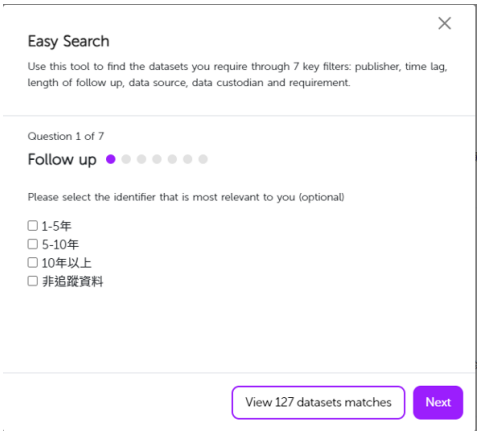


Fig. 5. UI of Easy Search

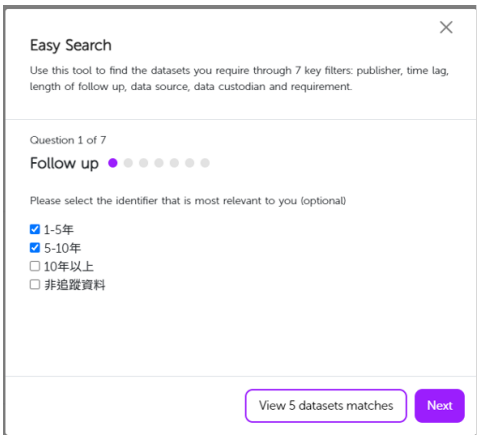


Fig. 6. Data screening example by Easy Search

For example, in the first indicator, ”Follow Up,” after selecting the options of 1-5 years and 5-10 years at the same time, the number of data sets is reduced from 127 to 5, indicating that screening such indicators can aid in the use of quick search.

3.7. Current Issues in the Data Analysis Process

In the data analysis research process, there are about 7 main steps, including ”finding data”, ”evaluating data use”, ”applying for data use”, ”obtaining”, ”data cleaning”, ”data analysis”, ”publishing research results or product development”, Most interviewees mention “applying” for the data use is currently the most difficult part of research process in Taiwan.

Calculated according to the burden points answered by the interviewees, the size of the red dot is shown in the Fig. 7. A larger red dot means more effort is required to perform the step. It is difficult to find the correct data, and the application after finding it is even more troublesome. It needs to go through many processes of review and waiting, and some data sets do not clearly explain how to apply.

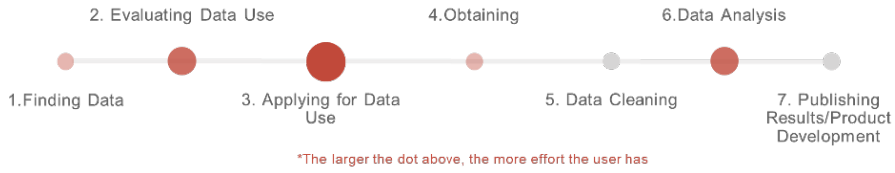


Fig. 7. Data analysis process and effort evaluate

Therefore, the design of the gateway should incorporate data and application-related information, and it is best to provide guidance and provide appropriate key reminders based on application specifications to effectively help demanders apply for information.

4. Discussion

This study performed in-depth interviews to identify some key points for assessing the availability and ease of use of public health data. The data custodian must improve public health data accessibility by ensuring it is readily available to relevant stakeholders [4]. Improving data accessibility involves addressing any barriers or restrictions that may hinder access, such as complex data formats, limited data sharing agreements, or outdated data systems.

The research results point out that High-Quality Datasets have several characteristics: 1) Data possesses high integrity and continuity, ensuring research quality; 2) Definition of data fields is standardized, with rules to facilitate collaboration and value-added applications; 3) Good data accessibility involves a user-friendly application process, minimal waiting time, and expected data collection frequency; 4) Data maturity entails a sufficient collection period to verify research hypotheses and meet regional research needs. Data quality assurance is a critical aspect of data content [9]. The gateway should establish mechanisms to assess the accuracy, reliability, and completeness of the public health data. This process involves implementing data validation processes, conducting regular audits, and promoting adherence to standardized data collection and reporting protocols. Standardizing data formats, coding systems, and terminologies across different sources is also essential for effective data integration and interoperability [7]. To facilitate data sharing and analysis, the gateway should support adopting common standards and ensure compatibility among disparate data sets.

For data management, The PIONEER Hub in UK [3] which is funded by UK HDR has a good practice reference. The PIONEER Hub is a data set and hub that includes primary, secondary, social care, and ambulance data, collects and curates acute care data from across the health economy. There are prudent, complete and clear requirements for data integration, management and release, including the use of international standard launched by International Organization for Standardization (ISO), including metadata catalog (ISO 11179)[11], data quality (ISO 8000)[10] and quality assurance (ISO 25012)[9]. Public health data often contains sensitive information, so it is crucial to prioritize privacy and security considerations[6]. There is a case that NIH TCGA (The Cancer Genome Atlas) is a large interdisciplinary initiative funded by the National Institutes of Health (NIH) in the United States, providing research scholars with access to de-identified data on specific research topics[15]. Its primary goal is to study the genomics alterations in various types of cancer, aiming to gain a deeper understanding of the molecular mechanisms of cancer and personalized treatment approaches. By analyzing genomics data from tumor samples, it reveals mutations, gene amplifications, gene deletions, and other variations present in different types of cancer. Allow researchers to obtain information honorably. TCGA secure data storage to protect individuals' privacy while enabling data access for authorized users. A well-defined governance framework is necessary to collect, store, share, and use public health data. This framework involves establishing clear roles and res possibilities, defining data ownership and stewardship, and adhering to ethical principles to ensure transparency, accountability, and responsible data management practices. At the same time, this type of data management architecture must be equipped with robust protocols for data anonymization, consent management, and encryption.

In terms of website design, the gateway to health data must have user-friendly interfaces to allow users to easily discover, access, and analyze the data[1]. This includes

designing intuitive search functionalities, providing data visualization tools, and offering user support and documentation for enhancing data usability and user experience. This is consistent with the research results. Research interviews pointed out that for a data search gateway, the interface design enables users to gradually adapt to using search engines to find datasets. Information design needs to display key information of across domains for search datasets. Website design strategy pointed out that users should be provided with information assistance for data application, which can effectively improve the difficulty of applying for data in the past, and help users understand data before applying through proper description of data and data content. Finally, the metadata of the data content of a dataset must have clear instruction documents enhance data freedom.

Taiwan's NHIS data contains complete information but only allows analysis in a controlled environment. Another example is the Taiwan government's open data platform allows for quick data browsing to grasp data overviews. The data custodian must emphasize the importance of comprehensive data documentation and metadata[14]. This process includes capturing relevant information about data sources, collection methods, variables, and any associated limitations or biases. Clear documentation helps users understand the context and quality of the data they are working with. Collaboration among various stakeholders, such as government agencies, research institutions, and the public, is vital for successful data utilization. The framework should encourage partnerships to promote data sharing, interdisciplinary research, and development of innovative solutions to public health challenges. The gateway should be designed in a way that it evolves over time and adapts to emerging technologies, changing data needs, and evolving best practices. Regular evaluations, feedback mechanisms, and a culture of continuous improvement are required to keep the gateway to health data remain relevant, effective, and updated. Promoting data literacy and providing user training opportunities is also crucial to maximizing public health data utilization. The gateway should support educational initiatives, capacity-building programs, and knowledge-sharing to equip users with the necessary skills to navigate, analyze, and interpret health data effectively.

5. Conclusion

Comprehensive interviews yielded crucial insights for designing an effective data gateway catering to researchers' needs. This study, involving six experienced interviewees from diverse fields, identified seven distinct categories of preferences and requirements among Taiwanese users regarding high-quality datasets.

Data integrity emerged as paramount, with an emphasis on completeness and continuity. Users stressed the importance of data being both complete, containing all necessary fields, and continuous over time. To address this, a robust data gateway enforcing stringent data standards is needed.

Collaboration and value-added applications were highlighted. Unified data field definitions and standardized formats were deemed essential for seamless data cleaning and analysis. Streamlining accessibility was another key consideration, emphasizing the need for a user-friendly application process with minimal waiting times. Data maturity, verified through extended data collection periods, was advocated to eliminate time-related inconveniences.

Experience-sharing underscored the importance of accessibility and comprehensibility, particularly through platforms like the Taiwan National Health Insurance System (NHIS) and the government's open data platform. User needs called for intuitive interfaces, keyword-based searches, and comprehensive dataset descriptions.

Metadata filter preferences, including data quality management and adherence to allowable uses, were identified as pivotal elements in dataset evaluation. The platform's prototype development focused on an "Easy Search" feature, streamlining dataset retrieval. Addressing challenges in data analysis, especially in the application process, was a key concern.

In summary, an effective data gateway aligned with these findings should prioritize data integrity, multi-party collaboration, streamlined accessibility, and dataset maturity. Additionally, it should cater to user needs through intuitive interfaces, comprehensive dataset descriptions, and efficient search mechanisms. Incorporating metadata filter preferences and addressing data analysis challenges will enhance the gateway's utility, bridging the gap between user expectations and high-quality dataset utilization for robust research outcomes.

Acknowledgments. This work was supported by the National Health Research Institutes, Taiwan. [grant number: No. CA 112-GP-09]

References

1. Aripiyanto, S., Agustin, F.E.M., Syakuro, A., Masruroh, S.U., Khairani, D., Sukmana, H.T.: User interface and user experience design using lean ux method on zakat ummat website. In: 2022 10th International Conference on Cyber and IT Service Management (CITSM). pp. 1–8
2. Director-General: Data and innovation: draft global strategy on digital health. Report, World Health Organization (23/12/2019 2019), https://apps.who.int/gb/ebwha/pdf_files/EB146/B146_26-en.pdf
3. Gallier, S., Price, G., Pandya, H., McCarmack, G., James, C., Ruane, B., Forty, L., Crosby, B.L., Atkin, C., Evans, R.: Infrastructure and operating processes of pioneer, the hdr-uk data hub in acute care and the workings of the data trust committee: a protocol paper. BMJ health & care informatics 28(1) (2021)
4. Hripcsak, G., Bloomrosen, M., FlatelyBrennan, P., Chute, C.G., Cimino, J., Detmer, D.E., Edmunds, M., Embi, P.J., Goldstein, M.M., Hammond, W.E.: Health data use, stewardship, and governance: ongoing gaps and challenges: a report from amia's 2012 health policy meeting. Journal of the American Medical Informatics Association 21(2), 204–211 (2014)
5. Intelligence, N.M.: Finding a role for ai in the pandemic. Nat Mach Intell 2, 291 (2020)
6. May, R., Denecke, K.: Security, privacy, and healthcare-related conversational agents: a scoping review. Informatics for Health and Social Care 47(2), 194–210 (2022)
7. de Mello, B.H., Rigo, S.J., da Costa, C.A., da Rosa Righi, R., Donida, B., Bez, M.R., Schunke, L.C.: Semantic interoperability in health records standards: a systematic literature review. Health and Technology 12(2), 255–272 (2022)
8. Sebire, N.J., Cake, C., Morris, A.D.: Hdr uk supporting mobilising computable biomedical knowledge in the uk. BMJ Health & Care Informatics 27(2) (2020)
9. Standardization, I.O.f.: Iso/iec 25012:2008 software engineering — software product quality requirements and evaluation (square) — data quality model (2008), https://www.iso.org/standard/35736.html
10. Standardization, I.O.f.: Iso 8000-1:2022 data quality — part 1: Overview (2022), https://www.iso.org/standard/81745.html

11. Standardization, I.O.f.: Iso/iec 11179-1:2023 information technology — metadata registries (mdr) — part 1: Framework (2023), https://www.iso.org/standard/78914.html
12. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M.: Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 12(3), e1001779 (2015)
13. UK, H.: Health data research uk (2023), https://www.hdruk.ac.uk/
14. Vardaki, M., Papageorgiou, H., Pentaris, F.: A statistical metadata model for clinical trials' data management. Computer Methods and Programs in Biomedicine 95(2), 129–145 (2009), https://www.sciencedirect.com/science/article/pii/S0169260709000601
15. Wang, Z., Jensen, M.A., Zenklusen, J.C.: A practical guide to the cancer genome atlas (tcga). Statistical Genomics: Methods and Protocols pp. 111–141 (2016)
16. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E.: The fair guiding principles for scientific data management and stewardship. Scientific data 3(1), 1–9 (2016)

Hsiu-An Lee – Senior researcher specializing in biomedical informatics and health data interoperability. Led the conceptualization and drafting of the study, focusing on user-centered design for health data search platforms.

Tung Lin – Researcher with expertise in information systems and usability engineering. Co-led the conceptualization and methodology design and contributed to the original draft.

Hsin-I Chen – UX designer and researcher experienced in human-computer interaction. Supported methodology development and iterative design testing.

Yi-Hsin Yang – Professor and expert in medical informatics and health data governance. Supervised the project, contributed to validation, and provided critical review and editing of the manuscript.

Wei-Chen Liu – Software engineer specializing in data platform development and integration. Participated in validation and system usability testing.

Yen-Ju Shen – Data analyst focused on health data quality and usability metrics. Contributed to validation and iterative improvement cycles.

Wen-Chang Tseng – System architect experienced in building scalable health information systems. Supported validation and technical review.

Chien-Yeh Hsu – Medical informatics researcher with expertise in health IT standards. Assisted in manuscript review and editing.

Received: December 04, 2024; Accepted: August 24, 2025.

Computer Science and Information Systems 22(4):1777–1796 https://doi.org/10.2298/CSIS241205070C

Elastic-Trust Hybrid Federated Learning

Yi-Cheng Chen[1], Lin Hui[2,⋆], and Yung-Lin Chu[3]

[1] Dept. of Information Management, National Central University, Taiwan
ycchen@mgt.ncu.edu.tw
[2] Dept. of Computer Science and Information Engineering, Tamkang University, Taiwan
121678@mail.tku.edu.tw
[3] Dept. of Information Management, National Central University, Taiwan
lexlie.yunglinchu@gmail.com

Abstract. Owing to the widespread application of machine learning, increasing attention has been focused on extensive data collection for learning model construction. Recently, with growing concerns about data privacy, private information protection has significantly increased the operation cost and difficulty of boosting model performance. The Federated Learning (FL) technique has been introduced to address this issue by keeping data on client devices and reducing the need to handle sensitive data directly. However, several challenging issues may arise when applying FL, such as data heterogeneity, efficient feature transmission, and additional computational demands. In this study, a novel FL model, Elastic-Trust Hybrid Federated Learning (ET-FL), is introduced with a dual federated learning framework. ET-FL incorporates the trust mechanism and differential aggregation strategy for model optimization and computation reduction. In addition, the proposed model is applied on real-world datasets to show the performance and practicability of promising results.

Keywords: machine learning, federated learning, decentralization, hybrid federated integration

1. Introduction

Over recent decades, machine learning has emerged as a prominent field characterized by rapid advancements and widespread adoption across various industries. Several machine learning techniques have transformed how businesses progress, empowering them to harness large volumes of data to gain insights and inform decision-making. Breakthroughs in algorithms and computational power have resulted in significant enhancements in areas such as predictive analytics, natural language processing, and computer vision, to name a few. The expansion of machine learning has also catalyzed the development of new applications, ranging from personalized recommendations in e-commerce to advanced diagnostics in healthcare.

Recently, with the growing concerns of data privacy regulations, safeguarding personal information and empowering individuals with more authority over private data have become important issues. These regulations, enforced by governments, necessitate businesses to be open about their privacy practices, and to adopt stringent security measures

⋆ Corresponding author

to protect their clients' data. Consequently, users are increasingly mindful of how organizations handle their sensitive information, leading to a heightened focus on data privacy and security. The shift in power dynamics, where individuals have more control over their data, has instilled a sense of security in the digital realm.

However, in traditional machine learning, data are centralized on a single server for training. Without any doubt, larger datasets typically improve model performance. In general, organizations aim to gather extensive data, which requires substantial storage and a high-performance server, making the process resource-intensive and time-consuming. Securing these centralized data, especially when sensitive user information is involved, adds further cost due to the necessary security measures. For example, in healthcare, patient data must be anonymized and encrypted, adding complexity and computation cost. Likewise, the financial sector must implement strict protocols to protect transaction data. Obviously, these measures are essential for preventing breaches and ensuring privacy, but also increase the cost and complexity of traditional machine learning operations.

Federated Learning (FL) is a promising solution to these problems. FL keeps the private data on each device, also called a client, thus removing the burden of implementing security measurements for organizations. Furthermore, while moving the data to each client, the training process could be done in parallel with each client, with less computing power and time. In this situation, the server orchestrates the training process across clients and maintains the consensus model. We use an application to show the significance of FL. Gboard [12] is a keyboard application installed on Android, one of the major operating systems used on mobile devices. It provides a wide range of input languages and has exceeded 1 billion installations. One of the main features of Gboard is that it suggests the next word according to the context that the user has typed in. To improve the recall of suggestions while protecting user privacy, Google has adopted an FL methodology, FedAvg [34], to complete the task successfully.

Nevertheless, transitioning from a single-server setup to a system with multiple instances may suffer several challenging issues when applying FL. These challenges mainly include data heterogeneity, feature transmission efficiency, and extra computing resource consumption. Data heterogeneity manifests as statistical imbalances, with individual clients possessing varying data distributions. In the context of FL, the model on the server acts as a collective representation of the entire system. While the system generally performs adequately in the presence of statistical imbalances, the performance of specific clients may suffer. Furthermore, feature transmission efficiency in decentralized FL encounters tradeoffs between communication efficiency and cost, with various structures such as line, ring, and mesh necessitating considerations about the optimal balance between communication efficiency and cost when spreading features to all clients. Finally, undoubtedly, the customized approach for adapting a shared model or weights requires additional computing resources to effectively complete the task at hand.

In this study, a novel hybrid framework, Elastic-Trust Hybrid Federated Learning (abbreviated as ET-FL), is proposed to tackle the aforementioned obstacles when applying FL in practical domains. We introduce a two-layer hierarchy including local and global tiers. The node in the local tier includes one server and multiple clients. The clients who have similar statistical distributions will be grouped into one node. With the hierarchy of each node (i.e., one server and multiple clients) in the local tier, we could directly apply the state-of-the-art centralized FL methodologies to learn the model and store in each

server. In the global tier, all servers in local nodes are extracted for further processing, using a decentralized approach. We introduce a novel elastic trust mechanism within the global tier to facilitate peer selection, and a merging weight concept to aggregate consensus models from other servers. The weights can be adjusted iteratively, allowing for precise calibration to extract specific features from various nodes at different iterations. Furthermore, we adopt a differential aggregation strategy on global iterations and local rounds to leverage global feature aggregation and resource consumption.

The contributions of this study are as follows:

- To the best of our knowledge, prior studies excluded emphasis on the integration of different FL methodologies. In this study, we developed a novel framework, ET-FL, a sophisticated two-layer architecture which comprises local and global tiers. The local tier learns models using the centralized FL approach, while the global tier utilizes the decentralized FL approach to integrate the learned models.
- Generally, the problem of performance downgrade in clients is mainly attributed to the presence of diverse and heterogeneous data. To address this challenging issue, we propose a strategy to organize clients with similar characteristics into groups. This contribution ensures that the system can maintain a high degree of personalized FL and also efficiently reduce the requirement of computation resources.
- We introduce a trust mechanism for client selection and aggregation weight control. The client selection process could strike a delicate balance between received models and transmission costs. The aggregation weight control ensures the necessary desired attributes throughout the process. With the proposed trust mechanism, ET-FL could optimize network performance and resource allocation.
- ET-FL equips a differential aggregation strategy bridging the global and local tiers. The proposed strategy allows the local consensus models to have ample time to aggregate features within the nodes before exchanging features in the global tier. The strategy effectively optimizes feature exchange efficiency and resource consumption.
- Finally, the proposed ET-FL framework is applied on several real-world datasets to show its performance and practicability.

The organization of the rest of this paper is as follows. Section 2 discusses the Related Work and Section 3 presents the proposed ET-FL framework in detail. We provide the experimental results in a performance study in Section 4, and conclude the paper in Section 5.

2. Related Work

2.1. Federated Learning

Concerning data privacy, the FL architecture was designed with two components: the server and the clients. The clients' private data are not transferred to the server for training; instead, they remain inside each client. To collect the features across clients, the server maintains a consensus model. In the training process, the server distributes the consensus model to the clients for client training and then collects the updated model weight from the clients, aggregating it into the new consensus model. McMahan et al. [34] were the first to propose architecture with the algorithm called FedAvg.

FedAvg suffers multiple challenges in practical use. One is the data heterogeneous problem, which indicates that FedAvg performs poorly when handling non-IID datasets. Researchers have utilized the data heterogeneous problem to improve performance. For example, Li et al. [27] proposed the Federated Proximal (FedProx) methodology which introduces an additional hyperparameter to limit the convergence direction from deviating too far from the consensus model of the client model optimization. Karimireddy et al. [21] proposed the SCAFFOLD methodology which adds two variates to the server and the client. The variates control the gradient direction to prevent the client model gradient from being directed to an optimal point far from the server consensus model. In[45], Wang et al. proposed FedNova, which normalizes the returned gradients by the numbers of the local updates to prevent the aggregated gradient from being pulled away by a larger dataset. Acar et al. [1] indicated that the client's minimal loss will not equal the global minimum loss. Therefore, they proposed the FedDyn methodology, which normalizes the calculated loss on the client to fit the global one. Duan et al. [8] proposed the Astraea framework which adds the role of mediator to manage a subset of training clients to have balanced data in the group view. Both [50] and [15] provided a concept of sharing few data on the clients over the system to resolve the data heterogeneous problem. In [18], Jeong et al. proposed a federated argumentation method using a generator to generate non-balanced data on the client side. However, the training of the generator requires clients to upload a few data to the server for the FL design.

Other approaches to improving centralized FL performance include applying optimizers to FL, resolving physical limitations, inquiring into the safety of transferred content, etc. [23,5,40,49] have tackled the communication challenges. Selecting the training client is another approach to improving FL performance; this approach was adopted by [37,20,36], while [35,29] defined new merging weight mechanisms, and [4,2,46] delved into security for aggregation. Aside from the aforementioned approaches, some researchers have adapted machine learning methodologies to FL. [39,28,31,44] implemented optimizers in FL, whereas [17,25,51,24,38] implemented knowledge distillation methodologies.

Instead of pursuing better performance using one consensus model on a system with heterogeneous data, some researchers have deployed different models on the client side. This approach is called personalized FL. Arvazagan et al. [3] proposed the FedPer algorithm. In addition to the shared consensus model across the system, FedPer adds an extra layer on top of the consensus model. Furthermore, this added layer does not attend the aggregation to enhance the client feature. Fallah et al. [10] proposed the Per-FedAvg model which splits the client training into two steps using two optimizers to generate the global and local gradients. The global gradient is aggregated to update the consensus model. Liang et al. [32] proposed the LG-FedAvg methodology. Each client maintains the global and local models and updates them by both models' loss in succession.

The concept of training one global model and fitting it to the different tasks of the different datasets in meta-learning and multi-task learning can help to solve the data heterogeneous problem by viewing each dataset as a different task. Smith et al. [42] introduced MOCHA, which identifies different clients as performing different tasks. To moderate the weight between clients, the server maintains a matrix that identifies the relationship between each client. The training process optimizes the client model and relationship matrix. In [9], Eicher et al. categorized the dataset by the time span in the day, such as midnight, morning, noon, etc. Different training rounds pick different time spans' data and clients

for training. In[22], Khodak et al. proposed ARUBA, which maintains an extra parameter, the learning rate, on the server side to better adapt the consensus model to different tasks, which is also adjusted in rounds. Li et al. [26] proposed MOON, a personalized methodology using contrastive learning to minimize the distance between the consensus and local models.

According to the aforementioned research, FL methodologies have diverse approaches to improving performance. However, the most frequently addressed issue is the data heterogeneous problem. The effect of data heterogeneity has even opened up a new branch of FL methodologies called personalized FL. Considering these studies, we divided the clients into different nodes according to the statistics. In addition, we adopted the centralized FL architecture inside nodes.

2.2. Decentralized Federated Learning

A decentralized network is structured without a central authority, relying instead on a distributed architecture where each node operates independently. In this network, control is not vested in a single entity, but is distributed among all participating nodes. Each node has the ability to make decisions, process data, and communicate with other nodes autonomously. The lack of a central control point means that there is no single point of failure, enhancing the network's fault tolerance and reliability.

With these decentralized network features, each node can gain control of its own model and dataset. Furthermore, while each client manages a personal dataset and model, it is a perfect situation to dive into the solution of dataset heterogeneity. These reasons encourage researchers to start their research in decentralized FL. Kalra et al. [19] proposed a model called ProxyFL, in which the consensus model is used only as a proxy model for exchanging the network features. Yue et al. [48]proposed the FedDCM method, which shares the same concept with [19], but uses distillation to exchange the feature between the proxy model and the local model. Gholami et al. [11] proposed a merging weight called trust, which was evaluated by the contribution in the previous round.

With the increased connections between clients in a decentralized network, carefully selecting peers is a new aspect of utilizing the framework. Tang et al. [43] provided a method that selects the peer by the connection bandwidth, while Masmoudi et al. [33] introduced OCD-FL, which includes a peer selection method determined by the energy consumption and knowledge gain.

Other than the research mentioned above, resolving communication efficiency and revising the existing centralized FL methodologies to a decentralized network are also available to improve the framework. Hu et al. [14] divided the model into segments and retained different segments from different peers to replace the aggregation and reduce the communication cost, whereas Li et al. [30] introduced a model using knowledge distillation in the decentralized setting.

In the practical usage of decentralized FL, the network designs vary. For example, [46-48] adopted decentralized FL in practical use in the medical category, but used different network architecture. Huang et al. [16] used line architecture, Chang et al. [6] used a ring network, while Xu et al. [47] used a mesh network in their decentralized FL network architecture.

From the above literature, which benefits from the feature of a decentralized network, decentralized FL has a natural advantage in handling heterogeneous data. In addition,

communication efficiency remains an approach to utilize the framework with some new research criteria, such as the peer selection problem. Considering these mentioned aspects, we adopted the decentralized FL setting in the global tier with the trust mechanism for controlling the merging weight and peer selection.

We propose a hybrid framework mixing the two architectures with client grouping and trust mechanisms. We present a detailed explanation of our proposed model, Elastic-Trust Hybrid Federated Learning, in the next section.

3. The Proposed Framework: ET-FL

Our first goal was to address the challenge of dealing with heterogeneous data and to avoid using extra computational resources on clients for a personalized approach. Instead, we proposed a client grouping strategy with a centralized server inside each node. To facilitate the exchange of features between nodes, we introduced a global tier to gather all the servers and build a network. We also aimed to remove control from any specific server and to form a decentralized network known as the global tier network. Additionally, we acknowledge that the optimal model weight derived from all the data may not always be the best for practical usage. To address this, we introduced the elastic Trust mechanism for each node to determine the preferred weight for global aggregation in different iterations. This approach also serves as a peer selection method. Finally, we observed that a synchronous setting for global tier training and local tier training leads to insufficient feature information exchange. As a solution, we proposed the differential aggregation strategy.

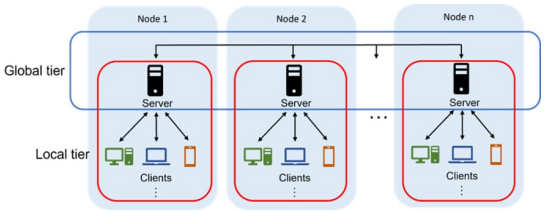


Fig. 1. The system architecture of ET-FL

Figure 1 displays the architecture of the system design. In the following section, we denote our nodes $\mathcal{N} = \{1, 2, 3, \dots\}$, and server $\mathcal{S} = \{s_n \mid n \in \mathcal{N}\}$ to better illustrate our architecture.

3.1. Local Tier

In the local tier, we group clients into clusters based on their statistical characteristics. Each cluster has a server assigned to manage the training process within the cluster. We denote $C_s = \{1, 2, 3, \dots\}$ as the clients that are grouped under a specific server. This organizational structure is illustrated in Figure 2.

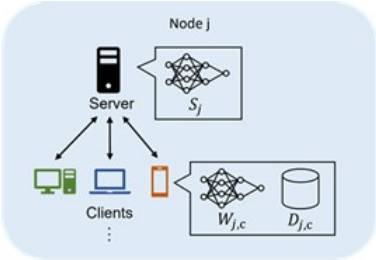


Fig. 2. The node structure

Each client device maintains a private dataset $\mathcal{D} = \{d_{s,c} \mid c \in \mathcal{C}_s,\ s \in \mathcal{S}\}$ as well as a local model $\mathcal{W} = \{w_{s,c} \mid c \in \mathcal{C}_s,\ s \in \mathcal{S}\}$. Meanwhile, the server stores a consensus model S and is responsible for coordinating the training process. In star network architecture, centralized FL methodologies such as the three most mentioned methodologies, FedAvg, FedProx, and SCAFFOLD, can be effectively implemented. Our objective function is given as follows:

$$w^{(t+1)} = \arg\min_{w} \mathcal{F}_L(w) \tag{1}$$

The FedAvg algorithm represents a pioneering approach to employing FL. The training begins with the server initializing the consensus model weight and randomly selecting a subset of clients for training. Subsequently, the server transmits the consensus model weight to the chosen clients to initialize the training process on their end. Upon reception of the model weight, the clients apply it to their respective models and commence the training procedure using their local datasets over a specified number of epochs. Upon completion of training, the client returns the model weight to the server. Post-receiving all the model weights from the selected clients, the server aggregates the model weights using the weight calculated from the training data size of the clients. The server then leverages the accumulated gradient to optimize the consensus model. These steps constitute one round, and a comprehensive training process encompasses several rounds. We denote the server weight $\mathcal{W}_f$ and the selected client $\mathcal{K}$ to express the aggregation as follows:

$$w_s = \frac{1}{M} \sum_{k \in \mathcal{K}} n_k w_k, \quad \text{where } M = \sum_{k \in \mathcal{K}} n_k \tag{2}$$

The FedAvg methodology has introduced a new avenue for research, prompting many researchers to explore its applications. FedProx has emerged as a key player in this field, with a particular focus on addressing the challenges posed by data heterogeneity. In order to mitigate the risk of highly diverse client datasets misleading the consensus model, FedProx incorporates a penalty mechanism that accounts for the discrepancy between the local model and the consensus model. This ensures that the client's search for the optimal model weight is guided by a refined equation, thereby enhancing the overall efficacy of the approach. The new equation of the client for searching for the optimal model weight is as follows:

$$w^{(t+1)} = \arg\min_{w} F_k(w) + \frac{\mu}{2} \|w - w^t\|^2 \tag{3}$$

Although we describe the client process as a search for the best model weight, it involves receiving the current round model weight $\mathcal{W}^{\sqcup}$ from the server, thereby controlling the client's distance from the server. Furthermore, the parameter $\Updownarrow\sqcap$ is crucial in determining how closely the client remains near the server. Essentially, a higher $\Updownarrow\sqcap$ level increases the client's difficulty finding an optimal space away from the server.

SCAFFOLD addresses the issue of data heterogeneity by introducing server and client control variates, which help control the client model stepping during training. These variates indicate the stepping direction in the previous round. As the client undergoes training, the gradient is adjusted using the gap between the server and client variates. The specific formula for this correction is provided as follows:

$$w_k^{(t+1)} = w_k^t - \eta_l \nabla \mathcal{L}(w_k^t) - c_k + c_s \tag{4}$$

where c_k and c_s refer to the client and server control variables. At the end of each round, these variables undergo updates based on a predefined formula. This update ensures that the variables accurately reflect the state of the client-server interaction. The formula is listed as follows:

$$c_k^{(t+1)} = c_k^t - c_s + \frac{1}{K\eta_l}\left(w_s - w_k^{(t+1)}\right), \quad \text{where } K \text{ stands for number of epochs} \tag{5}$$

$$c_s^{(t+1)} = c_s^t + \eta_g \left(\frac{1}{S}\sum_{i\in\mathcal{S}} w_i^{(t+1)} - w_s^t\right), \quad \text{where } S \text{ is the number of selected clients} \tag{6}$$

3.2. Global Tier

Within our local network infrastructure, clients are segmented into distinct nodes for organizational purposes. Each node's server is interconnected to facilitate the seamless exchange of features across the various nodes. This interconnected network, known as the global tier, plays a pivotal role in our operations. To ensure that server control is uniformly distributed and feature exchange is conducted with optimal efficiency, we have implemented a mesh network within the global tier. Utilizing this mesh network enables us to manage equal server controllability and to maximize the efficiency of feature exchange across nodes. Figure 3 depicts the intricate interactions between a server and other servers within the global tier, providing a comprehensive illustration of our network architecture.

Figure 3 illustrates how a server retrieves model weights from other servers, known as an aggregation step. We introduce a novel concept called "Trust" for merging weights during aggregation. Trust governs the feature weight gathered from the servers. Higher trust weight indicates the greater significance of the corresponding model's feature. Trust also serves as a method for selecting peers while adjusting the trust weight to zero. Moreover, trust is an independent setting that differs from the servers and is elastic over iterations. Trust (T) is represented as follows:

$$T \subseteq \left\{(x, y, t) \mid (x, y) \in \mathbb{N}^2,\ t \in [0, 1]\right\} \tag{7}$$

We denote the aggregated model in the global tier as S_con. We can express the relation of trust as follows:

$$\mathcal{S}_{\text{con}} = \left\{ \sum_{t \in T_s,\ w \in \mathcal{S}} t \cdot w \,\middle|\, s \in \mathcal{S} \right\} \tag{8}$$

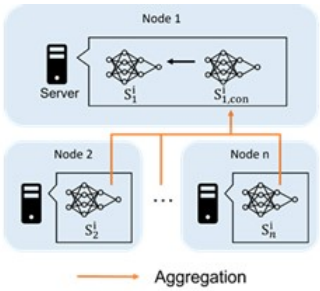


Fig. 3. Global tier aggregation

Algorithm 1 Elastic-Trust Hybrid Federated Learning

initialize the consensus models S
for i in Iterations I **do**
  for s in Servers S **in parallel do**
    $s \leftarrow LocalTierTraining(s)$
  end for
  $S_{con} \leftarrow \{\Sigma_{t \in T_s, w \in S} t * w | s \in S\}$
  $S \leftarrow S_{con}$
end for

LocalTierTraining(s):
for r in Rounds R **do**
  $s \leftarrow FL(s)$
end for
return s

// This method is capable of all centralized FL methods
// We take FedAvg as an example
FL(s):
for e in Epochs E **do**
  select a subset K from clients in C_s
  for k in K **in parallel do**
    $w_k \leftarrow w_k - \eta \nabla l(w_k, d)$
  end for
  $m \leftarrow \Sigma_{k \in K} n_k$
  $s \leftarrow \Sigma_{k \in K} \frac{n_k}{m} w_k$
end for
return s

Furthermore, we have put forward a differential aggregation strategy, which entails adjusting the pace of aggregation in the global and local tiers. In this strategy, we define

the global training cycle as an iteration and the local training cycle as a round. By implementing the differential aggregation strategy, multiple rounds of training are incorporated into a single iteration. Through the implementation of a differential aggregation strategy, our model effectively optimizes the sharing of features and minimizes communication costs. This allows for a harmonious balance between the utilization of shared features and the associated costs of communication within the model. Algorithm 1 details the main learning process in the ET-FL framework.

4. Experiments and Evaluation

To thoroughly evaluate the effectiveness of our proposed methodology, ET-FL, we undertook a comprehensive series of experiments involving three diverse datasets:

- Shakespeare [41]: This dataset is derived from "The Complete Works of William Shakespeare," including all the plays written by William Shakespeare. We obtained the preprocessed version from LEAF [50], which adopted the dataset for the task of next-letter prediction with an input sequence length of 80 characters. We adopt the LSTM model for the next-letter prediction task.
- Amazon Review [13]: This dataset is a collection of Amazon product details and customer reviews, encompassing both the rating and review text. To prepare the data for analysis, we truncated and tokenized the review text to a maximum length of 200 words for input while selecting the rating as the corresponding label. We chose to use a Transformer model to predict the rating
- EMNIST [7]: The EMNIST dataset, short for extended MNIST, is a comprehensive extension of the original MNIST dataset. It was derived from the NIST Special Database 19, which encompasses handwritten digits and both upper- and lower-case letters. The MNIST dataset comprises 70,000 images.

Table 1. Summary of Datasets

Dataset	# train	# test	# labels
Shakespeare [41]	606,277	202,103	80
Amazon Review [13]	418,811	139,613	6
EMNIST [7]	209,993	70,007	10

We divided these datasets into 20 non-IID sub-datasets to accommodate our specific needs. Each sub-dataset consists of a training set and a corresponding test set.

4.1. Baseline and Metrics

We selected the average training loss, test accuracy, test recall, test precision, and test F1-score for evaluating our model. The training loss is calculated as the difference between the model's predicted value (output logit) and the actual value, indicating how well the model fits the training dataset. Commonly used methods for calculating training loss are

mean square error and cross-entropy loss. In our experiment, we opted for cross-entropy loss. The cross-entropy loss formula follows, where $\mathrm{p}(x)$ represents the label encoded into a one-hot vector and $\mathrm{q}(x)$ represents the model's output logit. Applying a negative sign to the formula results in a positive value, facilitating easier comprehension. A lower value indicates effective model training, while a higher value suggests the opposite.

$$\text{Cross-Entropy Loss} = -p(x) \log q(x) \tag{9}$$

The concept of accuracy, recall, precision, and F1-score is a computed metric obtained from the confusion matrix depicted in the accompanying Figure. A true positive denotes a scenario where both the actual and predicted values are positive. On the other hand, false negatives and false positives refer to cases where the actual and predicted values are discordant, indicating either a misclassification of a positive actual value as negative or a misclassification of a negative actual value as positive, respectively. Finally, a true negative encompasses situations where the actual and predicted values are correctly identified as negative.

		Actual Class	
		Positive (**P**)	Negative (**N**)
Predicted Class	Positive (**P**)	True Positive (**TP**)	False Positive (**FP**)
	Negative (**N**)	False Negative (**FN**)	True Negative (**TN**)

Fig. 4. The node structure

Accuracy is the ratio of correctly predicted data to the overall data. It is the ratio of the true positive and the true negative data to the overall data. The formula is as follows:

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{10}$$

Recall is the ratio of corrected predicted positive data to the overall data whose true label is positive. It provides the performance on the true label side. The formula is as follows:

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

Precision is the ratio of corrected predicted positive data to the overall data whose predicted label is marked as positive. It provides the performance on the predicted output's side. The formula is as follows:

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

To identify a model that performs well, it should not only have good performance on either recall or precision, but on both. The F1 score therefore takes into account both recall and precision simultaneously. The formula is as follows:

$$\text{F1-score} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{13}$$

To mitigate the impact of heterogeneous datasets, we averaged the training loss and the accuracy achieved during the training iterations as the reference metrics for comparison.

As for the baseline methods, we carefully chose several FL algorithms to serve for evaluation, and compared these baseline models with our proposed methodology using specific evaluation metrics. The evaluation methods we chose are test loss and accuracy. The algorithms we selected for comparison are integral to our research and will provide valuable insights into the effectiveness of our proposed approach.

- ML (single): This baseline collects all the datasets into one server to undergo the training process. It represents the traditional machine learning training process, which provides a comparison of the traditional machine learning training process and the FL approaches.
- FedAvg [34]: FedAvg, which stands for Federated Averaging, is a foundational methodology in FL. This approach revolutionizes traditional training architectures by introducing two essential roles: the server and the client. The server orchestrates the collaborative model training process, while the client devices actively participate in model training, all while ensuring that data privacy is securely maintained.
- FedProx [10]: FedProx is a novel approach designed to deal with diverse and varying datasets. This is achieved by incorporating a hyperparameter that plays a crucial role in determining the direction of model optimization. Additionally, FedProx ensures that the model does not stray too far from the central server during the optimization process, thus maintaining stability and consistency.
- SCAFFOLD [21]: SCAFFOLD introduces the server and client variates in the training process. Every gradient will get a correction, which is the difference between the server and client variates. This prevents the client update from drifting away from the system's optimal point.
- Per-FedAvg [10]: Per-FedAvg adopts the Model-Agnostic Meta-Learning approach, which stands for the personalized approach in FL. The server holds a consensus model that is validated to handle all the tasks. However, applying the consensus model to any specific task will lead to poor performance. Therefore, the client model has to undergo an extra training process to fit the usage.

4.2. Performance Comparison

We evaluated our proposed method against baseline models using the Shakespeare, Amazon Review, and Loan datasets. For consistency, we reviewed each baseline's original research to apply the optimal settings. Our experimental setup included an optimizer with Stochastic Gradient Descent at a 0.01 learning rate, with 250 training rounds per server and four epochs per client per round. We adopted other specialized parameter settings from the existing methodologies.

To group clients effectively, we utilized a pre-trained Transformer model to extract data features from output hidden states. We then calculated client centroids as representative statistical data, applying K-means clustering to organize clients into distinct groups. Finally, we recalculated the centroids for each node and used the distance between nodes as a trust metric. Results are displayed in Tables 2 to 6.

Table 2 shows the cross-entropy loss metric, where centralized FL approaches face higher losses due to data heterogeneity. Per-FedAvg improves on the centralized methods, but our model surpasses even personalized models, achieving outstanding performance.

Table 2. Experiment Results of Averaged Cross-Entropy Loss

Model	Shakespeare	Amazon Review	EMNIST
ML (single)	1.798	0.232	0.000*
FedAvg	3.926	3.285	0.528
FedProx	3.982	3.475	0.243
SCAFFOLD	3.144	3.320	0.796
Per-FedAvg	1.302	0.258	0.018
ET-FL (FedAvg)	2.529	1.690	0.209
ET-FL (FedProx)	2.618	1.932	0.064
ET-FL (SCAFFOLD)	2.135	1.836	0.226

* less than 0.001

Table 3. Experiment Results of Averaged Accuracy

Model	Shakespeare	Amazon Review	EMNIST
ML (single)	0.466	0.710	0.996
FedAvg	0.208	0.457	0.882
FedProx	0.190	0.440	0.924
SCAFFOLD	0.201	0.449	0.878
Per-FedAvg	0.233	0.673	0.956
ET-FL (FedAvg)	0.323	0.679	0.936
ET-FL (FedProx)	0.315	0.644	0.978
ET-FL (SCAFFOLD)	0.359	0.658	0.930

Next, we examine the average accuracy results in Table 3, which show a pattern similar to the cross-entropy loss findings. Centralized FL methodologies perform below personalized FL methods, while our model surpasses centralized FL and achieves comparable accuracy to Per-FedAvg.

Table 4. Experiment Results of Averaged Recall

Model	Shakespeare	Amazon Review	EMNIST
ML (single)	0.466	0.710	0.996
FedAvg	0.208	0.457	0.882
FedProx	0.190	0.440	0.924
SCAFFOLD	0.201	0.449	0.878
Per-FedAvg	0.233	0.673	0.956
ET-FL (FedAvg)	0.323	0.679	0.936
ET-FL (FedProx)	0.315	0.644	0.978
ET-FL (SCAFFOLD)	0.359	0.658	0.930

From Tables 4, 5, and 6, these metrics further confirm that centralized FL methods generally underperform compared to personalized ones, with the exception of FedAvg and FedProx on the EMNIST dataset for the precision metric. These findings underscore our model's strengths. By effectively grouping clients, our approach successfully mitigates

Table 5. Experiment Results of Averaged Precision

Model	Shakespeare	Amazon Review	EMNIST
ML (single)	0.519	0.687	0.996
FedAvg	0.338	0.585	0.988
FedProx	0.297	0.547	0.994
SCAFFOLD	0.419	0.559	0.979
Per-FedAvg	0.303	0.753	0.969
ET-FL (FedAvg)	0.425	0.783	0.990
ET-FL (FedProx)	0.372	0.727	0.995
ET-FL (SCAFFOLD)	0.528	0.735	0.989

Table 6. Experiment Results of Averaged F1

Model	Shakespeare	Amazon Review	EMNIST
ML (single)	0.426	0.696	0.996
FedAvg	0.210	0.461	0.914
FedProx	0.181	0.442	0.952
SCAFFOLD	0.212	0.453	0.911
Per-FedAvg	0.218	0.676	0.958
ET-FL (FedAvg)	0.331	0.683	0.954
ET-FL (FedProx)	0.317	0.641	0.973
ET-FL (SCAFFOLD)	0.363	0.653	0.951

the data heterogeneity challenge faced in centralized FL. Its improved performance over centralized methods and its competitive standing with personalized approaches highlight its robustness and effectiveness.

4.3. Trust Weight Influence Analysis

As previously mentioned, the initial trust weight is derived from the distance between centroids of different nodes' data. It is important to note that the trust weight of each node consists of two key components: the weights for aggregating consensus models from other servers and the weight applied to the server's consensus model. In our calculations, we assigned a specific value to the latter weight, while the undistributed weight was determined based on the distance. A more considerable distance results in a lower weight, while a shorter distance leads to a higher weight. Consequently, we experimented by assigning different values to the latter weight to test the personalized level of nodes. To evaluate the personalized level, we performed the evaluation before and after the aggregation and subtracted the value. In cross-entropy loss, a higher value means the loss increases more after the aggregation, while a lower value means a low loss increase. This means less personalized and more personalized, respectively. In the accuracy, recall, precision, and F1-score metrics, the lower the reduction in performance after the aggregation, the more personalized the aggregated model will be. In contrast, higher reduction means the aggregated model receives more features from the other consensus model and, thus, is less personalized. The results are presented in Table 7.

Table 7. Personalized levels under different trust settings

	Loss	Accuracy	Recall	Precision	F1-score
0.5	0.676	-0.196	-0.196	-0.072	-0.145
0.6	0.477	-0.130	-0.130	-0.044	-0.089
0.7	0.293	-0.074	-0.074	-0.034	-0.053
0.8	0.143	-0.029	-0.029	-0.016	-0.020
0.9	0.042	-0.004	-0.004	-0.001	-0.002

According to the results, we can discover that a lower trust weight on the server's weight leads to a greater increase in cross-entropy loss and a decrease in the other metrics. In contrast, higher trust in the server's weight results in less cross-entropy increase and less performance decrease. The higher weight means the consensus model has a better-personalized level inside the group. In comparison, a lower weight means that the consensus model receives more information from another consensus model. In conclusion, our design of trust successfully provides users with a method to control performance, whether it is more personalized or adopts more global features.

4.4. Iteration and Round Ratio Analysis

In designing the differential aggregation strategy, we considered the number of features exchanged in aggregation and the resource cost, trying to find a balance point between rounds executed in one iteration. Therefore, we conducted this experiment to explore the influence of different ratio settings on the rounds and iterations. In the FL methodologies, the non-IID datasets lead to fluctuation in performance. To eliminate the variation factor, we divided the clients into 20 nodes. We randomly picked one node to be the observation target. The target node's model converges at around 20 epochs if trained using the traditional machine learning approach. As a result, we executed our framework with a total of 20 epochs, with one epoch in one round, and conducted 2, 5, 10, and 20 rounds in one iteration. In addition, we added a test set for executing 30 rounds in one iteration. Collecting these test scenarios allowed us to observe the difference between pre-convergence, convergence, and post-convergence situations. To evaluate the outcome of each node's collection of sufficient information for global tier aggregation, we used the feature gained from other nodes as the evaluation metric. To measure the feature gain in one iteration from other nodes, we evaluated the target node using other nodes' test datasets, subtracted the results before and after the aggregation, and averaged the value from different test datasets. For those settings that are executed over one iteration, we averaged the result by the aggregation times. The experiment results are shown in Table 8.

According to the results, we can discover that some values are negative. This is because the global consensus model mixes multiple models simultaneously. However, with the mixing in model weight, some features will be less significant, causing the performance to decrease when evaluating using the corresponding test dataset. In addition, we discovered that fewer rounds in one iteration led to better performance in feature gain. Therefore, doing a global aggregation more frequently is a better option. However, our differential aggregation strategy still provides users with an option to leverage the feature gain and the communication cost for tuning the best ratio for different users.

Table 8. Averaged feature gain on the different ratios between iterations and rounds

Rounds	Accuracy	Recall	Precision	F1-score
2	0.011	0.011	0.050	0.014
5	-0.003	-0.003	-0.014	-0.002
10	-0.006	-0.006	0.049	0.001
20	-0.007	-0.007	0.004	-0.001
30	-0.012	-0.012	0.073	-0.007

4.5. Ablation Study

We conducted a series of experiments to assess how well our model's various components performed when we turned off specific components. The situations are listed below:

- w/o global tier: We deemed all the nodes to have equal authority over other nodes and complete control of the model and dataset. Thus, we adopted the decentralized FL to eliminate the possibility of any client having control over others. By removing the global tier, our architecture retains the local tier. In addition, due to the absence of connection between different nodes for feature exchange, the clients should be grouped into one node, which is a centralized FL architecture.
- w/o local tier: We designed the client grouping strategy in the local tier and adopted centralized FL architecture inside nodes. By removing the local tier, our framework remains a decentralized network in the global tier. Thus, every client is equivalent to a node located in the global tier, aggregating different nodes' features according to the trust.
- w/o trust: The trust mechanism is designed for a user-controllable parameter to determine the weight to aggregate different consensus models. While removing the trust from our model, we used the weight calculated by the client train data size, a method mentioned in [2]. Within the group level, we calculate the weight from the total size of training data inside nodes instead of the selected clients' training data size.
- w/o differential aggregation strategy: The differential aggregation strategy indicates that the training process inside the global and local tiers is asynchronous. While one training iteration performs in the global tier, the local tier may perform several training rounds. Our orientation in designing the feature is to find a balance between the feature aggregation and the resource cost. If this strategy is disabled, only one round will be performed in one iteration.

Table 9. The component effectiveness in the ablation study

	Loss	Accuracy	Recall	Precision	F1-score
w/o global tier	3.926 (+1.397)	0.208 (-0.115)	0.208 (-0.115)	0.338 (-0.087)	0.210 (-0.121)
w/o local tier	0.841 (-1.688)	0.742 (+0.419)	0.742 (+0.419)	0.707 (+0.282)	0.707 (+0.376)
w/o trust	2.997 (+0.468)	0.268 (-0.055)	0.268 (-0.055)	0.391 (-0.034)	0.268 (-0.063)
w/o differential aggregation strategy	0.899 (-1.630)	0.735 (+0.412)	0.735 (+0.412)	0.695 (+0.270)	0.698 (+0.367)
ET-FL	2.529	0.323	0.323	0.425	0.331

According to the metrics, our model without a global tier loses the ability to personalize the model for each node, thus reducing performance. Our model without the local tier results in every client being a node in the global tier. This structure mitigates the loss inside the node due to only one client in each node. Thus, it can easily reach the system optimal inside the node. However, while leveling up the client tier, every client has to train every round, thus raising the computation resource cost in contrast to the random client selection strategy in centralized FL. While removing the trust mechanism, the model performance is affected by the size of the dataset nodes, losing the ability to control the convergence direction. This results in a subpar performance. The experiment without the differential aggregation strategy also results in better performance. However, exchanging the model weight in every round increases the transmission cost. In summary, our design of the hybrid framework with client grouping, the trust mechanism, and the differential aggregation strategy provides the optimal setting for resolving the data heterogeneous problem and the personalized approach with a more straightforward, user-controllable method.

5. Conclusion

ET-FL is an advanced two-layer hybrid Federated Learning (FL) framework that integrates both global and local tiers. It applies centralized FL methods at the local level and decentralized methods at the global level. To tackle data heterogeneity within nodes, we developed a unique grouping strategy that clusters clients by statistical similarity. Additionally, we introduced "Trust," a new aggregation weight that enhances both global aggregation and peer selection, enabling secure, reliable collaboration. A differential aggregation strategy further balances global feature aggregation with resource efficiency across tiers. We rigorously evaluated ET-FL on real-world datasets, comparing it to five baseline models with two primary metrics. Results showed that ET-FL consistently outperformed centralized FL methods and rivaled personalized FL approaches, achieving these outcomes with lower computational costs, making it highly efficient and cost-effective. ET-FL is especially promising for applications dealing with data heterogeneity or limited computational resources, as well as for decentralized setups organized by device ownership. This balanced, resource-efficient approach opens up new possibilities for Federated Learning applications.

Acknowledgments. The work of Yi-Cheng Chen was supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC 111-2628-H-008-005-MY4 and 113-2410-H-008 -065 -MY3.

References

1. Acar, D.A.E., Zhao, Y., Navarro, R.M., Mattina, M., Whatmough, P.N., Saligrama, V.: Federated learning based on dynamic regularization. arXiv abs/2111.04263 (2021)
2. Agarwal, N., Suresh, A.T., Yu, F., Kumar, S., McMahan, H.B.: cpsgd: Communication-efficient and differentially-private distributed sgd. arXiv abs/1805.10559 (2018)
3. Arivazhagan, M.G., Aggarwal, V., Singh, A.K., Choudhary, S.: Federated learning with personalization layers. arXiv abs/1912.00818 (2019)

4. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. pp. 1175–1191. ACM (2017)
5. Caldas, S., Konečný, J., McMahan, H.B., Talwalkar, A.: Expanding the reach of federated learning by reducing client resource requirements. arXiv abs/1812.07210 (2018)
6. Chang, K., Balachandar, N., Lam, C.K., Yi, D., Brown, J.M., Beers, A.L., Rosen, B.R., Rubin, D., Kalpathy-Cramer, J.: Distributed deep learning networks among institutions for medical imaging. Journal of the American Medical Informatics Association 25, 945–954 (2018)
7. Cohen, G., Afshar, S., Tapson, J., Schaik, A.v.: Emnist: an extension of mnist to handwritten letters. arXiv abs/1702.05373 (2017)
8. Duan, M., Liu, D., Chen, X., Tan, Y., Ren, J., Qiao, L., Liang, L.: Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications. arXiv preprint arXiv:1907.01132 (2019)
9. Eichner, H., Koren, T., McMahan, H.B., Srebro, N., Talwar, K.: Semi-cyclic stochastic gradient descent. arXiv abs/1904.10120 (2019)
10. Fallah, A., Mokhtari, A., Ozdaglar, A.: Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In: Advances in Neural Information Processing Systems. pp. 3557–3568 (2020)
11. Gholami, A., Torkzaban, N., Baras, J.S.: Trusted decentralized federated learning. In: 2022 IEEE 19th Annual Consumer Communications Networking Conference (CCNC). pp. 1–6. IEEE (2022)
12. Hard, A., Rao, K., Mathews, R., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., Ramage, D.: Federated learning for mobile keyboard prediction. arXiv abs/1811.03604 (2018)
13. Hou, Y., Li, J., He, Z., Yan, A., Chen, X., McAuley, J.: Bridging language and items for retrieval and recommendation. arXiv abs/2403.03952 (2024)
14. Hu, C., Jiang, J., Wang, Z.: Decentralized federated learning: A segmented gossip approach. arXiv abs/1908.07782 (2019)
15. Huang, L., Yin, Y., Zhang, Z.F., Deng, H., Liu, D.: Loadaboost: Loss-based adaboost federated machine learning with reduced computational complexity on iid and non-iid intensive care data. arXiv abs/1811.12629 (2020)
16. Huang, Y., Bert, C., Fischer, S., Schmidt, M., Dörfler, A., Maier, A., Fietkau, R., Putz, F.: Continual learning for peer-to-peer federated learning: A study on automated brain metastasis identification. arXiv abs/2204.13591 (2022)
17. Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., Kim, S.L.: Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. arXiv abs/1811.11479 (2018)
18. Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., Kim, S.L.: Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. arXiv abs/1811.11479 (2023)
19. Kalra, S., Wen, J., Cresswell, J.C., Volkovs, M., Tizhoosh, H.R.: Decentralized federated learning through proxy model sharing. Nature Communications 14 (2021)
20. Kang, J., Xiong, Z., Niyato, D.T., Yu, H., Liang, Y.C., Kim, D.I.: Incentive design for efficient federated learning in mobile networks: A contract theory approach. In: 2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS). pp. 1–5. IEEE (2019)
21. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. arXiv preprint arXiv:1910.06378 (2020)
22. Khodak, M., Balcan, M.F., Talwalkar, A.: Adaptive gradient-based meta-learning methods. arXiv abs/1906.02717 (2019)
23. Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. arXiv abs/1610.05492 (2016)

24. Lee, G., Jeong, M., Shin, Y., Bae, S., Yun, S.Y.: Preservation of the global knowledge by not-true distillation in federated learning. In: Advances in Neural Information Processing Systems. pp. 38461–38474 (2022)
25. Li, D., Wang, J.: Fedmd: Heterogenous federated learning via model distillation. arXiv abs/1910.03581 (2019)
26. Li, Q., He, B., Song, D.: Model-contrastive federated learning. arXiv abs/2103.16257 (2021)
27. Li, T., Sahu, A.K., Sanjabi, M., Zaheer, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. arXiv abs/1812.06127 (2018)
28. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Feddane: A federated newton-type method. In: 2019 53rd Asilomar Conference on Signals, Systems, and Computers. pp. 1227–1231. IEEE (2019)
29. Li, T., Sanjabi, M., Smith, V.: Fair resource allocation in federated learning. arXiv abs/1905.10497 (2019)
30. Li, X., Chen, B., Lu, W.: Feddkd: Federated learning with decentralized knowledge distillation. Applied Intelligence pp. 1–17 (2022)
31. Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the convergence of fedavg on non-iid data. ArXiv abs/1907.02189 (2019)
32. Liang, P.P., Liu, T., Ziyin, L., Salakhutdinov, R., Morency, L.P.: Think locally, act globally: Federated learning with local and global representations. ArXiv abs/2001.01523 (2020)
33. Masmoudi, N., Jaafar, W.: Ocd-fl: A novel communication-efficient peer selection-based decentralized federated learning. arXiv abs/2403.04037 (2024)
34. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629 (2016)
35. Mohri, M., Sivek, G., Suresh, A.T.: Agnostic federated learning. arXiv abs/1902.00146 (2019)
36. Nguyen, H.T., Sehwag, V., Hosseinalipour, S., Brinton, C.G., Chiang, M., Poor, H.V.: Fast-convergent federated learning. IEEE Journal on Selected Areas in Communications 39, 201–218 (2020)
37. Nishio, T., Yonetani, R.: Client selection for federated learning with heterogeneous resources in mobile edge. In: ICC 2019 - 2019 IEEE International Conference on Communications (ICC). pp. 1–7. IEEE (2018)
38. Qi, P., Zhou, X., Ding, Y., Zhang, Z., Zheng, S., Li, Z.: Fedbkd: Heterogenous federated learning via bidirectional knowledge distillation for modulation classification in iot-edge system. IEEE Journal of Selected Topics in Signal Processing 17, 189–204 (2023)
39. Reddi, S.J., Charles, Z.B., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., McMahan, H.B.: Adaptive federated optimization. arXiv abs/2003.00295 (2020)
40. Sattler, F., Wiedemann, S., Müller, K.R., Samek, W.: Robust and communication-efficient federated learning from non-i.i.d. data. IEEE Transactions on Neural Networks and Learning Systems 31, 3400–3413 (2019)
41. Shakespeare, W.: The Complete Works of William Shakespeare. Project Gutenberg (1994)
42. Smith, V., Chiang, C.K., Sanjabi, M., Talwalkar, A.: Federated multi-task learning. arXiv abs/1705.10467 (2017)
43. Tang, Z., Shi, S., Li, B., Chu, X.: Gossipfl: A decentralized federated learning framework with sparsified and adaptive communication. IEEE Transactions on Parallel and Distributed Systems 34, 909–922 (2023)
44. Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., Khazaeni, Y.: Federated learning with matched averaging. arXiv abs/2002.06440 (2020)
45. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. ArXiv abs/2007.07481 (2020)
46. Xu, G., Li, H., Liu, S., Yang, K., Lin, X.: Verifynet: Secure and verifiable federated learning. IEEE Transactions on Information Forensics and Security 15, 911–926 (2020)

47. Xu, J., Glicksberg, B.S., Su, C., Walker, P., Bian, J., Wang, F.: Federated learning for healthcare informatics. arXiv abs/1911.06270 (2020)
48. Yue, H., Lanju, K., Qingzhong, L., Baochen, Z.: Decentralized federated learning via mutual knowledge distillation. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 342–347. IEEE (2023)
49. Zhang, X., Hong, M., Dhople, S.V., Yin, W., Liu, Y.: Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. ArXiv abs/2005.11418 (2020)
50. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv abs/1806.00582 (2018)
51. Zhu, Z., Hong, J., Zhou, J.: Data-free knowledge distillation for heterogeneous federated learning. In: Proceedings of the 38th International Conference on Machine Learning. pp. 12878–12889. Proceedings of Machine Learning Research (2021)

Yi-Cheng Chen received his Ph.D. degree from the Department of Computer Science at National Chiao Tung University (NCTU), Taiwan, in 2012. Currently, he is a professor and chair of the Department of Information Management at National Central University (NCU), Taiwan. He has been active in international academic activities, as conference organizer and journal editor/reviewer. Dr. Chen has published a number of papers in several prestigious conferences and journals. His research interests include machine learning, social network analysis, data mining and cloud computing.

Lin Hui is currently a professor with the department of computer science and information engineering, Tamkang University, Taiwan. Her research interests include machine learning, multimedia applications, and mobile information systems. She has published some journal articles, book chapters, and conference papers related to these research fields. She had served as journal guest editor/reviewer, and program co-chair/chair for many international conferences and workshops.

Yung-Lin Chu received the M.S. degree from the Department of Information Management, National Central University, Taiwan.

Received: December 05, 2024; Accepted: May 18, 2025.

Computer Science and Information Systems 22(4):1797–1816 https://doi.org/10.2298/CSIS250126071L

Toward Key Factors in Travel Time Prediction for Sustainable Mobility and Well-Being

Chuang-Chieh Lin[1], Ming-Chu Ho[2], and Chih-Chieh Hung[2]

[1] Department of Computer Science and Engineering,
National Taiwan Ocean University,
Keelung City, 202301, Taiwan
josephcclin@mail.ntou.edu.tw
[2] Department of Management Information Systems,
National Chung Hsing University,
Taichung City, 402202, Taiwan
mingchuho310243@gmail.com
smalloshin@nchu.edu.tw

Abstract. Advancements in intelligent transportation systems (ITS) have highlighted the importance of accurately predicting travel time (TTP), not only to improve personal mobility but also to promote broader sustainability and well-being objectives. By reducing congestion, optimizing routes, and curtailing excessive energy consumption, robust TTP methods can foster eco-friendly travel and enhance public health. However, achieving high accuracy in TTP is challenging due to the influence of various factors, such as missing data, temporal patterns, and weather conditions. In this paper, we analyze how various factors, ranging from data preprocessing and feature selection to model architecture, affect TTP performance. Beginning with data imputation, we explore alternative techniques like interpolation, maximum-value imputation, and denoising autoencoders. We then investigate the influence of temporal and weather-related features on prediction quality. Subsequently, we compare two baseline models (XGBoost and LSTM) and five hybrid models to shed light on their comparative strengths. Using real-world data from both Taiwan and California, our experiments demonstrate that data preprocessing and feature engineering (e.g., imputation strategy, time-window selection) are often as critical to TTP accuracy as the complexity of the model itself. Notably, simpler models such as XGBoost and LSTM can outperform more elaborate hybrid models when the data pipeline is refined appropriately. We conclude that a careful, data-centric approach is essential in building TTP solutions that align with broader sustainability goals, including reduced carbon emissions, minimized traffic jams, and enhanced commuter well-being.

Keywords: Travel Time Prediction, Machine Learning, Sustainable Mobility.

1. Introduction

Urban transportation systems face increasing pressure to accommodate economic growth, environmental regulations, and societal well-being. Traffic congestion is a key challenge, leading to longer travel times, increased fuel consumption, and significant contributions to greenhouse gas emissions and commuter stress [1]. Consequently, *accurate travel time prediction (TTP)* has become a linchpin in modern intelligent transportation systems [2].

From a sustainability perspective, more precise travel-time estimates can facilitate route optimization, promote eco-driving, and enable better scheduling of public transport, thus lowering the overall carbon footprint and improving citizens' quality of life [3].

Despite its real-world impact, TTP is not straightforward. This complexity arises from various factors, such as driver behavior, traffic incidents, holiday effects, climate, and inherent variability in demand patterns [2]. Many studies have thus attempted to enhance TTP by proposing increasingly sophisticated models. Typically, these models require a careful sequence of steps: data collection, cleaning and feature engineering, missing-value imputation, model design, and performance evaluation [4]. Each stage may substantially affect the overall accuracy.

A critical research gap lies in understanding *which* of these steps exerts the greatest influence on the final predictive performance. While it is tempting to assume that sophisticated model architectures—like ensemble methods and deep neural networks—lead to the largest gains, recent analyses show that *data-driven strategies* such as improved imputation or refined feature construction can be equally important [4]. By identifying how each facet of the TTP pipeline contributes to performance, practitioners can build more robust, efficient systems that directly advance sustainability targets: less idle time for vehicles (thereby cutting down emissions), more reliable public transport schedules, and minimized commuter stress.

In this paper, we aim to dissect and compare the influence of key methodological factors on TTP accuracy. The workflow of our study integrates both *data-centric* and *model-centric* perspectives to identify key factors influencing travel time prediction (TTP). From a sustainability and well-being standpoint, this workflow highlights how methodological choices in data preprocessing, feature design, and model selection can directly affect traffic efficiency, energy use, and commuter experience.

- **data preprocessing.**
  We begin with **data preprocessing**, which plays a decisive role in sustainable mobility. Incomplete or unrealistic data (e.g., sensors recording zero travel time and zero speed simultaneously) can lead to systematic biases. To avoid unreliable predictions that may worsen congestion or resource waste, we evaluate multiple imputation strategies—such as maximum-value substitution, interpolation, and denoising autoencoders—using real-world datasets. Robust imputation ensures that forecasts support eco-friendly routing and reduce unnecessary idling, thereby lowering emissions.
- **temporal and contextual features.**
  Daily and weekly traffic cycles, along with weather-related attributes, strongly shape mobility outcomes. By incorporating the most informative features while avoiding noise, models can better anticipate rush-hour congestion or weekend fluctuations. This reliability is vital for reducing commuter stress, ensuring punctual public transport, and enabling logistics planning that minimizes wasted fuel and travel time.
- **model comparison.**
  Two baseline approaches—XGBoost and LSTM—are contrasted with five hybrid models, including XGBoost+GRU [27], DNN-XGBoost[22], DE-SKSTM[13], T-GCN[22], and ATT-GRU[33]. We have more detail discussion about the models in Section 2. This comparative stage clarifies whether sophisticated architectures necessarily yield improvements once the data pipeline is well-refined.

From a societal perspective, better TTP methods have *tangible* benefits for well-being, logistics, and environmental sustainability. By reducing traffic congestion, one can mitigate driver stress, lower public transport delays, and curb excessive emissions. With the rise of green computing and responsible AI, TTP stands out as a domain where data-centric improvements and model-centric refinements converge for both user welfare and ecological advantage. Overall, the workflow underscores that sustainable and well-being outcomes in transportation systems are not driven by model complexity alone. Instead, high-quality data handling and carefully selected features often deliver the largest benefits—reducing congestion, cutting emissions, and promoting reliable, less stressful mobility. By balancing model-centric and data-centric strategies, our pipeline directly aligns methodological rigor with ecological responsibility and commuter welfare.

Contributions

- **Comprehensive Factor Analysis:** We provide an end-to-end analysis of TTP pipelines, detailing how various steps—from imputation to feature design—affect accuracy. This clarifies which elements merit the most attention in practice.
- **Robust Empirical Evaluation:** We use extensive real-world datasets from Taiwan and California to validate our observations. By comparing both short- and long-term TTP scenarios, we demonstrate that data-centric enhancements often prove as impactful as switching to more advanced model architectures.
- **Data-Centric vs. Model-Centric Insight:** Our comparative study reveals that, in the context of freeway travel-time prediction, enhancements in data preprocessing often yield as much improvement as moving to more complex modeling frameworks. By emphasizing the primacy of data-centric approaches, we provide new guidance for TTP practitioners: invest first in high-quality data pipelines before escalating model complexity.
- **Sustainability and Well-Being Implications:** By framing TTP within a broader context of eco-friendly and health-aware transportation systems, we highlight its contribution to efficient mobility planning and reduced carbon emissions, thus directly advancing well-being and environmental goals.

The rest of this paper is organized as follows. Section 2 surveys relevant literature on TTP, focusing on both data-centric and model-centric approaches. Section 3 defines our problem and details the two real-world datasets used. Section 4 discusses the imputation techniques. Section 5 presents both base and hybrid models. Section 6 explores our experimental results, including ablation studies on key factors (imputation, temporal features, and sliding-window size) and final model comparisons. Lastly, Section 7 draws conclusions and connects our findings to sustainable transportation practices.

2. Related Work

Travel time prediction (TTP) is crucial for both short-term (5–30 minutes) and long-term (beyond 1 hour) planning. Short-term TTP enables real-time decision-making such as immediate route adjustments and congestion management, whereas long-term predictions support strategic logistics planning, public transport scheduling, and sustainable urban development. Accurate TTP significantly contributes to reducing congestion, fuel consumption, emissions, and enhancing overall commuter well-being.

2.1. Traditional Statistical Methods

Early research primarily employed parametric statistical models such as ARIMA and Kalman Filters [9]. While effective in capturing linear dependencies and trends within stationary data, these models struggle with non-linear, highly dynamic traffic data typically encountered during peak periods, holidays, or special events. Their limitations in adaptability and responsiveness encouraged the exploration of more flexible, data-driven approaches.

2.2. Machine Learning Approaches

Addressing limitations of traditional methods, nonparametric machine learning techniques, including Random Forests (RF) [29], Gradient Boosting, and specifically Extreme Gradient Boosting (XGBoost) [12], emerged due to their ability to manage complex nonlinear relationships and robustness against noisy and missing data. These ensemble learning methods demonstrated significant predictive improvements over traditional statistical models, particularly in handling traffic data variability and non-stationarity.

2.3. Deep Learning Methods

Recent advancements in deep learning techniques, notably recurrent neural networks (RNN) such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), further enhanced TTP accuracy by effectively modeling intricate temporal dependencies [21]. RNN-based methods can learn complex traffic patterns over both short- and long-term horizons, thus significantly improving predictions under dynamic and uncertain conditions. Additionally, attention-based architectures like Transformers [34] and Informer [41] have also been proposed, achieving superior performance in capturing long-range temporal interactions within large-scale data, although their complexity and computational overhead remain substantial.

2.4. Hybrid Models

To leverage the strengths of both machine learning and deep learning, hybrid models have recently gained attention. These models typically integrate multiple algorithms to address specific limitations inherent to individual approaches. For example, Ting et al. [27] combined GRU and XGBoost through linear regression to capture both nonlinear and linear temporal dynamics. Similarly, Ho et al. [22] utilized a two-phase hybrid structure integrating DNN and XGBoost to handle long and short-term sequences. Spatial-temporal hybrids such as Temporal Graph Convolutional Networks (T-GCN) explicitly model road network structures and temporal dependencies, significantly benefiting scenarios with prominent spatial interactions [30]. Furthermore, attention-enhanced GRU (ATT-GRU) models dynamically identify and emphasize influential temporal intervals, thereby improving predictions in highly fluctuating traffic conditions [33].

2.5. Sustainability and Well-Being Perspectives

Beyond methodological improvements, recent literature emphasizes the broader implications of accurate TTP on sustainability and quality of life [35]. Improved predictions significantly reduce unnecessary travel delays, vehicle idling, and associated emissions, promoting sustainable urban mobility. Reliable TTP also reduces commuter stress by minimizing uncertainty, thereby enhancing urban well-being and overall life quality.

Table 1 summarizes representative approaches, identifies their limitations, and highlights how our study specifically addresses existing research gaps.

Table 1. Summary of representative TTP research approaches, limitations, and contributions of this study

Category	Representative Studies	Limitations	Contribution of This Study
Traditional Statistical Methods	ARIMA, Kalman Filters [9]	Limited flexibility for nonlinear patterns; struggle with high variability	Investigates hybrid methods effectively managing complex nonlinear patterns and variability
Machine Learning Methods	RF [29], XGBoost [12]	Insufficient modeling of sequential and temporal dynamics	Emphasizes integration with deep recurrent models to explicitly capture temporal sequences
Deep Learning Methods	LSTM, GRU [21]; Transformers [34], Informer [41]	High complexity, limited interpretability, computational overhead	Examines simpler yet effective hybrid models that balance predictive performance with interpretability
Hybrid Models	Ting's Hybrid [27], Ho's Hybrid [22], T-GCN [30], ATT-GRU [33]	Lack of systematic comparison on how each modeling stage (data preprocessing, feature engineering) affects predictive outcomes	Provides comprehensive experimental evaluation of key methodological factors, emphasizing the critical role of data preprocessing and feature selection in achieving accurate predictions
Sustainability & Well-being	Sustainable mobility frameworks [35]	Limited direct linkage between predictive methodology and sustainability benefits	Explicitly connects predictive improvements to sustainability outcomes, demonstrating how enhanced TTP reduces congestion, emissions, and commuter stress

3. Preliminary

3.1. Problem Definition

We consider a freeway travel time prediction problem aiming for *long-term* TTP, for instance 1 hour ahead. Let t represent the current 5-minute time slot and $t^* = t + 12$ be the future target (i.e., 1 hour later). Our goal is to forecast the travel time T_{t^*} based on features observed in the sequence of preceding time slots. Specifically, denote x_t as the feature vector of time slot t. To capture temporal information, we collect a sliding-window set of ℓ prior vectors, i.e., $(x_{t-\ell}, x_{t-\ell+1}, \ldots, x_t)$. The task is to learn a function f such that:

$$f\Big(\bigcup_{i=0}^{\ell} x_{t-\ell+i}\Big) \;=\; T_{t+12}.$$

Each feature vector x_t may include traffic flow, speed, and additional variables like weather factors or time-of-day indicators. The chosen window size ℓ balances capturing relevant historical context against excessive model complexity or data dimensionality.

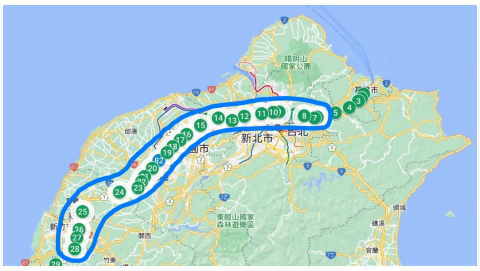


Fig. 1. The 69 routes in the Taiwan dataset. Each numbered green circle indicates a sensor-defined route segment

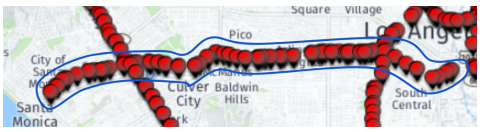


Fig. 2. The 81 routes in the California dataset (`https://pems.dot.ca.gov/`). Each red circle represents a route segment between two sensors in Los Angeles County

3.2. Data Preparation

We employ two real-world datasets, each using a 5-minute time interval:

1. **Taiwan Dataset (MOTC).** Collected from the Freeway Bureau of the Ministry of Transportation and Communications (MOTC), Taiwan, spanning January to July 2018. Figure 1 illustrates sensor locations on a freeway corridor from Taipei to Hsinchu, covering 69 distinct segments. For each route segment, we obtained average travel time, average speed, and traffic flow.
2. **California Dataset (PeMS).** Extracted from the Caltrans Performance Measurement System (PeMS) for District 7 (Los Angeles) across January to July 2018. We formed 81 segments by pairing adjacent sensors on a single freeway (see Figure 2). Distances are used with speeds to infer travel time. The dataset captures average flow, speed, and segment length.

Both datasets contain missing or zero values. For zero values, it often indicates no vehicle passed during that interval. Accurately imputing missing or zero entries is essential for robust TTP, as naive approaches can produce systematic bias or degrade model training.

3.3. Quantitative Sustainability Impact

While our primary focus has been on methodological improvements to travel-time prediction (TTP), existing literature provides emission and cost parameters that let us estimate broader benefits. For example, a typical U.S. gasoline passenger vehicle emits about 400 g CO_2 per mile (≈250 g/km) [50]. The average one-way commute in the U.S. is 27.6 minutes (≈22 km at 50 km/h) [51]. If a 10% reduction in mean absolute error (MAE)

of TTP translates to 2.8 minutes (≈2.3 km) fewer kilometers driven per trip, the CO_2 savings per vehicle per trip are:

$$\Delta m_{\mathrm{CO_2}} = 2.3\,\mathrm{km} \times 250\,\frac{\mathrm{g}}{\mathrm{km}} \approx 575\,\mathrm{g}\,.$$

Over 10 000 daily commuters, this yields roughly 5.75 t of CO_2 avoided each day.

Economically, the U.S. Department of Transportation values travel time at $18.80 per hour per driver [52]. A 2.8-minute time saving corresponds to

$$\Delta\mathrm{Value} = \frac{2.8}{60}\,\mathrm{h} \times \$18.80/\mathrm{h} \approx \$0.88 \text{ per trip.}$$

Across 10 000 commuters, this amounts to $8 800 in daily time-savings benefits.

These simple calculations, based entirely on established emission factors and economic valuations, demonstrate that even modest TTP accuracy improvements can yield sizeable environmental and economic gains without additional experiments.

4. Imputation Methods Description

Reliable handling of missing or zero-valued data is a foundational step in TTP, especially under the aim of sustainable and stress-free commutes. Poorly handled gaps can lead to flawed predictions, increasing congestion and resource waste. Here, we compare several approaches:

4.1. Interpolation with Historical Reference

Following Chou et al. [13], a multi-step interpolation is performed. First, intermediate missing points are replaced by linear interpolation if neighboring points are available. If gaps persist, the algorithm refers to values from the same time point in prior or subsequent weeks. Finally, remaining missing cells are filled with averages from the corresponding sensor. This multi-layered strategy is efficient and straightforward.

4.2. Max Imputation

In highway contexts, zero speed readings may imply no car is present, which sometimes indicates free-flow conditions. Thus, we can impute zero speed by the legal speed limit. Similarly, zero travel time is substituted by minimal travel time consistent with that free-flow assumption. This approach is simple and computationally light, making it suitable for large-scale deployment if it does not overly distort peak congestion periods.

4.3. Denoising Autoencoder (DAE)

Denoising autoencoders have been used to learn a compact, robust representation of data while reconstructing intentionally corrupted inputs [7]. One can randomly mask certain features (e.g., speed or flow) and train a DAE to reconstruct them. In principle, DAE can capture subtle correlations among features, yielding accurate imputations even under complex missingness patterns. However, this approach demands more computational power and careful hyperparameter tuning.

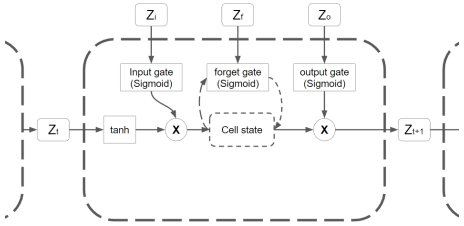


Fig. 3. Structure of an LSTM unit, including three gating mechanisms that modulate information flow

5. Base and Hybrid Model Frameworks

We examine both standard and composite models. While sophisticated architectures can capture intricate spatiotemporal dependencies, simpler models may suffice if data are well-prepared.

5.1. Base Models

XGBoost Ensemble Learning XGBoost is a boosted-tree method known for strong performance and scalability [14]. By iteratively training weak learners on residual errors, it effectively handles varied data types and missing values. The objective combines a loss function (e.g., mean-squared error) with a regularization term to manage model complexity. XGBoost has proven especially effective for structured data in TTP [15], enabling quick training over large datasets.

Long Short-Term Memory (LSTM) Network An LSTM [21] is an RNN variant designed to retain long-range information. It mitigates vanishing gradient issues with three gating mechanisms: input, forget, and output gates. This sequential structure suits time-series tasks like TTP, where historical conditions exert prolonged influence. Figure 3 outlines the LSTM cell.

5.2. Hybrid Models

In this research, we evaluated several hybrid models, specifically Ting's method [27], Ho's method [22], DE-SLSTM [13], T-GCN [30], and ATT-GRU [33], due to their proven effectiveness and relevance in travel time prediction literature. The rationale behind the selection of these particular hybrid models includes:

– **Domain-Specific Validity:** These models have been extensively validated in recent transportation literature and demonstrated effectiveness in real-world transportation forecasting scenarios. Specifically, Ting's method [27] and Ho's method [22] successfully combined robust traditional ensemble methods (XGBoost) and deep recurrent models (GRU/LSTM), making them ideal candidates for capturing both stable

and transient temporal patterns in freeway travel time data. Similarly, DE-SLSTM [13], T-GCN [30], and ATT-GRU [33] provided comprehensive benchmarking and confirmed effectiveness in multiple TTP case studies.

- **Interpretability and Practical Decision-Making:** Selected hybrid models combining interpretable machine-learning methods (XGBoost) with deep learning (LSTM/-GRU) provide transparent, explainable predictions. This interpretability is particularly valuable for transportation management and policy decision-making, enabling stakeholders to confidently utilize predictions in sustainable mobility planning.
- **Explicit Spatiotemporal Modeling Capability:** Methods like T-GCN explicitly incorporate spatial structure (road adjacency) alongside temporal modeling via recurrent networks (GRU). Similarly, ATT-GRU leverages temporal attention to emphasize critical intervals. Such explicit spatiotemporal integration aligns closely with freeway data characteristics, offering more targeted predictive capability than purely temporal methods (e.g., Transformers and Informer).

Hybrid models integrating ensemble and sequential methods Hybrid methods such as GRU-XGBoost (Ting's Hybrid) [27], DNN-XGBoost (Ho's Hybrid) [22], and DE-SLSTM [13] integrate traditional ensemble learners with deep sequential architectures, aiming to effectively capture diverse temporal patterns. Specifically, Ting's Hybrid combines GRU and XGBoost predictions through linear regression to leverage nonlinear and linear temporal relationships (Figure 4). Ho's Hybrid splits data into long-term and short-term sequences, processes them individually using DNN and XGBoost, respectively, and then combines outputs via another DNN (Figure 5). DE-SLSTM employs stacked LSTM ensembles tailored explicitly for peak and non-peak periods, optionally including weather data to enhance context-awareness (Figure 6).

Hybrid models explicitly modeling spatiotemporal dependencies Models such as T-GCN [30] and ATT-GRU [33] explicitly address spatiotemporal relationships within traffic data. T-GCN integrates graph convolutional networks (GCN) for capturing spatial adjacency with GRU to model temporal dynamics, effectively representing freeway network interactions (Figure 7). ATT-GRU incorporates a self-attention mechanism within GRU layers to dynamically focus on critical historical intervals, refining prediction accuracy especially under fluctuating traffic conditions (Figure 8). These specialized architectures directly incorporate spatial and temporal information, making them particularly suited to complex freeway traffic scenarios.

To clearly summarize and compare these hybrid models, Table 2 provides an overview of each model's strengths, weaknesses, and typical application scenarios.

6. Experiment

We now detail experimental setups and evaluations. Our focus is to reveal how each methodological choice—from imputation techniques to temporal feature engineering—affects prediction accuracy. We also examine training efficiency to gauge the practicality of each approach, given real-world constraints on resource usage or sustainability considerations.

Table 2. Summary comparison of hybrid travel-time prediction models

Model	Strengths	Weaknesses
GRU-XGBoost (Ting's Hybrid) [27]	Clearly integrates linear and nonlinear temporal patterns; straightforward linear regression combination.	Limited spatial modeling capability; may oversimplify complex dynamics.
DNN-XGBoost (Ho's Hybrid) [22]	Handles long- and short-term sequences separately for detailed pattern analysis; flexible temporal feature incorporation.	Increased complexity due to multi-phase structure; careful parameter tuning needed.
DE-SLSTM [13]	Explicitly designed for peak vs. non-peak conditions; incorporates optional weather indicators.	Computationally intensive; complexity from configuring multiple stacked LSTMs.
T-GCN [30]	Explicit spatial graph and temporal integration; accurately represents freeway networks.	Requires precise adjacency data; more complex model structure.
ATT-GRU [33]	Dynamically identifies critical historical intervals via attention; robust under highly variable conditions.	Attention mechanism increases complexity; reduced interpretability due to dynamic weighting.

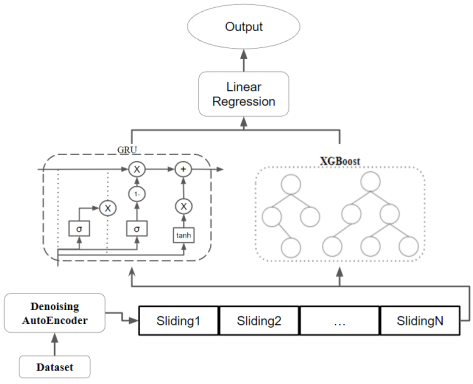


Fig. 4. Structure of Ting's GRU-XGBoost hybrid model [27]. Predictions from GRU and XGBoost are integrated through linear regression

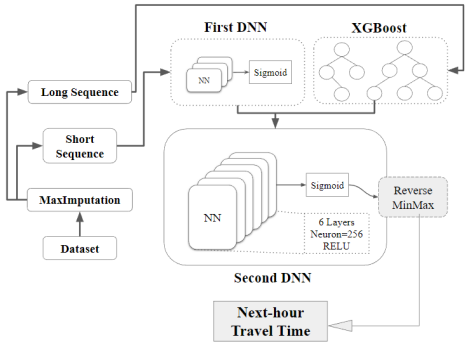


Fig. 5. Structure of Ho's Hybrid Model [22], leveraging DNN and XGBoost in tandem

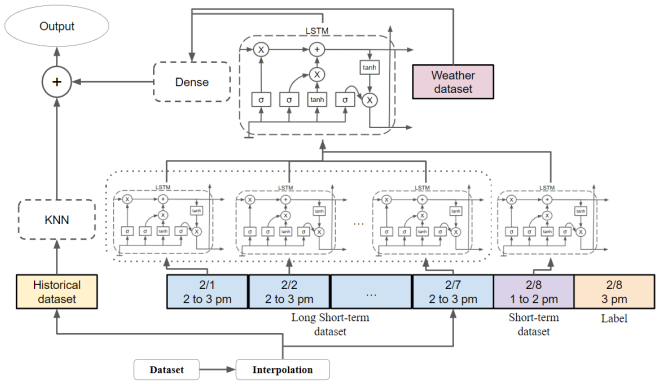


Fig. 6. Structure of DE-SLSTM [13]. Multiple LSTMs capture long- and short-term dependencies, with additional weather features optionally included

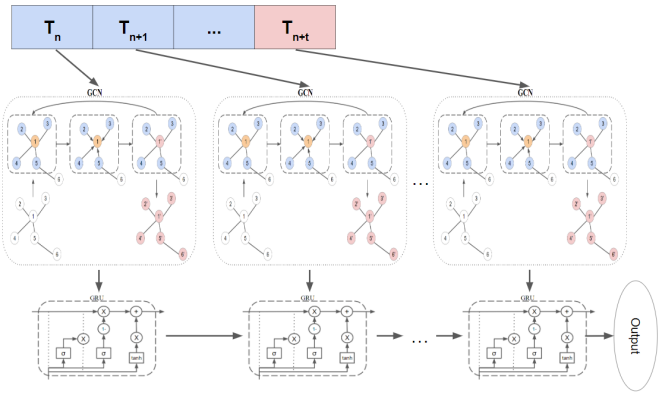


Fig. 7. Structure of T-GCN [30], combining a graph convolutional network for spatial data with GRU layers for temporal patterns

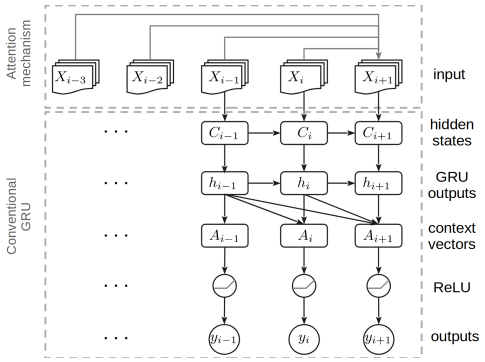


Fig. 8. ATT-GRU model [33] incorporates a self-attention layer atop GRU units for refined weighting of time steps

Table 3. TTP performance by XGBoost under different imputation methods

Taiwan	**Max**	**Chou's**	**DAE**
Mean MAE	16.766	**16.762**	16.965
Median MAE	14.183	**13.551**	14.244
Mean RMSE	47.422	**47.168**	47.646
Median RMSE	37.751	38.907	**37.748**
California	**Max**	**Chou's**	**DAE**
Mean MAE	**4.143**	**4.143**	4.150
Median MAE	3.152	3.152	3.152
Mean RMSE	8.244	**8.237**	8.250
Median RMSE	**6.318**	**6.318**	6.365

Table 4. TTP performance by LSTM under different imputation methods

Taiwan	**Max**	**Chou's**	**DAE**
Mean MAE	**16.940**	17.024	17.184
Median MAE	**12.767**	13.749	13.631
Mean RMSE	**45.034**	45.243	45.152
Median RMSE	**35.005**	35.835	35.027
California	**Max**	**Chou's**	**DAE**
Mean MAE	**4.420**	4.462	4.453
Median MAE	**3.290**	3.450	3.452
Mean RMSE	**8.453**	8.468	8.461
Median RMSE	6.608	**6.573**	6.654

6.1. Evaluation Metrics

We adopt Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for quantitative performance. Let y_i and $\hat{y}_i$ be the ground-truth and predicted values:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \hat{y}_i \,|, \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\, y_i - \hat{y}_i \,)^2}. \tag{2}$$

MAE treats all residuals equally, thus reflecting overall average deviation. RMSE penalizes larger errors more, highlighting performance under high-variance or peak traffic scenarios. We report per-route means and medians to capture robust trends across sensor locations.

6.2. Data Preprocessing Phase

Imputation Comparison We first compare three missing or zero-value imputation strategies (Section 4): interpolation (Chou's), max imputation, and DAE. Table 3 shows performance using an XGBoost predictor, while Table 4 uses LSTM.

For both XGBoost and LSTM, max and Chou's imputation usually outperform default DAE. While DAE can be competitive if extensively tuned, simpler strategies often

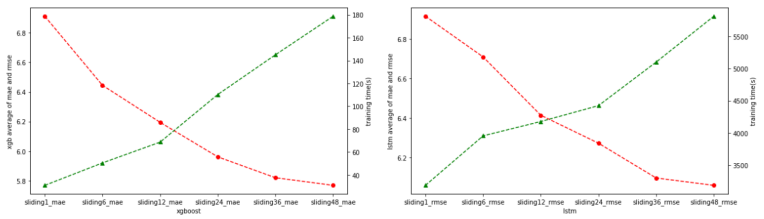


Fig. 9. Relation between (RMSE + MAE)/2, window size, and training time. Error improvement tapers off beyond 24 slots

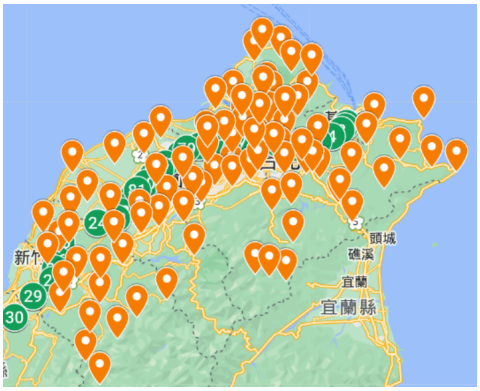


Fig. 10. Weather stations (orange) and freeway sensors (green) in northern Taiwan. Each sensor is mapped to its closest weather station

suffice in large-scale highway data. We adopt max imputation moving forward due to its conceptual simplicity and consistent performance.

Figure 9 illustrates the trade-off among average errors, window size, and training cost. A 24-slot window is a practical sweet spot for many real-world settings.

Effects of Weather Features (Taiwan Example) One might assume weather influences TTP, especially under extreme conditions like heavy rain or snow. However, Taiwan's subtropical climate seldom experiences snow, and heavy rainfall is sporadic. Figure 10 shows the distribution of weather stations. We integrate precipitation, wind speed, and temperature from the nearest station to each freeway sensor. Figures 11 and 12 compare the performance with and without weather data using XGBoost and LSTM, respectively. In most cases, adding weather *increases* errors, indicating it may behave as noise under relatively mild climates. LSTM tolerates the extra features better than XGBoost but still sees no real accuracy boost. Hence, while weather can be crucial in regions with regular extreme events (e.g., heavy snow), it may not always be beneficial. Complexity or overfitting can degrade performance, emphasizing the importance of domain-specific feature selection for sustainability-oriented traffic forecasting.

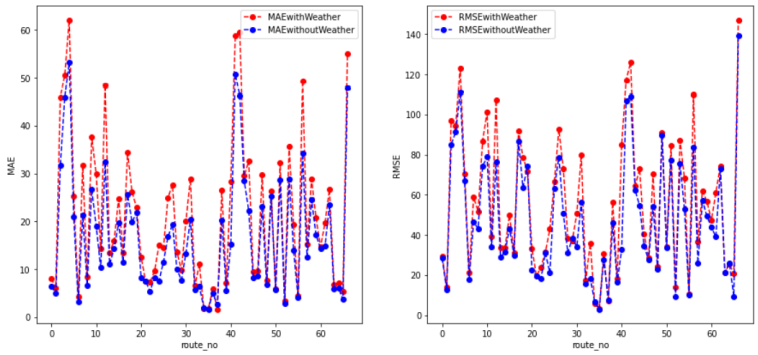


Fig. 11. Weather-feature comparison for XGBoost on the Taiwan dataset. Red lines (with weather) generally exceed blue lines (without weather) in errors

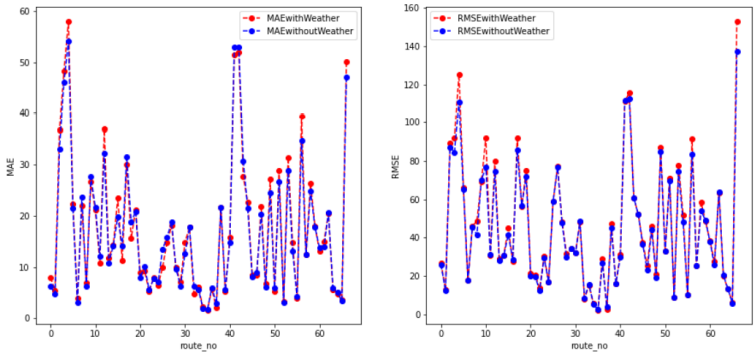


Fig. 12. Weather-feature comparison for LSTM on the Taiwan dataset. Additional features do not systematically improve prediction

Table 5. With or without extra temporal features. "All features" includes hour, minute, day-of-week, and national holiday

	Metric	None	Holiday	Minute	Hour	Day	All
Taiwan	MAE	16.892	16.793	16.011	**15.984**	16.662	15.693
	RMSE	45.997	46.063	45.509	**45.467**	45.820	45.973
California	MAE	4.144	4.125	**3.519**	3.526	4.012	3.963
	RMSE	8.067	8.062	**7.195**	7.219	7.855	7.802

Table 6. Model configurations before and after editing (EF). "-" means unspecified or not used in the original paper

Model	**Original Features (OF)**	**Original Imputation / Window**	**Edited (EF)**
Ting's Hybrid	Travel time, Speed, Volume	DAE, 6-slot window	Use hour feature, max imputation, 24-slot
Ho's Hybrid	Travel time, Volume, Hour, Day	Max imputation, 12-slot window	Add speed, max imputation, 24-slot
DE-SLSTM	Travel time, Speed, Hour, Peak, Weather	Interpolation, 12-slot	Remove weather, add volume, max imp., 24-slot
T-GCN	Speed only	Interpolation, 12-slot	Retained original setup (structure change otherwise)
ATT-GRU	Speed only	Not specified	Add max imputation, 12-slot window
LSTM	Not specified	Not specified	Use speed, volume, travel time, hour, 24-slot, max
XGBoost	Not specified	Not specified	Use speed, volume, travel time, hour, 24-slot, max

Temporal Features Finally, we examine cyclic features (e.g., hour of day, minute of hour), holiday labels, and day-of-week indicators. Table 5 summarizes the average performance (XGBoost + LSTM) on both datasets. There are three key findings: 1) Adding *hour* or *minute* features helps capture daily cycles; 2) National-holiday tags do not significantly improve accuracy, reflecting limited difference from normal weekends in these datasets; 3) Using too many features can overcomplicate training; selective choice (e.g., hour) can be more effective.

6.3. Model Comparison Phase

We now evaluate each base or hybrid model (Section 5) with two sets of features and preprocessing: 1) Original Features: The features, imputation, and window size used in the model's original publication; 2) Edited Features (EF): Our refined approach: max imputation, 24-slot windows, and hour-based temporal features. (T-GCN and ATT-GRU maintain their original structural assumptions.)

Table 6 outlines these adjustments, while Table 7 summarizes final results.

Observations:

Table 7. Comparison of all models on Taiwan (left) and California (right), with MAE and RMSE under both original features (OF) and edited features (EF). A dash "-" indicates the model's reference or default was not tested in that scenario

	Taiwan				**California**			
	MAE		RMSE		MAE		RMSE	
Model	OF	EF	OF	EF	OF	EF	OF	EF
Ting's Hybrid	19.743	16.871	52.205	49.024	4.723	3.687	9.290	7.598
Ho's Hybrid	23.844	21.800	55.150	54.972	**4.378**	4.284	8.627	8.505
DE-SLSTM	**17.235**	16.648	**44.902**	44.109	4.538	4.654	**8.022**	8.117
T-GCN	17.955	18.908	46.882	47.266	27.626	26.039	30.832	30.424
ATT-GRU	-	18.417	-	**42.207**	-	3.833	-	7.837
LSTM	-	16.041	-	44.128	-	3.577	-	**7.200**
XGBoost	-	**15.927**	-	46.807	-	**3.476**	-	7.239

1. *Importance of Data Editing.* Most models (Ting, Ho, DE-SLSTM) see improved accuracy after adopting our refined pipeline (EF), underscoring the high impact of data preprocessing.
2. *Base Models Often Excel.* Surprisingly, simpler base models—LSTM and XGBoost—deliver superior or comparable performance, despite the complexity of certain hybrid architectures (e.g., T-GCN). In particular, XGBoost yields the lowest mean MAE, whereas LSTM excels in RMSE.
3. *Context Matters.* T-GCN, which benefits from spatial adjacency, underperforms in California because we used data from a single freeway corridor with fewer complex interlinks. For large, well-connected networks, T-GCN might prove more advantageous.
4. *Sustainability Relevance.* When TTP is accurate, it not only enhances personal travel but also reduces wasted vehicle hours and emissions, contributing to environmental and well-being gains. Simpler models with well-prepared data may be especially useful for large-scale or resource-limited deployments (e.g., smaller city agencies).

7. Conclusion and Outlook

This work investigated how various data- and model-centric factors influence long-term travel time prediction (TTP) on freeways, with an eye toward environmental sustainability and human well-being. We compared multiple imputation methods, finding that simpler interpolation or max-value techniques can match or outperform more complex denoising autoencoders. We also demonstrated how carefully chosen temporal features (like hour-of-day) improve predictive accuracy, whereas additional signals (e.g., weather in Taiwan) may introduce noise. In the second phase, we examined seven modeling approaches—ranging from standard ensemble learners (XGBoost) and neural networks (LSTM) to hybrids (Ting's model, Ho's model, DE-SLSTM, T-GCN, ATT-GRU). Our results show that base models can rival or exceed more elaborate architectures if the data pipeline is well-structured. For practical deployment, these findings suggest that focusing effort on data cleanliness, the right feature set, and properly sized sliding windows is often more beneficial than adopting excessively complicated models.

For the future work, test the relative data-vs-model benefits on datasets from regions with markedly different traffic dynamics—such as cities with extreme weather (snowy or monsoon climates), high urban density, or rapidly developing suburbs—to assess the robustness and transferability of our "data over complexity" principle. On the other hand, given the demonstrated importance of accurate TTP for promoting sustainable mobility, we propose the following future directions for designing more accurate TTP algorithms that directly support sustainability goals:

1. **Dynamic and Context-Aware Feature Engineering for Sustainable Routing:** Future algorithms could incorporate adaptive feature selection mechanisms sensitive to real-time traffic conditions, weather patterns, and unusual events (e.g., accidents, large-scale gatherings). Such dynamic feature engineering approaches would enhance prediction reliability, enabling intelligent transportation systems to proactively alleviate congestion and lower emissions by directing traffic flows to more efficient routes during peak congestion periods or adverse weather conditions.

2. **Hybrid Attention-based Models for Eco-Friendly Commuting Decisions:** Integrating attention mechanisms (such as Transformer-based models or attention-enhanced GRU) with interpretable gradient-boosting methods could allow models to dynamically identify critical factors that most influence travel time variability. By clearly understanding these influential time intervals or external factors, city planners and travelers can make informed commuting decisions, effectively supporting eco-driving behavior, reducing fuel consumption, and lowering greenhouse gas emissions.
3. **Real-Time Spatiotemporal Graph Models for Enhanced Public Transport Efficiency:** Developing advanced graph-based neural network algorithms (e.g., enhanced Temporal Graph Convolutional Networks, T-GCN) that integrate real-time multimodal data (traffic sensors, GPS trajectories, social media incident reports, dynamic weather updates) could substantially improve prediction accuracy. Such improvements would directly support efficient scheduling and management of public transportation, reduce unnecessary idling times, and enhance reliability, thereby encouraging commuters to prefer public transit options over private vehicles.

In summary, future TTP algorithm research should explicitly integrate these predictive enhancements with sustainable mobility goals. Improved accuracy in TTP will reduce vehicle idling and emissions, optimize route planning, foster eco-friendly driving habits, and boost the reliability of public transit services. Consequently, such advancements will directly contribute to sustainable urban mobility, enhancing environmental outcomes and public well-being.

References

1. Schrank, D., Eisele, B., and Lomax, T. (2019). Urban Mobility Report 2019. *Texas A&M Transportation Institute*.
2. Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43, 3-19.
3. Li, Y., Zheng, Y., Zhang, H., Chen, L., Liu, Z., and Lu, X. (2015). Trend-based prediction of traffic congestion in urban areas. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 167-176.
4. Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F. Y. (2015). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2), 865-873.
5. J. Flum and M. Grohe, *Parameterized Complexity Theory*, Springer, New York, 2006.
6. C. Àlvarez, J. Gabarro and M. Serna, "Equilibria problems on games: Complexity versus succinctness," *Journal of Computer and System Sciences*, vol. 77, no. 6, pp. 1172–1197, 2011.
7. N. Abiri, B. Linse, P. Edén, and M. Ohlsson, "Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems," *Neurocomputing*, vol. 365, no. 6, pp. 137–146, 2019.
8. I. Ahmed, I. Kumara, V. Reshadat, A. S. M. Kayes, W. J. van den Heuvel, and D. A. Tamburri, "Travel time prediction and explanation with spatio-temporal features: A comparative study," *Electronics*, vol. 11, no. 1, article 106, 2021.
9. D. Billings and J. S. Yang, "Application of the ARIMA models to urban roadway travel time prediction-a case study," in *2006 IEEE International Conference on Systems, Man and Cybernetics (ICSMC 2006)*, vol. 3, Oct. 2006, pp. 2529–2534, doi:10.1109/ICSMC.2006.385244.
10. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

11. L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
12. C. M. Chen, C. C. Liang, and C. P. Chu, "Long-term travel time prediction using gradient boosting," *Journal of Intelligent Transportation Systems*, vol. 24, no. 2, pp. 109–124, 2020.
13. C. H. Chou, Y. Huang, C. Y. Huang, and V. S. Tseng, "Long-term traffic time prediction using deep learning with integration of weather effect," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2019)*, vol. 3, Apr. 2019, pp. 123–135.
14. T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, pp. 785–794, 2016, doi:10.1145/2939672.2939785.
15. Z. Chen and W. Fan, "A Freeway Travel Time Prediction Method Based on an XGBoost Model," *Sustainability*, vol. 13, no. 15, article 8577, 2021.
16. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
17. S. Das, R. N. Kalava, K. K. Kumar, A. Kandregula, K. Suhaas, S. Bhattacharya, and N. Ganguly, "Map enhanced route travel time prediction using deep neural networks," *arXiv preprint arXiv:1911.02623*, 2019.
18. Y. Duan, L. Yisheng, and F. Y. Wang, "Travel time prediction with LSTM neural network," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC 2016)*, pp. 1053–1058, 2016.
19. J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.
20. Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the 13th International Conference on International Conference on Machine Learning (ICML 1996)*, pp. 148–156, 1996.
21. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
22. M. C. Ho, Y. C. Chen, C. C. Hung, and H. C. Wu, "Deep ensemble learning model for long-term travel time prediction on highways," in *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE 2021)*, pp. 129–130, 2021.
23. I. Islek and S. G. Ögüdücü, "Use of LSTM for short-term and long-term travel time prediction," in *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*, 2018.
24. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Use of LSTM for short-term and long-term travel time prediction," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS 2017)*, pp. 3146–3154, 2017.
25. W. Qiao, A. Haghani, and M. Hamedi, "Short-term travel time prediction considering the effects of weather," *Transportation Research Record*, vol. 2308, no. 1, pp. 61–72, 2012.
26. B. Qiu and W. Fan, "Machine learning based short-term travel time prediction: Numerical results and comparative analyses," *Sustainability*, vol. 13, no. 13, article 7454, 2021.
27. P.-Y. Ting, T. Wada, Y.-L. Chiu, M.-T. Sun, K. Sakai, W.-S. Ku, A. A.-K. Jeng, and J.-S. Hwu, "Freeway Travel Time Prediction Using Deep Hybrid Model – Taking Sun Yat-Sen Freeway as an Example," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8257–8266, 2020.
28. J.-S. Yang, "Travel time prediction using the GPS test vehicle and Kalman filtering techniques," in *Proceedings of the 2005, American Control Conference, 2005*, pp. 2128–2133 vol. 3, 2005.
29. B. Yu, H. Wang, W. Shan, and B. Yao, "Prediction of bus travel time using random forests based on near neighbors," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 4, pp. 333–350, 2004, doi:10.1109/TVT.2020.2999358.
30. L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3848–3858, 2020.

31. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3146–3154, 2017.
32. X. Ran, Z. Shan, Y. Fang, and C. Lin, "An LSTM-based method with attention mechanism for travel time prediction", *Sensors*, vol. 19(4), pp. 861, 2019.
33. J.-R. Chughtai, I. U. Haq, and M. Muneeb, "An attention-based recurrent learning model for short-term travel time prediction", *PLoS One.*, vol. 17(12), pp. 1–20, 2022.
34. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6000–6010, 2017.
35. A. Kandiri, R. Ghiasi, M. Nogal, and R. Teixeira, "Travel time prediction for an intelligent transportation system based on a data-driven feature selection method considering temporal correlation", *Transportation Engineering*, vol. 18, pp. 100272, 2024.
36. K. Fu, F. Meng, J. Ye, Z. Wang, "Compacteta: A fast inference system for travel time prediction", *The 26th ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, pp. 3337–3345, 2020.
37. M. Abdollahi, T. Khaleghi, K. Yang, "An integrated feature learning approach using deep learning for travel time prediction", *Expert Systems with Applications*, vol. 139, pp.112864, 2020.
38. H. Yuan, G. Li, Z. Bao, L. Feng, "Effective Travel Time Estimation: When Historical Trajectories over Road Networks Matter", *ACM SIGMOD International Conference on Management of Data*, pp. 2135–2149, 2020.
39. Y. Shen, C. Jin, J. Hua, "TTPNet: A neural network for travel time prediction based on tensor decomposition and graph embedding", *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, pp.4514–4526, 2020.
40. Boris Nikolaevich Oreshkin, Dmitri Carpov, N. Chapados, and Yoshua Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," *arXiv preprint arXiv:1905.10437*, 2019.
41. Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," *arXiv preprint arXiv:2012.07436*, 2021.
42. Bryan Lim, Sercan O. Arık, N. Loeff, and Tomas Pfister, "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting," *arXiv preprint arXiv:1912.09363*, 2020.
43. Haixu Wu, Jianmin Xu, Jian Wang, and Mingsheng Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *arXiv preprint arXiv:2106.13008*, 2021.
44. Slawek Smyl, "Hybrid methods and ensembles for time series forecasting," *International Journal of Forecasting*, vol. 36, no. 1, pp. 11–20, 2020.
45. Valentin Flunkert, David Salinas, and Jan Gasthaus, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *arXiv preprint arXiv:1704.04110*, 2017.
46. Huimin Han, Zehua Liu, Mauricio Barrios, Jiuhao Li, and Zhixiong Zeng, "Time Series Forecasting Model for Non-Stationary Series Pattern Extraction Using Deep Learning and GARCH Modeling," *Journal of Cloud Computing*, vol. 13, no. 3, pp. 45–61, 2024.
47. Hanjia Jiang, Leilei Song, Yu Zhang, Yue Yang, and Chao Li, "Graph Neural Networks for Traffic Forecasting: A Survey," *arXiv preprint arXiv:2007.01626*, 2021.
48. Kostas Benidis, Syama Sundar Rangapuram, and others, "Deep learning for time series forecasting: A survey," *arXiv preprint arXiv:2004.13408*, 2020.
49. John A. Miller, Mohammed Aldosari, Subas Rana, and Ninghao Liu, "A Survey of Deep Learning and Foundation Models for Time Series Forecasting," *arXiv preprint arXiv:2401.13912*, 2024.
50. U.S. Environmental Protection Agency, *Greenhouse Gas Emissions from a Typical Passenger Vehicle*, EPA-420-F-18-024, 2018. [Available: https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle]

51. U.S. Census Bureau, *Census Bureau Estimates Show Average One-Way Travel Time to Work Rises to All-Time High*, 2021. [Available: https://www.census.gov/newsroom/press-releases/2021/one-way-travel-time-to-work-rises.htm]
52. U.S. Department of Transportation, *Guidance on Treatment of the Economic Value of a Statistical Life in U.S. Department of Transportation Analyses*, January 2020.

Chuang-Chieh Lin received his Ph.D. degree in Computer Science and Information Engineering from National Chung Cheng University in 2011. He received his master degree in Computer Science and Information Engineering from National Chi Nan University in 2004, and his bachelor degree in Mathematics from National Cheng Kung University in 2002. In 2007–2008, he won the DAAD-NSC Sandwich Program Scholarship and visited RWTH Aachen University for the whole year. In 2011–2014, Chuang-Chieh served his Research and Development Alternative Service as a postdoctoral research fellow in Genomics Research Center of Academia Sinica. In 2014–2018, he joined the Computation Theory and Algorithms Group of the Institute of Information Science at Academia Sinica as a postdoctoral research fellow. In 2018–2020, he devoted himself into the fin-tech industry and focused on quantitative analysis of market micro-structures. In 2021–2024, he served as the faculty of Department of Computer Science and Information Engineering of Tamkang University. Started from August of 2024, he joined the faculty of Department of Computer Science and Engineering of National Taiwan Ocean University. His research interests include design and analysis of algorithms, game theory, machine learning, bioinformatics and quantitative finance.

Min-Chu Ho received the B.S. degrees in the Department of Management Information Systems and the Department of Foreign Languages and Literatures at National Chung Hsing University, Taiwan, in 2022. His interdisciplinary background combines technical training in data science with cross-cultural communication skills. His research interests include machine learning, deep learning, and their applications in intelligent information systems.

Chih-Chieh Hung received the Ph.D. degree in computer science from National Chiao Tung University, Taiwan, in 2010. He is currently an Associate Professor in the Department of Management Information Systems at National Chung Hsing University, Taiwan. He has held leadership positions including Chair (2024–2027) and Vice Chair (2021–2024) of the IEEE Computational Intelligence Society, Tainan Section, as well as Director and Secretary General of the Taiwan Association for Artificial Intelligence (TAAI). His research interests include data mining, spatiotemporal data analytics, intelligent transportation systems, and artificial intelligence. Dr. Hung has received several international awards, including Best Paper Awards at TAAI 2024 and 2023, the ACM Workshop on Location-Based Social Networks (2009), and a Gold Award and a Silver Award in Artificial Intelligence at the International Conference on Frontier Computing 2024 and 2025.

Received: January 26, 2025; Accepted: June 12, 2025.

Computer Science and Information Systems 22(4):1817–1834 https://doi.org/10.2298/CSIS250311072C

Cultural Pragmatics and Causal Connectives: A Contrastive Study of Korean and English Using the AI-Hub Parallel Corpus

Sujeong Choi[1] and Sin-hye Nam[2,⋆]

[1] KDI School of Public Policy and Management,
263 Namsejong-ro, Sejong-si, 30149, Republic of Korea
schoi@kdischool.ac.kr
[2] Kyung Hee University,
26 Kyungheedae-ro, Dongdaemun-gu, Seoul, 02447, Republic of Korea
namsh@khu.ac.kr

Abstract. This interdisciplinary study analyzes Korean and English causal connective expressions using the AI-Hub Korean-English parallel corpus. The primary objective is to identify the unique linguistic and cultural features of Korean causal connectives by comparing them with their English counterparts. Korean includes a wide range of causal connectives, many of which exhibit additional pragmatic features such as [+negative], [+uncertainty], and [+plurality]. From both linguistic and cultural perspectives, this study investigates whether these features are exclusive to Korean and explores the cultural factors contributing to their prevalence.

To extend the analysis into a computational framework, the study defines a formal task for evaluating the preservation of pragmatic meaning in translation. Specifically, each Korean-English sentence pair is annotated for the pragmatic features expressed in Korean, and the extent to which those features are retained in the English translation is assessed. Two task formulations are proposed: (1) a binary classification indicating full preservation vs. loss or shift, and (2) a continuous "pragmatic shift score" ranging from 0.0 to 1.0. This enables future implementation of rule-based or learning-based models to detect pragmatic mismatches in translation.

The analysis of seven Korean causal connectives reveals that the additional pragmatic features are specific to Korean and rarely appear in English. These features are culturally grounded: the [+negative] feature aligns with Korean speakers' tendency to express disapproval indirectly to preserve politeness; [+uncertainty] reflects a cultural preference for hedging and softening assertions; and [+plurality] indicates an avoidance of definitive statements, consistent with indirect and euphemistic communication strategies common in Korean discourse.

Ethical considerations regarding data licensing and cultural bias are addressed. This research offers practical implications for computational linguistics, translation studies, and Korean language education. By uncovering culturally embedded differences in how causality is expressed, the study enhances cross-cultural understanding and contributes to improved communication in multilingual contexts.

Keywords: Parallel Corpus, Contrastive Study, Causal Connective Expressions, Interpretive Ethno-grammar

⋆ Corresponding author

1. Introduction

1.1. Linguistic and Cultural Motivation

Computer science, particularly in the field of natural language processing (NLP), has significantly contributed to the development and analysis of linguistic corpora. Technological advances have enabled efficient collection, organization, and large-scale processing of corpora, facilitating various applications in language research and education. Among different types of corpora, the parallel corpus, which consists of aligned translations in two or more languages, offers especially rich resources for contrastive linguistic analysis. By providing a one-to-one mapping of corresponding expressions across languages, parallel corpora are instrumental in examining syntactic, semantic, and pragmatic differences.

In the context of Korean-English comparison, one particularly fruitful area of inquiry involves causal connective expressions, which are more diverse and nuanced in Korean than in many other languages. This richness is not merely grammatical but reflects culturally embedded communicative preferences. For instance, the Korean language often employs a variety of causal expressions to convey subtle shades of intention, obligation, and emotion. Such phenomena are similar to the lexical diversity observed in the Hanunoo language of the Philippines regarding rice types (Conklin, 1957), or the range of friendship-related terms in Russian that contrast with the single English word ”friend” (Wierzbicka, 1997). These cases exemplify how cultural values shape language and influence not only vocabulary but also the structure and use of expressions.

Understanding this connection between language and culture is essential for effective cross-cultural communication. As Saville Troike (2003) notes, language proficiency includes not only mastery of phonology, vocabulary, and grammar but also the ability to communicate appropriately in various social and cultural contexts. In this light, the present study seeks to explore how causal meaning is expressed and translated across Korean and English, aiming to shed light on underlying cultural differences that shape linguistic expression.

1.2. Research Objectives and Scope

This study conducts a contrastive linguistic analysis of Korean and English causal connective expressions using a large-scale Korean-English parallel corpus. The primary goal is to identify how specific Korean causal connectives are translated into English and to evaluate whether the pragmatic meanings embedded in the Korean expressions are preserved, altered, or lost during translation.

By doing so, the study aims to uncover both linguistic and cultural asymmetries between the two languages. This contrastive analysis has practical implications for second language education, particularly for English-speaking learners of Korean. A better understanding of how Korean causal connectives function—both linguistically and culturally—can help learners avoid misinterpretations and improve pragmatic competence. Ultimately, the findings aim to support cross-cultural communication by enhancing awareness of how causality and intention are linguistically encoded in different language communities.

1.3. Overview of the Computational Task

To extend the linguistic analysis into a computational framework, this study proposes a formal task that quantitatively evaluates the preservation of pragmatic meaning in Korean-English translations of causal connectives. Using the AI-Hub Korean-English parallel corpus, the task involves sentence pairs in which a Korean sentence containing a specific causal connective (e.g., *Ast-(eun) tase, Vst-neun tase, Ast/Vst-(eun) nameoji, Vst-neurago, Vst-neun barame, Vst-neun tonge, Ast/Vst-a/eoseo geureonji, Ast/Vst-go haeseo*) is aligned with its English translation. Each Korean sentence is annotated for pragmatic features such as [+negative], [+uncertainty], or [+plurality].

The objective is to determine the extent to which these pragmatic features are retained in the English translation. Two task formulations are proposed: (1) a binary classification, where a label of 1 indicates full preservation of meaning and 0 indicates loss or change, and (2) a continuous scoring system, where a "pragmatic shift score" ranges from 0.0 (fully preserved) to 1.0 (completely lost or altered). This task definition paves the way for developing rule-based or machine learning models capable of detecting pragmatic mismatches in translation.

2. Related Works

2.1. Parallel Corpus and Linguistic Research

A parallel corpus is a corpus consisting of pairs of texts in two or more languages. Since the proposal of the first parallel corpus alignment system by Kay and Röscheisen (1988), it has been widely utilized in linguistics and NLP (Natural Language Processing) fields. The application areas of parallel corpora include machine translation, Cross-language Information Retrieval (CLIR), lexicography, language education, contrastive linguistics, and more. Over the years, various studies using parallel corpora have been conducted in Korean language and culture research. Recently, with the global spread of Korean cultural content, there has been a growing interest in Korean language education and the development of automatic translation technologies, leading to an increasing demand for language resources.

In response to this demand, institutions like the National Institute of Korean Language and AI-Hub, which are focusing on Korean language data construction, are actively involved in building Korean-foreign parallel corpora. Research on parallel corpora can be broadly categorized into studies in the field of Natural Language Processing (NLP) and studies in the fields of linguistics and applied linguistics. In the NLP field, discussions mainly revolve around the construction of parallel corpora. For instance, Vu et al. (2020) reported on the development of the Ulsan Parallel Corpora (UPC), which includes Korean-English and Korean-Vietnamese datasets. Park et al. (2022) reported on the construction of seven types of parallel corpora, including Korean, by AI Hub, and evaluated their performance. Additionally, Bang et al. (2022) constructed the English-Korean speech parallel corpus (EnKoST-C) and presented the evaluation results of its performance.

On the other hand, in the fields of linguistics and applied linguistics, the majority of research reports are based on the utilization of parallel corpora. Studies conducted using parallel corpora, with Korean as the focus, mostly involve Korean-Chinese (Sim, 2015; Li 2021; Yu 2022; Shim, 2023), Korean-English (Seo, 2008; Park, 2017; Park & Lim,

2020), and Korean-Japanese (O & Takiguchi, 2015; Kim & Jang, 2011) corpora. Particularly, parallel corpora have been utilized for contrastive linguistic studies, and a few specific examples are as follows: Seo (2008) described the English equivalents of the Korean sentence-ending suffix *'-get-'* using a Korean-English parallel corpus; Yu (2002) used a Korean-Chinese parallel corpus to describe the Chinese counterparts of the Korean expression *'-(eu)leo.'*; Xue (2021) conducted a contrastive analysis of honorific expressions based on a Korean-Chinese parallel corpus; Wilczynski (2021) conducted a contrastive study on the *'ida'* (to be) construction based on a trilingual Korean-Polish-English parallel corpus.

Although we have examined examples of such research, the parallel corpora traditionally utilized by linguists have had limitations in terms of their size and balance, which in turn might have imposed certain constraints on the research outcomes. However, as noted earlier, due to the recent demand in the fields of artificial intelligence and natural language processing, high-quality Korean-foreign language parallel corpora have been rapidly constructed on a large scale. This allows for the utilization of these new language resources to conduct contrastive linguistic studies of even higher quality than previous research. Especially, this is true for Korean-English parallel corpora. Therefore, in this study, our aim is to take advantage of the latest parallel corpora, which offer high precision, large scale, and balance, driven by the achievements in the field of NLP, to conduct applied linguistic research with improved precision.

2.2. Korean Causal Connective Expressions

According to Choi (2022), Korean has the highest number of grammatical connective expressions used to indicate causation. Choi (2022) conducted an analysis of grammatical items from ten Korean textbooks and three grammar references, organizing them according to their meanings. This study identified a total of 68 pragmatic categories, with the [cause] category exhibiting the most synonymous expressions. Specifically, 24 synonymous grammar items that express the concept of [cause] in Korean were identified. Although these expressions share a common function of indicating the cause for the subsequent clause, they differ in terms of syntactic conditions, contextual formality, and pragmatic nuances. For instance, the following causal grammar items from Korean textbooks are noted for their additional pragmatic features compared to the more neutral causal items.

(1) *Ast-(eun) tase, Vst-neun tase*: This expression conveys a negative cause or reason.
(2) *Ast/Vst-(eun) nameoji*: This is used when an action or situation in the first clause deteriorates, leading to negative consequences in the subsequent clause.
(3) *Vst-neurago*: This indicates a cause, reason, or purpose that results in a negative outcome.
(4) *Vst-neun barame*: This expression denotes a cause or reason where the preceding context negatively impacts the subsequent action.
(5) *Vst-neun tonge*: This indicates a cause or reason that leads to a negative situation or results in the following clause.
(6) *Ast/Vst-a/eoseo geureonji*: This is used when a preceding action or situation appears to be a cause but remains uncertain.

(7) *Ast/Vst-go haeseo*: This suggests that the preceding clause represents one of several reasons for what follows.

As underlined above, the causal connective expressions in examples (1) to (7) feature additional pragmatic attributes: the grammatical items in (1) to (5) include a [+negative] feature, the item in (6) denotes a [+uncertainty] feature, and the item in (7) implies a [+plurality] aspect. Thus, Korean causal connective expressions not only indicate the cause for the subsequent clause but also convey additional pragmatic nuances and complexities. As we have seen above, there are a variety of grammar items in Korean that express cause. Due to the diversity of connectives expressing cause, there has been a lot of research on the similarities and differences between Korean causal connectives expressions(Chin, 2005; Ahn, 2007; Park, 2008; Li, 2011; Yoo, 2015; Park, 2018; Jeon, 2021, etc.) However, there is a lack of comparative linguistic research on whether these diversities and characteristics are unique to Korean or whether they are also present in other languages. If these features are unique to Korean, it would be possible to explore their causes from a linguistic and cultural perspective, and in terms of language education, it would be a particularly important topic to focus on. To explore this, this study uses a parallel Korean-English corpus to compare seven causal connectives that contain pragmatic qualities additional to the meaning of cause with their counterparts in English. Through the analysis of these findings, this study aims to interpret the reasons for the substantial number of causal connective expressions in Korean from a cultural standpoint, and to explore the pedagogical implications of these differences in the context of language education.

3. Data and Methodology

The data for this study were sourced from the Korean-English parallel corpora created by *AI-Hub*, an AI integration platform managed by Korea's Ministry of Science and ICT and the National Information Society Agency. These corpora were produced as part of a project aimed at constructing data for artificial intelligence training.

3.1. Corpus Overview and Genre Distribution

The *AI-Hub* Korean-English corpus includes 1.6 million sentence pairs across three styles: literary, colloquial, and conversational. From this dataset, we selected a subset of 300,000 sentence pairs, distributed as follows.

In total, 300,000 sentences (3,438,086 words) from the Korean-English parallel corpora were analyzed to investigate the English expressions that correspond to Korean reason/causal connective expressions.

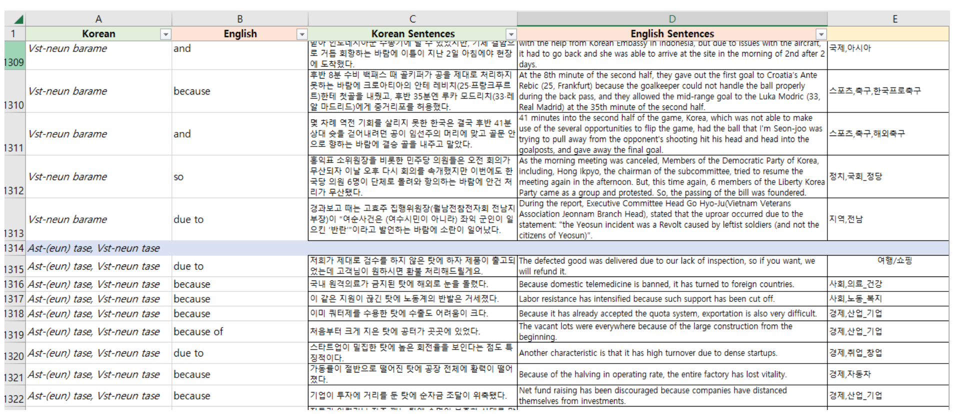
3.2. Preprocessing and Data Filtering

The data were processed using a hybrid approach that combined automatic filtering and manual validation. First, automatic filtering was conducted based on keyword matching to identify sentences containing one or more of seven pre-defined Korean causal connective expressions. Following this, manual validation was performed to remove false

Table 1. Information of Korean-English Parallel Corpora

No.	Style	Context	The Number of Sentences/Words	
1	Literary	News Articles	200,000 sentences (2,658,545 words)	300,000 sentences (3,438,086 words)
2	Colloquial (Conversation)	Conversation in a meeting, shopping, school, restaurant, etc.	100,000 sentences (779,541 words)	

positives and ensure contextual appropriateness. The overall preprocessing procedure included sentence tokenization, normalization, filtering based on the presence of the target expressions, and manual verification of the extracted sentence pairs. Figure 1 provides an illustrative example of this process. A total of 479 sentences containing the seven target Korean causal connective expressions were extracted from the Korean-English parallel corpus. For each sentence pair, the Korean causal connective was identified (as shown in Column A), and its corresponding English expression was extracted (Column B). Column C displays the original Korean sentence, and Column D presents the corresponding English translation. For example, in row 310, the Korean connective expression *Vst-neun barame* corresponds to the English connective because. While some expressions showed a direct causal correspondence, others were translated with non-causal expressions. Based on this analysis, each English expression was manually categorized as either a causal or non-causal expression.



	A	B	C	D	E
1	Korean	English	Korean Sentences	English Sentences	
1309	*Vst-neun barame*	and	로 거듭 회항하는 바람에 이틀이 지난 2일 아침에야 현장에 도착했다.	with the help from Korean Embassy in Indonesia, but due to issues with the aircraft, it had to go back and she was able to arrive at the site in the morning of 2nd after 2 days.	국제,아시아
1310	*Vst-neun barame*	because	후반 8분 수비 백패스 때 골키퍼가 공을 제대로 처리하지 못하는 바람에 크로아티아의 안테 레비치(25·프랑크푸르트)한테 첫골을 내줬고, 후반 35분엔 루카 모드리치(33·레알 마드리드)에게 중거리포를 허용했다.	At the 8th minute of the second half, they gave out the first goal to Croatia's Ante Rebic (25, Frankfurt) because the goalkeeper could not handle the ball properly during the back pass, and they allowed the mid-range goal to the Luka Modric (33, Real Madrid) at the 35th minute of the second half.	스포츠,축구,한국프로축구
1311	*Vst-neun barame*	and	몇 차례 역전 기회를 살리지 못한 한국은 결국 후반 41분 상대 슛을 걷어내려던 공이 임선주의 머리에 맞고 골문 안으로 향하는 바람에 결승 골을 내주고 말았다.	41 minutes into the second half of the game, Korea, which was not able to make use of the several opportunities to flip the game, had the ball that I'm Seon-joo was trying to pull away from the opponent's shooting hit his head and head into the goalposts, and gave away the final goal.	스포츠,축구,해외축구
1312	*Vst-neun barame*	so	홍익표 소위원장을 비롯한 민주당 의원들은 오전 회의가 무산되자 이날 오후 다시 회의를 속개했지만 이번에도 한국당 의원 6명이 단체로 몰려와 항의하는 바람에 안건 처리가 무산됐다.	As the morning meeting was canceled, Members of the Democratic Party of Korea, including, Hong Ikpyo, the chairman of the subcommittee, tried to resume the meeting again in the afternoon. But, this time again, 6 members of the Liberty Korea Party came as a group and protested. So, the passing of the bill was foundered.	정치,국회_정당
1313	*Vst-neun barame*	due to	경과보고 때는 고효주 집행위원장(월남전참전자회 전남지부장)이 "여순사건은 (여수시민이 아니라) 좌익 군인이 일으킨 '반란'"이라고 발언하는 바람에 소란이 일어났다.	During the report, Executive Committee Head Go Hyo-Ju(Vietnam Veterans Association Jeonnam Branch Head), stated that the uproar occurred due to the statement: "the Yeosun incident was a Revolt caused by leftist soldiers (and not the citizens of Yeosun)".	지역,전남
1314	*Ast-(eun) tase, Vst-neun tase*				
1315	*Ast-(eun) tase, Vst-neun tase*	due to	저희가 제대로 검수를 하지 않은 탓에 하자 제품이 출고되었는데 고객님이 원하시면 환불 처리해드릴게요.	The defected good was delivered due to our lack of inspection, so if you want, we will refund it.	여행/쇼핑
1316	*Ast-(eun) tase, Vst-neun tase*	because	국내 원격의료가 금지된 탓에 해외로 눈을 돌렸다.	Because domestic telemedicine is banned, it has turned to foreign countries.	사회,의료_건강
1317	*Ast-(eun) tase, Vst-neun tase*	because	이 같은 지원이 끊긴 탓에 노동계의 반발은 거세졌다.	Labor resistance has intensified because such support has been cut off.	사회,노동_복지
1318	*Ast-(eun) tase, Vst-neun tase*	because	이미 쿼터제를 수용한 탓에 수출도 어려움이 크다.	Because it has already accepted the quota system, exportation is also very difficult.	경제,산업_기업
1319	*Ast-(eun) tase, Vst-neun tase*	because of	처음부터 크게 지은 탓에 공터가 곳곳에 있었다.	The vacant lots were everywhere because of the large construction from the beginning.	경제,산업_기업
1320	*Ast-(eun) tase, Vst-neun tase*	due to	스타트업이 밀집한 탓에 높은 회전율을 보인다는 점도 특징적이다.	Another characteristic is that it has high turnover due to dense startups.	경제,취업_창업
1321	*Ast-(eun) tase, Vst-neun tase*	because	가동률이 절반으로 떨어진 탓에 공장 전체에 활력이 떨어졌다.	Because of the halving in operating rate, the entire factory has lost vitality.	경제,자동차
1322	*Ast-(eun) tase, Vst-neun tase*	because	기업이 투자에 거리를 둔 탓에 순자금 조달이 위축됐다.	Net fund raising has been discouraged because companies have distanced themselves from investments.	경제,산업_기업

Fig. 1. An Example of Data Analysis

This classification was conducted using a hybrid approach combining automatic keyword-based filtering and manual validation. The annotation was performed by two trained annotators, and the inter-annotator agreement, measured using Cohen's Kappa, was 0.87, indicating substantial consistency. Due to licensing restrictions, the preprocessing scripts

and annotation guidelines are not publicly available; however, they can be provided upon reasonable request for academic and research purposes. (see Section 3.7).

3.3. Train/Dev/Test Split

As this study is not aimed at model training but rather focuses on the qualitative and quantitative analysis of linguistic patterns, the dataset was not divided into separate training, development, or test subsets. Instead, a total of 479 sentence pairs containing Korean causal connectives were extracted and analyzed. This annotated subset can be made available for evaluation or further research upon reasonable request.

3.4. Annotation Guidelines and Procedure

The annotation process was carried out by two trained annotators with backgrounds in linguistics. Their task was to categorize the corresponding English expressions as either causal or non-causal. To ensure consistency, detailed annotation guidelines were developed and followed. These guidelines included definitions of causal and non-causal relations, illustrative examples, descriptions of edge cases, and clearly defined decision criteria for handling ambiguous expressions. Before annotating the main dataset, the annotators participated in a training session using a pilot dataset of 50 sentence pairs. They then independently annotated the full dataset, after which discrepancies were reviewed and resolved collaboratively, leading to refinements in the annotation protocol.

3.5. Inter-Annotator Agreement

To evaluate the reliability of the annotation, inter-annotator agreement was measured using Cohen's Kappa (κ). The resulting κ value was 0.87, which indicates a high level of agreement and suggests substantial consistency between the two annotators.

3.6. Data Analysis Procedure

The overall procedure for analyzing the Korean-English sentence pairs is illustrated in Figure 2. First, Korean sentences containing one or more of the seven target causal connective expressions were identified within the parallel corpus. The corresponding English sentences were then extracted, and the English expressions aligned with the Korean causal markers were manually examined. Based on pragmatic criteria, each English expression was categorized into either a causal expression group or a non-causal expression group. Expressions that conveyed explicit causal meaning were included in the causal group, while others that did not indicate causality were classified as non-causal.

Additionally, the study investigated the specific English equivalents of each Korean causal connective. For instance, in the case of the Korean expression *'-(eun) tase'*, 145 sentence pairs were retrieved from the corpus. Among the corresponding English sentences, some included causal expressions while others did not. These were divided into the two aforementioned categories and further analyzed to determine the frequency and proportion of each English counterpart. The results of this frequency analysis, including representative expressions and their distribution, are presented in Tables 2 through 4 in Section 4.

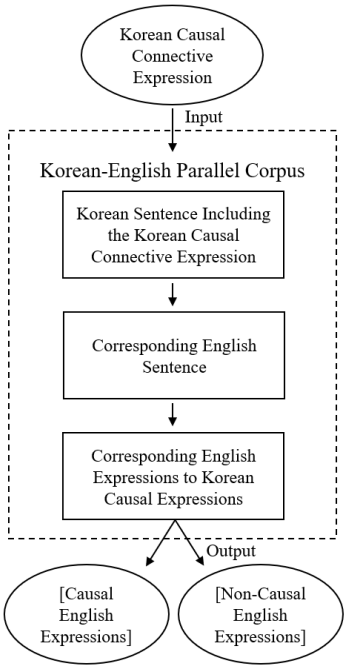


Fig. 2. Procedure of Data Analysis

3.7. Data Licensing and Access

The original corpus used in this study is distributed by *AI-Hub* under specific licensing terms that restrict public redistribution. Researchers who wish to access the original dataset may do so by applying through the *AI-Hub* website and agreeing to the platform's usage policies. To support reproducibility and facilitate further research, supplementary materials—including preprocessing scripts, annotation guidelines, and a sample of annotated data with licensing-safe examples—can be made available upon request to qualified researchers for non-commercial academic purposes.

4. Contrastive Analysis on Causal Expressions in Korean and English

As described in Chapter 2, the Korean causal connective expressions *Ast-(eun) tase, Vst-neun tase, Ast/Vst-(eun) nameoji, Vst-neurago, Vst-neun barame, Vst-neun tonge* connote [+negative] meaning, the expression *Ast/Vst-a/eoseo geureonji* has [+uncertainty] meaning, and the expression *Ast/Vst-go haeseo* contains [+plurality] meaning additionally. In this chapter, the study investigates whether the additional pragmatic features of Korean causal connective expressions are reflected in their English counterparts, through a contrastive analysis of the results from a Korean-English parallel corpus. This section analyzes the degree of pragmatic preservation between Korean causal connectives and their

English translations. While the computational task is defined in Section 1.3 as a classification or scoring problem, this study implements a manual and qualitative evaluation, rather than applying an automated classifier. Each connective is examined to assess whether the pragmatic features are retained or lost in translation, providing a foundational dataset for future computational modeling.

4.1. [+Negative] feature in causal connective expressions

First, the English expressions corresponding to the Korean causal connective expressions *Ast-(eun) tase, Vst-neun tase, Ast/Vst-(eun) nameoji, Vst-neurago, Vst-neun barame, Vst-neun tonge*, which imply the [+negative] feature, were analyzed. Table 2 presents the frequency and proportion of these Korean causal connective expressions with the [+negative] feature in the Korean-English parallel corpus, along with their corresponding English expressions.

As shown in Table 2, in English no causal expressions exhibiting the [+negative] feature observed in Korean expressions. Instead, only basic or neutral causal expressions such as *because, due to, as, since* were used in English as their corresponding counterparts. The following (a) (e) in Figure 3 are the examples of Korean sentences containing the causal connective expressions with [+negative] feature and the corresponding English sentences.

In (a) (e) in Figure 3, the causal connective expressions used in the Korean sentences all indicate that the cause of the negative outcome in the second clause is in the first clause, while implying a negative attitude towards the cause described in the first clause. On the other hand, the corresponding English causal expressions are neutral, indicating that the speaker is conveying a neutral view of the causal event without a negative attitude. Rather, the analysis revealed a strong tendency in English to use direct negative vocabulary when the speaker intends to convey a negative meaning. In English, words with negative connotations are employed directly to express negative intentions. In Korean, however, grammatical expressions that imply negativity are often used in a more indirect manner. This can also be seen in the example that some of the sentences that use causal connective expressions that connote the [+negative] feature in Korean are not expressed by grammatical expressions of cause in English, but rather verbs with negative connotations. For example, in some cases, the speaker's negative attitude expressed through *'Vst-neun tonge'* in Korean was corresponded to by the negative verb *'confuse'* in English, and in other cases, the negative attitude expressed through *'Vst-neun barame'* in Korean was corresponded to by the negative verb *'disturb'* in English. This means that in Korean, even when speakers do not use direct vocabulary to convey negative intentions, negative nuances can still be communicated indirectly through grammatical expressions. Searle (1969, 1979) argued that the primary motivation for indirect speech is politeness. The discussion of politeness in linguistics was developed by Lakoff (1972) and culminated in Brown& Levinson (1987) with the concept of personal politeness strategy. Since individuals have a sense of face, and any act that threatens their desire for face tends to be viewed as a face-threatening act, speakers and listeners will adopt politeness communication strategies to avoid face-threatening acts as much as possible and minimize conflict. In Korean, the speaker's negative attitude toward the situation can be interpreted as an intention to increase politeness and minimize status threats by using causal connectives that have negative connotations rather than using direct negative vocabulary.

Table 2. Korean Causal Connective Expressions Containing the [+negative] feature and Corresponding English Expressions

No.	Korean	N	Category	English	N	Rate
1	Ast-(eun) tase, Vst-neun tase	145	Causal Expressions (N=138, R=95.1)	because (of)	57	39.31
				due to	32	22.07
				as	25	17.24
				since	14	9.66
				cause	4	2.76
				so	2	1.38
				by	1	0.69
				for	1	0.69
				therefore	1	0.69
				so … that	1	0.69
				after	2	1.38
				and	1	0.69
				as a result	1	0.69
				result in	1	0.69
				lead	1	0.69
				while	1	0.69
			Non-causal Expressions (N=7, R=4.82)	because	5	20.83
				as	3	12.50
2	Ast/Vst-(eun) nameoji	24	Causal expressions (N=13, R=54.16)	because	5	20.83
				as	3	12.50
				so … that	3	12.50
				for	2	8.33
			Non-causal expressions (N=11, R=45.84)	and	5	20.83
				∅	5	20.83
				by ing	1	4.17
3	Vst-neurago	13	Causal expressions (N=6, R=46.16)	because	4	30.77
				due to	1	7.69
				as	1	7.69
				ing	2	15.38
			Non-causal expressions (N=7, R=53.84)	with	1	7.69
				from	1	7.69
				and	1	7.69
				bring, on	2	15.38
4	Vst-neun barame	153	Causal expressions (N=122, R=79.74)	because	61	39.87
				so	24	15.69
				as	14	9.15
				since	11	7.19
				due to	6	3.92
				cause	6	3.92
				and	12	7.84
				∅	9	5.88
			Non-causal expressions (N=31, R=20.26)	, which	2	1.31
				after	2	1.31
				when	2	1.31
				from, drive, disrupt, get	4	2.60
5	Vst-neun tonge	4	Causal expressions (N=2, R=50)	since	1	25
				due to	1	25
			Non-causal expressions (N=2, R=50)	and	1	25
				confuse	1	25

a. Korean: 기업이 투자에 거리를 *둔 탓에* 순자금 조달이 위축됐다.
kiʌbi tʰudʑɑɛ gʌɾiɾɯɾ *dun tʰɑsɛ* sundʑɑgɯm dʑodɑɾi witɕʰukt*wɛtt*ɑ.
English: Net fund raising has been discouraged *because* companies have distanced themselves from investments.

b. Korean: 한때 우리나라 농업은 생산성을 너무나도 중시*한 나머지* 농약과 비료를 너무 많이 써서 생명력을 잃은 논이 많다.
hɑnt*ɛ uɾinɑɾɑ noŋʌbɯn sɛŋsɑnsʌŋɯɾ nʌmunɑdo dʑuŋsi*an namʌdʑi* noŋjakk*wa biɾjoɾɯɾ nʌmu mɑni s*ʌsʌ sɛŋmjʌŋnjʌgɯɾ iɾɯn noni mɑnʰɑ.
English: There are many rice paddies that have lost their vitality due to excessive use of pesticides and fertilizer *because* productivity was much too important for Korean agriculture.

c. Korean: 진짜? 나도 출장 다녀오*느라고* 전혀 몰랐지.
tɕintɕ*ɑ? nɑdo tɕʰuldʑɑŋ dɑnjʌo*nɯɾago* dzʌnjʌ mollattɕ*i.
English: Really? I didn't know *because* I was travelling for work.

d. Korean: 지하철이 고장으로 멈추*는 바람에* 하마터면 지각할 뻔했어.
tɕiatɕʰʌɾi godʑɑŋɯɾo mʌmtɕʰu*nun barame* ɑmɑtʰʌmjʌn dʑigɑkʰɑɾ p*ʌnɛs*ʌ.
English: I was almost late *because* the subway stopped due to a failure.

e. Korean: 대형마트나 홈쇼핑, 백화점에서 이벤트를 벌이며 가격 할인행사까지 하*는 통에* 소상공인들은 더욱 찬바람만 느끼고 있다.
tɛjʌŋmɑtʰɯnɑ omsjopʰiŋ bɛkʰwadʑʌmɛsʌ ibɛntʰɯɾɯɾ bʌɾimjʌ gɑgjʌg ɑɾinɛŋsɑk*adʑi ɑ*nun tʰoŋɛ* sosɑŋgoŋindɯɾɯn dʌug tɕʰambɑɾɑmmɑn nɯk*igo itt*ɑ.
English: Small business owners are feeling the chill of the wind more *since* the large supermarkets, home shopping and department stores are holding events and even price discount events.

Fig. 3. Example 1: Korean sentences and corresponding English sentences

4.2. [+Uncertainty] feature in causal connective expression

The English expressions corresponding to the Korean causal connective expression *Ast/Vst-a/eoseo geureonji*, which indicates the [+uncertainty] feature, were examined. Table 3 presents the frequency and proportion of Korean causal connective expressions with the [+uncertainty] feature in the Korean-English parallel corpus, along with their corresponding English expressions.

Table 3. Korean Causal Connective Expression Containing the [+uncertainty] feature and Corresponding English Expressions

No.	Korean	N	Category	English	N	Rate
1	Ast/Vst-a/eoseo geureonji	124	Causal expressions (N=113, R=90.4)	because	31	24.80
				maybe (it's) because	28	22.40
				so	13	10.40
				probably because	11	8.80
				perhaps because	7	5.60
				since	7	5.60
				due to	3	2.40
				maybe that's why	2	1.60
				maybe the reason	2	1.60
				not sure if it is because	2	1.60
				with	2	1.60
				guess it's because	1	0.80
				perhaps … so	1	0.80
				perhaps due to	1	0.80
				whether it's from	1	0.80
				so … that	1	0.80
			Non-causal expressions (N=12, R=9.6)	∅	8	6.40
				and	2	1.60
				as to whether	1	0.80
				no wonder	1	0.80

Conjectural terms such as *"maybe," "probably,"* and *"perhaps"* were absent in over 50% of the English expressions corresponding to *Ast/Vst-a/eoseo geureonji* within the causal expression category. This observation suggests that instances of speculation in causal expressions are significantly more frequent in Korean.

The Korean sentences in (a) (c) in Figure 4, all use causal expressions that imply the [+uncertainty] feature. However, in the corresponding English sentences, only (a) contains *'probably'*, which expresses [+uncertainty]. The English sentence in (b) uses only the causal connective expression *'since'* without the [+uncertainty] feature, and (c) has no corresponding causal connective expression at all. As noted, this aligns with the phenomenon of frequent use of hedging expressions in Korean. When expressing causation, there is a tendency to convey the message in a mild rather than strong or explicit manner, which is closely related to demonstrating politeness toward the listener. This distinction is particularly apparent in examples (b) and (c). In (b), if the hearer is struggling with an event while the speaker is fine, even though the speaker does not have to use the [+uncertainty] feature because *"I'm used to it because I go there every day"* is a recurring event for the speaker and the speaker's personal belief, to show consideration for the hearer, the speaker softens her stance by adding [+uncertainty] to the explanation of her ease with the

a. Korean: 시험 기간 *이라서 그런지* 도서관에 사람이 엄청 많네.
sɪʌm gigɑn*irasʌ gurʌndʑi* dosʌgwanɛ sɑrɑmi ʌmtɕʰʌŋ mɑnnɛ.
English: There are a lot of people in the library *probably because* it's the exam season.

b. Korean: 나는 매일 다*녀서 그런지* 익숙한데, 어떤 게 가장 힘드니?
nɑnɯn mɛiɾ dɑ*njʌsʌ gurʌndʑi* iks*ukʰɑndɛ ʌt*ʌn gɛ gɑdʑɑŋ imdɯni?
English: I'm used to it *since* I walk there every day, but what is most difficult?

c. Korean: 제가 일한 지 얼마 안 *돼서 그런지* 업무량이 조금 버거운데, 혹시 비법 같은 것 없어요?
tɕɛgɑ iɾɑn dʑi ʌlmɑ ɑn *ɗwɛsʌ gurʌndʑi* ʌmmuɾjaŋi dʑogɯm bʌgʌundɛ oks*i bibʌb gɑtʰɯn gʌs ʌps*ʌjo?
English: It hasn't been long since I've worked here. I feel like the workload is a little heavy. Is there a secret you can tell me?

Fig. 4. Example 2: Korean sentences and corresponding English sentences

situation. In the case of (c), the speaker is reducing the burden on the listener by adding [+uncertainty] to the reason that she is feeling overwhelmed, instead of saying that her workload is heavy.

4.3. [+Plurality] feature in causal connective expression

The English expressions corresponding to the Korean causal connective expression *Ast/Vst-go haeseo*, which denotes the [+plurality] of the reason, were examined. Table 4 presents the frequency and proportion of Korean causal connective expressions featuring the [+plurality] aspect in the Korean-English parallel corpus, along with their corresponding English expressions.

Table 4. Korean Causal Connective Expression Containing the [+plurality] feature and Corresponding English Expressions

No.	Korean	N	Category	English	N	Rate
1	*Ast/Vst-go haeseo*	9	Causal expressions (N=7, R=77.78)	because	2	22.22
				so	2	22.22
				since	1	11.11
				due to	1	11.11
				as	1	11.11
			Non-causal expressions (N=2, R=22.22)	∅	2	22.22

As shown in Table 4, [+plurality] feature of the causal expression is included only in Korean. In English, there is no expression which indicate the [+plurality] feature at all. In that, the [+plurality] feature of cause is the unique characteristic of Korean causal expressions. The following (a) and (b) in Figure 5 are the examples of Korean sentences containing the causal connective expressions with [+plurality] feature and the corresponding English sentences.

a. Korean: 아직 못 했습니다. 새로운 바이어 분이라 취향도 모르*고 해서* 아직 고민 중입니다.
ɑdʑig motʰɛts*tɯmnidɑ sɛɾoun bɑiʌ bunirɑ tɕʰwijaŋdo moɾɯ*go ɛsʌ*
ɑdʑig gomindʑuŋimnidɑ.
English: Not yet. I don't know really know what they like since it's our first time meeting them.

b. Korean: 어린 아이도 있*고 해서* 집에서 먹으려고요.
ʌrinɑido it*k*o ɛsʌ* dʑibɛsʌ mʌgɯrjʌgojo.
English: I'm going to eat at home *since* I have children.

Fig. 5. Example 3: Korean sentences and corresponding English sentences

In both (a) and (b) in Figure 5, *Ast/Vst-go haeseo* is employed to indicate that the preceding clause represents one of several reasons for the information presented in the subsequent clause in Korean sentences. However, in the corresponding English sentences, this connotative meaning was not expressed. Unlike in English, the [+plurality] feature, which indicates that the presented reason is just one among several possible reasons, frequently occurs in Korean. This allows speakers to avoid definitive expressions by emphasizing that the reason given is one of multiple factors, rather than asserting it as the sole reason. For example, the speaker in (a) is in a situation where they must admit that they have not yet completed a task and must explain why. In this case, they adopt a defensive attitude by employing *Ast/Vst-go haeseo*, which implies that there are various reasons, to avoid explicitly stating the reason that is causing concern. The use of *Ast/Vst-go haeseo* also hints that the reason they have not yet finished is due to the many considerations they must consider. Should they provide a single, clear reason for their indecision, and that reason is countered, they would lose any further opportunity to defend their position. However, by establishing a situation of indecision and implying that there are various reasons, not just one, the speaker ensures they will have additional opportunities to defend themselves, even if the listener objects to the reason they have given.

This can also be interpreted as a strategy to avoid expressing one's intentions too assertively. It is often utilized in contexts where the speaker seeks to refrain from stating a definitive reason. For instance, when declining someone's request or suggestion, a speaker may opt to provide nuanced explanations to prevent the other person from losing face or may choose not to reveal their true reasons for rejecting the request. In (b), for example, the speaker faces the challenge of declining an invitation to dine outside with a large group. Instead of directly expressing a desire to refuse the listener's proposal, the speaker aims to minimize any potential damage to the proposer's face. They achieve this by im-

plicitly conveying, using *Ast/Vst-go haeseo*, that there are multiple reasons prompting their decline, rather than simply a personal intention to reject the offer.

While the current analysis does not directly implement the computational task defined in Section 1.3, it demonstrates its theoretical feasibility through manual, feature-based evaluation. The findings serve as preliminary evidence and a foundational dataset for developing future automated classifiers or scoring systems that assess pragmatic preservation in translation.

5. Conclusion

5.1. Summary of Findings

This study is an interdisciplinary investigation that utilizes the AI-Hub Korean-English parallel corpus to contrastively analyze causal connective expressions in Korean and English. The primary objective was to identify the distinctive characteristics of Korean causal connectives, particularly those that exhibit additional pragmatic features such as [+negative], [+uncertainty], and [+plurality]. These features were analyzed both linguistically and culturally to determine their uniqueness to the Korean language and their prevalence in Korean discourse.

Through the analysis of seven specific Korean causal connectives that carry these features, the study revealed that such expressions are largely absent in English. The additional pragmatic layers found in Korean are closely tied to cultural tendencies toward indirectness, politeness, and the mitigation of speaker stance. For instance, Korean speakers often employ causal connectives with [+negative] connotations to express disapproval or refusal indirectly, a strategy aligned with cultural norms that emphasize maintaining social harmony and avoiding face-threatening acts. Similarly, connectives with [+uncertainty] and [+plurality] allow speakers to hedge their intentions, introduce speculative reasoning, or diffuse responsibility, all of which align with Korean cultural values surrounding modesty and relational sensitivity.

In contrast, English tends to favor direct and explicit expressions of causality, reflecting cultural norms that value clarity, individual agency, and unambiguous communication. English causal connectives thus typically lack the additional semantic and pragmatic layers found in their Korean counterparts.

By comparing how causality is linguistically and culturally encoded in Korean and English, this study offers meaningful insights into the intersection of language, culture, and communication. The findings not only contribute to contrastive linguistics but also have practical implications for language education, particularly in the field of Korean as a foreign language. Recognizing these culturally embedded linguistic patterns can support more effective cross-cultural communication and foster greater intercultural understanding, which are essential for promoting social cohesion in a globalized society.

5.2. Ethics and Limitations

This study also acknowledges several ethical and methodological considerations. The data used for analysis come from the publicly available Korean-English parallel corpus developed by *AI-Hub*. Due to licensing constraints set by the National Information Society

Agency (NIA), the corpus itself cannot be openly redistributed. However, qualified researchers may apply for access through the *AI-Hub* platform, and, upon request, we can provide preprocessing scripts, annotation guidelines, and representative samples that comply with licensing terms for academic use.

From an ethical standpoint, we recognize the potential for cultural bias embedded in the corpus. Causal connectives often reflect nuanced pragmatic intentions that are deeply influenced by cultural norms. Therefore, any computational or rule-based interpretation of pragmatic equivalence across languages must be approached with caution. The indirectness and politeness strategies observed in Korean causal expressions may not have direct equivalents in English, and modeling such shifts automatically can oversimplify or misrepresent these meanings.

Additionally, there is a risk of overgeneralization when applying pragmatic classification tasks to other language pairs or datasets. Cultural and linguistic variability means that tools trained on one corpus may not perform reliably in other contexts. The pragmatic shift classification task proposed in this study is meant as a theoretical and methodological foundation, not as a comprehensive solution.

In terms of methodological limitations, although we developed detailed annotation guidelines and achieved high inter-annotator agreement (Cohen's $\kappa = 0.87$), pragmatic interpretation is inherently subjective. Ambiguities remain, particularly in context-dependent cases that are difficult to resolve without deeper discourse information. Moreover, while this study defines a computational task and proposes potential formulations for pragmatic shift evaluation, we have not implemented or tested these models empirically within this paper. Future work should include experimental evaluation using classification metrics such as Accuracy or F1-score, as well as error analysis to better understand the impact of cultural and linguistic nuances on model performance.

Despite these limitations, we believe this study contributes to a deeper understanding of pragmatic shifts in translation and provides a meaningful step toward computational modeling of culturally embedded language use.

References

1. Ahn, Heeyoung. *A Study of the Hierarchy of Modern Korean Connective Endings*. Master's dissertation, University of Seoul (2007).
2. Bang, J.-U., Maeng, J.-G., Park, J., Yun, S., and Kim, S.-H. English–Korean speech translation corpus (EnKoST-C): Construction procedure and evaluation results. *ETRI Journal*, 45, 18–27 (2023).
3. Chin, Chongnan. *A Study on the Discourse Grammar of Korean Causative Expressions*. Doctoral dissertation, Hankuk University of Foreign Studies (2005).
4. Choi, Sujeong. *A Study of the Meaning-Based Categorization of Grammar Items for Synonymous Grammar Education of Korean Language*. Doctoral dissertation, Yonsei University (2022).
5. Conklin, Harold. *Hanunoo Agriculture*. Rome: Food and Agriculture Organization of the United Nations (1957).
6. Jeon, Hye-Young. *A Study on the Teaching of Connective Endings for Reason·Cause*. Master's dissertation, Hansung University (2021).
7. Kay, M., and Röscheisen, M. *Text-translation alignment*. Technical Report. Xerox Palo Alto Research Center (1988).

8. Kim, Sun Hye, and Jang, Seung Hye. The Method of Describing Micro-structure in Bilingual Dictionaries Using Parallel Corpus - Focusing on Description of Grammatical Elements. *Language Facts and Perspectives*, 28, 263–292 (2011).
9. Li, Yunhui. *A Study on the Spiral Education of Korean's Connective Ending '-aseo' and '-nikka' for Foreign Students*. Master's dissertation, Sejong University (2011).
10. Li, Wenhua. A Study on the Corresponding Form of the Chinese "Zai" in the Parallel Corpus. *The Journal of Learner-Centered Curriculum and Instruction*, 21(19), 59–73 (2021).
11. O, Seon Yeong, and Keiko, Takiguchi. Study on Korean "eopsi" adverbs and their Japanese equivalent patterns based on examples in the Korean-Japanese parallel corpus. *Korean Language and Literature*, 65, 37–69 (2015).
12. Park, Chan-Jun, and Lim, Heui Seok. A Study on the Performance Improvement of Machine Translation Using Public Korean-English Parallel Corpus. *Journal of Digital Convergence*, 18(6), 271–277 (2020).
13. Park, Chanjun, Shim, Midan, Eo, Sugyeong, Lee, Seolhwa, Seo, Jaehyung, Moon, Hyeonseok, and Lim, Heuiseok. Empirical Analysis of Parallel Corpora and In-Depth Analysis Using LIWC. *Applied Sciences*, 12 (2022).
14. Park, Dae Beom. *Connection Expression Education Research of the Reason Cause for the Korean Language Studying Person*. Master's dissertation, Sangmyung University (2008).
15. Park, Minsin. *Study of Semantic Category-based Korean Grammar Education, Focusing on Causative Expressions*. Doctoral dissertation, Seoul National University (2018).
16. Park, Myongsu. Investigation into English-Korean Parallel Corpus with ParaConc. *The Journal of Translation Studies*, 18(5), 29–57 (2017).
17. Saville-Troike, Muriel. *The Ethnography of Communication: An Introduction*. 3rd edn. Oxford: Blackwell (2003).
18. Seo, Se Jung. A Contrastive Study of the Korean Precedent Ending "-ket-" and its Corresponding Forms in English. *Language Facts and Perspectives*, 22, 193–215 (2008).
19. Shen, Lanji. A Study on the Korean Expressions Corresponding to the Chinese Adverbs '(hai)' - focused on Chinese-Korean parallel corpus. *Language Facts and Perspectives*, 36, 247–277 (2015).
20. Sim, Ji-Young. A Study on the Corresponding Expressions of '' Causative Using Parallel Corpus - Focused on the Teaching of A→B Translation between Chinese and Korean. *The Journal of the Korea Contents Association*, 23(5), 475–483 (2023).
21. Vu, Van-Hai, Nguyen, Quang-Phuoc, Shin, Joon-Choul, and Ock, Cheol-Young. UPC: An Open Word-Sense Annotated Parallel Corpora for Machine Translation Study. *Applied Sciences*, 10 (2020).
22. Wierzbicka, Anna. English Causative Constructions. In N. J. Enfield (Ed.), *Ethnosyntax*. New York: Oxford University Press (2002).
23. Wierzbicka, Anna. *Understanding Cultures through Their Key Words: English, Russian, Polish, German, Japanese*. New York: Oxford University Press (1997).
24. Wilczynski, Tomasz. *A Contrastive Study on the Korean 'ida' Constructions and Their Polish and English Counterparts - Focused on Bible Parallel Corpus*. Doctoral dissertation, Keimyung University (2021).
25. Xue, Wenying. *Study on the Contrast Between Korean-Chinese Honorific Expression Based on Parallel Corpus*. Doctoral dissertation, Keimyung University (2021).
26. Yoo, Hae-jun. Contents Construction Method for Teaching Korean Grammar Based on Language Functions. *The Journal of Language & Literature*, 61, 515–538 (2015).
27. Yu, Xiangxin. A Study on Chinese Corresponding Expressions of '-(eu)leo' Based on Parallel Corpus. *Teaching Korean as a Foreign Language*, 66, 161–192 (2022).

Sujeong Choi is an Assistant Professor of Practice in the Korean Language Program at KDI School. She received her Ph.D. in Korean Linguistics with a specialization in

Korean language education from Yonsei University. Her research interests include Korean pragmatics and AI-integrated Korean language education.

Sin-hye Nam Sin-hye Nam is an Assistant Professor in the Department of Korean Language and Literature at Kyung Hee University. She received her Ph.D. in Korean Linguistics from Yonsei University. Her research interests include corpus linguistics and applied linguistics.

Received: Month DD, 20YY; Accepted: Month DD, 20YY.