Journal of Applied Economic Sciences
Volume XX, Fall, Issue 3(89), 2025

# A SHAP-Based Comparative Analysis of Machine Learning Model Interpretability in Financial Classification Tasks

Chia-Pang CHAN
https://orcid.org/0000-0003-0016-8674
Department of Information Management
Cheng Shiu University[1], Taiwan
1013@gcloud.csu.edu.tw

Chiung-Hui TSAI
https://orcid.org/0000-0002-2250-4715
Department of Computer Science and Information Engineering
Da-Yeh University[2], Taiwan
filamike@mail.dyu.edu.tw

Fang-Kai TANG
https://orcid.org/0009-0001-8200-6101
Department of E-Sports Technology
Cheng Shiu University, Taiwan
0692@gcloud.csu.edu.tw

Jun-He YANG
https://orcid.org/0000-0002-4008-4785
Department of Mass Communication
The Open University of Kaohsiung[3], Taiwan
blackwhale1853@gmail.com

Article's history:

*Received* 20th of June, 2025; *Revised* 18th of July, 2025; *Accepted* 15th of August, 2025; *Available online*: 30th of September, 2025. *Published* as article in the Volume XX, Fall, Issue 3(89), 2025.



Suggested citation:

Chan, C.-P., Tsai, C.-H., Tang, F.-K. & Yang, J.-H. (2025) A SHAP-Based Comparative Analysis of Machine Learning Model Interpretability in Financial Classification Tasks. *Journal of Applied Economic Sciences*, Volume XX, Fall, 3(89), 385 - 400. https://doi.org/10.57017/jaes.v20.3(89).03

Abstract

As artificial intelligence technologies become increasingly prevalent across the financial sector, the interpretability of machine learning models has become a critical concern for regulatory authorities and financial institutions. This study employs SHAP (SHapley Additive exPlanations) to systematically compare the predictive performance and interpretability of five mainstream machine learning models in financial classification tasks. Using a real financial dataset containing 24 financial indicators to train logistic regression, five machine learning models - logistic regression, random forest, XGBoost, LightGBM, and support vector machine - are trained on this dataset. SHAP is then applied to analyse the feature importance patterns

[1] Cheng Qing Road, Kaohsiung City, 833301, Taiwan
[2] University Road, Dacun, Changhua 515006, Taiwan
[3] Daye North Road, Kaohsiung City, 812008, Taiwan

across models. Empirical results demonstrate that LightGBM achieves the best predictive performance (accuracy 95.90%, Area Under the Curve (AUC) 99.18%), while XGBoost shows advantages in terms of interpretability. SHAP analysis identifies those prior earnings per share is the most critical feature, and the Top-K overlap analysis reveals a high degree of consistency among tree-based models in feature importance recognition. This study provides scientific basis for financial institutions to select appropriate explainable AI models, and holds significant importance for enhancing transparency and trustworthiness in financial AI applications.

Keywords: SHAP, explainable artificial intelligence, financial classification, machine learning, feature importance.

JEL Classification: C45, C52, G28, G32.

## Introduction

The advancement of artificial intelligence (AI) technologies within the financial sector has introduced transformative changes in areas such as risk management, credit assessment, and investment decision-making. However, the increasing complexity of machine learning models has amplified their "black box" nature, raising significant concerns among regulatory bodies and financial industry practitioners regarding transparency and interpretability. The European Union's General Data Protection Regulation (GDPR) explicitly mandates that automated decision-making systems provide "meaningful logic disclosure," while the Financial Stability Board (FSB) has cautioned that insufficient interpretability in AI models could pose systemic risks (Khan et al., 2025).

In response to these challenges, Explainable Artificial Intelligence (XAI) has emerged as an essential framework for reconciling high-performance machine learning with regulatory and ethical requirements. Among the various XAI approaches, SHAP (SHapley Additive exPlanations) has become one of the most influential model-agnostic interpretability methods. Rooted in the Shapley value concept from cooperative game theory, SHAP assigns fair and consistent contribution scores to each input feature, enabling a transparent assessment of model behaviour. Its methodological rigor and adaptability have led to its widespread adoption in diverse financial applications (Lundberg & Lee, 2017; Yeo et al., 2025).

This study aims to address the following three main research questions (RQ):

- RQ1: What are the differences in predictive performance among various machine learning models in financial classification tasks? Which model is most suitable for financial classification applications?
- RQ2: Based on SHAP analysis, how do different models identify important features? Which features are most critical to financial classification decisions?
- RQ3: What is the degree of consistency among different models in feature importance identification? How to quantify the interpretive consensus across models?

The main contributions of this study include: (1) systematically comparing the predictive performance and interpretability of five mainstream machine learning models; (2) conducting in-depth analysis of key features in financial classification tasks based on the SHAP method; (3) proposing a Top-K overlap rate metric to quantify interpretive consistency across models; (4) providing empirical evidence for financial institutions in selecting interpretable AI models.

## 1. Literature Review and Theoretical Foundation

### Theoretical Foundation of SHAP Method

The SHAP method originates from the Shapley value concept in cooperative game theory, first applied to machine learning model interpretation by Lundberg & Lee (2017). This method satisfies four axioms - efficiency, symmetry, dummy, and additivity - ensuring fairness and uniqueness in feature contribution allocation. In a recent systematic review, Khan et al. (2025) conducted an in-depth analysis of model-agnostic interpretable AI methods in the financial domain, highlighting the theoretical strengths and practical value of the SHAP method.

Chen et al. (2020) explored the philosophical foundation of SHAP explanations in their seminal study on " true to the model or true to the data", highlighting the importance of balancing model faithfulness and data realism in high-dimensional financial datasets. Benoumechiara et al. (2019) further developed Shapley effects sensitivity analysis algorithms under input-dependent conditions. It provides theoretical support for addressing feature correlation issues commonly encountered in financial data.

### Applications of SHAP in the Financial Domain

In recent years, the application of SHAP methods in the financial sector has rapid growth. Yeo et al. (2025) highlighted in their comprehensive review of explainable AI in finance that SHAP has become the most widely adopted explanation method among financial institutions, particularly playing a crucial role in credit risk assessment, fraud detection, and investment decision-making.

In the area of credit risk assessment, Nallakaruppan et al. (2024) developed a SHAP-based decision support system integrating decision tree and random forest models, achieving a prediction accuracy of 93% on a peer-to-peer (P2P) lending platform. Zhou et al. (2023) proposed a user-centered explainable AI framework that employs SHAP methods to provide both local and global explanations for financial fraud detection, effectively meeting the interpretability needs of diverse stakeholders.

Thanathamathee et al. (2024) innovatively combined SHAP instance weighting with anchor explanation methods to enhance the interpretability of XGBoost in financial fraud detection. Their research demonstrated the effectiveness of SHAP methods in handling imbalanced financial datasets, providing significant insights for practical applications.

### SHAP Analysis Methods and Technical Development

The SHAP method is based on the Shapley value concept from game theory, assigning an importance score to each feature. For a given predictive model f and input feature set N, the SHAP value of feature *i* is defined as Equation (1):

$$\phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup i) - f(S)] \quad (1)$$

*where*: S represents a feature subset that does not contain feature i, and f(S) represents the model prediction value using only feature subset S. This formula ensures four important properties: efficiency, symmetry, dummy, and additivity.

The technical development of SHAP methods primarily focuses on two dimensions: computational efficiency and explanation quality. Lundberg et al. (2020) proposed the TreeSHAP algorithm, which is specifically optimized for tree-based models to enhance computational efficiency. This improvement makes SHAP analysis feasible on large-scale financial datasets. For such models, the computational complexity of TreeSHAP is $O(TLD^2)$, where T is the number of trees, L is the number of leaf nodes, and D is the maximum depth.

Covert and Lee (2021) further refined the KernelSHAP method by employing linear regression techniques to enhance the practicality of Shapley value estimation. KernelSHAP approximates Shapley values through weighted linear regression as Equation (2):

$$min\phi \sum z' \in Z[\pi x(z')(f(h_x(z')) - \phi_0 - \sum i = 1^M z_i' \phi_i)]^2 \quad (2)$$

*where* πx(z') is the weighting function, hx(z') is the mapping function, and Z represents the simplified input space.

Saarela & Podgorelec (2024) found in their systematic literature review of explainable AI applications that SHAP has become the most favoured local explanation method due to its stability and mathematical guarantees. Their study analysed 512 relevant publications, confirming SHAP's extensive application value across domains including finance, healthcare, and industry.

### Applications of Machine Learning Models in Financial Classification

Gradient boosting methods have demonstrated superior performance in financial classification tasks. Nguyen & Ngo (2025) compared the performance of AdaBoost, XGBoost, LightGBM, and CatBoost in personal default prediction, finding that LightGBM achieved optimal performance when handling large and complex datasets. Their study, based on a sample of 7,542 customers from a Vietnamese commercial bank, identified five key factors influencing default risk: monthly liability, credit balance, credit history length, max credit limit, and yearly income.

Alghamdi & Alqithami (2025) proposed a robust machine learning framework for stock market classification, integrating traditional models (logistic regression, random forest, gradient boosting) with deep learning architectures (Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), and Multilayer Perceptrons (MLP)). The results showed that deep learning models improved classification accuracy by 6-12% compared to traditional methods, with LSTM and GRU demonstrating superior capability in capturing temporal dependencies.

Ding et al. (2023) investigated the time-varying characteristics of feature importance in predicting corporate financial distress before and after the COVID-19 pandemic. Their findings revealed that profitability indicators were the dominant factors pre-pandemic, while financial leverage became the most significant predictor post-pandemic. The study, based on an XGBoost-genetic programming framework, provided valuable insights into feature selection under different economic environments.

### Theoretical Development and Practical Challenges of Explainable AI

The theoretical foundation of Explainable AI (XAI) covers multiple academic disciplines. In a comprehensive study on the key elements of trustworthy AI, Ali et al. (2023) identified explainability as one of the core components for achieving trustworthy AI systems, emphasizing that it must be developed in coordination with other critical elements such as fairness, robustness, and privacy protection. Arrieta et al. (2020) proposed a conceptual taxonomy for XAI, providing a theoretical framework for understanding the applicable scenarios of different explanation methods.

Angelov et al. (2021), in their analytical review of Explainable AI, highlighted that the financial sector imposes particularly stringent requirements on explanation methods. These methods should simultaneously satisfy multiple demands including regulatory compliance, risk control, and business decision-making. In a study on explainable FinTech lending, Babaei et al. (2023) explored the practical challenges of applying explanation techniques to real-world financial products, discussing issues such as the accuracy, consistency, and comprehensibility of the generated explanations.

## 3. Research Methodology

### Research Design

This study adopts a systematic comparative analysis framework, employing multi-level empirical analysis to address three core research questions.

First, to investigate model performance comparison, five representative machine learning models are selected for systematic evaluation. These models include linear approaches (logistic regression), ensemble learning methods (Random Forest), gradient boosting techniques (XGBoost and LightGBM), and Kernel-based methods (Support Vector Machines), ensuring methodological comprehensiveness and representativeness.

Second, to examine feature importance identification, the SHAP method is employed to analyse the decision mechanisms of each model in depth. Through the use of TreeSHAP and KernelSHAP algorithms, the contribution of each feature to model predictions is quantitatively assessed, and its economic interpretation is provided in the context of financial theory.

Finally, to address the issue of explanation consistency across models, a Top-K overlap rate evaluation framework is proposed. This framework calculates the degree of overlap in feature importance rankings across different models. By quantifying this overlap, it measures the level of consensus among models and provides a credibility assessment for multi-model decision-making.

### Dataset Description

This study employs a real-world financial dataset containing 24 financial indicators, categorized into five major dimensions: profitability, cash flow, growth capability, solvency, and operational capability. The dataset includes 1,340 samples and has been standardized through pre-processing to ensure comparability across indicators of different scales.

Details of the dataset composition are as follows:

- Profitability indicators (7): Earnings per share (EPS), return on assets (ROA – comprehensive income), return on equity (ROE – comprehensive income), gross margin, net income margin, operating margin, and previous EPS.
- Cash flow indicators (3): Cash flow ratio, cash flow per share, and cash dividend yield.
- Growth capability indicators (4): Revenue growth rate, gross profit growth rate, net income growth rate, and after-tax net income growth rate.
- Solvency indicators (4): Current ratio, quick ratio, debt ratio, and interest coverage ratio.
- Operational capability indicators (3): Accounts receivable turnover, total asset turnover, and fixed asset turnover.
- Other Important indicators (3): Dividend yield, long-term capital adequacy ratio (A), and Tobin's Q ratio.

The dataset also includes company identifiers and time variables, which were processed through Label Encoding before being incorporated into model training. All numerical features underwent standardization using the StandardScaler, with the standardization formula as Equation (3):

$$x_{scaled} = \frac{x-\mu}{\sigma} \tag{3}$$

*where*: μ is the mean and σ is the standard deviation of the feature. Standardization ensures fair and comparable feature contributions during model training, preventing features with larger numerical ranges from dominating the learning process.

### Model Configuration and Evaluation Metrics

All models used a consistent training-testing split ratio (80:20) and cross-validation strategy. The evaluation metrics include accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) values to ensure comprehensive and objective assessment of model performance.

The mathematical definitions of the core evaluation metrics are shown as Equation (4) - (8) as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{7}$$

The AUC value is obtained by calculating the area under the ROC curve:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \quad (8)$$

*where*: TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively; TPR refers to the true positive rate, and FPR to the false positive rate.

### Implementation of SHAP Analysis

To ensure the accuracy and consistency of explanatory results, this study employed the TreeSHAP algorithm for analysing tree-based models and the KernelSHAP algorithm for linear models. Various visualization tools, including the Summary Plot and Beeswarm Plot, were generated to provide an intuitive analysis of feature importance.

### Top-K Overlap Rate Evaluation Method

This study proposes the Top-K overlap rate metric to quantify the degree of consistency between different models in feature importance identification. For two models A and B, their Top-K overlap rate is defined as Equation (9):

$$Overlap_{A,B}(K) = \frac{|TopK_A \cap TopK_B|}{K} \quad (9)$$

*where*: TopKA and TopKB represent the sets of the TopK most important features for model A and model B, respectively, and $|\cdot|$ denotes the cardinality of the set. The overlap rate ranges from [0, 1], *where* a higher value indicates greater consistency between the two models in identifying feature importance.

To evaluate the overall consistency among multiple models, the average pairwise overlap rate is used as Equation (10):

$$Overlap(K) = \frac{2}{M(M-1)} \sum i = 1^{M-1} \sum j = i + 1^{M} Overlap_{i,j}(K) \quad (10)$$

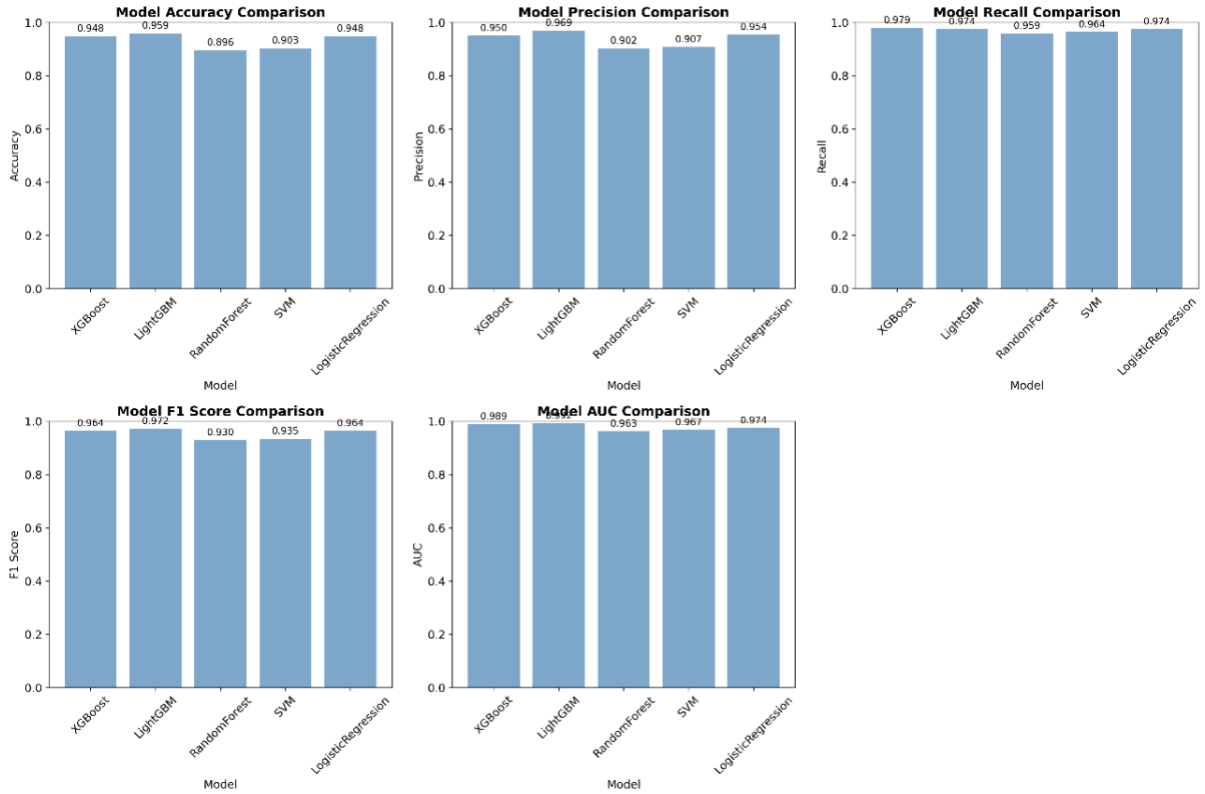
*where*: M is the total number of models. This metric can quantify the degree of consensus among a group of models in feature selection, providing reliability assessment for multi-model decision-making.

## 4. Empirical Results and Analysis

The empirical results reveal significant differences in the performance of five machine learning models on financial classification tasks. LightGBM achieved the highest performance with an accuracy of 95.90% and an AUC of 99.18%, demonstrating its strong capability in handling complex financial data. XGBoost followed closely with an accuracy of 94.78% and an AUC of 98.94%, thereby confirming the effectiveness of gradient boosting methods in financial classification.

It's noteworthy that Logistic Regression, as a linear model, also achieved an accuracy of 94.78%, tying with XGBoost for second place. This suggests the presence of important linear relationships in the financial classification task. Random Forest (89.55% accuracy) and Support Vector Machine (90.30% accuracy) performed relatively weaker but still maintained acceptable levels.

Figure 1: Performance comparison of five machine learning models




Source: Authors

Overall, the empirical evidence indicates that LightGBM offers the best predictive performance and is particularly well-suited for financial classification tasks. Its superior results can be attributed to its leaf-wise tree growth strategy and optimized memory efficiency, making it especially effective for high-dimensional financial data. These findings provide valuable guidance for financial institutions in model selection, particularly in applications where high predictive accuracy is critical.

## In-Depth Analysis of SHAP Feature Importance

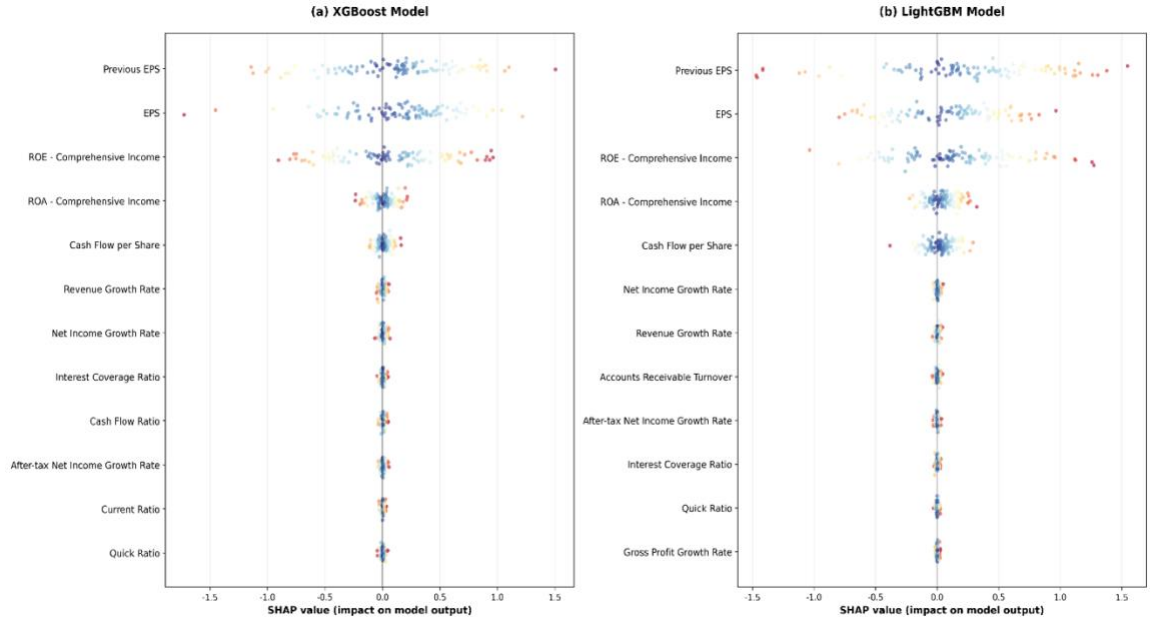
The SHAP analysis reveals key feature patterns in financial classification decisions. As shown in Figure 2(a), in the XGBoost model, Previous EPS emerges as the most important feature, with an average SHAP value of 1.72, highlighting the central role of earnings persistence in financial classification. Other profitability indicators such as EPS, ROE-Comprehensive Income, and ROA-Comprehensive Income also rank highly, confirming the dominant influence of profitability in model decision-making.

In Figure 2(b), the SHAP analysis of the LightGBM model demonstrates a highly consistent feature importance pattern with that of XGBoost. Previous EPS again ranks first, while indicators such as interest coverage ratio, accounts receivable turnover, and long-term capital adequacy ratio highlight the critical roles of solvency and operational efficiency in financial classification.

Comparing Figures 2(a) and 2(b), both models reveal high consistency in feature importance identification. The findings suggest that previous EPS is the most influential factor, with profitability indicators playing a leading role in model decisions. The distribution pattern of SHAP values shows that these features have strong and consistent impacts on model outputs, highly aligning with the earnings persistence hypothesis and efficient market theory. This finding not only validates the effectiveness of traditional financial theories but also provides important guidance for feature engineering in practice.

Figure 2: Comparative Analysis of SHAP Feature Importance for XGBoost and LightGBM Models

(a) SHAP Feature Importance Analysis of XGBoost Model; b) SHAP Feature Importance Analysis of LightGBM Model

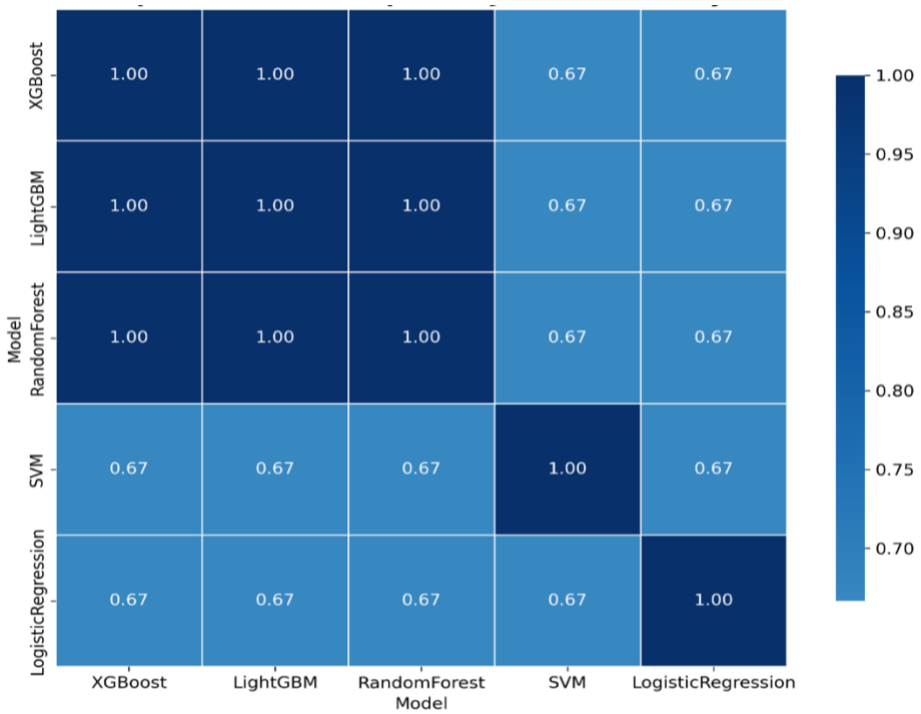



Source: Authors

## Top-K Feature Overlap Rate Analysis

The heatmap of Top-5 feature overlap (Figure 3) illustrates the degree of consistency in feature importance identification across different models. The overlap between XGBoost and LightGBM reaches 100%, indicating complete agreement in feature selection between these two gradient boosting methods. Similarly, the overlap between Random Forest and the gradient boosting models is also 100%, reflecting a high level of consensus among tree-based models.

Figure 3: Top-5 feature overlap



In contrast, logistic regression shows an overlap rate of 66.67% with other models, reflecting the differences in feature importance perception between linear and non-linear models. Support Vector Machine similarly exhibits an overlap rate of 66.67% with other models, highlighting its unique feature selection mechanism.

Figure 3: Heatmap of Top-5 feature overlap rate

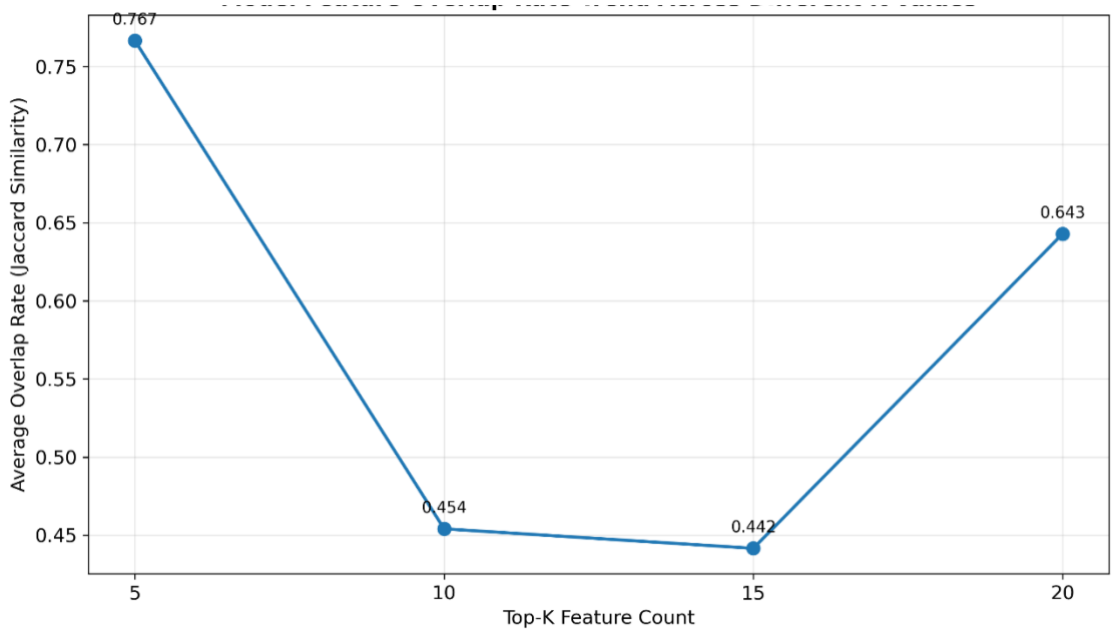
Source: Authors

## Trend Analysis of Top-K Feature Overlap Rate

The trend analysis of Top-K feature overlap rate (Figure 4) reveals the hierarchical structure of feature importance consistency. As the value of K increases, the overlap rate between models exhibits a U-shaped pattern: it decreases from 76.67% at K=5 to a minimum of 44.16% at K=15, and then rebound to 64.29% at K=20. This phenomenon indicates that models demonstrate high consistency in identifying the most important features, exhibit the greatest divergence in ranking features of moderate importance, and show improved consistency again when more features are included.

Empirical analysis shows that tree-based models exhibit high consistency in feature importance identification (100% overlap rate), whereas significant differences exist between linear and non-linear models (66.67% overlap rate). The observed U-shaped trend of Top-K overlap rates uncovers the hierarchical characteristics of feature importance perception, which is of significant importance for understanding the feature selection mechanisms of different algorithms.

Figure 4. Trend of Top-K feature overlap rate




Source: Authors

## Comprehensive Evaluation of Model Interpretability

Considering both predictive performance and interpretability, XGBoost demonstrates the best performance in balancing these two aspects. Although LightGBM slightly outperforms in terms of pure predictive accuracy, XGBoost presents clearer patterns of feature importance and more stable interpretability results in SHAP analysis. Furthermore, while logistic regression has the strongest interpretability, it is limited in its ability to deal with complex non-linear relationships.

## 5. Discussion

### 5.1 Theoretical Implications of the Findings

The findings of this study hold significant theoretical implications. First, the identification of Previous EPS as the most critical feature (mean SHAP value: 1.72) supports the earnings persistence hypothesis, which posits that past earnings performance serves as an important indicator for predicting future financial conditions. This result aligns with the theoretical framework of Fama and French's three-factor model, emphasizing the central role of profitability in asset pricing and risk assessment.

From an economic perspective, the importance of Previous EPS is grounded in the robust theoretical underpinnings of earnings persistence. The hypothesis suggests that past earnings reflect a company's business model sustainability, managerial execution capabilities, and competitive positioning. A study by Fatma & Hidayat (2020), published in the Asian Journal of Accounting Research, empirically examined the effects of earnings persistence and profitability on equity valuation. They found that firms with higher earnings persistence receive valuation premiums, providing direct support for the predictive value of Previous EPS in financial models.

Moreover, according to signalling theory, Previous EPS also carries substantial informational value. In his seminal work published in The Bell Journal of Economics, Bhattacharya (1979) highlighted the role of signalling in corporate finance, arguing that in environments with asymmetric information, investors and creditors rely on observable financial metrics to assess a company's true value and risk profile. A consistent record of earnings serves as a credible signal of operational quality, helping explain why machine learning models consistently identify Previous EPS as a key predictive factor.

The prominence of return on equity (ROE) and return on assets (ROA) in the SHAP analysis can be further understood through the DuPont analytical framework. This framework decomposes ROE into three core components, i.e., net profit margin (reflecting operating efficiency), asset turnover (indicating asset utilization), and equity multiplier (capturing financial leverage), thereby revealing the multidimensional drivers of profitability. Soliman (2008), in a study published in The Accounting Review, provided empirical evidence that investors and analysts actively employ the DuPont analysis to evaluate profitability drivers, and that it serves as a reliable predictor of future stock returns.

Additionally, the importance of the interest coverage ratio revealed through SHAP analysis can be interpreted through the lens of the financial distress cost theory. This ratio measures a company's ability to meet interest obligations using operating income and is directly linked to the risk of financial distress. Altman et al. (2010), in their research published in the Journal of Credit Risk, analysed the value of non-financial information in small and medium enterprise risk management, identifying interest coverage ratio as a key indicator for predicting corporate financial distress.

Second, the high consistency of tree-based models in feature importance identification provides new empirical support for ensemble learning theory. Although XGBoost and LightGBM employ different optimization strategies, their complete consistency in feature selection demonstrates that gradient boosting methods can stably identify genuine signals within the data, thereby providing theoretical justification for model selection.

The non-linear patterns revealed by SHAP analysis reflect complex economic realities, which can be primarily categorized as threshold effects, diminishing returns, and interaction effects. A theoretical basis for the threshold effect can be traced back to the foundational research on bankruptcy costs by Warner (1977) in the Journal of Finance. The study's key insight is that bankruptcy costs increase disproportionately after a company's financial standing weakens beyond a critical point. This non-linear escalation of costs directly causes certain financial metrics to exhibit a sudden change in their predictive impact when they approach this threshold.

Third, the U-shaped trend in Top-K overlap rates reveals the hierarchical characteristics of feature importance recognition, enriching feature selection theory. The agreement among models on extremely important features suggests a shared recognition of strong signals. In contrast, their divergence on moderately important features highlights the differing sensitivities of various algorithms to signal strength.

### 5.2 Practical Application Value

The research findings may offer meaningful practical insights for financial institutions. In terms of model selection, financial institutions can make trade-offs between predictive performance and interpretability based on their specific requirements: LightGBM should be chosen when pursuing the highest prediction accuracy, XGBoost when balancing performance with interpretability, and logistic regression when emphasizing regulatory compliance.

Different financial classification tasks are characterized by distinct economic challenges, which further inform model design and application strategies. In credit scoring, a core tool for risk management, the primary economic challenge is information asymmetry. In credit markets, borrowers possess superior knowledge about their repayment ability and intent compared to lenders. This asymmetry gives rise to problems such as adverse selection and moral hazard. The seminal work by Stiglitz and Weiss (1981), published in the American Economic Review, systematically analysed information asymmetry in credit markets and highlighted how it leads to credit rationing. This market failure underscores the critical role of credit scoring systems in mitigating asymmetric information.

The economic significance of financial distress prediction lies in its capacity to identify and prevent company failures that could trigger market-wide disruptions or systemic risk. The core challenge extends beyond individual company default to understanding its potential contagion effect. Research by Lang & Stulz (1992) in the Journal of Financial Economics provided key insights into this phenomenon, showing how a company's distress can spread through industry affiliations, supply chains, and shared financial exposures, potentially leading to systemic-level consequences. Effective prediction models are therefore vital tools for maintaining financial stability.

Fraud detection presents a unique adversarial economic environment, where fraudsters continuously evolve their strategies to avoid detection mechanisms. The foundational work by Becker (1968) in the Journal of Political Economy laid the groundwork for the economics of crime, positing that fraudulent behaviour is the result of rational cost-benefit analysis. When the expected gains from fraud outweigh the anticipated costs, rational actors may choose to engage in such activities - underscoring the need for adaptable and intelligent detection models.

Regarding feature engineering, profitability indicators such as Previous EPS, ROE, and ROA should be prioritized as core features. Solvency and operational indicators—such as interest coverage ratio and accounts receivable turnover, also provide valuable information and should be included in the feature selection process.

In terms of risk management, the feature contribution analysis provided by SHAP can be utilized to construct early warning systems that issue timely alerts when key features exhibit abnormal changes. Furthermore, Top-K overlap rate analysis facilitates the establishment of model ensemble strategies, enhancing decision reliability through multi-model consensus.

### 5.3 Regulatory Compliance Considerations

In the increasingly stringent financial regulatory environment, this research provides important reference for regulatory compliance. The SHAP method fulfils the GDPR's requirements for explaining automated decisions by providing a detailed analysis of feature contributions for each prediction. The visualization tools demonstrated in the study, such as Beeswarm Plots and Summary Plots, can be directly applied to present model decision logic to regulatory authorities.

Article 22 of the GDPR grants individuals the right to explanation when confronted with automated decision-making processes, establishing explicit interpretability requirements for financial institutions' AI systems. Casey et al. (2019), in their research published in International Data Privacy Law, comprehensively analyzed the legal requirements of GDPR's right to explanation, demonstrating that financial institutions must provide "meaningful information" to explain the logic underlying automated decisions. SHAP methodology effectively fulfills these regulatory mandates by delivering feature importance analyses and individual prediction explanations that meet the substantive requirements for regulatory transparency.

Furthermore, the consistency analysis of feature importance supports the development of a model validation framework. When significant discrepancies in key feature identification are observed across different models, further investigation is warranted to ensure model robustness. This multi-model validation mechanism aligns with the Basel III Accord's requirements for model risk management.

Interpretability also plays a critical role in fostering trust among customers and investors. In their influential study published on arXiv, Doshi-Velez and Kim (2017) explored the need for interpretability in machine learning and provided empirical evidence that users are more likely to trust and adopt model predictions when they can understand the underlying decision-making logic. In the context of financial services, such trust is essential for user acceptance and business success.

### 5.4 Methodological Considerations

Despite the solid theoretical foundations of the SHAP method, several inherent limitations persist in practical applications. In a recent systematic literature review published in Artificial Intelligence Review, Černevičienė and Kabašinskas (2024) identified fundamental issues in SHAP's approach to feature importance estimation. Their findings indicate that when correlations exist among input features, SHAP values may yield misleading attributions, potentially contradicting the true causal relationships. This has significant implications for the current study, given the high degree of multicollinearity commonly found in financial data. The research by Kumar et al. (2020), presented at the International Conference on Machine Learning, further quantifies this limitation of SHAP explanations. Their findings demonstrate that SHAP values fundamentally reflect the statistical associations learned by a model, rather than the true causal relationships in the data-generating process. This distinction between correlation and causation assumes heightened importance within the context of financial analysis. This is because numerous financial indicators possess intrinsic economic connections, creating complex webs of interdependency. Consequently, a high SHAP value for one feature could be an artifact of its correlation with another, truly causal feature, thereby confounding the interpretation and leading to flawed insights. Moreover, financial classification tasks frequently involve high-dimensional feature spaces, which poses additional challenges for SHAP analysis. As stated by Chen et al. (2020), conventional SHAP algorithms face two problems in high-dimensional settings: computational intractability due to the exponential complexity of calculating exact Shapley values, and the inherent complexity of interpreting explanations involving a vast number of features.

To objectively evaluate the advantages and limitations of the SHAP method, this study compares it with other mainstream interpretability techniques. The LIME method, proposed by Ribeiro et al. (2016), offers superior computational efficiency and can rapidly generate local explanations. However, due to the lack of consistency and stability guarantees, LIME explanations are often sensitive to small perturbations in input samples, resulting in potentially unreliable outputs. In contrast, the SHAP method, grounded in cooperative game theory, possesses rigorous theoretical properties such as consistency and completeness, which contribute to more stable and trustworthy explanations. In practical financial applications, SHAP has demonstrated significant value across different use cases. A study by Misheva et al. (2021) showed that its dual capability for global and local explanations offers a more holistic approach to managing credit risk in peer-to-peer lending.

Similarly, Tyagi (2022) underlined the reliability of SHAP for credit scoring and investment decisions. His work showed that SHAP's feature attributions are more consistent and robust when compared to the unstable nature of LIME's explanations. Overall, across contexts that demand model transparency, regulatory alignment, and explanation stability, SHAP demonstrates a more advantageous interpretability profile than LIME (Ribeiro et al., 2016; Misheva et al., 2021; and Tyagi, 2022).

## Conclusions and Recommendations

This study systematically compared the performance of five machine learning models in financial classification tasks, and yielded the following key findings:

- *Finding* 1: LightGBM demonstrated the best predictive performance (accuracy: 95.90%, AUC: 99.18%). XGBoost showed strengths in balancing performance and interpretability, while logistic regression remains valuable in regulatory compliance scenarios.
- *Finding* 2: Previous EPS was identified as the most critical predictive feature (SHAP value: 1.72). Profitability indicators played a dominant role in model decisions. This finding aligns closely with the earnings persistence hypothesis in financial theory. A deeper theoretical analysis highlights the signalling theory that underpins the importance of Previous EPS. It also elucidates the analytical role of ROE and ROA within the DuPont framework, and underscores the relevance of the financial distress cost theory in supporting the predictive value of the interest coverage ratio.
- *Finding* 3: Tree-based models (XGBoost, LightGBM, and Random Forest) demonstrate high consistency in feature importance identification (100% overlap rate), whereas significant differences were observed between linear and non-linear models (66.67% overlap rate).
- *Finding* 4: SHAP analysis reveals nonlinear relationships with strong economic foundations, including threshold effects, diminishing returns, and interaction effects. These patterns reflect the inherent complexity and nonlinearity of financial systems.

This study's theoretical contributions are demonstrated in four aspects: (1) providing systematic empirical analysis for the application of SHAP method in the financial domain, and explores the underlying economic theories associated with each key feature, such as the earnings persistence hypothesis, signalling theory, the DuPont analysis framework, and the theory of financial distress costs; (2) proposing the Top-K overlap rate metric to quantify interpretability consistency across models, thereby enriching the evaluation methodology of explainable AI; (3) revealing the hierarchical characteristics of feature importance cognition, offering novel insights for feature selection theory; (4) resenting a systematic analysis of the regulatory compliance value of SHAP in financial applications, addressing aspects such as the “right to explanation” under GDPR and its role in trust-building.

Based on the research findings, the following recommendation are proposed for financial institutions:

(1) Model Selection Strategy: Select appropriate models based on the specific application context. LightGBM is recommended for high-risk decision-making scenarios due to its excellent predictive performance. XGBoost is suitable for regulatory reporting settings where a balance between performance and interpretability is required. Logistic regression should be employed in contexts with strict compliance demands, where transparency is essential. In tasks such as credit scoring, financial distress prediction, and fraud detection, model choice should reflect the specific economic challenges and regulatory constraints of each application.

(2) Feature Engineering Guidance: Focus on profitability indicators, particularly core features such as Previous EPS, ROE, and ROA, while simultaneously considering solvency and operational capability indicators. These features, grounded in economic theory, not only exhibit statistical predictive power but also offer strong theoretical justification.

(3) Risk Management Applications: Develop SHAP-based risk early warning systems to identify potential risks through changes in feature contribution values, and enhance decision credibility through multi-model consensus.
(4) Regulatory Compliance Framework: Utilize the SHAP methodology to construct a comprehensive model interpretability framework that aligns with regulatory requirements such as the General Data Protection Regulation (GDPR), thereby fostering greater trust among clients and investors.

This study has the following limitations: (1) The dataset sources are relatively singular, and future research could expand to multiple financial markets and product categories; (2) Only static feature importance was considered, and future work could explore the variation of feature importance in dynamic environments; (3) SHAP analysis primarily focused on local explanations, while global explanation methods warrant further investigation; (4) Although this study provides an in-depth analysis of the inherent limitations of the SHAP method, including issues such as the confounding of correlation and causality, as well as challenges associated with high-dimensional data, fully addressing these limitations requires further methodological innovations; and (5) A crucial area for future research is the quantitative assessment of how specific biases inherent in financial datasets affect the fidelity of SHAP-based explanations, such as sample selection, survivorship, industry concentration, and temporal effects.

Future research in this area could focus on several promising directions. One avenue is the exploration of interpretability methods for deep learning models, particularly when applied to high-dimensional financial data. Another is the investigation of explanation techniques for multimodal financial datasets, such as text, images, and time series, to support the development of more comprehensive risk assessment frameworks. Additionally, researchers could work on creating customized explanation frameworks designed to meet specific regulatory requirements and compliance needs across different legal jurisdictions. Efforts may also be directed toward constructing standardized evaluation systems for explainable AI, emphasizing quantifiable indicators such as explanation quality, consistency, and reliability. Further progress could be achieved by advancing hybrid interpretability approaches that integrate SHAP with other explanation methods, thereby addressing the limitations inherent in individual techniques. Finally, exploring the use of causal inference techniques in interpretable financial AI could help distinguish between mere statistical correlations and genuine causal relationships.

## Credit Authorship Contribution Statement

All authors contributed to the conceptualization of this study. Dr. Chan was responsible for data curation, formal analysis, investigation, methodology, project administration, validation, and writing. Dr. Tsai contributed to formal analysis, methodology, and validation. Dr. Tang contributed to investigation, visualization, and writing. Dr. Yang contributed to investigation, resources, and software.

## Acknowledgments/Funding

The authors received no financial or material support that could have influenced the results or their interpretation.

## Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

## References


Alghamdi, H., & Alqithami, S. (2025). A robust machine learning framework for stock market classification. *Expert Systems with Applications*, 241, 128573. https://doi.org/10.1016/j.eswa.2025.128573

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Altman, E. I., Sabato, G., & Wilson, N. (2010). The value of non-financial information in SME risk management. *Journal of Credit Risk*, 6(2), 95-127. https://doi.org/10.21314/JCR.2010.110

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(6), e1424. https://doi.org/10.1002/widm.1424

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion,* 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Babaei, G., Giudici, P., & Raffinetti, E. (2023). Explainable FinTech lending. *Journal of Economic Business*, 125-126, 106126. https://doi.org/10.1016/j.jeconbus.2023.106126

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2), 169-217. https://doi.org/10.1086/259394

Benoumechiara, N., & Elie-Dit-Cosaque, K. (2019). Shapley effects for sensitivity analysis with dependent inputs: bootstrap and kriging-based algorithms, *ESAIM: Proceedings and Surveys*, Volume 65, 266–293. https://doi.org/10.1051/proc/201965266

Bhattacharya, S. (1979). Imperfect information, dividend policy, and "the bird in the hand" fallacy. *The Bell Journal of Economics*, 10(1), 259-270. https://doi.org/10.2307/3003330

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking explainable machines: The GDPR's "right to explanation" debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*, 34(1), 143-188. https://doi.org/10.15779/Z38M32N986

Chen, H., Janizek, J. D., Lundberg, S., & Lee, S. I. (2020). True to the model or true to the data? *arXiv Preprint*. https://doi.org/10.48550/arXiv.2006.16234

Covert, I., & Lee, S. I. (2021). Improving KernelSHAP: Practical Shapley value estimation using linear regression. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 3457-3465. Available from https://proceedings.mlr.press/v130/covert21a.html

Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, 57(8), 216. https://doi.org/10.1007/s10462-024-10854-8

Ding, S., Cui, T., Bellotti, A. G., Abedin, M. Z., & Lucey, B. (2023). The role of feature importance in predicting corporate financial distress in pre and post COVID periods: Evidence from China. *International Review of Financial Analysis*, 90, 102851. https://doi.org/10.1016/j.irfa.2023.102851

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv Preprint*. https://doi.org/10.48550/arXiv.1702.08608

Fatma, N., & Hidayat, W. (2020). Earnings persistence, earnings power, and equity valuation in consumer goods firms. *Asian Journal of Accounting Research*, 5(1), 3-13. https://doi.org/10.1108/ajar-05-2019-0041

Khan, F. S., Mazhar, S. S., Mazhar, K., AlSaleh, D. A., & Mazhar, A. (2025). Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review*, 58(232), 1-45. https://doi.org/10.1007/s10462-025-11215-9

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491-5500. https://doi.org/10.48550/arXiv.2002.11097

Lang, L.H.P., & Stulz, R. (1992). Contagion and competitive intra-industry effects of bankruptcy announcements: An empirical analysis. *Journal of Financial Economics*, 32(1), 45-60. https://doi.org/10.1016/0304-405X(92)90024-R

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.48550/arXiv.1705.07874

Misheva, B.H., Osterrieder, J., Kulkarni, O., & Lin, S.F. (2021). Explainable AI in credit risk management. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2103.00949

Nallakaruppan, M. K., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., Meena, V. P., & Hameed, I. A. (2024). Credit Risk Assessment and Financial Decision Support Using Explainable Artificial Intelligence. *Risks*, 12(10), 164. https://doi.org/10.3390/risks12100164

Nguyen, N., & Ngo, D. (2025). Comparative analysis of boosting algorithms for predicting personal default. *Cogent Economics & Finance*, 13(1), 2465971. https://doi.org/10.1080/23322039.2025.2465971

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144. https://doi.org/10.1145/2939672.2939778

Saarela, M., & Podgorelec, V. (2024). Recent applications of explainable AI (XAI): A systematic literature review. *Applied Sciences*, 14(19), 8884. https://doi.org/10.3390/app14198884

Soliman, M. T. (2008). The use of DuPont analysis by market participants. *The Accounting Review*, 83(3), 823-853. https://doi.org/10.2308/accr.2008.83.3.823

Stiglitz, J. E., & Weiss, A. (1981). Credit Rationing in Markets with Imperfect Information. *The American Economic Review*, 71(3), 393–410. http://www.jstor.org/stable/1802787

Thanathamathee P., Sawangarreerak S., Chantamunee S., & Nizam D. N. (2024). SHAP-instance weighted and anchor explainable AI: enhancing XGBoost for financial fraud detection. *Emerging Science Journal*, 8(6), 2404–2430. https://doi.org/10.28991/ESJ-2024-08-06-016

Tyagi, S. (2022). Analysing machine learning models for credit scoring and investment decisions: Interpretability matters. *arXiv Preprint.* https://doi.org/10.48550/arXiv.2209.09362

Warner, J. B. (1977). Bankruptcy costs: Some evidence. *The Journal of Finance*, 32(2), 337-347. https://doi.org/10.1111/j.1540-6261.1977.tb03274.x

Yeo, W. J., van der Heever, W., Mao, R., Cambria, E., Satapathy, R., & Mengaldo, G. (2025). A comprehensive review on financial explainable AI. *Artificial Intelligence Review*, 58, 189. https://doi.org/10.1007/s10462-024-11077-7

Zhou, Y., Li, H., Xiao, Z., & Qiu, J. (2023). A user-centered explainable artificial intelligence approach for financial fraud detection. *Finance Research Letters*, 58, 104309. https://doi.org/10.1016/j.frl.2023.104309