Computer Science and Information Systems 23(3):969–1000 https://doi.org/10.2298/CSIS250616028M

Model Parameter-Based Transfer Learning for ESG Score Prediction in Developing Markets

Ivana Marković[1], Adela Ljajić[2], Jelena Z. Stanković[1], Miloš Košprdić[2], and Jovica Stanković[1]

[1] Faculty of Economics, University of Niš
Trg kralja Aleksandra Ujedinitelja 11, 18000 Niš
ivana.markovic@eknfak.ni.ac.rs (corresponding author),
jelenas@eknfak.ni.ac.rs,jovica.stankovic@eknfak.ni.ac.rs
[2] The Institute for Artificial Intelligence Research and Development of Serbia
Fruškogorska 1, 21000 Novi Sad
adela.ljajic@ivi.ac.rs, milos.kosprdic@ivi.ac.rs

Abstract. While ESG (Environmental, Social, and Governance) assessment plays a key role in sustainable finance, data scarcity and noise in emerging economies hinder robust model development. To address this, we propose a model parameter-based transfer learning with random forest (MPBTL-RF) approach for domain adaptation situations where source data are not available. The proposed model is evaluated using three traditional learning approaches: Random Forest (RF), eXtreme Gradient Boosting (XGB), and Feedforward Neural Networks (FNN). Cross-validation is used to assess model generalizability, and domain adaptation is tested through in-domain and out-of-domain settings. The proposed MPBTL-RF approach achieves competitive performance compared to traditional baselines in scenarios with limited training data, offering time advantages with predictive efficiency and stability. This work demonstrates how machine learning pipelines can adapt to data-constrained, real-world domains, fostering the synergy between AI (Artificial Intelligence) and business.

Keywords: Machine Learning, Domain Adaptation, Transfer Learning, Small Datasets, ESG Score, Developing Markets, Corporate Financial Performance.

1. Introduction

Small data samples are challenging for machine learning across various domains, including finance, where databases are frequently limited by numerous factors such as the number of available firms, reporting restrictions, or institutional privacy policies. As noted by Kokol et al. [11], "using machine learning on small-sized datasets presents a problem because, in general, the 'power' of machine learning in recognizing patterns is proportional to the size of the dataset; the smaller the dataset, the less powerful and accurate the machine learning algorithms".

Transfer learning has emerged as a crucial paradigm in machine learning, enabling knowledge learned in a source domain to improve predictive performance in a related target domain under data-scarce conditions [15]. This approach addresses the critical challenge of data scarcity by allowing models to improve performance on target tasks or domains through knowledge transfer from related domains with larger labeled datasets.

Recently, source-free unsupervised domain adaptation, also known as unsupervised model adaptation, has attracted significant attention, enabling effective generalization of pretrained models to target domains without labeled data. This setting is particularly relevant in real-world applications, where organizations often share only trained models rather than raw data due to privacy, security, or scalability constraints [8].

ESG scores, widely used as indicators of corporate sustainability, benefit from advances in artificial intelligence (AI), which plays a crucial role in finance and is an influential factor in ESG investing [12]. The existing work on this topic can be divided into eight categories: Trading and Investment, ESG Disclosure, Measurement and Governance, Firm Governance, Financial Markets and Instruments, Risk Management, Forecasting and Valuation, Data, and Responsible Use of AI [14]. Distinct AI and machine learning techniques are employed across these categories. Among these, the categories of Data and Forecasting include either predicting the ESG score or exploring novel approaches for measuring ESG performance [14].

However, most existing studies focus on developed economies, while ESG score prediction in developing markets remains particularly challenging due to data scarcity, heterogeneous reporting standards, and incomplete coverage. The annual publication cycle of ESG scores complicates the data collection necessary for model training. Furthermore, rigid privacy policies increase challenges in data sharing within economics, especially in the insurance and finance sectors. To the best of our knowledge, no previous studies have addressed ESG score prediction in developing markets.

The research presented here fills this gap by proposing and empirically validating the model parameter-based domain adaptation approach to facilitate transfer learning based on RF (MPBTL-RF). We compare our approach with strong baselines (RF, XGB and FNN) and evaluate its performance in both in-domain and out-of-domain settings, including a real-world case study on the Belgrade Stock Exchange (BELEX). Additionally, in contrast to most research focused on the transfer learning approach within classification tasks [23, 24, 10], this study explores regression problems.

The contributions of this paper are threefold:

- We design a computationally efficient and practical algorithm and formalize a model parameter-based transfer learning pipeline for RF (MPBTL-RF).
- We conduct a comprehensive experimental evaluation comparing MPBTL-RF against state-of-the-art baselines (RF, XGB, FNN), analyzing both in-domain and out-of-domain scenarios, together with a statistical validation of the obtained results.
- We discuss robustness and out-of-domain transferability, positioning ESG score prediction as a benchmark task that highlights broader machine learning challenges in small-data and heterogeneous environments.

The remainder of the paper is structured as follows: Section 2 reviews related work, including machine learning for ESG prediction and transfer learning. Section 3 describes the MPBTL-RF model and baseline models. Section 4 presents in-domain and out-of-domain results with comparative analysis. Section 5 applies the model to the BELEX dataset. Section 6 provides insights into the practical application of the proposed method, while Section 7 defines its limitations. Finally, Section 8 concludes with key findings and directions for future research.

2. Background

This section summarizes key research contributions in this field, highlighting the methodologies and datasets utilized for ESG score prediction in developed markets. It concludes with a literature review that motivates the proposed research.

2.1. Related work

A significant amount of literature now employs machine learning techniques for ESG score estimation, demonstrating the potential of these approaches to provide valuable insights into corporate sustainability performance.

D'Amato et al. [5] explained the determinants of the ESG score by performing the RF algorithm. They used financial statement information, such as profitability indicators, liquidity, and solvency ratios, from a subset of 109 companies listed in the STOXX Europe 600 Index between 2014 and 2018. The study found that financial statement items are powerful tools for explaining and predicting ESG scores, with performance measured by Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2) metrics.

Krappel et al. [12] proposed a heterogeneous ensemble model to predict ESG ratings using fundamental data. The model combines FNN, CatBoost, and XGB ensemble members. Their dataset included 7,413 companies with annual observations between 2002 and 2019, amounting to 57,310 observations. They utilized 475 numerical and 44 categorical features and measured performance using MAE and R^2 metrics. The study highlighted the importance of a comprehensive dataset and advanced machine learning techniques for accurate ESG score prediction.

Garcia et al. [7] developed a rough set model to relate ESG scores to key corporate financial performance measures from the investor's perspective. Their dataset comprised 1,688 observations from 2013 to 2018, with seven numerical features. They suggested that the industry sector and financial variables reveal significant differences across firms regarding ESG, but the model's significance diminishes when examining small differences in ESG performance.

Sokolov et al. [20] proposed an approach to automatically convert unstructured text data into ESG scores using deep learning for Natural Language Processing (NLP). They incorporated Bidirectional Encoder Representations from Transformers (BERT) to improve the accuracy of assessing the relevance and content of documents in an ESG context using social media data. This approach emphasizes the potential of automating ESG scoring and constructing ESG portfolios through advanced NLP techniques.

D'Amato et al. [6] predicted ESG scores using data from 401 companies that are constituents of the STOXX Europe 600 index. They used seven numerical features and employed the RF algorithm. The performance was measured using RMSE and MAPE metrics. Their findings indicate that the RF model can better grasp the nonlinear aspects of ESG score prediction than a classical generalized linear model.

Chowdhury et al. [4] applied six machine learning algorithms, including Artificial Neural Networks Classifier (ANNC), Bagging Classifier (BGC), k-Nearest Neighbors Classifier (KNNC), Naive Bayes Classifier (NBC), Random Forest Classifier (RFC), and Support Vector Machines Classifier (SVMC) on a global data sample of 6,166 firms in 73 countries from 2005 to 2019. They found that the RFC provided the highest accuracy

(78.50%) among the six algorithms. The study revealed that the lagged ESG score had the highest contribution to the model, with performance assessed using accuracy, Kappa, the area under the curve, receiver operating characteristic, and logLoss metrics.

The available literature analyzes the problem of predicting ESG scores from the point of view of developed markets, but no literature refers to the same problem in developing markets.

2.2. Motivations for the work

We found motivation for this research in the methods proposed in Tan et al. [22] transitive transfer learning, as well as in the Pinto et al. [16], Jin et al. [9] and active transfer learning model in Shim et al. [19].

Inspired by the human ability for transitive inference, where seemingly unrelated concepts can be connected through intermediate bridges using auxiliary knowledge, [22] introduced a novel learning paradigm called Transitive Transfer Learning (TTL). They proposed a framework that emulates this human-like learning process through two main components: intermediate domain selection and knowledge transfer. Extensive empirical evidence shows that the framework yields state-of-the-art classification accuracies on several classification datasets.

In [16] the authors suggested that model parameters or hyperparameters generated for similar tasks would also be similar, and that the information collected from the source task could be sent to another task in the form of shared model weights. According to their experiments, where different neural networks share the same feature and label space, they investigated a homogeneous inductive problem using model-based transfer learning. The authors conducted experiments that leveraged 250 data-driven models based on a synthetic dataset of a building archetype and studied, among other things, the influence of data availability for the transfer process of thermal dynamics.

Authors in [19] predicted reaction conditions from limited data through active transfer learning. They showed that specifically tuned machine learning models based on RF classifiers improved the applicability of Pd-catalyzed cross-coupling reactions to nucleophile types unknown to the model. They stated that in active transfer learning, simple models that are composed of a small number of decision trees with limited depths are key to ensuring generalizability, interpretability, and performance.

According to Jin et al. [9], model-based transfer learning is particularly efficient because it leverages the source domain model to directly transfer high-level knowledge, eliminating the need to reprocess training data or engage in complex relational reasoning, thereby enhancing its ability to generalize insights from the source domain to the target domain.

Let D_S represent the source domain and D_T the target domain with their corresponding feature spaces as X_S and X_T and let *P(X)* defines the probability distribution of the feature space. Denote the learning task τ_T and τ_S, and target the predictive learning function as $f_T(.)$. According to [16], transfer learning can be classified based on label availability, domain and task similarity, and the technique used for knowledge transfer.

According to label availability, there is inductive transfer learning, where both D_S and D_T have labeled data but $\tau_T \neq \tau_S$; transductive transfer learning where $\tau_T = \tau_S$, but $D_S \neq D_T$ while there are labeled data only in D_S; and unsupervised transfer learning where $D_S = D_T$ and $\tau_T \neq \tau_S$ but the tasks are related and there are no labeled data in

both domains. Within transductive transfer learning, we differentiate between cases where $X_S \neq X_T$ and the second case where $X_S = X_T$, but $P(X_S) \neq P(X_T)$ [16].

Regarding domain and task similarity, there is homogeneous transfer learning, where both the feature and labeled spaces in both source and target domains are the same, and heterogeneous transfer learning, where either the feature or label space is different.

Relative to a strategy that is adopted to share knowledge, according to [15], there are four different transfer learning approaches: instance-based methods, feature-based approaches, model-based transfer methods, and relational knowledge transfer. Instance-based techniques are often employed when X_T=X_S, allowing certain source data to be reweighted and utilized as training data in the target domain. Feature-based methods focus on uncovering a hidden feature space from the source domain to enhance performance in the target domain. Model-based transfer methods, on the other hand, transfer knowledge by sharing parameters or prior distributions of model hyperparameters.

This approach assumes that similar tasks should have similar models or hyperparameters, enabling knowledge transfer from the source task to the target task through shared parameters. Lastly, relational knowledge transfer relies on the assumption that the source and target domain data exhibit similar relationships, enabling this group of methods to transfer relational structures between the two domains.

Also, a common way of transferring is with a setting where X_S=X_T in the source and target domain, which minimizes the differences in the data distribution [21]. A traditional machine learning model relies solely on data from the target building, whereas a transfer learning model reuses knowledge from a source building to lower implementation costs, accelerate training, and improve performance [16].

However, because in this work, the source and target tasks share the same feature space and there are no labeled data in the target domain, a hybrid method, as a homogeneous transductive approach using model parameter-based transfer learning, was explored. This setting is also, according to [15], related to domain adaptation.

In the next section, we define how the proposed approach can expand the applicability of ESG score prediction.

3. Methodology

This section outlines two approaches to machine learning: model parameter-based transfer learning and traditional machine learning.

We address scenarios in which a large transitive dataset with target variables is not available. We applied parameter transfer from a source domain to perform model parameter-based transfer learning with RF, inspired by D'Amato et al. [5]. This approach leverages model parameters from the source domain to adapt to the target domain, despite limited target data in the transitive dataset.

A limitation of the proposed workflow concerns the availability of transitive domain data. In domains with the same data restriction problems but with multiple candidate transitive domains or transitive datasets, a selection algorithm according to the domain and dataset properties should be provided. The presented model tackles problems where transitive domain data samples are limited to provide source learning, so that knowledge transfer on a hyperparameter-based approach from a model trained on larger data is provided for training a model on a transitive dataset. The proposed approach is not specific to

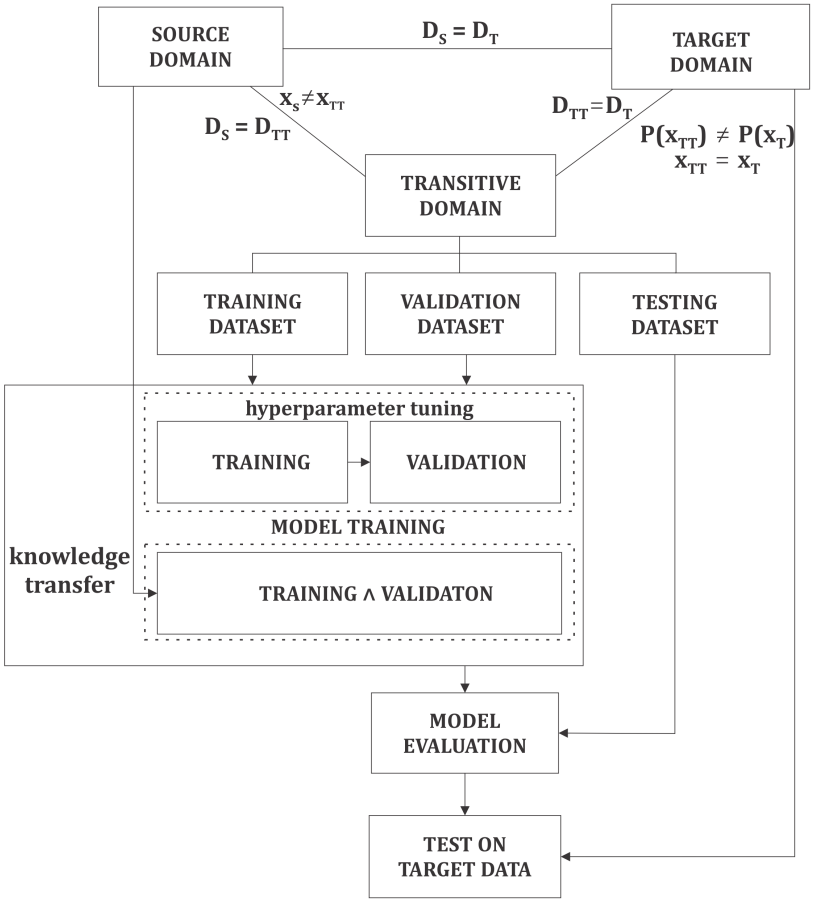


Fig 1. Proposed transfer learning workflow

any traditional machine learning algorithm, although the RF is used as a prediction model in this paper.

We compared our MPBTL-RF with traditional machine learning approaches, using the transitive dataset for training. We experimented with RF, XGB, and FNN for traditional model training and analyzed how our model performs relative to them under varying data availability conditions in the transitive dataset. Figure 1 presents the complete flow for all model trainings.

3.1. Datasets

We define two datasets: a transitive dataset and a target dataset. In the application scenario, there is no available labeled data in the target dataset, while the transitive dataset has a limited number of labeled data. The approach assumes that the learning tasks across the source, transitive, and target domains are the same. While the feature space between the

transitive and target domains is the same, X_{TT}=X_T, there is a difference in data distribution $P(X_{TT}) \neq P(X_T)$.

Transitive dataset. As a transitive dataset in this work, we used the Euronext Tech Leaders (TECLP) index, which covers the high-growth and leading companies in Europe from the Technology super-sector according to the Industry Classification Benchmark (ICB) industry classification.

The technology sector has exhibited slower ESG score development compared to other industries [10], due in part to challenges such as energy consumption and evolving regulatory expectations. However, recent trends show an increasing integration of ESG practices driven by regulatory pressures, stakeholder demands, and the sector's pivotal role in innovation. This makes the technology sector a relevant focus for studying ESG adoption and prediction in developing markets, where transparency and sustainability efforts are still emerging.

The available TELCP dataset consists of 94 companies with annual observations between 2018 and 2022, with some missing data for certain years. In our experiments, we used financial statements and ESG Refinitiv scores from this dataset for the companies included in the index.

We partitioned the transitive dataset into training, validation, and test subsets with a 60%-20%-20% split to prepare it for model training. Each model was assessed using five different fold splits to ensure robust evaluation, as detailed in the results section.

Considering the panel nature of the data, similar to [12], which shows a high correlation between scores for the same company across years, the train-validation-test split (60%-20%-20%) was done in a way that included each company in only one part of the split. This approach ensures the independence of the splits, preventing any sample from being included in more than one subset. This approach is crucial for maintaining the integrity of our proposed methodology, as it prevents data leakage between the training and validation sets. This precaution ensures that the machine learning model can generalize effectively to unseen data, thereby providing reliable predictions on the target dataset. Furthermore, consistent with the proposed transfer learning workflow, the feature space remains the same for both transitive and target data, X_{TT}=X_T. The test part of the transitive dataset is subsequently utilized for in-domain model evaluation to compare the performance of the machine learning and transfer learning aproaches.

Target datasets. As part of the model evaluation process, we created two target datasets. The first target dataset (holdout) in this study is based on available real data and consists of 14 blue-chip stocks traded in European markets classified as frontier according to the Financial Times Stock Exchange (FTSE) equity country classification[3]. According to the FTSE classification scheme from September 2024, there are 11 countries whose equity markets are classified as frontier markets, namely Bulgaria, Croatia, Cyprus, Estonia, Latvia, Lithuania, Malta, the Republic of North Macedonia, Serbia, the Slovak Republic, and Slovenia.

We examined the main stock market indices from these markets and selected the stocks according to the availability of the companies' ESG scores. Observed stocks are

[3] https://www.lseg.com/en/ftse-russell//equity-country-classification

constituents of the leading stock exchange indices - 8 stocks from the Bucharest Exchange Trading (BET) index, 2 from the Ljubljana Stock Exchange index (SBITOP), 2 from the Bulgarian Stock Exchange index (SOFIX), 1 from the Slovak share index (SAX), and 1 from the Cyprus Stock Exchange (CSE) index. The stocks come from different industries: Financials (35.7%), Industrials (28.6%), Technology (21.4%), Utilities (7.14%), and Energy (7.14%). The dataset, representing diverse sectors from developing markets, is used as an out-of-domain evaluation set relative to the technology-focused Euronext dataset, allowing the assessment of model generalization across markets with differing maturity and sectoral composition. Since this dataset is, in effect, cross-domain relative to the technology-focused transitive dataset, it provides a suitable setting for evaluating model generalization.

The second target dataset (synthetic) was created by expanding the first target dataset with additional artificially generated samples produced using the Gaussian Copula Synthesizer from the Python Synthetic Data Vault (SDV) library. This augmentation preserves the statistical characteristics of the original data while increasing its size, allowing for a more comprehensive evaluation of model performance in cross-domain scenarios.

Case study dataset. The case study dataset contains real-world data derived from the indices of the regulated market, the Belgrade Stock Exchange. In this experiment, the case study dataset consists of 14 companies listed on the BELEX, whose stocks were continuously traded on the regulated market and served as constituents of the BELEXline Index from 2018 to 2022 – a total of seventy data instances. Through this case study, we contribute to bridging the gap between theoretical advancements and practical applications in ESG score prediction, fostering the synergy between AI and business.

3.2. Model parameter-based transfer learning with random forest (MPBTL-RF)

Among machine learning models, RF is frequently utilized in transfer learning studies due to its ease of use and its well-documented high performance across a wide range of tasks [21, 19, 18]. Following the approach in [19], we hypothesize that overly complex models tend to overfit and struggle to generalize to dissimilar data and that simplifying model architecture can enhance predictive accuracy in the target domain. According to [17], the implementation of the RF requires setting two main parameters: the number of trees and the number of randomly selected predictor variables. In our experiment, we utilized the RF regression model for parameter transfer, as detailed by [5], recognizing that these parameters were tuned on a dataset from the STOXX® Europe Index.

Source model. Small transitive datasets make it challenging to train a model effectively, and parameters from a source domain model can be inherited for training. Conversely, when the transitive dataset is large enough, it serves as the source model itself. In both cases, the source model trained on a larger and well-sampled dataset transfers its parameters to facilitate learning on the target dataset.

Briefly explained, this model is in our work referred to as the source model and is used only for the transfer of model parameters, given that the available transitive data are from the Euronext Tech Leaders Index. The experiments in [5] utilized financial statement items and Bloomberg ESG scores collected for the constituents of the STOXX Europe 600

Index, representing large, mid, and small-capitalization companies across 17 countries in the European region. They selected a sample of 109 companies listed in the STOXX Europe 600 Index between 2014 and 2018, representing 21% of the entire set of companies included in the index. The selected companies belong to the Communications, Energy, Technology and Utilities industry sectors [5].

We applied a RF regressor with transferred parameters from the source model by setting: minimum samples per leaf to 1, maximum features per split to 5, and the number of trees in the forest to 500 to optimize predictive performance on the target dataset. Setting minimum samples per leaf to 1 allows each leaf node to capture finer patterns, which is beneficial for smaller datasets. Setting the maximum features per split parameter to the value of 5 limits the features considered at each split, balancing variance and interpretability, while the number of trees in the forest is set to 500 to create a large forest of trees to enhance stability and accuracy. These parameters should enable the model to leverage source knowledge effectively, yielding substantial generalization and performance in our target domain.

Feature selection. According to the proposed workflow in Figure 1, our study also employs the same financial features except the categorical sector feature presented in the source domain dataset, as there are no sector differences in our transitive dataset. The indicators used, listed in Table 1, represent liquidity, solvency, profitability, operating efficiency, and the company's market value. Variables used in the formulas are defined as follows: Net sales – the total revenue generated by the company from the sale of goods or services during the year, after deducting sales returns, allowances, and discounts; Total assets – the sum of current and long-term assets owned by the company; EBIT – a company's operating profit, calculated as earnings before deducting interest expenses and income taxes; Annual dividend per share – the total dividends paid divided by the number of outstanding shares over the year; Share price – the closing price of a company's share on the last trading day of the year; Net income – the company's total profit after deducting all expenses and taxes from total revenue; Equity – the residual owner's interest in a company's assets after all liabilities have been deducted; Total liabilities – the sum of all current and long-term financial obligations owed by the company; Cash – cash and cash equivalents readily available to the company; Cash flow – operating cash flow calculated by adjusting net income for non-cash items like amortization and changes in working capital; Total current assets – assets expected to be sold, consumed, or converted to cash within one year or the company's operating cycle; Total current liabilities – financial obligations and debts the company is expected to settle within one year or its normal operating cycle; Total debt – the sum of all interest-bearing debt, including current borrowings and long-term liabilities; Earnings per share – net income divided by the average number of shares outstanding. All financial values are based on the company's financial statements for the relevant fiscal year.

3.2.1. Parameter transfer procedure and overfitting control

In this study, MPBTL-RF was implemented through *structural parameter inheritance* between domains using the RF regressor. Model parameters optimized on the STOXX

Table 1. Financial Indicators

Label	Variable Name	Formula
Sales_to_Assets	Asset Turnover Ratio	Net sales / Total assets
EBIT_to_Sales	Operating Margin	EBIT / Net sales
DY	Dividend Yield	Annual dividend per share / Share price
NI_to_Sales	Net Profit Margin	Net income / Net sales
Rating	Kralicek QuickTest[1]	0.25 × (Equity/Total liabilities)
		+0.25 × ((Total liabilities - Cash)/Cash flow)
		+0.25 × (EBIT/Total assets)
		+0.25 × (Cash flow/Net sales)
LR	Current Liquidity Ratio	Total current assets / Total current liabilities
SR	Solvency Ratio	Total debt / Total assets
P/E	Price to Earnings Ratio	Share price / Earnings per share

[1] Results expressed in points from 1 to 5.

Europe 600 dataset [5] were transferred to the Euronext Tech Leaders dataset and subsequently applied for prediction on the target dataset. The transferred parameters included: (i) the number of trees ($n\_estimators = 500$), (ii) the number of features considered at each split ($max\_features = 5$), and (iii) the minimum number of samples per leaf ($min\_samples\_leaf = 1$).

This approach follows the *homogeneous transductive* parameter-transfer paradigm [15, 16], in which the model structure–rather than learned weights–is reused to regularize learning across domains. By constraining model complexity, the inherited parameters function as a form of inductive bias, stabilizing learning in the transitive and target domains where labeled data are limited. The procedure is summarized on Figure 2.

```
# Source: STOXX Europe 600
source_model = RandomForestRegressor(
    n_estimators=500, max_features=5, min_samples_leaf=1
)
source_model.fit(X_source, y_source)

# Transitive: Euronext Tech Leaders
transitive_model = RandomForestRegressor(
    n_estimators=source_model.n_estimators,
    max_features=source_model.max_features,
    min_samples_leaf=source_model.min_samples_leaf
)
transitive_model.fit(X_transitive, y_transitive)

# Target: BELEX (Frontier Markets)
y_pred = transitive_model.predict(X_target)
```

Fig 2. Transitive Model Training and Prediction

To prevent overfitting and maintain generalization, the model was trained using a 80%–20% train–test split, ensuring that each company appears in only one subset, thereby eliminating data leakage. Additionally, a 5-fold cross-validation procedure was employed to assess model stability. Fixing the structural parameters from the source model serves as a regularization mechanism, reducing the risk of overfitting and enhancing robustness in small-sample target domains.

3.3. Traditional ML with parameter-tuning

To evaluate traditional machine learning models, we employed a systematic hyperparameter-tuning methodology using 5-fold cross-validation. This involved splitting the data into five folds and training a separate model on each fold, using a 60%-20%-20% split for training, validation, and testing within each fold to ensure robust performance comparison. Consistency was maintained by using the same split as in the previous experiment with model parameter-based transfer learning in Section 3.2 for the test set, but reallocated 25% from each training split (previously 80%) to create a validation set for each fold, achieving a final 60%-20%-20% split.

A comprehensive hyperparameter grid search was conducted for each model. This process entailed evaluating the models across the five validation sets to identify the optimal parameter configurations for predicting ESG, E, S, and G targets. The best-performing parameters were then applied to the test set for final evaluation across all targets (ESG, E, S, G).

For this evaluation, we employed three different machine learning approaches:

- **RF**: An ensemble learning method that builds multiple decision trees and outputs the mode of the classes (classification) or the mean prediction (regression).
- **XGB**: Unlike RF, XGB builds trees sequentially, with each tree focusing on correcting the errors of the previous one. This iterative error-correction process results in a highly accurate model by gradually reducing bias. This approach is especially valuable with small datasets, where limited information can benefit from stepwise refinement [3].
- **FNN**: Although typically requiring large datasets, FNN can still perform well on smaller datasets if structured carefully to avoid overfitting. Their flexibility and capacity to model non-linear relationships make them effective even with limited data and features, as in this study. Additionally, FNN offers advantages in terms of training ease and interpretability due to the relatively small number of parameters. This efficiency, combined with their ability to model complex relationships, justifies the use of FNN for predicting ESG scores on the available dataset.

These approaches were selected for their established effectiveness in predictive modeling and their ability to handle the challenges posed by the small dataset and the complexity of ESG score prediction. Additionally, these networks enable efficient hyperparameter tuning, which was performed with various combinations, as detailed in Table 2.

3.4. Baseline method

In assessing our models, we also benchmarked our results against values from [12]. In their study, three individual models were used – FNN, CatBoost (CB), and XGB, along

Table 2. Grid Search Parameter Ranges for Evaluated Models

Model	Parameter	Value Combinations
RF	Number of Estimators	100, 500, 1000
	Max Features	auto, sqrt, log2
	Max Depth	10, 30, 50, None
	Min Samples Split	2, 5, 10
	Min Samples Leaf	1, 2, 4
	Bootstrap	True, False
XGB	Early Stopping Rounds	None, 10
	Learning Rate	10^{-2}, 10^{-3}, 10^{-4}
	Max Depth	3, 5, 10, 15
	Min Child Weight	0.5, 0.7, 1
	Number of Estimators	100, 200, 300
FNN	Number of Layers	5, 6, 7
	Number of Neurons	1000, 600, 300, 200
	Activation	sigmoid, relu, leaky_relu
	Dropout	0, 0.3, 0.5, 0.7
	Learning Rate	10^{-3}, 10^{-4}, 10^{-5}
	Batch Normalization	True, False
	Epochs	100, 200, 300
	Early Stopping	True, False

with two types of ensemble models – NN Ensemble (NNE) and Heterogeneous Ensemble (HE) to predict ESG and pillar scores using data from companies in the S&P 500 index. Their results, along with their simple mean baseline model (BL), are presented in Table 3.

Table 3. Baselines for different models from [12]

	HE	BL	FNN	CB	XGB	NNE
ESG	11.2	16.5	12.1	11.3	11.4	12.1
E	14.9	22.7	16.3	15.0	15.2	16.2
S	13.5	19.0	14.6	13.5	13.6	14.5
G	16.7	18.7	17.4	16.7	16.8	17.4

3.5. Models training, validation, and evaluation parameters

The training subset was used to train all traditional machine learning models. To ensure that the experiments are not influenced by random chance, we implemented a 5-fold cross-validation strategy to optimize model performance and reduce overfitting. We performed hyperparameter tuning across different fold splits of our transfer dataset for each of the pillars: ESG, E, S, and G. To ensure that the model treated each feature equally and to

facilitate faster convergence during training, we standardized the dataset by scaling the features so that they had a mean of 0 and a standard deviation of 1.

Fine-tuning parameters for FNN were performed on the National Platform for AI of Serbia, utilizing a single NVIDIA A100 GPU with 40GB using PyTorch. After evaluating 5,184 models per fold, we identified the optimal parameters, totaling 25,920 models across all folds. The process took up to 50 hours for all five folds. The training was performed in batches of 32. All other fine-tuning, model training and inference were done on local computers.

Validation was performed for the traditional ML models with parameter-tuning in Section 3.3. The validation was performed through cross-validation and cross-sample evaluation (test subset) across multiple folds. A test subset is considered for in-domain evaluation. At the same time, both target datasets were utilized for out-of-domain evaluation to assess the generalization of the best-performing models across different contexts not seen during training.

Furthermore, to ensure methodological coherence with the workflow illustrated in Figure 1, we conducted statistical analyses to validate the suitability of the selected transitive dataset and to assess feature distribution differences across all datasets used in the study. The results of the Kruskal-Wallis test presented in Table 4 show no statistically significant differences in the distribution of ESG scores – both overall and across its individual Environmental (E), Social (S), and Governance (G) components–among the observed datasets. This suggests that ESG adoption levels do not significantly vary among the companies from the technology sector in the transitive dataset and those in the holdout and synthetic datasets.

Table 4. Kruskal-Wallis test results for TECLP, holdout, and synthetic datasets

Indicator	Test Statistic	p-value
ESG	0.8367	0.6581
E	0.3176	0.8532
S	4.6039	0.1001
G	1.2305	0.5405

Note: Significance level (α) set at 0.05.

On the other hand, the Kruskal-Wallis test results followed by the Dunn-Bonferroni post hoc pairwise comparison among TECLP, holdout and BELEX datasets yielded mixed results as presented in Table A1 in the Appendix. Several variables, including Sales_to_Assest, Rating and LR, did not show statistically significant differences after adjustment, indicating insufficient evidence to reject the null hypothesis of equal distributions among these groups. Statistically significant differences (adjusted $p < 0.05$) were observed for EBIT_to_Sales and NI_to_Sales in all group comparisons. Conversely, variables such as DY, P/E, and SR showed significant differences only in selected group comparisons, indicating heterogeneous effects depending on the specific group pairing.

The obtained results correspond to the experimental framework shown in Figure 1.

Additionally, for the purpose of a comprehensive analysis of model robustness and properties, traditional machine learning models were implemented in conjunction with the BORUTA feature selection algorithm.

During cross-validation, we used grid search (Table 2) to find the best combination of hyperparameters on the training dataset, aiming to minimize the value of the scorer function, which calculates the mean squared error (MSE) between predicted and actual values. The best model was used to predict outcomes on the validation and test subsets. The best model was also used to predict values for the out-of-domain evaluation, holdout and synthetic dataset. Similarly, in models with BORUTA feature selection implementation, feature selection was performed for each fold separately.

The evaluation was performed in similar regression tasks as in [12], using MAE, as defined in Equation 1.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{1}$$

Finally, it is worth noting that the proposed approach is versatile and not limited to any specific traditional machine learning algorithm – it can be applied across various algorithms and data domains.

4. Results

This section introduces the results obtained through empirical analysis of different approaches used in this study. We conducted experiments for ESG score prediction and for the Environmental, Social, and Governance pillars separately. Thus, the obtained results provide comprehensive insight into the possibility of transfer learning in developing markets. As previously described in section 3.1 in order to further verify the robustness of the proposed model, a synthetic dataset was created and an out-of-domain evaluation was performed on it, as well as on the holdout dataset. Figure 3 illustrates the experimental setup.

4.1. In-domain evaluation

We first present the evaluation results on the test subset of the transitive dataset (in-domain), as defined in Section 3.1. The analysis includes the proposed MPBTL-RF approach and traditional machine learning approaches (RF, XGB), with and without feature selection.

MPBTL-RF in-domain.

The results obtained using the proposed MPBTL-RF are shown in Table 5. Based on the average values displayed, it can be observed that the G pillar score is the most challenging to predict, as it has the highest error. In contrast, the ESG component is predicted with the lowest error. Analyzing the results across folds, the difference between the highest and lowest errors is smallest for the ESG score (2.48), while the largest difference is observed in the E component (6.39 absolute error difference). The variation in error levels suggests stability in the predictions obtained with this model, considering the dataset split defined in Section 3.1.

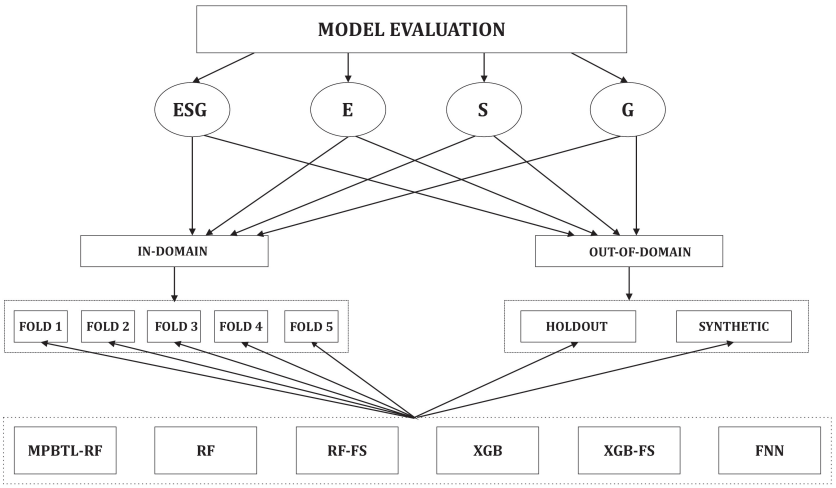


Fig 3. Experimental settings

Table 5. MPBTL-RF in-domain evaluation

MAE	**ESG**	**E**	**S**	**G**
Fold 1	14.61	17.08	18.12	21.40
Fold 2	15.41	18.78	15.94	22.26
Fold 3	13.05	15.03	15.01	17.71
Fold 4	15.53	21.42	15.60	21.13
Fold 5	15.01	20.48	15.44	19.24
Avg MAE	14.72	18.56	16.02	20.34

RF in-domain. Table 6 summarizes the in-domain performance of the RF model. The results exhibit a consistent pattern across folds, with relatively stable MAE values for all ESG components. The ESG score is associated with the lowest prediction error, whereas the G pillar remains the most difficult to predict, yielding the highest MAE. In most cases, validation errors are slightly lower than those on the test set, which is expected given that model tuning is performed on the validation data. The limited discrepancy between validation and test performance indicates that the model does not suffer from pronounced overfitting.

When feature selection is introduced (Table 7), the RF-FS model attains very similar MAE values across all targets. This indicates that restricting the feature space does not adversely affect predictive accuracy. The results suggest that the selected subset of features retains the essential information required for prediction, while less informative variables can be excluded without loss of performance. Overall, the RF model preserves stable behavior under feature reduction in the in-domain setting.

XGB In-domain. Table 8 presents the in-domain results for the XGB model. The model shows consistent performance across folds, with relatively stable MAE values for

Table 6. Traditional RF in-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	13.14	16.57	16.27	19.94	14.45	17.15	20.96	20.66
Fold 2	14.38	20.33	15.22	18.51	16.43	19.87	18.00	22.12
Fold 3	16.13	18.90	16.67	19.31	13.01	14.76	14.32	18.32
Fold 4	15.62	17.98	16.32	19.62	14.78	21.13	15.67	20.31
Fold 5	14.66	16.30	16.75	21.55	15.70	21.21	15.02	17.91
Avg MAE	14.78	18.02	16.25	19.79	14.87	18.83	16.07	19.86

Table 7. Traditional RF -FS in-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	13.45	17.00	16.90	19.89	14.01	17.15	17.77	19.93
Fold 2	13.62	20.21	14.97	17.96	16.59	19.34	17.65	21.86
Fold 3	16.68	19.50	16.92	20.07	12.57	14.83	15.73	19.31
Fold 4	14.63	18.38	15.48	19.44	15.01	20.45	15.78	19.85
Fold 5	15.21	16.68	17.21	21.38	15.44	21.20	16.33	16.22
Avg MAE	14.72	18.35	16.30	19.05	12.72	18.59	16.65	19.43

all ESG components. The ESG score is predicted with the lowest error, while the G pillar remains the most challenging target, exhibiting the highest MAE. Across most components, the average MAE is lower on the validation set than on the test set, which is expected given that hyperparameter tuning is performed on the validation data. The relatively small differences between validation and test errors suggest that the model does not exhibit severe overfitting. However, for the S component, a slight increase in test error (difference of 0.647) indicates mild overfitting.

When feature selection is applied (Table 9), the XGB-FS model achieves comparable MAE values across all targets, indicating that reducing the feature space does not negatively impact predictive performance. This suggests that the selected subset captures the most relevant information, while redundant features can be removed without loss of accuracy. The results further indicate that XGB remains robust under feature reduction, although minor overfitting effects may still occur for certain components in small-sample settings.

Compared to the results in Table 3, where XGB was trained on a much larger dataset and reported MAE errors of 11.4, 15.2, 13.6, and 16.8 for the ESG score and the E, S, and G components, respectively, the higher MAE errors observed with XGB (15.19, 19.51, 16.29, and 19.81) highlight the significant influence of training set size on prediction accuracy.

FNN In-domain. The results in Table 10 show that, although the FNN achieves the lowest MAE on the validation set, its performance on the test set is substantially worse and exhibits the largest variance across folds. This behavior indicates strong overfitting,

Table 8. XGB in-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	14.16	17.44	16.97	19.86	14.61	18.74	18.04	18.67
Fold 2	14.66	20.41	16.43	17.17	17.13	20.74	18.29	22.21
Fold 3	15.80	18.45	17.07	18.06	12.90	15.02	15.77	18.18
Fold 4	15.43	18.12	16.46	19.33	15.32	21.61	13.92	20.77
Fold 5	15.79	17.05	17.75	21.22	15.99	21.48	15.42	19.23
Avg MAE	15.17	18.29	16.94	19.13	15.19	19.51	16.29	19.81

Table 9. XGB-FS in-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	14.14	18.24	18.28	19.40	14.63	18.71	17.76	17.75
Fold 2	14.96	20.54	15.28	17.25	17.65	20.75	17.87	21.47
Fold 3	15.54	18.45	16.65	17.94	13.02	15.02	15.98	18.27
Fold 4	15.43	18.30	16.43	19.00	15.29	21.18	13.91	19.45
Fold 5	16.03	18.52	16.67	20.92	16.68	23.08	15.05	18.76
Avg MAE	15.22	18.81	16.66	18.90	15.45	19.75	16.11	19.14

which is expected when applying high-capacity neural networks to small and heterogeneous datasets. Due to their expressive nature, FNNs are sensitive to variations in training-validation splits, often converging to different local minima and producing unstable predictions across folds.

In contrast, ensemble tree-based models (RF, XGB) benefit from averaging mechanisms that provide more stable and consistent predictions. The observed behavior of the FNN therefore reflects the trade-off between flexibility and stability in small-sample deep learning settings [11]. Although the model captures non-linear relationships between financial indicators and ESG scores, its lack of generalization limits its practical applicability.

This discrepancy between validation and test performance is not observed in the other models, and data splitting procedures ensured strict separation across folds, indicating that the observed overfitting is inherent to the model rather than a consequence of data leakage or methodological bias. Consequently, the FNN model is excluded from further comparative analysis.

In-domain comparative analysis. The comparison of the results for all previously described models is presented in Figure 4.

ESG score. The proposed MPBTL-RF achieves an average MAE of 14.72, which is comparable to the MAE value of 11.3 presented in Table 3. At the same time, the proposed transfer learning approach has the smallest standard deviation value (1.00), compared to

Table 10. FNN In-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	12.01	13.89	13.41	17.08	24.69	24.46	26.69	23.00
Fold 2	12.97	17.91	11.47	16.93	100.08	86.60	102.75	106.75
Fold 3	13.29	16.04	14.89	17.56	62.02	51.96	77.92	68.13
Fold 4	11.11	15.18	12.08	16.51	21.58	26.53	25.56	24.65
Fold 5	12.77	14.21	12.49	20.16	26.53	28.59	29.99	24.51
Avg MAE	12.43	15.45	12.868	17.65	46.98	43.63	52.58	49.41

the value of 1.30 obtained with RF, 1.65 obtained with baseline models, and 1.58 obtained with the XGB algorithm.

The MPBTL-RF model exhibits the smallest variation in MAE values across folds, with a difference of just 2.48 between the highest and lowest values. This suggests that the model is the most robust and least affected by changes in data splits.

In relation to the models with feature selection in term of ESG predictions, the proposed MPBTL-RF approach exhibit somewhat weaker but still comparative results.

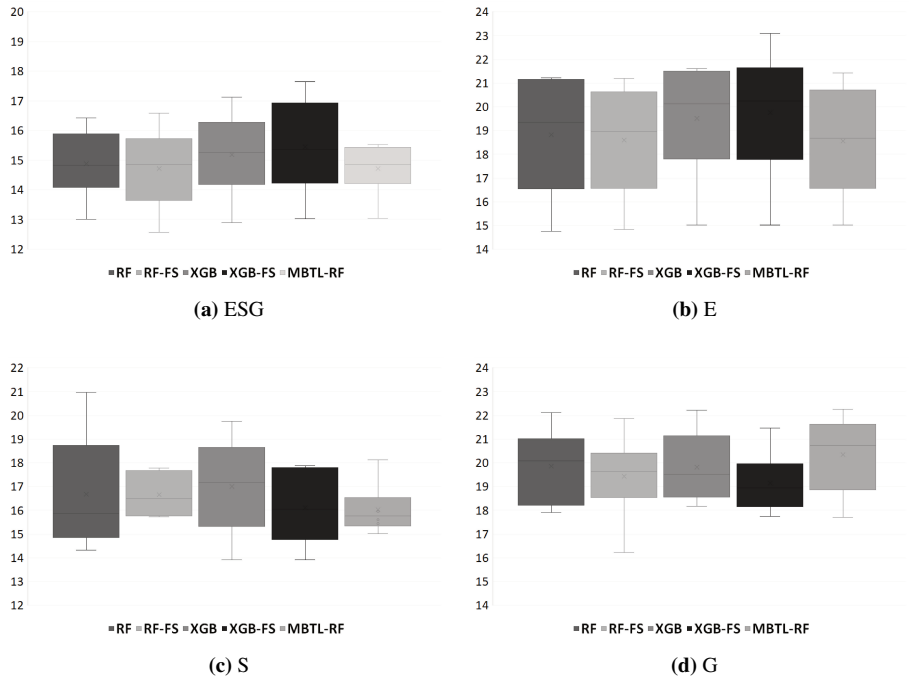
E pillar score. For the E pillar score, the proposed MPBTL-RF results in average MAE to 18.56 which is comparable to the MAE value of 14.9 obtained in Table 3. Additionally, in the case of the E pillar, the proposed transfer learning approach has the smallest standard deviation. The lowest MAE value obtained across folds was in Fold 3 using the RF model. The smallest difference between the highest and lowest MAE values is observed with the MPBTL-RF model, amounting to 6.39, suggesting stability in this model across different splits (train, validation, and test). In relation to the RF and XGB models with feature selection in term of average MAE for E pillar predictions, the proposed MPBTL-RF approach outperformed feature selection approaches.

S pillar score. The proposed MPBTL-RF results in average MAE to 16.02, which is comparable to the MAE value of 13.5 obtained in Table 3. Additionally, for the S pillar, the proposed transfer learning approach exhibits the smallest standard deviation. The smallest difference between the highest and lowest MAE values across folds is observed with the MPBTL-RF model, amounting to 3.11.

For S pillar predictions, the proposed MPBTL-RF approach again has the best average MAE values in comparison to RF and XGB with feature selection.

G pillar score. For the G pillar score, the proposed MPBTL-RF brings the MAE to 20.34, which is comparable to the MAE value of 16.1 obtained in Table 3. Additionally, in Fold 3, the MPBTL-RF approach achieves the lowest MAE among the models studied, indicating a stronger generalization capability. However, the smallest range between the highest and lowest MAE values across models is observed with the traditional RF model.

Regarding standard deviation, the XGB model shows the smallest variation for the G component, which, along with the MAE level, suggests slightly more favorable predictions using traditional ML and more complex models on challenging problems. However, it is important to note that for all models in the comparison, the average MAE is derived from MAE values across five folds. This approach indirectly represents a homogeneous ensemble in traditional ML models, as the average MAE is the mean prediction of five



(a) ESG **(b)** E

(c) S **(d)** G

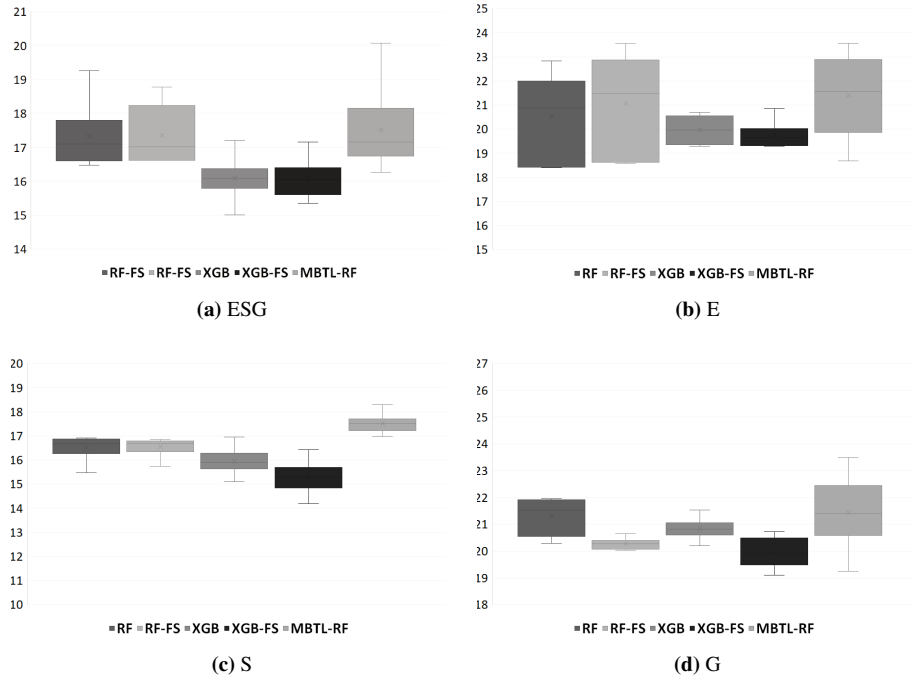
Fig 4. In-domain comparative analysis.

distinct models using the same algorithm. Compared to other models, MPBTL-RF uses the same parameters for each fold but has a larger training set available. Thus, even though XGB achieves a lower MAE, the difference in average MAE between these two models (0.53) still supports the effectiveness of the proposed model, even for highly complex tasks. In relation to the models with feature selection in terms of average MAE for G pillar predictions, the model exhibits somewhat weaker but still comparative results.

4.2. Out-of-domain evaluation

To evaluate model generalization, we conducted out-of-domain experiments using the holdout and synthetic datasets described in Section 3.1. The results are presented in Figures 5 and 6.

Figure 5 presents the prediction results of the MPBTL-RF, XGB, XGB-FS, RF and RF-FS models on the holdout target dataset. The results indicate that the average MAE values for the observed labels are similar to those obtained on the transitive (in-domain) dataset's test subset, highlighting the approach's effectiveness for complex predictive tasks. The standard deviation of the average MAE across both test sets (for ESG, E, S, and G) is 2.47. Additionally, the MAE values achieved on the target dataset further confirm that dividing the data into independent sets during model training is suitable for enhancing model generalization capabilities. The G component remains the most challenging to predict. By examining all the results in Figure 5, the MAE values show slightly better



(a) ESG **(b)** E

(c) S **(d)** G

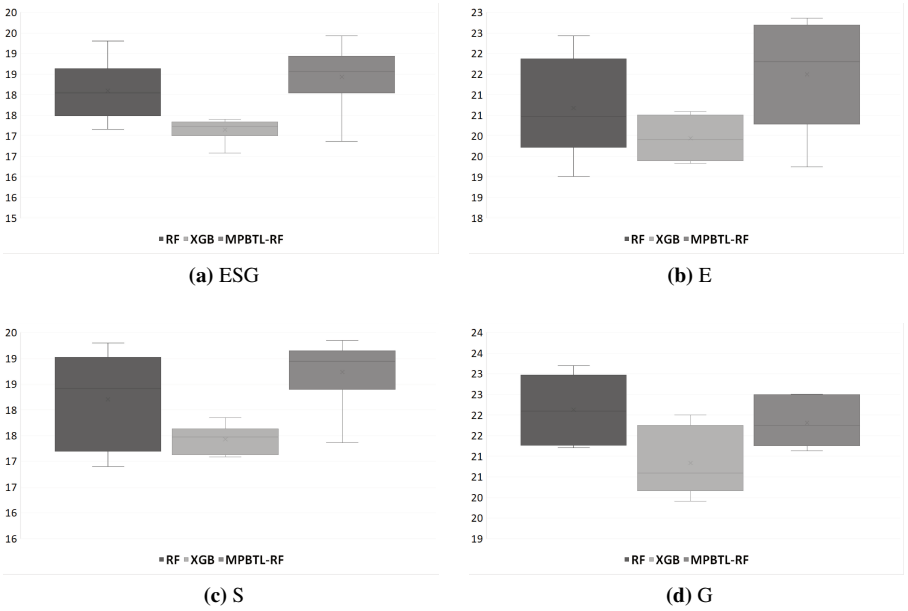
Fig 5. Out-of-domain (holdout dataset) Comparative Analysis

predictions for traditional ML models. However, it is important to note that for all models in the comparison, the average MAE is derived from 5-fold MAE values, which can be considered a homogeneous ensemble across all traditional machine learning models. This is because the average MAE represents the mean prediction of five different models using the same algorithm. Unlike other models, the MPBTL-RF model uses the same parameters for each fold but benefits from a larger training set. Even though traditional learning models achieve slightly lower MAE values, this small difference supports the effectiveness of the proposed transfer learning approach, especially for highly complex tasks.

ESG score. The proposed MPBTL-RF for ESG results in an average MAE of 17.51. The average MAE for the MPBTL-RF and RF models are nearly identical (17.51 versus 17.32), which directly supports the validity of applying the MPBTL-RF.

E pillar score. The same conclusion applies to predictions for the E pillar score. Although the resulting MAE might appear high, they remain comparable to values found in the literature for the specific domain of ESG score prediction and its components. It is particularly important to note that the target dataset was created using data from four countries. Given this diversity, the selection of learning features and the predictive model can effectively address such complex problems.

S pillar score. For the S pillar score the difference in error compared to the RF and XGB models are 1.00 and 1.52, respectively, indicating that the MPBTL-RF model can also be considered satisfactory in this case.



(a) ESG **(b)** E

(c) S **(d)** G

Fig 6. Out-of-domain (synthetic dataset) Comparative Analysis

G pillar score. Finally, when predicting the G pillar score, which is considered the most complex based on the MAEs, the MPBTL-RF model remains competitive, even though other traditional machine learning models achieved slightly better average values.

Applying the trained MPBTL-RF, XGB and RF on the synthetic target dataset yields results shown in Figure 6.

ESG score. By examining all the results in Figure 6 for the ESG score on the synthetic dataset, the MAE values show slightly better predictions for traditional ML models. Notably, for all models in the comparison, the average MAE is derived from 5-fold MAE values, which can be considered a homogeneous ensemble across all traditional machine learning models.

E pillar score. The same conclusion applies to predictions for the E pillar score. Although the resulting MAE might appear high, it remains comparable to values found in the literature for the specific domain of ESG score prediction and its components.

S pillar score. For the S pillar score, the difference in MAE value compared to the RF and XGB models is 0.53 and 1.32, respectively, indicating that the MPBTL-RF model can also be considered satisfactory in this case.

G pillar score. Finally, when predicting the G pillar score, which is considered the most complex based on the MAEs, the MPBTL-RF model remains competitive and achieved slightly better average values than the RF model.

To conclude, in the in-domain analysis, the MPBTL-RF model achieves second-best average results for the ESG score, and only the RF-FS approach has a better average MAE value. For the E and S components, the proposed MPBTL-RF approach outperforms other models, while the XGB-FS approach performs best for the G component. To compare multiple models on multiple datasets, we applied Friedman's test to check

whether the compared models have significant general differences in performance among folds. Thus, based on the Friedman statistical test conducted and shown in the Appendix, Table A2, there is no statistically significant difference in the performance of the observed approaches, except for the E pillar score. But even the null hypothesis is rejected in the second stage by applying the Nemenyi test, which did not find a statistically significant difference between any of the group pairs post-hoc test.

However, in the out-of-domain evaluation on the holdout dataset, XGB-FS delivers the best performance across the ESG score and all individual components. This indicates that transfer capabilities could be improved with a larger and higher-quality transfer dataset, potentially leading to better MAE values overall with XGB or another algorithm.

Based on the performed statistical test shown in the Appendix, Table A2, it can be concluded that there is no significant difference in MAE values across approaches for ESG scores and E pillar scores, while a statistically significant difference was observed for S and G pillars between the XGB-FS model and the MPBTL-RF approach. However, it should be pointed out that here too the comparison is where the XGB model is at an advantage in the selection of attributes for model training compared to the MPBTL-RF approach, which always generates results based on the same model parameters. Thus, even observed statistically significant differences between the models do not diminish the importance of the proposed model.

In relation to the synthetic dataset, it can be seen that the XGB model gives slightly less variation in precision when it comes to ESG scores prediction, while the MPBTL-RF model has the least variation when it comes to the prediction of the G component, which indicates model robustness according to all observations about G pillar score prediction. To test whether there is a statistically significant difference between the obtained MAE values among folds in the observed models, we conducted the Friedman test in the first stage and the Nemenyi test as the post-hoc test. All of the tests have been performed at 5% significance level. According to the test results available in the Appendix, Table A2, there is no statistically significant difference between MPBTL-RF and the other models obtained on the synthetic dataset. Thus, a statistically significant difference was confirmed between the RF and XGB models on the synthetic dataset.

In terms of computational efficiency, the differences in total training time were substantial. MPBTL-RF, which does not require hyperparameter fine-tuning, completed a fold in only 4.46 seconds. By contrast, the remaining models required full one-fold hyperparameter tuning: the standard RF model needed 239.86 seconds, XGB required 26.818 seconds, and the FNN model was by far the most time-consuming at 20,639.10 seconds.

The observed results highlight the computational efficiency and practicality of the proposed model. Overall, the obtained results demonstrate the potential of machine learning models, specifically MPBTL-RF, for predicting ESG scores in developing financial markets. Despite the challenges of predicting the G component, MPBTL-RF exhibited strong generalization capabilities across in-domain and out-of-domain data. This supports the idea that transfer learning can be a valuable tool for improving ESG score predictions, especially when dealing with limited or less-represented data in developing markets.

Considering that all models in this study were trained using only nine attributes, the results can be regarded as highly favorable, highlighting the simplicity, time efficiency, and applicability of the proposed approach.

5. Model application: case study of the BELEX

This section presents the experimental results and discusses the application of the proposed model to datasets derived from indices of the regulated market of the BELEX.

The case study in this experiment consists of 14 companies listed on the BELEX, whose stocks were continuously traded on the regulated market and served as constituents of the BELEXline Index from 2018 to 2022. Given that ESG scores can significantly influence investor decisions, the results are presented anonymously, as recommended by [12]. Table 11 shows the ESG scores for the selected blue-chip companies listed on the BELEX.

Table 11. One-year ESG rating predictions using MPBTL-RF on BELEX case study dataset

Company	ESG	Class	Company	ESG	Class
A	32.15	1	H	43.03	1
B	56.20	2	K	49.72	1
C	58.02	2	L	49.65	1
D	66.88	2	M	53.71	2
E	52.85	2	N	44.35	1
F	55.05	2	O	52.44	2
G	47.62	1	P	52.96	2

The selected companies represent various industry sectors: (1) Industrials (50%), (2) Financials (28.57%), (3) Utilities (7.14%), (4) Energy (7.14%), and (5) Real estate (7.14%). Figure 7 shows the ESG scores for the selected blue-chip companies listed on the BELEX per industry sector.

The box plot for ESG scores in Figure 7 predicted over five years indicates that the lowest average ESG score is evidenced in real estate (39.26) and utilities industry (48.89) with a wide interquartile range and no significant outliers, while the companies in the energy industry have the highest average ESG score (60.32) with most data clustered in a narrow interquartile range (5.09) but with outliers in both directions. The average ESG scores in the financial and industry sectors show a similar level of MAE (53.44 and 52.42 respectively), but ESG scores of companies in the financial sector are more dispersed, while in the industry sector, they display concentration in a tight interquartile range with some notable outliers. The obtained ESG score values were used as a basis for further analysis of financial performance. We clustered the observed stocks based on the predicted ESG scores into the following categories, corresponding to different performance levels:

- Class 0: ESG score ranging from 0 to 25 (first quartile) – Poor ESG performance.
- Class 1: ESG score ranging from 25 to 50 (second quartile) – Satisfactory ESG performance.
- Class 2: ESG score ranging from 50 to 75 (third quartile) – Good ESG performance.
- Class 3: ESG score ranging from 75 to 100 (fourth quartile) – Excellent ESG performance.

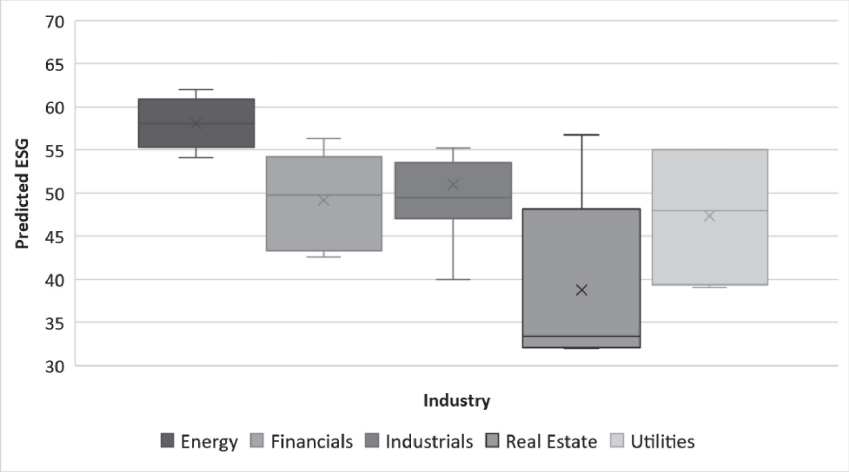


Fig 7. BELEX predicted ESG

Table 11 presents the obtained cluster categories. The predicted ESG scores were clustered into two groups: Class 1 – containing 39 observations, and Class 2 – with 31 observations. Predicted ESG-based clustering provides stakeholders with a predictive framework to incorporate sustainability considerations early in the decision-making process, thereby enhancing proactive risk management and promoting long-term value creation, despite the absence of direct ESG data.

To assess whether the proposed clusters exhibit significant differences in financial parameters, we conducted statistical tests using the non-parametric Mann-Whitney U test, which compares two independent groups based on a single continuous variable. Table 12 presents the descriptive statistics for these parameters within each cluster, along with the significance of the differences between the variables in these groups.

The analysis reveals significant differences in the following variables across the clusters: Dividend Yield (DY), Rating, and Price to Earnings (P/E) ratio. In line with [5], these financial indicators can be considered significant determinants of the ESG score.

- Dividend Yield (DY): Companies in Class 1 showed a clear tendency not to pay dividends during the observed period, with an average dividend yield of 3% and a standard deviation of 5%. Conversely, more than half of the companies in Class 2 paid dividends above 4%, with a standard deviation of 26%, indicating better financial capacity to support higher ESG scores.
- Rating: The overall operating performance rating showed significant differentiation between the groups. Companies in Class 1, with an average rating of 2.56, exhibited stable but lower capacities for improving ESG practices. In contrast, Class 2 companies, with a lower average rating of 2.24 but a higher standard deviation of 0.90, demonstrated greater sensitivity to business dynamics, suggesting more flexibility and potential for growth in ESG practices.
- The P/E ratio, which indicates market expectations about a company's future earnings, showed considerable differences between groups. Companies in Class 1 had an

Table 12. Descriptive statistics of financial indicators of the companies in the BELEX target dataset

Indicators[1]	Class	Descriptive Statistics				
		Mean	Median	SD	Min	Max
Sales_to_Assets	1	0.69	0.59	0.51	0.01	2.26
	2	0.85	0.71	0.50	0.10	1.92
EBIT_to_Sales	1	-0.03	0.03	0.46	-2.48	0.54
	2	0.30	0.06	1.22	-0.12	6.83
DY*	1	0.03	0.00	0.05	0.00	0.17
	2	0.09	0.04	0.26	0.00	1.48
NI_to_Sales	1	-0.07	0.03	0.45	-2.60	0.17
	2	0.23	0.05	1.04	-0.28	5.77
Rating*	1	2.56	2.75	0.58	1.00	3.50
	2	2.24	2.00	0.90	1.00	4.00
LR	1	2.95	1.49	3.23	0.22	10.71
	2	1.93	1.52	1.57	0.28	6.61
SR	1	0.46	0.50	0.26	0.10	0.97
	2	0.42	0.39	0.23	0.07	0.90
P/E*	1	90.74	8.15	475.18	-718.39	2751.72
	2	1.60	5.06	24.04	-108.31	48.54

[1] Indicators explanations: Sales_to_Assets – Asset Turnover Ratio, EBIT_to_Sales – Operating Margin, DY – Dividend Yield, NI_to_Sales – Net Profit Margin, Rating – Kralicek QuickTest, LR – Current Liquidity Ratio, SR – Solvency Ratio, P/E – Price to Earnings Ratio.

[*] Test significance (α) is set at 0.05. p-values are denoted as follows: * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

average P/E ratio of 90.74, indicating high investor expectations and potential overvaluation. On the other hand, Class 2 companies had a significantly lower average P/E ratio of 1.60 with a standard deviation of 24.04, reflecting more reasonable investor expectations and stable performance, with greater capacity to manage ESG challenges effectively.

These results showed that overall business rating and market value indicators are key determinants in predicting ESG scores for companies listed on the BELEX. Companies with higher dividend yields were more likely to have higher ESG scores, while the price-to-earnings ratio demonstrated sensitivity to economic conditions, further highlighting the importance of financial indicators in ESG predictions.

6. Practical application

The proposed hybrid learning approach applies to problems with no label data in the target domain and limited label data in the transitive domain for satisfactory machine learning training. By proposing a computationally efficient and practical model, this research offers valuable insights and several practical implications for companies, investors, and financial market players in developing markets.

Primarily, it enables firms to better align their strategies with ambitious sustainability and climate goals (the EU Green Deal and the Green Agenda for the Western Balkans). In developing markets where sustainability reporting practices are limited, ESG prediction assists companies in anticipating compliance requirements associated with sustainability reporting standards. This proactive model facilitates timely adaptation to regulatory demands embedded within existing frameworks.

Furthermore, predictive ESG analytics enable early identification of climate-related and transition risks, thereby supporting innovations and transition. By providing insights into ESG performance for companies lacking current scores, predictions facilitate the incorporation of sustainability factors into investment decision-making. This aligns capital flows with the principles for responsible investment and advances sustainable finance in emerging markets.

Finally, this approach enhances access to ESG ratings for external stakeholders. By making ESG predictions publicly available, it fosters greater transparency and accountability, enabling a broader audience–including consumers, regulators, and civil society–to better understand and actively engage with corporate sustainability initiatives.

This paper makes a significant contribution to ESG research by introducing a data-driven methodology designed to improve the reliability and consistency of ESG ratings. Given the current challenges in ESG evaluation — including methodological inconsistencies and divergent scoring criteria across rating agencies — our approach offers regulators and rating agencies an automated tool that facilitates standardized, transparent, and objective assessment of ESG scores.

7. Limitations

A key limitation of the proposed model concerns the availability of transitive-domain data. In cases where similar data constraints exist but multiple candidate transitive domains are available, a selection algorithm based on domain characteristics should be developed. From a computational perspective, several additional limitations should be acknowledged. First, the parameter-transfer procedure assumes static model parameters and does not include adaptive fine-tuning in the transitive domain, which may restrict the model's flexibility in evolving market conditions. Second, the proposed model does not incorporate explicit uncertainty estimation, which could enhance the robustness of predictions under noisy or incomplete data. Third, the absence of a formal bias and fairness assessment represents a limitation, as ESG data may embed regional or structural biases that influence model outcomes.

Furthermore, some limitations are associated with the selected feature set. Macroeconomic conditions significantly influence the ESG scores of companies in emerging markets by shaping both their external operating environment and their internal capacity for ESG integration. Future research should therefore examine the effects of key economic variables—such as GDP growth, exchange rate volatility, interest rates, and inflation—since these factors have been shown to affect ESG performance in diverse ways [2].

Additionally, ESG ratings from different agencies often exhibit systematic bias arising from variations in rating objectives, the number of input variables, and methodological frameworks. Each agency applies distinct evaluation criteria, weighting schemes, and data

sources, which leads to inconsistent assessments across providers [1, 13]. Ongoing regulatory efforts in the European Union aim to address these issues by introducing greater transparency and standardization in ESG rating activities. The forthcoming framework is expected to enhance the reliability and comparability of ESG scores, strengthen their alignment with firms' financial indicators, and support more consistent evaluations of corporate sustainability.

Nonetheless, given the complex and context-dependent relationships between business operations and ESG performance — particularly in developing markets — future research should prioritize the identification and refinement of indicators that better capture market dynamics and macroeconomic trends relevant to ESG outcomes.

8. Conclusion

Sustainability-related data is often qualitative and challenging to compare, producing inconsistent information for key stakeholders. Therefore, independent ratings have become the most widely used method for evaluating ESG performance. Companies without an external ESG score, especially in developing markets, are disadvantaged, lowering trading volume and their ability to attract financial resources. Therefore, this study proposed an approach to determining initial ESG scores in developing capital markets.

The results of this study demonstrate that MPBTL-RF is an effective and computationally efficient model for predicting ESG scores in data-constrained environments. Across five-fold in-domain evaluations, the model achieved an average MAE of 14.72 for the overall ESG score, comparable to the best-performing baseline (MAE = 14.87). In out-of-domain testing, MPBTL-RF achieved an average MAE of 17.51 on the holdout dataset and 18.43 on the synthetic dataset, showing robust generalization despite substantial domain shifts. Compared to traditional models such as RF and XGB, MPBTL-RF required significantly less training time (4.5 s vs. $>$ 26.8 s) while delivering comparable predictive accuracy.

These findings confirm that parameter transfer from a well-trained source model can enhance performance in small-sample and heterogeneous market conditions without extensive retraining. The success of this approach underscores the importance of using high-quality transitive data sets and suggests that other machine learning approaches, such as XGB, may benefit from similar approaches to improve prediction accuracy further.

Applying this model to the BELEX case study further demonstrated its practical value by generating consistent ESG estimates for frontier-market companies. This study contributes to the growing knowledge on ESG prediction in underrepresented regions and provides a foundation for future research and applications in similar markets.

The results should be interpreted as demonstrating effectiveness under data-constrained settings rather than universal superiority across domains.

Future work will extend this model through adaptive parameter tuning and the integration of macroeconomic indicators to improve prediction accuracy and interpretability across diverse economic contexts. Furthermore, future studies will explore the integration of source-free domain adaptation and unsupervised domain adaptation approaches to improve model predictions as well as the integration of transfer and active learning methods.

References

1. Berg, F., Kölbel, J.F., Rigobon, R.: Aggregate confusion: The divergence of esg ratings. Review of Finance 26(6), 1315–1344 (2022)
2. Bilivogui, P., Iqbal, M.A.: Do esg scores matter? an empirical analysis of corporate financial performance in brics economies. Environmental Research Communications (2025)
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
4. Chowdhury, M.A.F., Abdullah, M., Azad, M.A.K., Sulong, Z., Islam, M.N.: Environmental, social and governance (esg) rating prediction using machine learning approaches. Annals of Operations Research pp. 1–25 (2023)
5. D'Amato, V., D'Ecclesia, R., Levantesi, S.: Fundamental ratios as predictors of ESG scores: A machine learning approach. Decisions in Economics and Finance 44(2), 1087–1110 (2021)
6. D'Amato, V., D'Ecclesia, R., Levantesi, S.: Esg score prediction through random forest algorithm. Computational Management Science 19(2), 347–373 (2022)
7. García, F., González-Bueno, J., Guijarro, F., Oliver, J.: Forecasting the environmental, social, and governance rating of firms by using corporate financial performance variables: A rough set approach. Sustainability 12(8), 3324 (2020)
8. Huang, J., Guan, D., Xiao, A., Lu, S.: Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. Advances in neural information processing systems 34, 3635–3649 (2021)
9. Jin, Y., Acquah, M.A., Seo, M., Han, S.: Short-term electric load prediction using transfer learning with interval estimate adjustment. Energy and Buildings 258, 111846 (2022), elsevier
10. Karnyoto, A.S., Sun, C., Liu, B., Wang, X.: Transfer learning and gru-crf augmentation for covid-19 fake news detection. Computer Science and Information Systems 19(2), 639–658 (2022)
11. Kokol, P., Kokol, M., Zagoranski, S.: Machine learning on small size samples: A synthetic knowledge synthesis. Science Progress 105(1) (2022)
12. Krappel, T., Bogun, A., Borth, D.: Heterogeneous ensemble for ESG ratings prediction (2021)
13. Larcker, D.F., Pomorski, L., Tayan, B., Watts, E.M.: Esg ratings: A compass without direction. Rock Center for corporate governance at Stanford University working paper forthcoming (2022)
14. Lim, T.: Environmental, social, and governance (esg) and artificial intelligence in finance: State-of-the-art and research takeaways. Artificial Intelligence Review 57(4), 76 (2024)
15. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22(10), 1345–1359 (2009), iEEE
16. Pinto, G., Messina, R., Li, H., Hong, T., Piscitelli, M.S., Capozzoli, A.: Sharing is caring: An extensive analysis of parameter-based transfer learning for the prediction of building thermal dynamics. Energy and Buildings 276, 112530 (2022), elsevier
17. Sahin, E.K.: Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using xgboost, gradient boosting machine, and random forest. SN Applied Sciences 2(7), 1308 (2020)
18. Segev, N., Harel, M., Mannor, S., Crammer, K., El-Yaniv, R.: Learn on source, refine on target: A model transfer learning framework with random forests. IEEE transactions on pattern analysis and machine intelligence 39(9), 1811–1824 (2016), iEEE
19. Shim, E., Kammeraad, J.A., Xu, Z., Tewari, A., Cernak, T., Zimmerman, P.M.: Predicting reaction conditions from limited data through active transfer learning. Chemical science 13(22), 6655–6668 (2022), royal Society of Chemistry
20. Sokolov, A., Mostovoy, J., Ding, J., Seco, L.: Building machine learning systems for automated esg scoring. The Journal of Impact and ESG Investing 1(3), 39–50 (2021)

21. Sukhija, S., Krishnan, N.C.: Supervised heterogeneous feature transfer via random forests. Artificial Intelligence 268, 30–53 (2019), elsevier
22. Tan, B., Song, Y., Zhong, E., Yang, Q.: Transitive transfer learning. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1155–1164 (2015)
23. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data 3(1), 9 (2016)
24. Wu, B., Zhang, T., Mao, L.: Large-scale image classification with multi-perspective deep transfer learning. Computer Science and Information Systems 20(2), 743–763 (2023)

APPENDIX

Table A1. Kruskal-Wallis test with a Dunn-Bonferroni post hoc analysis results for TECLP, holdout, and BELEX dataset

Indicator	Group 1	Group 2	Statistics	p-value	p-value adj.
Sales_to_Assets	BELEX	TECLP	-1.28	0.199	0.598
Sales_to_Assets	BELEX	holdout	-2.35	0.019	0.056
Sales_to_Assets	TECLP	holdout	-1.92	0.055	0.163
EBIT_to_Sales	BELEX	TECLP	4.67	0.000	0.000
EBIT_to_Sales	BELEX	holdout	4.43	0.000	0.000
EBIT_to_Sales	TECLP	holdout	2.55	0.011	0.032
DY	BELEX	TECLP	-0.909	0.364	0.999
DY	BELEX	holdout	3.88	0.000	0.000
DY	TECLP	holdout	4.61	0.000	0.000
NI_to_Sales	BELEX	TECLP	3.65	0.000	0.001
NI_to_Sales	BELEX	holdout	4.39	0.000	0.000
NI_to_Sales	TECLP	holdout	3.00	0.003	0.008
P/E	BELEX	TECLP	7.27	0.000	0.000
P/E	BELEX	holdout	0.048	0.962	0.999
P/E	TECLP	holdout	-3.41	0.001	0.002
Rating	BELEX	TECLP	-2.08	0.038	0.113
Rating	BELEX	holdout	0.316	0.752	0.999
Rating	TECLP	holdout	1.33	0.184	0.552
LR	BELEX	TECLP	-0.291	0.771	0.999
LR	BELEX	holdout	-0.989	0.323	0.968
LR	TECLP	holdout	-0.927	0.354	0.999
SR	BELEX	TECLP	-7.42	0.000	0.000
SR	BELEX	holdout	-4.76	0.000	0.000
SR	TECLP	holdout	-1.61	0.108	0.325

Note: Significance level (α) set at 0.05.

Table A2. Friedman test results over folds on transitive, holdout and synthetic datasets

Model	Test Statistic	p-value
Transitive dataset - IND evaluation		
ESG	8.687	0.069
E*	10.122	0.038
S	2.400	0.663
G	6.240	0.182
Target dataset holdout - ODD evaluation		
ESG	6.586	0.160
E	5.479	0.242
S*	16.160	0.003
G*	11.360	0.023
Target dataset Synthetic - ODD evaluation		
ESG	5.200	0.074
E	2.800	0.247
S	5.200	0.074
G*	6.400	0.048

Note: Significance level (α) set at 0.05.

Ivana Marković is an Assistant Professor in the Department of Accounting, Mathematics, and Informatics at the Faculty of Economics, University of Niš, specializing in Business Informatics. Her research focuses on machine learning methods, feature and instance selection algorithms and their applications in economics, as well as the application of advanced algorithms in optimization problems, mainly computer modelling in economics. Her work also explores transfer learning, business information systems, business intelligence, and generative AI.

Adela Ljajić is a Research Associate at the Institute for Artificial Intelligence Research and Development of Serbia. Her research spans machine learning, natural language processing, generative AI, retrieval-augmented generation, and AI evaluation. Much of her work focuses on adapting AI methods to low-resource, multilingual, and data-constrained settings, with an emphasis on reliable outputs and practical evaluation. Her broader interests include trustworthy AI, transfer learning, and real-world AI applications.

Jelena Z. Stanković is an Associate Professor at the University of Niš, Faculty of Economics, in the narrow scientific field of Accounting, Auditing, and Financial Management. Her research interests include sustainable investing, corporate financing, and risk management in finance and insurance. Her work focuses on integrating sustainable finance principles with quantitative risk assessment in corporate finance and insurance systems.

Miloš Košprdić was a Researcher at the Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, working in natural language processing (NLP) and large language models (LLMs), with a focus on semantic text search and scientific question answering. He is also a Senior Associate at the Linguistics Department of Petnica Science Center, where he previously served as Head of the Department. His work spans claim verification, natural language inference, and sentiment analysis of Serbian-language texts.

Jovica Stanković is an Assistant Professor in the Department of Accounting, Mathematics, and Informatics at the Faculty of Economics, University of Niš, specializing in Business Informatics. His research focuses on information systems, data science, data analytics, and visualization. His interests include information technology, business information systems, business intelligence, and AI.

Received: June 16, 2025; Accepted: February 28, 2026.