Computer Science and Information Systems 22(3):859–879 https://doi.org/10.2298/CSIS240930019Z

Identification and Detection of Illegal Gambling Websites and Analysis of User Behavior

Zhimin Zhang[1], Dezhi Han[1], Songyang Wu[2,⋆], Wenqi Sun[2], and Shuxin Shi[1]

[1] College of Information Engineering, Shanghai Maritime University,
201306 Shanghai, China
zhangzhimin@stu.shmtu.edu.cn
dzhan@shmtu.edu.cn
shishuxin@stu.shmtu.edu.cn
[2] Network Security Center, The Third Research Institute of the Ministry of Public Security,
200031 Shanghai, China
wusongyang@stars.org.cn
sunwenqi@gass.ac.cn

Abstract. Illegal gambling websites use advanced technology to evade regulations, posing cybersecurity challenges. To address this, we propose a machine learning method to identify these sites and analyze user behavior accurately. The method extracts key data from post messages in a real-world network environment, generating word vectors via Word2Vec with TF-IDF, which are then downscaled and feature-extracted using a Stacked Denoising Auto Encoder (SDAE). Next, this paper uses Agglomerative Clustering, improved through a combination of distance caching and heap optimization, to initially cluster post-template websites of the same type by clustering them into the same cluster. Then, multiple algorithms are integrated within each website cluster to cluster users' different operational behaviors into different clusters based on the cosine similarity consensus function voting secondary clustering. Results show improved detection of illegal gambling sites and classification of user activities, offering new insights for combating these sites.

Keywords: Gambling websites, post messages, feature extraction, illegal website identification, cluster analysis.

1. Introduction

At present, the rapid development of gambling websites has brought great harm to society. These websites not only jeopardize the financial security of individuals but also undermine social order and contribute to the spread of criminal activities. Due to their hidden and transnational nature, gambling websites are complicated to regulate and combat. Gambling websites usually utilize advanced network technologies to hide their true intentions through sophisticated encryption and camouflage means, making it difficult for traditional regulatory means to be effective. By attracting users to engage in gambling activities, these websites reap substantial illegal profits and legalize these profits through money laundering and other means. Users who participate in illegal gambling often face significant financial risks and may even lose all their money as a result of indulgence in

⋆ Corresponding author

gambling, bringing severe negative impacts on families and society. In addition, illegal gambling may be closely linked to other criminal activities such as fraud, extortion, and violence, further contributing to social instability[1]. With technological advances, these websites have become increasingly covert and highly adaptable, making it difficult for traditional regulatory measures to cope. Users' operational behavior on these websites not only affects their financial and psychological health but also threatens the security of the entire online environment. In the face of these challenges, there is an urgent need to develop and optimize new identification and detection methods. Traditional detection methods, however, often struggle to cope with the complexity of illegal gambling websites due to the constant changes in technology and strategies. Therefore, it is crucial to effectively combat illegal gambling activities by exploring more accurate and efficient identification and detection methods using advanced data analytics and machine learning techniques to gain insights into users' operational behaviors on illegal gambling websites and reveal their potential commonalities and differences. In this paper, we propose a machine learning-based gambling website identification method that can accuriately identify and detect illegal gambling websites and analyze the different operating behaviors of users in different types of gambling websites.The main research and contributions of this paper are summarized below:

(1)By crawling post messages and extracting critical information in a real-world network environment, critical information such as cookie parameters, request body parameters, request line parameters, or keywords are grouped into two datasets, DataSet1 and DataSet2, through WebName and Host links. The critical information is reduced and features extracted using the Term Frequency-Inverse Document Frequency(TF-IDF) weighted Word2Vec method as well as through an improved Stacked Denoising Auto Encoder (SDAE) for dimensionality reduction and feature extraction to obtain stable features of critical information in post messages.

(2)The DataSet1, i.e., Cookie parameter is initially clustered with the same type of websites using a cohesive Agglomerative Clustering algorithm improved by combining distance caching and heap optimization, and the identification and detection of different types of illegal gambling websites is achieved through the stable features of different types of post message template websites obtained. The experimental results show that the adopted method performs well in terms of accuracy and stability and can successfully realize the identification and detection of illegal gambling websites.

(3)Further clustering of DataSet2, i.e., request body parameters and keywords in the request line, within each website clustering cluster, K-means, DBSCAN, Agglomerative Clustering, OPTICS, and Gaussian Mixture Models clustering algorithms are selected, and by evaluating the performance of each method as well as its adaptability to the present experimental By evaluating the performance of each method and its adaptability to the data of this experiment, the consensus function based on cosine similarity votes for integrated clustering to further classify the different behaviors of the same type of websites.

The remainder of the paper is organized as follows: in Section. 2, the paper reviews current research advances and technical approaches in the field of illegal gambling website detection. Section. 3 describes the overall architecture of this paper and the process of dataset production, including data sources, data cleaning, and feature extraction methods, and describes in detail the clustering methods and algorithms used, specifically including preliminary cohesive clustering of the same type of post message template websites using

distance caching combined with heap optimization for improved Agglomerative Clustering, and multiple algorithms integrated clustering of user behaviors based on the voting of consensus functions.Section. 4 shows the experimental results and analysis to evaluate the performance of the methods proposed for identifying and detecting illegal gambling websites in comparison with different models, focusing on profile coefficients, accuracy, and recall and showing the overall results obtained in this paper.Section. 5 summarizes the main contributions of this paper and proposes future research directions, suggesting further optimization of the model to improve its effectiveness in practical applications.

2. Related work

In recent years, many research results have been published on the identification and detection of illegal gambling websites and their user behavior. These researches mainly focus on network traffic analysis, deep learning, multi-view clustering, text analysis, behavioral pattern detection, etc., which promote developing and applying technologies in this field.

In network traffic analysis, Kong et al. (2020) proposed a hybrid model based on convolutional neural network (CNN) and long short-term memory network (LSTM) for network traffic classification. This method combines CNN's extraction of spatial features and LSTM's processing of time-series data, and especially performs well in encrypted traffic detection, effectively distinguishing between legitimate and illegitimate traffic[2].Mu et al. (2022) further developed this idea by proposing a hybrid intrusion detection model based on CNN-LSTM and attention mechanism, which combines the attention mechanism with an enhanced feature extraction capability, making the model high accuracy and robustness when dealing with complex intrusion detection tasks[3].Alshingiti et al. (2022), on the other hand, developed a phishing website detection system based on CNN, LSTM, and LSTM-CNN, which efficiently categorizes phishing website URLs with an accuracy of more than 99% through deep learning[4].

In terms of multi-view clustering, Alnemari and Alshammari (2023) proposed a feature extraction based phishing domain name classification system using algorithms such as Support Vector Machines (SVMs) and decision trees[5]. The method significantly improves the detection accuracy of phishing domain names by analyzing domain name features such as length and character structure.Chen et al. (2023), on the other hand, explored the application of graph convolutional networks (GCN) and feature fusion techniques in multiview learning, which can efficiently extract multiview features and improve the clustering accuracy[6].Huang et al. (2023) proposed a depth-weighted multiview clustering method based on self-supervised graphical attention networks. weighted multiview clustering method, which especially performs well when dealing with complex data structures[7].

In the field of text analytics, Chen et al. (2020) proposed an automated detection system that combines visual and textual content for identifying pornographic and gambling websites, extracting web page HTML text features through Doc2Vec and combining them with visual content for categorization, which ultimately achieves more than 99% accuracy[8].Wang et al. (2022) proposed a multimodal data fusion framework that improves the accuracy of gambling website detection by combining text features extracted from images and OCR and fusing multiple data using a self-attention mechanism improves the accuracy of gambling website detection[9].Sun et al. (2023) apply domain cer-

tificate information and textual analysis to enhance gambling domain name recognition, achieving improved accuracy in domain classification[10].

In terms of user behavior clustering, Singh et al. (2021) proposed a clustering-based e-commerce webpage recommendation system, which improves the personalization and accuracy of recommendation by analyzing user behavior data[11].Li et al. (2021) combined K-mean clustering and hybrid particle swarm optimization algorithms to segment e-commerce users and optimize the effect of precision marketing[12].Liu (2022) used machine learning to to develop a personalized recommendation system based on user behavioral data, which significantly improves recommendation accuracy and user experience[13].

These research results demonstrate the application of various technical means in identifying gambling websites and their user behaviors, covering a wide range of areas from network traffic analysis to behavioral pattern detection and providing valuable references for improving detection accuracy and effectiveness. However, illegal gambling websites are becoming more and more covert and adaptable, and the technical means continue to evolve. Therefore, exploring new techniques, such as machine learning and data mining, to effectively detect and cluster analyze gambling websites has become a hot and challenging topic in current research. This paper aims to explore the characteristics and laws of illegal gambling websites by effectively clustering the POST messages of illegal gambling websites and analyzing them in depth to provide strong technical support for the regulatory authorities and help them better identify and respond to illegal behaviors.

3. Methodology

3.1. Overall architecture

In this paper, by crawling illegal gambling website post messages in the real-world network environment for data cleaning and extracting critical information, cookie parameters are divided into one group, request body parameters and keywords in the request line are another group, which are compiled into two datasets through WebName and Host links. Respectively, we use the Word2Vec method with TF-IDF weighting and SDAE for feature extraction to obtain stable features of critical information in post messages. Agglomerative Clustering, an improved cohesive hierarchical clustering algorithm combining distance caching and heap optimization, was used to initially cluster the same type of post-message template websites for DataSet1, and further Clustering was performed for DataSet2 within each cluster through OPTICS, Gaussian Mixture Models, Agglomerative Clustering, K-means, and DBSCAN multiple algorithms based on the cosine similarity of the consensus function voting integrated clustering, through the obtained different types of websites and different behaviors of the stability of the characteristics of the different types of illegal gambling websites to achieve different types of illegal gambling websites identification and detection, and further to obtain the different behaviors of the user in different types of illegal gambling websites. The overall architecture of this paper is depicted in Fig. 1.

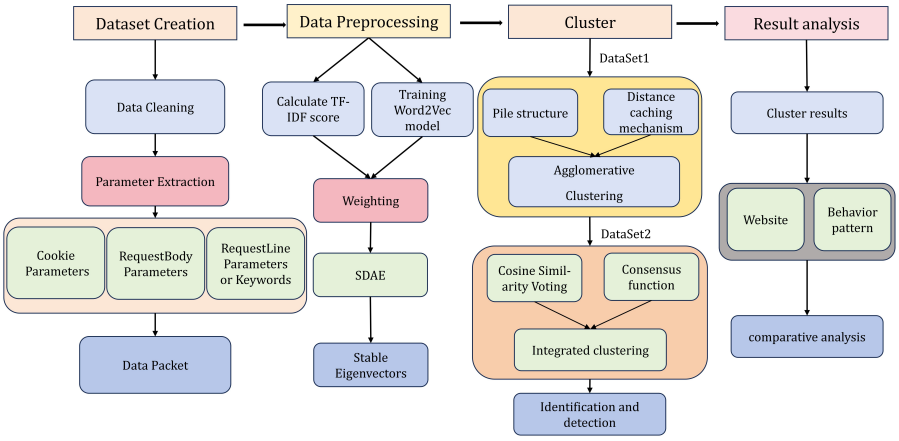


Fig. 1. Overall Architecture Diagram

3.2. Dataset production and data preprocessing

Dataset production Based on the collected website information, this paper simulates the gambling operation behaviors of real users at illegal gambling websites in a real-world network environment. For these gambling websites, the usual operation behaviors include registering, logging in, binding bank cards, recharging, placing bets, consulting customer service, and participating in website activities. In this paper, we collect the POST request messages generated by each behavior and mark the collected data with unique tags for subsequent analysis.POST request message is a method used in the HTTP protocol to send data to the server. Unlike GET requests, POST requests include the data in the request body rather than passing it through a URL. A typical POST request message consists of three parts: the request line (which specifies the method, destination URL, and HTTP protocol version), the request header (which contains information describing the request and the client, such as Host, Content-Length, User-Agent, Content-Type, etc.), and the request body (which contains the data to be sent to the server and is typically used to submitting form data, uploading files, etc.).The POST request message is essential for communication between illegal gambling websites and users. It contains rich request data, such as form information submitted by users, files, JSON data, etc. The gambling user's specific operation and behavioral mode can be identified by analyzing the parameters and data content in the POST request, such as the user's betting records and query records on the illegal gambling website. These data are used for feature extraction. Key features that reflect user behavior and website characteristics are extracted by analyzing the parameters and contents in different requests to provide a basis for subsequent analysis. Fig. 2 shows the original format of a particular post message collected.

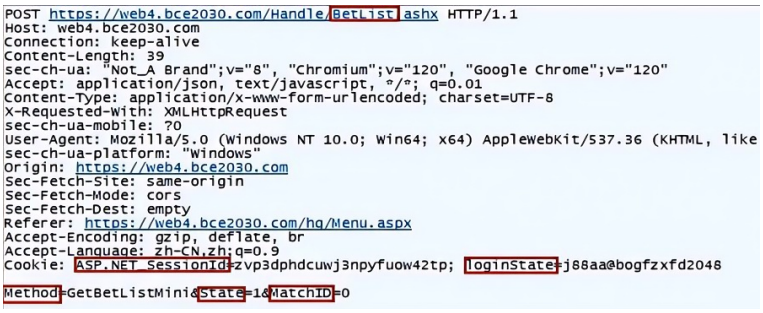


Fig. 2. Post message format and information of each part

In a POST message, the cookie parameter is transmitted in the header of the HTTP message, which may contain the user's login information, identifiers for tracking user activity, session tokens, and so on, which can help identify user behavior patterns or find the mechanisms used by these sites to track users, such as the above figure labeled 'ASP.NET_SessionId'. The request body of a POST request is the data contained in the HTTP request and is usually used to submit a form or upload data. In illegal gambling sites, the parameters in the POST request body may include sensitive data such as the user's betting amount, selected stakes, and payment information. These parameters can reveal how the user interacts with the gambling platform, and analyzing these parameters can help to understand the user's behavior, such as the 'Method', 'State' and 'MatchID' labeled in the figure above. Keywords in the request line usually refer to specific paths in the URL or query string parameters that represent actions performed by the user or application. The keywords in the request line can indicate specific functions accessed by the user, such as logging in, placing bets, withdrawing money, and so on. By analyzing these keywords, it is possible to determine the intent of the user's action and identify the specific interaction patterns of a gambling website, such as the 'BetList' labeled in the figure above.

Among all the collected POST request messages, this paper tags each message with the WebName, Host, and behavior. It extracts critical data, including cookie parameters, parameters in the request body, and representative keywords in the request line. Since in the real-world network environment, the exact behavior of the same website may generate multiple identical POST request messages, in order to ensure the clarity and rationality of the data, these messages are first de-duplicated, and the garbled parameter problem generated by some post request messages is handled specially. According to the characteristics of the post message and the data form characteristics of this dataset, this paper will be data grouping[14], the same site under the cookie parameters merged and de-emphasized data for a group, all the different behaviors of the site parameters in the request body and the request line with a representative of the keyword for another group, the two sets of data according to the WebName and Host links, made into illegal gambling sites The two sets of data are linked according to WebName and Host to create two datasets of illegal gambling websites, DataSet1 and DataSet2, which are convenient for subsequent work.

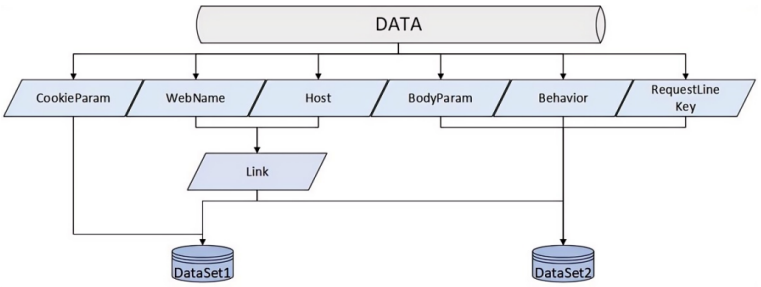


Fig. 3. Schematic representation of each part of the dataset

Word2vec feature extraction with TF-IDF weighting TF-IDF is a weighting technique commonly used in text analysis to measure the importance of a term in a document. The core idea is that the combination of the Term Frequency (TF) of a term in a particular document and the rarity of its occurrence in all documents (Inverse Document Frequency (IDF)) can effectively differentiate representative words from common irrelevant words.Word2Vec is a word embedding model that maps words to a Word2Vec, which maps words into a high-dimensional vector space. It can learn the semantic similarity between words through training.Word2Vec learns the vector representations of words in two ways: CBOW (Continuous Bag of Words) and the Skip-gram model. In this paper, we mainly use the Skip-gram model, and the goal of the training is to capture the co-occurrence probability of words between the target word and its context by maximizing the co-occurrence probability of words. The training objective is to capture the semantic relationship between words by maximizing the co-occurrence probability between the target word and its context words. In order to obtain the weighted word vectors, the Word2Vec vectors of each word are weighted and averaged according to their TF-IDF values[15],as shown in Eqs. 1, 2, 3.

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)\,. \tag{1}$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c<j<c, j\neq 0} \log p(w_{t+j} \mid w_t)\,. \tag{2}$$

$$v_d = \sum_{t\in d} \text{TF-IDF}(t,d)\cdot v_t\,. \tag{3}$$

Eqs. 1 represents the TF-IDF value of the word in the document, $\text{TF}(t,d)$ is the weight of the term t in the document d, and $\text{IDF}(t)$ is the inverse document frequency of the term t. The co-occurrence probability of the target word in its context is maximized by Eqs. 2, where w_t is the target word, w_{t+j} is the target word's contextual word, and c is the size of the contextual window; and Eqs. 3 represents the final word vector of the word, v_t is the Word2Vec vector of the word t, and v_d is the TF-IDF-weighted Word2Vec word vector of the document.

In addition, to eliminate the effects between different feature magnitudes and transform the data to the same scale, the two parts of the data are normalized. The normal-

ization step ensures that the features are within the same magnitude range, improving the model's performance and the comparability of the results. Through the above method, statistical and semantic features can be comprehensively utilized to extract feature vectors with high differentiation and representativeness so that data analysis and model training can be carried out more effectively[16].

Obtaining stable features by SDAE dimensionality reduction In this paper, refer to Xin et al. to add the attention mechanism to the model[17], and weight the word vectors obtained from the Word2vec model with TF-IDF through the SDAE model, so as to further obtain the stable features of the above data.The SDAE model is initialized with specified input dimensions and his. Eden layer dimensions and the Leaky ReLU activation function are used. The SDAE model is trained by performing forward propagation, loss computation, backpropagation, and weight updating through the training function. The tensor is moved from the GPU to the CPU and converted to a Numpy array to obtain more stable and representative feature vectors[18]. The SDAE model self-encoder consists of two parts: encoder and decoder, and the specific implementation process is shown in Eqs. 4, 5, 6.

$$h^{(k)} = f(W_e^{(k)} h^{(k-1)} + b_e^{(k)}) \,. \tag{4}$$

$$z^{(k)} = g(W_d^{(k)} h^{(k)} + b_d^{(k)}) \,. \tag{5}$$

$$L(x, z) = \frac{1}{n} \sum_{i=1}^{n} (x_i - z_i)^2 \,. \tag{6}$$

Eqs. 4 computes the implied representation h of the input data compressed by the encoder, W_e is the weight matrix of the encoder, b_e is the bias vector of the encoder, and f is the Leaky ReLU activation function; Eqs. 5 computes the reconstructed z of the decoder that returns the implied representation h back to the original input, W_d is the weight matrix of the decoder, b_d is the bias vector of the decoder, and g is the activation function. The SDAE measures the difference between the reconstructed output z and the original input x via the loss function $L(x, z)$ in Eqs. 6.

SDAE stacks multiple denoising self-encoders layer by layer to build a deep network. After the layer-by-layer pre-training, all layers are connected and fine-tuned using labeled data to optimize the whole network[19]. In this paper, through the SDAE model, the noise in the dataset is effectively removed to obtain a purer feature representation. The complex features of the data are captured through a multilayer nonlinear transformation, and both parts of the data are downscaled to one-tenth of the original, respectively, while retaining the essential features to reduce the computational overhead.

3.3. Clustering methods

Website clustering methods Agglomerative Clustering is a bottom-up clustering method. The method starts by treating each data point as a separate cluster and then gradually

merges the most similar clusters until all data points have been merged into a single cluster or a predetermined number of clusters has been reached. The traditional cohesive hierarchical clustering algorithm needs to recalculate the distance between clusters each time the clusters are merged, and this process has a significant computational overhead, which makes the algorithm inefficient, especially when the dataset size is large.Therefore, several improvements to the algorithm are proposed, as illustrated in Fig. 4.

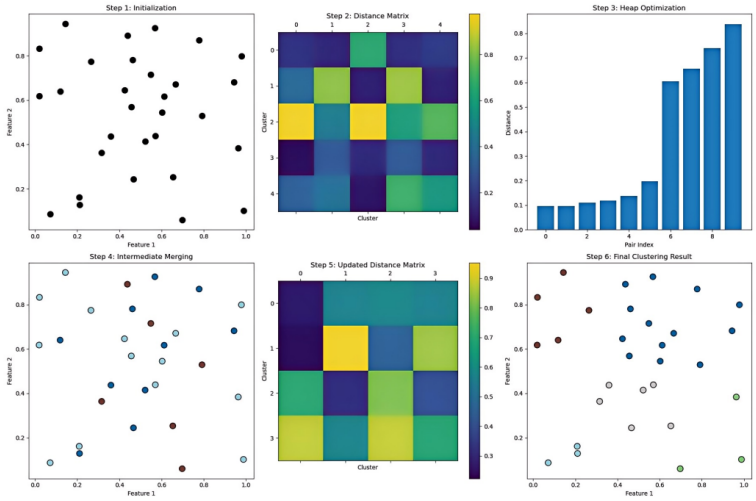


Fig. 4. Implementation of Improved Agglomerative Clustering Algorithm Combining Distance Caching and Heap Optimization

In order to improve the efficiency of the algorithm, this paper introduces a heap structure to store the distances between clusters based on traditional Agglomerative Clustering. Heap structure is an efficient data structure, and this paper uses the minimal heap to store the distance between clusters. During each cluster merging process, the heap structure can quickly retrieve the cluster pair with the smallest distance with a complexity of O(1) and only needs to update the distance associated with the merged clusters after the merger, thus avoiding recalculating the distances between all clusters. The complexity of updating the heap is (logN) (where N is the number of clusters), which greatly reduces the computation time. The introduction of the heap structure makes the merging clusters more efficient each time. In addition, a distance caching mechanism is introduced in this paper to reduce further the repetitive distance computation in the cluster merging process. In traditional methods, the distances between the new cluster and other clusters need to be recalculated after each cluster merge, and these distances may be calculated multiple times in subsequent iterations. To avoid this problem, this paper uses a distance caching mechanism to store the computed inter-cluster distances in a cache. Each time the distance needs to be calculated, it first checks whether the corresponding distance value already exists in the cache; if it does, the cached value is used directly; if not, it is calculated and stored. Through this mechanism, the overhead of repeated computation is greatly reduced, espe-

cially when dealing with large-scale datasets, which can significantly improve the overall efficiency of the algorithm.

This paper combines a further distance caching mechanism with heap optimization; firstly, the distances between all cluster pairs are computed in the initialization phase and stored in the distance cache. After each cluster merger, the new cluster pairs' distances are updated. Then, a minimal heap is used to store the initial clusters and their distance values, and the cluster pairs with the smallest distances are extracted from the heap each time a merge occurs; the distance cache is updated, and the distance values of the new clusters are reinserted into the heap. This approach reduces repeated computations and accelerates the cluster pair-finding process through the heap's prioritization mechanism. The algorithm complexity is reduced from $O(n^3)$ of the original cohesive clustering to $O(n^2 \log n)$, which improves the efficiency of the cohesive hierarchical clustering algorithm, makes it more suitable for large-scale datasets, and enhances the speed of the clustering process and resource[20,21].

Through the improved cohesive hierarchical clustering algorithm combining distance caching and heap optimization, the DataSet1 data, i.e., the cookie parameters of all websites, are clustered to cluster websites of the same type into the same cluster and to design a cluster label for each cluster, i.e., WebsetClusterLabel.The pseudocode is shown in Algorithm 1.

Algorithm 1 Agglomerative Clustering with Distance-Based Heap

Require: Dataset1, num_clusters
Ensure: WebsetClusterLabel
Normalize vectors using `StandardScaler` to obtain `vectors_scaled`.
Compute `dist_matrix` for all points using the pairwise Euclidean distance.
for each pair of points (i, j) in `vectors_scaled` **do**
  Push $(distance, i, j)$ to heap.
end for
Define a cache `dist_cache` to store computed distances.
while the number of clusters is reduced to `num_clusters` **do**
  Merge cluster j into cluster i and delete cluster j from clusters
  Update `dist_cache` and heap with new distances:
  for each remaining cluster k **do**
    if distance between (i, k) is not in `dist_cache` **then**
      Compute the minimum distance between points in clusters i and k
      Update `dist_cache` with the computed distance
    end if
    Push $(distance, i, k)$ to heap
  end for
end while
for each cluster i in clusters **do**
  Assign label i to all points in the cluster.
end for

Behavioral clustering methods After completing website clustering on DataSet1, based on the same data in DataSet1 and DataSet2, i.e., WebName and Host, the WebsetClusterLabel, i.e., the same type of website obtained above, is mapped to DataSet2 through the WebName and Host's Uniqueness is mapped to DataSet2 and clustered again in the same type of website through WebsetClusterLabel identity.

Due to different URLs and behavioral data within each cluster, the request body parameters and keyword data in the request line may exhibit varying distribution characteristics and patterns, often resulting in a disorganized state. This paper performs a finer cluster analysis to refine the data in these clusters[22]. This paper tries to design multiple clustering algorithms after many experiments to ensure the accurate categorization of the data. K-means, DBSCAN, Agglomerative Clustering, OPTICS, and Gaussian Mixture Models (GMM) clustering algorithms are chosen through careful data analysis.The advantages and disadvantages of several clustering algorithms are shown in Table. 1.

Table 1. Advantages and disadvantages of different clustering algorithms

Algorithm	Advantages	Disadvantages
K-means	1. Simple and easy to understand. 2. Efficient for large datasets. 3. Fast computation.	1. Requires specifying the number of clusters (K) in advance. 2. Sensitive to outliers. 3. Assumes spherical clusters.
DBSCAN	1. No need to specify the number of clusters. 2. Handles noise and outliers. 3. Can find clusters of arbitrary shapes.	1. Sensitive to parameters (distance threshold and min points). 2. Struggles with high-dimensional data.
Agglomerative Clustering	1. No need to specify the number of clusters. 2. Can capture hierarchical relationships.	1. High computational complexity. 2. Sensitive to noise and outliers.
OPTICS	1. Handles clusters of varying densities. 2. No need to specify the number of clusters. 3. Reveals clustering structure in the data.	1. High computational complexity. 2. Sensitive to parameter settings.
GMM	1. Capable of modeling clusters with different shapes. 2. Provides soft clustering (probability of membership). 3. Offers a well-interpretable model.	1. Requires specifying the number of clusters. 2. Sensitive to initial conditions. 3. Computationally intensive.

K-means is a center-of-mass-based clustering method suitable for dealing with spherical distributions through iterative optimization to classify the data points into the nearest clusters. K-means is a center-of-mass-based clustering method that divides data points into the nearest clusters through iterative optimization. It is suitable for dealing with globally distributed data, with the advantages of fast computation speed and simple implementation. It realizes clustering by defining core, boundary, and noise points and is suitable for

dealing with non-uniformly distributed data. Agglomerative Clustering gradually merges the most similar clusters by constructing a hierarchical tree that is able to deal with clusters of different shapes and sizes and is suitable for scenarios requiring hierarchical structural analysis.OPTICS is similar to DBSCAN but is able to deal with clusters of different densities better by generating clusters of different shapes. Clusters of different densities reveal the clustering structure in the data by generating ordered reachable graphs.GMM clusters the data again using the Expectation Maximization (EM) algorithm, which is suitable for dealing with data characterized by Gaussian distribution.

In this paper, based on the clustering results generated within each website clustering cluster, the cosine similarity $\mathrm{cosine\_similarity}(i,j)$ between different clusters is calculated, and the cosine similarity between every two clusters is calculated to construct the cosine similarity matrix $\mathrm{similarity\_matrix}[i,j]$ [23]. Based on the inter-cluster similarity matrix $\mathrm{similarity\_matrix}[i,j]$, a "weighted" vote is assigned for each data point according to a threshold value by Eqs. 7, and if the data points are assigned to similar clusters (i.e., the similarity is above a threshold value of 0.5) in both clustering algorithms, a weighted vote is assigned according to Eqs. 8 to get a final voting score $\mathrm{vote}[j]$.

$$\mathrm{cosine\_similarity}(i,j) = \frac{i \cdot j}{\|i\|\|j\|}\,. \tag{7}$$

$$\mathrm{vote}[j] += \mathrm{similarity\_matrix}[i,j] \quad \text{if} \quad \mathrm{similarity\_matrix}[i,j] > 0.5\,. \tag{8}$$

Algorithm 2 Clustering with Multiple Methods and Voting Consensus

Require: Dataset2, WebsetClusterLabel
Ensure: BehaviorLabel
1: Extract ClusterIndices of points with the same WebsetLabel
2: Normalize the data points in ClusterIndices to obtain `CombinedScaled`
3: **for** each MethodName, MethodClass, ParamGrid in ClusteringMethods **do**
4: **for** each parameter set `params` in `ParameterGrid`(PparamGrid) **do**
5: Configure `method` using `params`
6: Apply `method` to `CombinedScaled` to obtain `labels`.
7: **end for**
8: **end for**
9: Initialize `SimilarityMatrix` with shape (`NumPoints`, `NumPoints`)
10: **for** each (MethodName, labels) in `ClusterLabelsList` **do**
11: Compute pairwise cosine similarity.
12: Update `SimilarityMatrix` with similarity values
13: **end for**
14: **for** each point `i` in `CombinedScaled` **do**
15: Calculate votes based on similarities in `SimilarityMatrix`
16: Obtain the final `BehaviorLabel`
17: **end for**

In this paper, according to the function and advantages and disadvantages of each clustering method, concerning the dataset used in the experiments of this paper, the algorithm with optimal performance and its parameter combinations are selected, and through the several clustering methods mentioned above, each cluster of clustering of websites is in-

tegrated through the consensus function based on the cosine similarity and the voting[24], and finally clusters out the different behaviors of users in different types of websites. The pseudocode is shown in Algorithm 2.

By integrating the clustering method through the consensus function based on similarity voting[25], based on the mapping WebsetClusterLabel of the results clustered from the DataSet and then set a small label for each behavior, i.e., BehaviorLabel, to cluster the same operation behaviors of the user in the same type of website, and finally, based on the obtained results WebsetClusterLabel and BehaviorLabel through the mapping of a certain type of user behavior under a certain type of website, so that the results are more precise, but also to reduce the data due to the different distribution characteristics and patterns of the use of a single clustering method of the results of the randomness brought about by the impact of the results of the final clustering results to improve the stability of the final clustering results.

4. Experiments and results

4.1. Experimental Configuration

The configuration of the experimental environment in this paper is shown in Table. 2.

Table 2. Experimental environment configuration

Name	Configuration Information
Development Language	Python 3.11
Framework	Pytorch 2.1.0 + cuda 12.3
GPU	NVIDIA GeForce RTX 4060 Laptop
CPU	13th Gen Intel(R) Core(TM) i7-13700H 2.40 GHz
Memory	16.0 GB

4.2. Evaluation indexes of experimental effect

This paper evaluates the effectiveness of the experimental clustering algorithm through the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Eqs. 9 calculates the Silhouette Score (S), which is used to interpret and verify the consistency of the data points in the clusters, where a is the average distance of the sample from the other points in the same cluster, and b is the average distance of the sample from all the points in the nearest cluster. The contour coefficient combines the two aspects of Cohesion within clusters and Separation between clusters. The contour coefficient of each sample is calculated and averaged to evaluate the effect of the whole cluster.Eqs. 10measures the quality of clustering by calculating the similarity between individual clusters and the dispersion of data points within clusters, i.e., Davies-Bouldin Index (DBI), where N is the number of clusters, σ_i, σ_j are the average distances of all the points in clusters i and j from the centers of their respective clusters, and d_{ij} is the distance between the centers of clusters in clusters i and j. The smaller the DBI is, the better the effect of the clustering is

indicated.Eqs. 11 measures the quality of clustering by calculating the ratio of the sum of squares of the inter-cluster and intra-cluster dissociations, i.e., the Calinski-Harabasz Index (CHI), also known as the Variance Ratio Criterion (VRC), where k is the number of clusters, $\mathrm{Tr}(B_k)$ is the trace of the inter-cluster discretization matrix, and $\mathrm{Tr}(W_k)$ is the trace of the intra-cluster discretization matrix. n is the total number of samples. A larger CHI indicates better clustering[26].

$$S = \frac{b-a}{\max(a,b)}\,. \tag{9}$$

$$DBI = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\left(\frac{\sigma_i+\sigma_j}{d_{ij}}\right)\,. \tag{10}$$

$$CHI = \frac{\mathrm{Tr}(B_k)/(k-1)}{\mathrm{Tr}(W_k)/(n-k)}\,. \tag{11}$$

In this paper, the performance of the experimental model is evaluated by Accuracy and Recall. Accuracy is used to evaluate the correctness of the clustering results, and Recall is a measure of the model's ability to recognize samples of the positive class, that is, among all the samples that are actually positive, the proportion of samples correctly recognized as positive by the model, specifically, as in Eqs. 12 and Eqs. 13. TP represents the positive sample predicted to be positive by the model, which can be called the accuracy rate judged to be true. TN represents the negative sample predicted to be negative by the model, which can be referred to as the percentage of correct judgments that are false. FP represents the negative sample predicted by the model to be positive, which can be referred to as the false alarm rate. FN represents the positive sample predicted to be negative by the model, which can be referred to as the underreporting rate.

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\,. \tag{12}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\,. \tag{13}$$

4.3. Analysis of results

The overall results obtained from the experiment are shown in Fig. 5. For the improved Agglomerative Clustering clustering method using distance caching in combination with heap optimization in DataSet1, based on the difference in clustering performance and effectiveness depending on the number of clusters, several experiments were conducted to compare the S obtained by the number of clusters clustered out of the data, the DBI and the CHI to select the clustering result with the most appropriate separation for each cluster with the best effect. After clustering the cookie parameters in DataSet1, roughly the same type of post-message templates from illegal gambling sites are clustered in the same cluster. According to the analysis of the obtained experimental results, the current illegal gambling websites are roughly board game websites, virtual lottery and scratch-off websites, sports betting websites, social gambling websites, and virtual sports betting websites. Each type can be further categorized into multiple post-message templates.

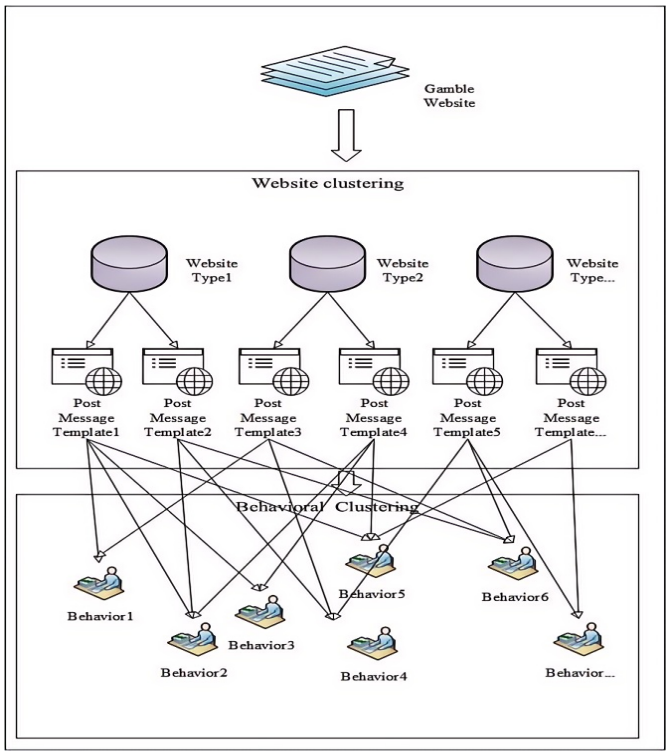
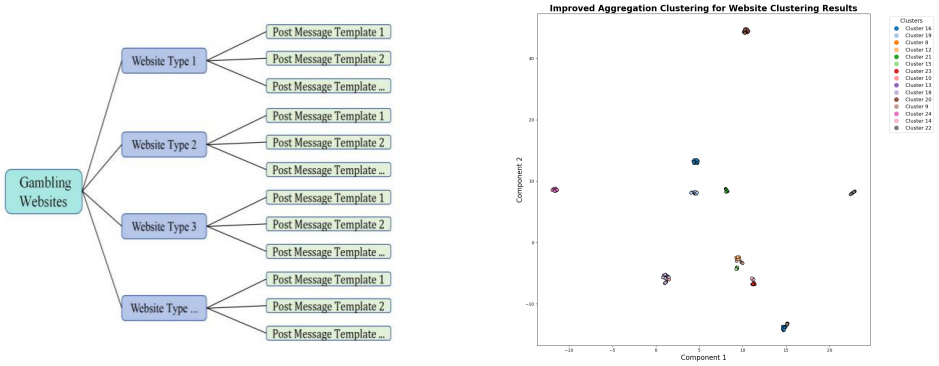


Fig. 5. Overall results of the experiment



(a) Website Message Clustering **(b)** Agglomerative Clustering Result

Fig. 6. Improved Agglomerative Clustering Result Classification and Scatter Plot

Through the above methods, this paper clusters, identifies, and detects illegal gambling websites and successfully detects illegal gambling websites of the same type. Current research commonly uses single clustering methods such as DBSCAN, K-means, and GMM to cluster and analyze the parameters obtained from POST messages. In contrast to the direct use of the above methods for clustering and detecting the parameters in the collected POST messages, in this paper, we use an improved cohesive hierarchical clustering algorithm that combines distance caching and heap optimization for clustering. This method not only improves the clustering of different types of websites, but also significantly improves the accuracy and recall of recognition and detection, shows higher effectiveness, and excels in several performance metrics.The following are the comparison results of several methods with the method mentioned in this paper for clustering detection, and their performance on the three metrics of S, Accuracy, and Recall, respectively, are shown in Table. 3.

Table 3. Comparison results of clustering detection

METHOD	SILHOUETTE SCORE	ACCURACY	RECALL
K-means	0.42	0.96368	0.96896
DBSCAN	0.48	0.95185	0.97533
GMM	0.52	0.97815	0.98907
Ours	0.57	0.98216	0.98942

From the comparison results in Table. 3, the model in this paper can significantly improve the clustering effect by further optimizing and adjusting the clustering model and method by comparing with the previous methods, showing better cluster separation and tightness. The classification and recognition effects are improved. According to the comparison results, the clustering analysis through phased and multi-algorithm can be more effective in identifying and detecting illegal gambling websites.

Since the two data sets of DataSet1 and DataSet2 are linked according to WebName and Host during the previous processing of the data set, the data are clustered in the same cluster under the same WebName and Host after the above clustering. According to DataSet1 and DataSet2 public data WebName and Host, the data of DataSet2 through the above clustering generated WebsetClusterLabel through the WebName and Host grouping, so that the same WebsetClusterLabel, that is, the same type of post message template illegal gambling WebsetClusterLabel, i.e., the same type of post message template illegal gambling website is grouped within the same group. Then, for each group of data, the same type of post message template illegal gambling websites are integrated and clustered according to the consensus function based on cosine similarity of multiple algorithms of K-means, DBSCAN, Agglomerative Clustering, OPTICS, and GMM clustering.

According to Fig.7, the percentage of Silhouette Score, DBI, and CHI effects for integrated clustering within clusters of the same type of post message templates on illegal gambling websites shows that the integrated clustering through multiple algorithms based on the cosine similarity of the consensus function is more effective to further categorize

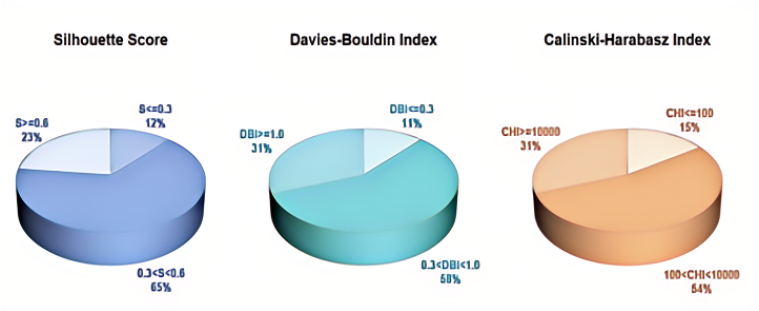


Fig. 7. Integration clustering effect percentage

the same post message templates of the same type of post message templates of illegal gambling websites with different behavioral approaches. In the real-world network environment, with different types of websites, users can realize different operations and behaviors. In this paper, we first cluster out the same type of post message template illegal gambling websites, and then cluster the different types of user operation behaviors together in the cluster through the integrated clustering of consensus functions based on cosine similarity of multiple algorithms and finally obtain the results of a certain type of website under a certain type of user operation of a certain behavior.

Based on the analysis of the experimental results in this paper, users of sports betting websites often place real-time bets while events are in progress. They analyze them based on past performance to make betting decisions, possibly exchanging betting advice and strategies through social interactions. Users of board game sites participate in various poker games and tournaments and may use statistical tools to analyze their performance. On lottery and instant game sites, users typically make random bets or select specific combinations of numbers; their participation is usually less frequent, the betting process is more straightforward, and users frequently check lottery results and manage winning information. Users interact with friends in games on social gambling sites, share game progress and strategies, and usually bet using virtual currency rather than real money. In contrast, users of virtual sports betting sites bet on simulated sporting events, such as virtual soccer or horse racing. They make quick bets and track the progress of the virtual event in real-time.

5. Conclusion and Outlook

In recent years, with the rapid development of the Internet, the rampant illegal gambling websites pose a severe challenge to the socio-economic and legal order of countries. In order to effectively curb online gambling behaviors, this paper examines website and behavioral characteristics through critical information from post messages obtained in a real-world network environment. The experimental results show that the different operational behaviors of users under different types of websites are obtained by first clustering all gambling websites using an improved cohesive hierarchical clustering algorithm combining distance caching and heap optimization to obtain the same type of post message template websites, and then integrating clustering of the same type of websites by

similarity-based consensus function. This multilevel clustering method shows high computational efficiency and classification accuracy when dealing with large-scale data and excels in indicators such as Silhouette Score, accuracy, and recall. It can effectively identify and detect gambling websites and user operating behaviors. By comparing various clustering methods and parameter combinations, this multilevel clustering method not only improves the accuracy of clustering but also enhances the robustness of the model to better adapt to changes in data. The approach in this paper can help identify and monitor illegal activity patterns, optimize risk assessment and alert systems, improve anti-fraud strategies, and support compliance and legal enforcement. In this way, standard features of the website and user behaviors can be revealed, and the efficiency of data analysis can be improved, thus effectively combating illegal gambling activities, protecting users' rights and interests, and maintaining the security of the online environment.

Although this study has made some progress in identifying and detecting illegal gambling websites, it still faces many challenges. First, it is not easy to acquire and label the data. Due to gambling websites' hidden nature and dynamic changes, obtaining high-quality datasets becomes a significant challenge. Second, regarding feature extraction and selection, effectively extracting and selecting features to improve the clustering effect still needs to be continuously optimized. Finally, the existing clustering algorithms have efficiency problems when dealing with large-scale and high-dimensional data, which need further optimization to improve performance. Future research should pay more attention to the practicality and efficiency of the technology, Such as blockchain and dynamically searchable encryption based data storage and sharing solutions provide secure data storage and privacy protection [27,28],to cope with the increasingly complex online gambling behavior and ensure the healthy development of cyberspace. Further development of automated data acquisition and cleaning tools, exploration of more advanced feature extraction techniques, and optimization of the computational performance of existing clustering algorithms are recommended to address these challenges better [29,30].

Acknowledgments. This work was Supported by Key Lab of Information Network Security, Ministry of Public Security,the National Natural Science Foundation of China under Grants 61672338, the Natural Science Foundation of Shanghai under Grant 21ZR1426500. .

References

1. Ghelfi, M., Scattola, P., Giudici, G., Velasco, V.: Online gambling: A systematic review of risk and protective factors in the adult population. In: Proceedings of the Journal of Gambling Studies. vol. 39, pp. 1–27 (2023)
2. Kong, X., Wang, C., Li, Y., Hou, J., Jiang, T., Liu, Z.: Traffic classification based on cnn-lstm hybrid network. In: International Forum on Digital TV and Wireless Multimedia Communications. pp. 401–411. Springer Singapore, Singapore (2021)
3. Mu, J., He, H., Li, L., Pang, S., Liu, C.: A hybrid network intrusion detection model based on cnn-lstm and attention mechanism. In: International Conference on Frontiers in Cyber Security. pp. 214–229. Springer Singapore, Singapore (2021)
4. Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q.E.U., Saleem, K., Faheem, M.H.: A deep learning-based phishing detection system using cnn, lstm, and lstm-cnn. Electronics 12(1), 232 (2023)
5. Alnemari, S., Alshammari, M.: Detecting phishing domains using machine learning. Applied Sciences 13(8), 4649 (2023)

6. Chen, Z., Fu, L., Yao, J., Guo, W., Plant, C., Wang, S.: Learnable graph convolutional network and feature fusion for multi-view learning. Information Fusion 95, 109–119 (2023)
7. Huang, Z., Ren, Y., Pu, X., Huang, S., Xu, Z., He, L.: Self-supervised graph attention networks for deep weighted multi-view clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 7936–7943 (2023)
8. Chen, Y., Zheng, R., Zhou, A., Liao, S., Liu, L.: Automatic detection of pornographic and gambling websites based on visual and textual content using a decision mechanism. Sensors 20(14), 3989 (2020)
9. Wang, C., Zhang, M., Shi, F., Xue, P., Li, Y.: A hybrid multimodal data fusion-based method for identifying gambling websites. Electronics 11(16), 2489 (2022)
10. Sun, G., Ye, F., Chai, T., Zhang, Z., Tong, X., Prasad, S.: Gambling domain name recognition via certificate and textual analysis. The Computer Journal 66(8), 1829–1839 (2023)
11. Singh, H., Kaur, P.: An effective clustering-based web page recommendation framework for e-commerce websites. SN Computer Science 2(4), 339 (2021)
12. Li, Y., Chu, X., Tian, D., Feng, J., Mu, W.: Customer segmentation using k-means clustering and the adaptive particle swarm optimization algorithm. Applied Soft Computing 113, 107924 (2021)
13. Liu, L.: e-commerce personalized recommendation based on machine learning technology. Mobile Information Systems 2022(1), 1761579 (2022)
14. Qiao, M., Wei, L., Han, D., et al.: Efficient multi-party psi and its application in port management. Computer Standards & Interfaces 91, 103884 (2025)
15. Jiang, T., Jia, L., Wan, C.M., et al.: The text modeling method of tibetan text combining word2vec and improved tf-idf. In: Proceedings of 2020 4th International Conference on Electrical, Mechanical and Computer Engineering (ICEMCE 2020). vol. 3, p. 8. IOP Publishing (2020)
16. Zhang, T., Wang, L.: Research on text classification method based on word2vec and improved tf-idf. In: Advances in Intelligent Systems and Interactive Applications: Proceedings of the 4th International Conference on Intelligent, Interactive Systems and Applications (IISA2019). pp. 199–205. Springer International Publishing (2020)
17. Xin, X., Han, D., Cui, M.: Daaps: A deformable-attention-based anchor-free person search model. Computers, Materials & Continua 77(2) (2023)
18. Ni, Q., Fan, Z., Zhang, L., Nugent, C.D., Cleland, I., Zhang, Y., Zhou, N.: Leveraging wearable sensors for human daily activity recognition with stacked denoising autoencoders. Sensors 20, 5114 (2020)
19. Fernández-García, M.E., Sancho-Gómez, J.L., Ros-Ros, A., Figueiras-Vidal, A.R.: Complete stacked denoising auto-encoders for regression. Neural Processing Letters 53, 787–797 (2021)
20. Shkaberina, G., Verenev, L., Tovbis, E., Rezova, N., Kazakovtsev, L.: Clustering algorithm with a greedy agglomerative heuristic and special distance measures. Algorithms 15, 191 (2022)
21. Ali, M.A., PP, F.R., Abd Elminaam, D.S.: An efficient heap based optimizer algorithm for feature selection. Mathematics 10, 2396 (2022)
22. Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I., Akinyelu, A.A.: A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence 110, 104743 (2022)
23. Iffath, N., Mummadi, U.K., Taranum, F., Ahmad, S.S., Khan, I., Shravani, D.: Phishing website detection using ensemble learning models. In: AIP Conference Proceedings. vol. 3007, p. 1. AIP Publishing (2024)
24. Chen, G.: Scalable spectral clustering with cosine similarity. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 314–319. IEEE (2018)
25. Pho, K.H., Akbarzadeh, H., Parvin, H., et al.: A multi-level consensus function clustering ensemble. Soft Computing 25, 13147–13165 (2021)

26. Alizade, M., Kheni, R., Price, S., Sousa, B.C., Cote, D.L., Neamtu, R.: A comparative study of clustering methods for nanoindentation mapping data. Integrating Materials and Manufacturing Innovation 13, 526–540 (2024)
27. Li, J., Han, D., Wu, Z., et al.: A novel system for medical equipment supply chain traceability based on alliance chain and attribute and role access control. Future Generation Computer Systems 142, 195–211 (2023)
28. Li, J., Han, D., Weng, T.H., et al.: A secure data storage and sharing scheme for port supply chain based on blockchain and dynamic searchable encryption. Computer Standards & Interfaces 91, 103887 (2025)
29. Han, D., Pan, N., Li, K.C.: A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection. IEEE Transactions on Dependable and Secure Computing 19(1), 316–327 (2022)
30. Han, D., Zhu, Y., Li, D., Liang, W., Souri, A., Li, K.C.: A blockchain-based auditable access control system for private data in service-centric iot environments. IEEE Transactions on Industrial Informatics 18(5), 3530–3540 (2022)

Zhimin Zhang is currently pursuing the M.S.degree with the School of Information Engineering, Shanghai Maritime University, Pudong, China. Her main research topic is network security.

Dezhi Han received the B.S. degree in applied physics from the Hefei University of Technology, Hefei, China, in 1990, and the M.S. and Ph.D. degrees in computing science fromthe Huazhong University of Science and Technology, Wuhan, China, in 2001 and 2005,respectively. He is currently a Professor with the Department of Computer, Shanghai Mar-itime University, Pudong, China, in 2010. His current research interests include cloud and outsourcing security, blockchain, wireless communication security, network, and information security.

Songyang Wu is a researcher and director of the Cyber Security Center of the Third Research Institute of the Ministry of Public Security (MPS) and also serves as the deputy director of the National Engineering Research Center for Network Security Level Protection and Security Technologies and the executive deputy director of the Key Laboratory ofthe Ministry of Public Security of the Ministry of Information Network Security, etc. He received his B.S. degree in Computer Science and Technology from Tongji University in2005 and his Ph.D. in Computer Application from Tongji University in 2011. He joined the Center of the Ministry of Public Security in the same year. He received his PhD in Computer Application from Tongji University in 2011. He joined the Network Security Center of the Third Research Institute of the Ministry of Public Security in the same year.His research interests include cybercrime investigation, electronic data forensics, ampledata security, and artificial intelligence security.

Wenqi Sun, Associate Researcher and Research Engineer at the Cyber Security Center of the Third Research Institute of the Ministry of Public Security; she received her B.S.degree in Computer Science and Technology from Northeastern University in 2010 and her Ph.D. degree in Computer Science and Technology from Tsinghua University in 2016. She joined the Cyber Security Center of the Third Research Institute of the Ministry of Public Security in 2018; her current research interest is in cybercrime investigation.

Shuxin Shi received an M.S. in Computer Science and Technology from Shanghai Maritime University, Pudong, China, in 2024 and is currently pursuing a Ph.D. in the Schoolof Information Engineering. His current research interest is network security.

Received: September 30, 2024; Accepted: January 17, 2025.