Research on Asset Return Rate Prediction of Listed Companies Based on Random Forest and XGBoost Hybrid Artificial Intelligence Algorithm

Yong Xiong[1,2], Zhiming Wu[3,*], Liu'an He[1], Xiaoyan Zhang[1], Zhu Zheng[4], Fang Lan[1], Hui Mi[1], Yijia Liu[1], Zongjun Lan[1], Jinglin Huang[1], Rui Li[1], Meiyan Pang[1]

[1] Department of Accounting, Guangzhou College of Technology and Business, Guangzhou, Guangdong, China

[2] Seokyeong University, Seoul, South Korea

[3] Department of Accounting, Anhui Business and Technology College, Hefei, Anhui, China

[4] Hainan Vocational University of Science and Technology, Haikou, Hainan, China

**Corresponding Author*

Abstract: To improve the prediction accuracy of return on assets (ROA) for listed companies and better support financial performance evaluation, risk early warning, and investment decision-making, this study proposes a hybrid artificial intelligence prediction model that combines Random Forest and Extreme Gradient Boosting (XGBoost). Based on 76,567 financial panel data of Chinese listed firms from 2001 to 2022 covering real estate, manufacturing, transportation, and comprehensive industries, we select 44 financial indicators reflecting solvency, operation capacity, profitability, and growth ability as input features. After rigorous preprocessing including column name standardization, IQR outlier detection, industry-median missing value imputation, and lagged feature construction, 70,845 valid observations are finally obtained. The two-stage model uses Random Forest for base prediction and feature selection, and adopts XGBoost to fit residuals and reduce prediction errors. Empirical results show that the hybrid model achieves a test R^2 of 0.9279, MAE of 0.004960, and RMSE of 0.011964, significantly outperforming traditional decision tree and linear regression models. With strong nonlinear fitting and autonomous learning ability, this model provides reliable technical support for financial data analysis and intelligent decision-making of listed companies.

Keywords: Artificial Intelligence; Integrated Learning; Random Forest; XGBoost; Return on Assets; Financial Forecasting

1. Introduction

Return on Assets (ROA) serves as a key metric for evaluating corporate asset utilization efficiency and profitability. Accurate forecasting of ROA is crucial for business management, investor decision-making, and financial risk mitigation.

In China's capital market, the characteristics of information asymmetry and frequent financial fluctuations make accurate prediction of corporate profitability particularly important. Traditional financial forecasting often employs linear regression, time series analysis, and single decision tree models. These methods assume stringent conditions, have limited nonlinear fitting capabilities, and struggle to capture complex correlations between financial indicators. Moreover, single decision tree models are prone to overfitting issues and are not strictly artificial intelligence algorithms.

With the deepening application of artificial intelligence technology in the financial sector, ensemble learning algorithms represented by Random Forest and Extreme Gradient Boosting (XGBoost) have become mainstream methods in financial forecasting research [1-4]. These algorithms leverage advantages such as self-learning capabilities, strong noise resistance, and adaptability to high-dimensional data. Random Forest constructs multiple decision trees through self-sampling and random feature selection to effectively mitigate overfitting risks. XGBoost, based on the gradient boosting framework, incorporates regularization and second-order Taylor expansion techniques to significantly enhance model convergence speed and prediction accuracy.

This study integrates Random Forest and

XGBoost to develop a residual fusion hybrid AI prediction model, conducting empirical research using financial data from listed companies. The objective is to provide more efficient and accurate AI solutions for financial indicator forecasting.

2. Literature Review

In the field of financial forecasting, traditional statistical methods rely on linear assumptions and are unable to accommodate the nonlinear and non-stationary characteristics of financial data. Superficial models such as decision trees and logistic regression exhibit simple structures and strong interpretability, but their prediction accuracy and generalization capabilities are insufficient to meet the demands for high-precision forecasting.

In recent years, artificial intelligence ensemble learning algorithms have emerged as a research hotspot [5-7]. The ensemble learning framework integrates multiple base learners, which can significantly enhance model stability and prediction effectiveness. In corporate financial performance forecasting studies, random forests can automatically output feature importance to realize key indicator screening [4], while XGBoost performs well in credit risk assessment and financial distress early warning scenarios [2,3]. Existing studies have verified the effectiveness of gradient boosting algorithms when modeling financial data for listed companies [2,3,8].

Previous studies predominantly employed single integrated algorithms for prediction, with relatively few investigations combining random forest and XGBoost to construct high-order hybrid models through residual fusion [8-9]. Building upon this foundation, this study further enhances the prediction accuracy of ROA.

3. Research Design

3.1 Sample Selection

The data is sourced from the annual financial reports of China's listed companies from 2001 to 2022. The raw data includes stock codes, reporting years, industry classifications, and 44 financial indicators. To ensure data quality, we excluded samples with missing values or anomalies. The final sample covers real estate, manufacturing, transportation, comprehensive sectors, and other major industries, representing the overall situation of China's listed companies.

3.2 Data Preprocessing

Prior to model training, the data preprocessing workflow adhered to standardized scientific research protocols, as illustrated in Figure 1. The specific steps are as follows:

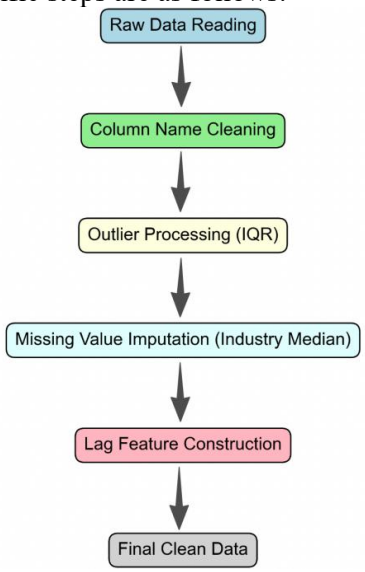

Figure 1. Data Preprocessing Flowchart

(1) Raw data reading: Import original financial panel data from listed companies.

(2) Column name cleaning: Remove special characters such as spaces, line breaks, and full-width spaces from column names to ensure standardized feature names.

(3) Outlier handling: The interquartile range (IQR) method is employed to detect outliers, with the median of this metric used for replacement to eliminate interference from extreme values. This approach is widely applied in financial data processing and effectively preserves the distribution characteristics of the data.

(4) Missing value imputation: By industry grouping, missing values are filled using the median of intra-group indicators to preserve the industry heterogeneity in financial data, mitigate the impact of outliers, and ensure data stability.

(5) Lagged feature construction: First-period lagged features of return on assets (ROA), total asset turnover ratio, and debt-to-asset ratio are constructed to capture the temporal patterns of corporate performance.

(6) Data screening: Invalid samples were excluded, resulting in 70,845 valid observations with a data cleaning rate of 7.47%.

(7) Feature standardization: Standardize all financial features to eliminate dimensionality

issues and enhance model training efficiency.

Through the aforementioned steps, we can maximize the retention of data distribution characteristics and eliminate noise interference, thereby providing high-quality modeling data for subsequent artificial intelligence models. Table 1 presents descriptive statistics of the main financial variables.

Table 1. Descriptive Statistics of the Main Financial Variables

Variable Name	Sample Count	Mean	Std. Dev.	Min	25%	Median	75%	Max
Return on Assets	70845	0.049830	0.044634	-0.084413	0.020955	0.046250	0.073363	0.185831
Total Asset Turnover	70845	0.577653	0.315163	-0.128290	0.348522	0.546240	0.750293	1.565852
Debt-to-Asset Ratio	70845	0.437234	0.202632	-0.126999	0.281994	0.436044	0.583837	1.049286
Return on Equity	70845	0.091934	0.072602	-0.131207	0.045870	0.085941	0.127287	0.316883

3.3 Variable Definition

This study uses return on assets (ROA) as the dependent variable and selects 44 financial indicators covering solvency, operational efficiency, profitability, and growth potential as independent variables. Specific definitions are detailed in Table 2.

Table 2. Variable Definition Table

Variable	Definition	Type
ROA	Return on Assets = Net Income / Total Assets	Dependent
Total Asset Turnover	Total Operating Revenue / Average Total Assets	Independent
Quick Ratio	(Current Assets - Inventory) / Current Liabilities	Independent
Current Ratio	Current Assets / Current Liabilities	Independent
Cash Ratio	Cash and Cash Equivalents / Current Liabilities	Independent
Debt-to-Asset Ratio	Total Liabilities / Total Assets	Independent
Long-term Debt to Assets	Long-term Liabilities / Total Assets	Independent
Equity Multiplier	Total Assets / Total Equity	Independent
Accounts Receivable Turnover	Revenue / Average Accounts Receivable	Independent
Inventory Turnover	Cost of Goods Sold / Average Inventory	Independent
Current Assets Turnover	Revenue / Average Current Assets	Independent
Fixed Assets Turnover	Revenue / Average Fixed Assets	Independent
Gross Profit Margin	Gross Profit / Total Revenue	Independent
EBIT	Earnings Before Interest and Taxes	Independent
ROE	Return on Equity = Net Income / Total Equity	Independent
Operating Revenue Growth Rate	Growth Rate of Operating Revenue	Independent
Sustainable Growth Rate	Sustainable Growth Rate Indicators	Independent

4. Construction of Hybrid Artificial Intelligence Prediction Model

4.1 Random Forest Model

Random forest is an ensemble learning algorithm based on the Bagging framework [1,4]. It generates multiple subsets through random sampling and trains multiple decision trees in parallel, outputting predictions via averaging. Its advantages include strong resistance to overfitting, automatic computation of feature importance, and suitability for high-dimensional financial data, making it a typical artificial intelligence-based learner.

4.2 XGBoost model

XGBoost (Extreme Gradient Boosting) is a high-performance artificial intelligence algorithm based on gradient boosting decision trees [3]. It iteratively trains weak learners to progressively reduce prediction errors, introduces regularization terms to control model complexity, and employs second-order Taylor expansion to enhance optimization accuracy. The algorithm demonstrates strong nonlinear fitting capabilities and autonomous learning capacity.

4.3 Mixed Model Structure

This study proposes a two-stage residual fusion hybrid artificial intelligence model, with its architecture illustrated in Figure 2.

The hybrid model framework comprises four core steps:

(1) Data input: Financial indicators and return on assets (ROA) of listed companies are collected as input data.

(2) First-stage prediction: Random Forest is employed to perform baseline prediction and

feature importance ranking, generating preliminary asset return forecasts and calculating residuals.

(3) Second-stage prediction: Targeting residual prediction, extreme gradient boosting (XGBoost) is employed for residual fitting and error optimization.

(4) Final prediction output: By combining the base learner's predicted values with residual correction values, we obtain high-precision asset return rate predictions, thereby leveraging the complementary strengths of ensemble learning algorithms.

(5) Model evaluation: The predictive performance of the models was assessed by comparing the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE).

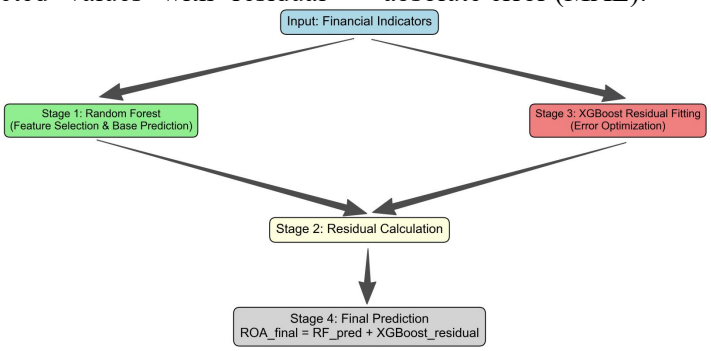

Figure 2. Flowchart of Hybrid Model Structure for Random Forest and XGBoost

Specifically, the first stage employs random forest to perform baseline asset return prediction and screen key features, laying the foundation for the model. The second stage utilizes extreme gradient boosting (XGBoost) to fit the residuals generated in the first stage, thereby correcting the prediction errors of the base model. The final prediction formula is as follows:

ROAfinal = RFpred + XGBoostresidual.

This hybrid architecture combines the strengths of random forests in feature selection and overfitting prevention with those of extreme gradient boosting (XGBoost) in error optimization and nonlinear fitting, significantly improving the accuracy of asset return rate predictions. Figure 3 presents the residual fitting performance of the XGBoost model.

The residuals exhibit a uniform distribution around zero, demonstrating that XGBoost effectively captures unexplained residual information from random forests and significantly improves final prediction accuracy. The study develops a hybrid model combining Random Forest and XGBoost residual fusion, using preprocessed financial indicators as input features and asset returns as the prediction target. The Random Forest performs initial prediction to generate baseline forecasts, then calculates residuals between these baseline values and actual measurements. These residuals serve as new targets for XGBoost fitting, resulting in final predictions calculated as Random Forest baseline forecasts plus XGBoost residual forecasts.

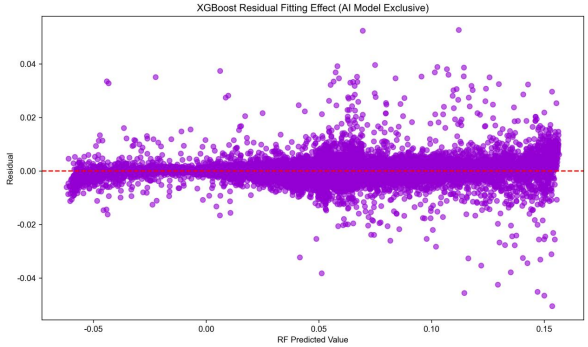

Figure 3. XGBoost Residual Fitting Performance

4.4 Model Training and Evaluation

The study employed a 7:3 ratio to partition the training and testing datasets, with stratified sampling conducted by industry categories to ensure consistent data distribution and enhance experimental reliability. Model parameters were optimized through a combination of grid search (GridSearchCV) and 3-fold cross-validation, using R^2 as the evaluation metric to automatically identify optimal parameter combinations. For Random Forest, the optimal parameters were set at decision trees ranging from 50 to 100 with maximum depth between 5 and 9. For XGBoost, the optimal parameters included learning rates between 0.05 and 0.1, with maximum depth ranging from 3 to 5.

The model performance evaluation employs four metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Determination Coefficient (R^2), and Adjusted Determination

Coefficient (R^2_adj). MAE and RMSE quantify the absolute magnitude of prediction errors, while R^2 measures the model's explanatory power for data variability—with values closer to 1 indicating better model fit. Adjusted R^2 eliminates the influence of feature quantity on model assessment, making it particularly suitable for comparing models with high-dimensional financial datasets.

Compared to traditional decision tree models, the hybrid model developed in this study demonstrates significant advantages: First, random forests mitigate overfitting risks and enhance model stability by integrating multiple decision trees. Second, XGBoost further explores nonlinear relationships through gradient boosting and regularization techniques, improving prediction accuracy. Third, the residual fusion strategy enables the model to correct fundamental prediction biases, achieving dual optimization. This architecture fully leverages the self-learning capabilities and feature extraction potential of ensemble learning, establishing it as a hallmark framework for artificial intelligence prediction systems.

5. Experimental Results and Analysis

5.1 Core Code

The following code demonstrates the core architecture of the proposed hybrid prediction model. The model initially employs a random forest as the base learner to generate preliminary predictions, followed by the application of the XGBoost algorithm to fit and correct prediction residuals. Ultimately, the model outputs comprehensive prediction results, with performance evaluated using R^2 and RMSE metrics. The detailed core code is shown in Figure 4.

```
# Hybrid Model: Random Forest + XGBoost for ROA Prediction
target = 'Return on Assets'
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Train Random Forest base model with grid search
rf = RandomForestRegressor(random_state=42, n_jobs=-1)
rf_opt = GridSearchCV(rf, param_grid_rf, cv=3).fit(X_train, y_train)
rf_pred_train = rf_opt.best_estimator_.predict(X_train)
residual = y_train - rf_pred_train

# Train XGBoost for residual fitting
xgb = XGBRegressor(random_state=42, n_jobs=-1)
xgb_opt = GridSearchCV(xgb, param_grid_xgb, cv=3).fit(X_train, residual)

# Final prediction of the hybrid model
xgb_pred = xgb_opt.best_estimator_.predict(X_test)
final_pred = rf_opt.best_estimator_.predict(X_test) + xgb_pred

# Model evaluation
r2 = r2_score(y_test, final_pred)
rmse = np.sqrt(mean_squared_error(y_test, final_pred))
```

Figure 4. Core Implementation of the Proposed Hybrid Model

5.2 Model Performance Analysis

Research findings indicate that the model achieves an R^2 coefficient of 0.9279 on the total test set, demonstrating its ability to explain approximately 92.79% of ROA variance. This performance significantly outperforms traditional models in prediction accuracy, showcasing exceptional nonlinear fitting capabilities and generalization performance. Figure 5 illustrates the predictive outcomes of the hybrid AI model on the test set. The dense clustering of data points around the diagonal highlights the model's high degree of fit.

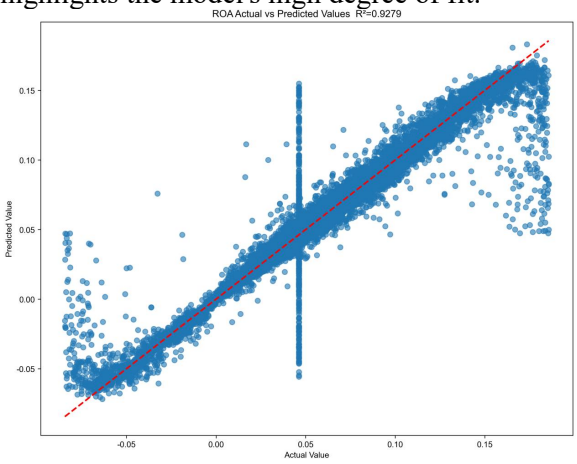

Figure 5. Comparison of True Values and Predicted Values of ROA for the test Set

The study found that the prediction error approximately follows a normal distribution, as shown in Figure 6, with a concentration around 0 and no significant skewness or fat-tail phenomenon, indicating stable model predictions and controllable errors.

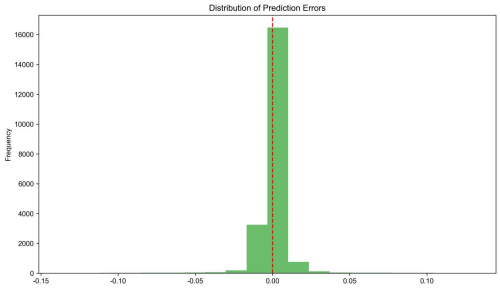

Figure 6. Histogram of Model Prediction Error Distribution

5.3 Feature Importance Analysis

The feature importance ranking derived from the random forest model indicates that key factors influencing listed companies' Return on Assets (ROA) include total asset return rate, current asset return rate, and net asset return rate, which aligns closely with financial theories.

Figure 7 presents the top ten core financial characteristics identified in this random forest model test.

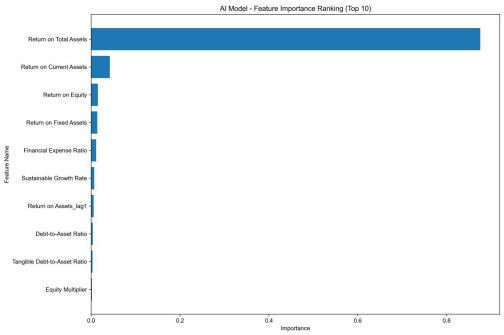

Figure 7. Top 10 Importance Rankings of Core Financial Characteristics

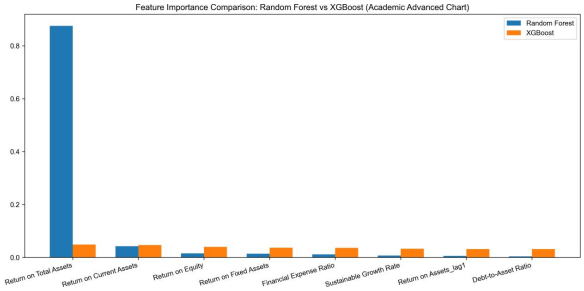

Figure 8. Comparison of Feature Importance between Random Forest and XGBoost

Comparative analysis of XGBoost revealed that both models exhibited consistent trends in identifying core features, indicating stable and consistent judgment of key influencing factors by the models. Figure 8 presents a comparison of feature importance between random forest and XGBoost.

Based on the aforementioned research findings, this study identifies the core indicators influencing the return on assets of listed companies as follows: return on total assets, return on current assets, return on equity, return on fixed assets, and financial expense ratio, as shown in Table 3. These results align with financial theories, thereby validating the model's rationality.

Table 3. Lists the top 10 Important Features Obtained by the Random Forest Model.

Rank	Feature Name	Importance
1	Return on Total Assets	0.875762
2	Return on Current Assets	0.042377
3	Return on Equity	0.015343
4	Return on Fixed Assets	0.013857
5	Financial Expense Ratio	0.011352
6	Sustainable Growth Rate	0.007047
7	Return on Assets_lag1	0.005723
8	Debt-to-Asset Ratio	0.003933
9	Tangible Debt-to-Asset Ratio	0.003080
10	Equity Multiplier	0.001851

5.4 Model Comparison Analysis

To validate the model's superiority, this study established a traditional single decision tree as the benchmark model, as shown in Table 4. Comparative results demonstrate that while the decision tree model achieved an R2 score of 0.8847 on the test set, our hybrid artificial intelligence model improved the R2 score to 0.9279 with concomitant reductions in all error metrics.

Table 4. Comparison between Hybrid Artificial Intelligence Model and Traditional Decision Tree Model

Model	R^2	MAE	RMSE
Proposed Random Forest + XGBoost Hybrid	0.927920	0.004960	0.011964
Traditional Decision Tree	0.884700	0.007735	0.015133

All experimental data were derived from the same dataset, following identical partitioning rules and preprocessing procedures, ensuring strict comparability of results. The hybrid model demonstrated significant advantages in nonlinear relationship mining, generalization capability, and prediction stability, validating the efficacy of residual fusion-based artificial intelligence algorithms [8-10].

6. Conclusion

This study develops a hybrid AI prediction model integrating Random Forest and XGBoost to achieve high-precision asset return forecasting for listed companies. Experimental results demonstrate the model's robust nonlinear fitting capabilities and self-learning capacity, achieving an R^2 value of 0.9279 on the test set. With exceptional predictive accuracy and strong generalization performance, the model effectively supports corporate financial performance evaluation, risk early warning systems, and investment decision-making processes.

Future research could integrate deep learning algorithms with multi-source data including macroeconomic indicators and market sentiment to develop multimodal fusion prediction models, thereby expanding their application scenarios and enhancing predictive performance.

Acknowledgements

This work was funded by the research project, Guangzhou Business School Federation of

Social Sciences Humanities and Social Sciences "Data Element-driven Financial Sharing Transformation and Business Environment Optimization of Small and Medium-sized Enterprises - Guangzhou Sample" (SKKYYB202515); "Research on the Cultivation of New Quality Productivity of Strategic Emerging Industries in Anhui Province from the Perspective of Resource Allocation Efficiency" (2025AHGXSK30388); "Anhui Province University Outstanding Talents Support Program Key Project" (JNFX2025178); Anhui Vocational and Adult Education Association Education Research Planning Project "Research on Inquiry-based Learning Mode of Artificial Intelligence in Higher Vocational Business Education" (AZCJ2025123); Anhui Provincial Humanities and Social Sciences Research Project "Research on How Fiscal and Taxation Policies Drive the High-quality Development of New Quality Productive Forces (SK2024A011); Anhui Province University Humanities and Social Sciences Key Research Project "Research on the Digital Empowerment of Anhui New Energy Vehicle Enterprises to Enhance Green International Competitiveness under the Background of "Double Carbon" (2025AHGXSK30221); AnHui Business and Technology College Teaching and Research Project "Smart Inquiry Learning: Research on a New Model of Conversational Learning for Taxation Based on Deepseek" (2024xjjy10); China Institute of Business Accounting Research Project: Research on a New Model of Conversational Learning for Accounting Based on Generative Artificial Intelligence (2025ZJ001).

References

[1] Braga, A., & Neto, P. (2022). Firm performance prediction using ensemble learning: A comparative study. Journal of Business Research, 142, 496-507.

[2] Wen, X., Li, Y., & Zhang, H. (2023). Financial distress prediction based on XGBoost and Random Forest ensemble. Expert Systems with Applications, 213, 119210.

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[5] Du, X., & Tan, Y. (2024). Forecasting ROA using machine learning: Evidence from listed firms. Emerging Markets Finance & Trade, 60(3), 786-804.

[6] Hastie, T., Tibshirani, R., & Friedman, J. H. (2021). Ensemble learning in finance: A review. Annual Review of Financial Economics, 13, 31-55.

[7] Louzada, F., Araujo, M., & Fernandes, P. (2022). Machine learning for corporate profitability prediction. Applied Economics, 54(18), 2069-2083.

[8] Li, Y., Wang, Z., & Chen, L. (2025). Hybrid ensemble model for financial performance forecasting. Sustainability, 17(4), 1892.

[9] Wang, X., & Li, Y. (2023). Hybrid ensemble learning for financial performance forecasting. Computational Economics, 61(4), 1123-1145.

[10] Kim, S., & Park, J. (2020). Machine learning vs. traditional regression in profitability forecasting. Journal of Business Research, 115, 342-350.

