

Article

Financial Statement Fraud Detection Through an Integrated Machine Learning and Explainable AI Framework

Tsolmon Sodnomdavaa  and Gunjargal Lkhagvadorj * Department of Finance and Economics, Mandakh University, Ulaanbaatar 16061, Mongolia;
tsolmon@mandakh.edu.mn

* Correspondence: gunjargal.l@mandakh.edu.mn

Abstract

Financial statement fraud remains a substantial risk in environments marked by weak regulatory oversight and information asymmetry. This study develops a decision-centric framework that integrates machine learning, explainable artificial intelligence, and decision curve analysis to improve fraud detection under severe class imbalance. Using 969 firm-year observations from 132 Mongolian firms (2013–2024), we evaluate 21 financial ratios with models including Random Forest, XGBoost, LightGBM, MLP, TabNet, and a Stacking Ensemble trained with SMOTE and class-weighted learning. Performance was assessed using PR-AUC, F1-score, Recall, and DeLong-based significance testing. The Stacking Ensemble achieved the strongest results (PR-AUC = 0.93; F1 = 0.83), outperforming both classical and modern baseline models. Interpretability analyses (SHAP, LIME, and counterfactual explanations) consistently identified leverage, profitability, and liquidity indicators as dominant drivers of fraud risk, supported by a SHAP Stability Index of 0.87. Decision curve analysis showed that calibrated thresholds improved decision efficiency by 7–9% and reduced over-audit costs by 3–4%, while an audit cost simulation estimated annual savings of 80–100 million MNT. Overall, the proposed ML–XAI–DCA framework offers a transparent, interpretable, and cost-efficient approach for enhancing fraud detection in emerging-market contexts with limited textual disclosures.

Keywords: financial statement fraud; machine learning; explainable artificial intelligence; decision curve analysis; audit cost simulation; theoretical consistency

JEL Classification: C45; C55; M41; G32; G38; K42

Academic Editors: Alessio Faccia,
Pythagoras Petratos and
Khaled Hussainey

Received: 21 October 2025

Revised: 11 December 2025

Accepted: 17 December 2025

Published: 24 December 2025

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Financial statement fraud (FSF) refers to the deliberate manipulation of financial information intended to mislead stakeholders, resulting in trillions of dollars in global losses each year and posing a serious threat to investor confidence and economic stability ([Association of Certified Fraud Examiners, 2024](#)). High-profile corporate scandals such as Enron and Wirecard have vividly exposed the weaknesses of manual auditing and rule-based detection systems, which remain particularly vulnerable in developing economies characterized by high information asymmetry and weak institutional capacity ([Jan, 2021](#); [Riskiyadi, 2024](#)).

The primary challenge in detecting financial statement fraud (FSF) lies in its adaptive or memory-based nature. Fraud perpetrators deliberately manipulate financial indicators to evade conventional detection signals, rendering static, rule-based systems increasingly

ineffective. Moreover, because fraudulent cases constitute less than 1% of all financial observations, FSF is a rare-event problem, reducing detection sensitivity and compromising the stability of statistical models (Craja et al., 2020; C.-H. Cheng et al., 2021). Traditional models, such as the Beneish (1999) M-Score, have yielded valuable insights; however, their explanatory capacity remains limited because they cannot fully capture nonlinear dependencies, domain-specific heterogeneity, and variations in data quality (Perols, 2011). Theoretically, this study is grounded in agency theory (Jensen & Meckling, 1976) and signaling theory (Spence, 1973). According to these frameworks, information asymmetry between managers, owners, and investors creates incentives for financial manipulation, whereas financial statements serve as signals of corporate integrity and financial soundness.

In recent years, machine learning (ML) techniques have been increasingly adopted for detecting financial statement fraud, demonstrating superior performance compared with traditional econometric approaches. Ensemble algorithms such as Random Forest, XGBoost, Support Vector Machines, and LightGBM consistently outperform rule-based systems (Kirkos et al., 2007; Ravisankar et al., 2011; Achakzai & Peng, 2023). Systematic reviews further highlight that hybrid and ensemble ML models deliver substantial improvements in predictive accuracy across diverse settings (Ali et al., 2022; Shahana et al., 2023; Hernandez Aros et al., 2024).

Despite these advancements, several limitations persist in current FSF research. First, most empirical datasets remain heavily U.S.-centric, limiting the generalizability of model performance to emerging markets. Second, many jurisdictions lack standardized, machine-readable textual disclosures such as Management Discussion and Analysis (MD&A) statements, which restricts the applicability of text-based and multimodal fraud detection models. Third, although various imbalance-handling techniques exist, challenges posed by extreme rarity and instability in fraud labels continue to affect predictive calibration and decision-level evaluation (Bertomeu et al., 2021; Rahman & Zhu, 2023; Li et al., 2024).

To address these limitations in a context-appropriate manner, this study develops an integrated framework that combines machine learning, explainable artificial intelligence, and decision curve analysis to evaluate predictive performance, interpretability, and decision-level utility within a unified structure. Rather than positioning this combination as a methodological innovation, the framework emphasizes calibrated decision-analytic evaluation, an area that remains comparatively underdeveloped in the financial statement fraud detection literature.

Although the integration of machine learning and explainable AI has become increasingly common since the mid-2010s (Ribeiro et al., 2016), the present study's contribution lies in incorporating decision curve analysis-based cost-benefit evaluation and audit cost simulation into the modeling workflow. This approach allows explicit examination of net benefit, threshold-sensitive gains, and practical audit efficiency—dimensions that have rarely been quantified in prior FSF studies.

The proposed framework is not intended to replace emerging multimodal or large-language-model-based fraud detection systems, which represent an important trajectory for future research. Instead, it is grounded in empirical constraints characteristic of environments where structured numerical data constitute the only reliable form of disclosure. In such settings, multimodal baselines cannot be meaningfully implemented due to the absence of consistent narrative records, such as management discussion and analysis sections; this constraint is explicitly reflected in the study's design and scope.

The study applies the framework to Mongolia, an emerging economy characterized by information asymmetry, limited audit coverage, and evolving regulatory institutions, providing a unique empirical setting for examining FSF detection under constrained data

conditions. While the present analysis focuses on structured financial data, the broader approach remains extensible to multimodal and textual disclosures in future research.

Recent advances in FraudLLM, multimodal document intelligence, and transformer-based tabular architectures (2024–2025) offer promising benchmarks for subsequent methodological development. Although these models cannot be evaluated within the current study due to the absence of paired structured–textual disclosures in Mongolia, acknowledging their relevance situates the framework within contemporary research directions. It delineates a clear trajectory for future extensions.

The primary objective of this study is threefold: to enhance the performance of machine learning models under severe class imbalance, to validate the theoretical consistency of key financial indicators, and to evaluate the cost–benefit implications of detection outcomes for decision-making. The empirical framework employs Random Forest, Support Vector Machine, XGBoost, and LightGBM models, and evaluates them using PR-AUC, Recall, and F1-score. Among 21 financial ratios, the most influential indicators were identified through SHAP and LIME analyses. Decision Curve Analysis was then applied to assess practical decision utility and to simulate potential reductions in audit-related costs.

The methodological design is operationalized through an integrated structure termed the Decision-Centric FSF Detection Framework. This framework sequentially integrates theoretical foundations, data preprocessing, machine learning models, explainable artificial intelligence interpretation, and decision-level evaluation to enhance performance, interpretability, and decision optimization. Through this integration, the study advances theoretical, methodological, and practical perspectives by linking agency and signaling theories with contemporary approaches in XAI and decision analytics. The resulting framework supports a transparent, evidence-based, and context-adaptive decision-support paradigm suited for both emerging and developed markets (Cai & Xie, 2024; Duan et al., 2024; Nguyen Thanh & Phan Huy, 2025).

2. Literature Review

Traditional ratio-based analytical approaches have long provided the foundation for detecting financial statement fraud (FSF). However, their static nature limits their ability to capture evolving behavioral patterns, severe data imbalance, and the complex multidimensional interactions that characterize modern financial reporting environments (Perols, 2011; Shahana et al., 2023). In this context, machine learning (ML) has emerged as a data-driven paradigm that surpasses rule-based systems and expert judgment by enabling dynamic pattern recognition, predictive modeling, and cost-sensitive decision-making (Ali et al., 2022; Ozbayoglu et al., 2020).

Recent studies increasingly emphasize that ML-based fraud detection frameworks, particularly those relying on tree-based ensembles and widely used explainable AI tools, have become methodological norms since the mid-2010s. As a result, the combination of machine learning and explainable AI is no longer inherently novel unless it is integrated into a broader decision-focused structure that incorporates audit costs, threshold optimization, or net benefit evaluation.

Between 1999 and 2025, FSF detection has progressed through several stages of methodological development. These include traditional statistical models, hybrid ensemble frameworks, deep learning architectures, explainable AI-based systems, and, most recently, decision analytic approaches such as decision curve analysis. Building on this progression, the present study brings together contemporary methodological advances, including imbalance correction techniques, stacking ensembles, LightGBM, LIME, counterfactual explanations, the SHAP Stability Index, and Decision Curve Analysis (DCA), into a unified

analytical design that explicitly links predictive performance with interpretability and decision-level utility.

2.1. Traditional and Early Machine Learning Approaches

The Beneish (1999) M-Score model established the first systematic ratio-based framework for detecting financial statement fraud (FSF) by identifying abnormal variations in accounting indicators associated with managerial incentives and compensation structures. Kirkos et al. (2007) employed Bayesian networks and decision trees, achieving accuracy levels above 90 percent and outperforming traditional logistic regression models. Perols (2011) demonstrated that Support Vector Machines (SVMs) and Neural Networks (NNs) can capture nonlinear dependencies more effectively, revealing complex data patterns that extend beyond the capabilities of classical theoretical models. Throughout the 2010s, research combining numerical indicators with natural language processing (NLP) gained momentum, contributing to deeper analytical perspectives. Studies by Humpherys et al. (2011), Glancy and Yadav (2011), and Ravisankar et al. (2011) integrated textual credibility analysis, rule-based logic, and probabilistic neural network feature selection to produce strong detection performance even under imbalanced data conditions, as further supported by Huang et al. (2014) and Kim et al. (2016).

2.2. Hybrid and Machine Ensemble Learning Frameworks

To overcome the limitations of single algorithms, hybrid machine learning (ML) approaches have become increasingly prevalent in FSF detection. S. Chen (2016) proposed a combined model that integrated decision trees and Support Vector Machines (SVMs), reducing the Type I error rate to below 8%. Lin et al. (2015) compared ML-based models with auditor judgment and found that the former exhibited greater objectivity. Dutta et al. (2017) applied a combination of Random Forest (RF) and SVM, with feature engineering, to detect financial restatements and reported an accuracy above 90%. Y.-J. Chen et al. (2019) identified risks arising from intra-group business linkages through ML modeling, and Hajek and Henriques (2017) demonstrated the effectiveness of SVM for extracting patterns from annual reports. C.-H. Cheng et al. (2021) enhanced the robustness of imbalanced data using synthetic feature generation and fuzzy logic techniques, a development further supported by Rahman and Zhu (2023).

To address class imbalance, oversampling techniques such as SMOTE and ADASYN (Chawla et al., 2002) have been widely used. However, recent studies show that these methods alone may not ensure stable generalization. Contemporary FSF research increasingly combines imbalance correction with probability calibration, feature stability assessment, and cost-sensitive evaluation. This trend highlights the need for integrated analytical frameworks rather than incremental adjustments to individual models.

Ensemble learning methods such as Random Forest (RF), Gradient Boosting Machine (GBM), and XGBoost play a critical role in managing the precision-recall trade-off under imbalanced data conditions (Kim et al., 2016; Achakzai & Peng, 2023; Duan et al., 2024). Bertomeu et al. (2021) showed that the GBM model achieved an AUC above 0.69 when predicting the probability of financial misreporting. In recent years, stacking ensemble approaches have emerged as powerful strategies that combine base learners, such as RF, XGBoost, and SVMs, via a meta-learner such as LightGBM. These frameworks enhance predictive performance while enabling global and local explanations through explainable AI techniques (Bao et al., 2020).

Recent methodological assessments indicate that tree-based ensembles and stacking frameworks have become standard baselines in financial statement fraud detection research. Consequently, new frameworks must be benchmarked against more advanced deep

tabular architectures, including TabNet, TabTransformer, and FT Transformer, as well as against relational models such as graph neural networks, to substantiate methodological contributions. Measuring feature stability using Permutation Importance and the SHAP Stability Index has also become important for validating model reliability. In this study, these techniques are used to examine the balance between feature stability and sensitivity, ensuring alignment between interpretability and model robustness.

2.3. Explainable and Advanced AI

Deep learning has significantly advanced the detection of financial statement fraud (FSF) by integrating and processing diverse semantic signals from heterogeneous data sources (Craja et al., 2020; Ozbayoglu et al., 2020; Peng & Yan, 2021). Fischer and Krauss (2018) employed an LSTM architecture to identify abnormal sequential patterns in financial time series and captured temporal dynamics associated with fraudulent behavior. Mai et al. (2019) and Matin et al. (2019) applied Convolutional Neural Networks and LSTM methods to analyze the tone and linguistic structures of annual reports, achieving accuracy rates of 80–85%. Kraus and Feuerriegel (2017) used transfer learning to facilitate knowledge transfer across domains, while Brown et al. (2020) utilized topic modeling of MD and A sections to identify tendencies toward misreporting.

To address the opacity of deep learning models, a range of explainable artificial intelligence (XAI) techniques, such as SHAP, LIME, counterfactual explanations, and Permutation Importance, have been developed to improve interpretability (Ribeiro et al., 2016; An et al., 2020; Cai & Xie, 2024). In this study, the XAI component is validated using the SHAP Stability Index and analyses of global and local consistency, introducing a structured validation approach that has not previously been applied in FSF research.

Recent FSF studies published from 2022 to 2025 have expanded these developments through knowledge graph-based learning (Cai & Xie, 2024), supply chain GNN architectures (D. Cheng et al., 2025), structured data ensemble frameworks (Achakzai & Peng, 2023), and hybrid regulatory and financial models (Luo et al., 2025). These studies show that the methodological frontier of FSF detection is moving toward relational and graph-based inference, as well as advanced deep tabular architectures such as TabNet and FT Transformer. Because the Mongolian dataset lacks relational, textual, or multimodal disclosures, these models cannot serve as empirical baselines. For this reason, the present study positions MLP and TabNet as the strongest feasible baselines for structured numerical data, while recognizing GNN and LLM-based multimodal systems as important directions for future research. The legal environment, transparency level, ownership structure, and cultural heterogeneity significantly affect FSF detection outcomes (Riskiyadi, 2024). Nguyen Thanh and Phan Huy (2025) found that when the auditor's opinion is used as a proxy variable, model performance varies considerably across different national contexts. Rahman and Zhu (2023) reported that family ownership concentration directly increases vulnerability to FSF. Proxy variables such as earnings quality and linguistic sentiment also influence model sensitivity. Deep learning models tend to perform better under heterogeneous conditions because they capture subtle contextual nuances that traditional models often overlook (Jan, 2021; Mai et al., 2019). In recent years, the application of Decision Curve Analysis, introduced by Vickers and Elkin (2006), has expanded rapidly as a method for estimating a model's net benefit and assessing audit cost effectiveness. This approach aligns FSF detection more closely with real-world decision-making by explicitly incorporating economic trade-offs between detection gains and auditing costs (Duan et al., 2024).

Despite these advances, recent methodological critiques emphasize that tree-based models alone are insufficient to demonstrate novelty unless they are benchmarked directly

against the most recent architectures introduced in 2024 and 2025. These include TabTransformer, graph neural network-based relational models, and FinBERT-style frameworks that combine textual and numerical information. In this context, such modern architectures are acknowledged as theoretically relevant comparators and are identified as important directions for future extension.

In recent years, research has increasingly emphasized the need for explainable and ethically grounded artificial intelligence in financial auditing and fraud detection. Yeo et al. (2025) provide a comprehensive review of explainable AI in finance and outline key challenges and opportunities in deploying transparent models. Nugrahanti et al. (2025) empirically demonstrate that integrating explainable AI improves the quality of auditor judgment and strengthens accountability mechanisms. Aljunaid et al. (2025) propose a federated learning based explainable AI framework that enhances audit transparency and data security in banking applications. Complementing these findings, Černevičienė and Kabašinskas (2024) examine the ethical dimensions of explainable AI and emphasize the importance of transparency, fairness, and auditability in financial analytics. Together, these interdisciplinary perspectives reinforce the conceptual foundation of the integrated ML, XAI, and DCA framework by linking technical interpretability with ethical audit assurance.

In addition to explainability and ethical assurance, recent research from 2024 and 2025 highlights the transformative role of large language models and multimodal architectures in financial fraud detection. Nie et al. (2025) propose a multimodal detection framework that integrates large language models with interpretable machine learning to jointly process textual disclosures and numerical indicators, thereby achieving improved precision and interpretability. Yang et al. (2025) introduce Fraud R1, a benchmark designed to assess the robustness of large language models under augmented fraud and phishing scenarios. Kamlakshya and Hota (2025) develop a multimodal document intelligence framework that integrates textual, visual, and tabular information to support fraud prevention. These developments indicate that future FSF systems will need to benchmark against large language models and multimodal baselines to substantiate claims of novelty. However, such models cannot be implemented in the present study because structured textual paired disclosures are not available in Mongolia.

Table 1 shows that recent state-of-the-art FSF models increasingly rely on multimodal or relational information, including knowledge graphs, supply chain networks, regulatory supervision signals, and advanced tabular architectures. Approaches based on knowledge graph learning (Cai & Xie, 2024; D. Cheng et al., 2025) and hybrid regulatory and financial modeling (Luo et al., 2025) require data types that are not available in the structured numerical dataset used in this study. Multimodal and text-driven models cannot be implemented either because complete, machine-readable MD and A disclosures are lacking. For these reasons, the most empirically appropriate baselines in the present study are deep tabular models such as MLPs and TabNet, as well as classical structured-only baselines, including the Beneish M Score. This selection ensures that the framework is benchmarked against the strongest feasible models while accurately reflecting the disclosure constraints typical of emerging markets.

Table 1. Positioning of recent FSF SOTA models (2022–2025).

Study	Data Type	Model Family
Cai and Xie (2024)	Ratios + Knowledge Graph	Two-layer KG + DL (Explainable)
D. Cheng et al. (2025)	Ratios + Supplier–customer (supply chain) network	Graph-enhanced ML/network-based fraud detection
Achakzai and Peng (2023)	Structured ratios	Dynamic Ensemble ML
Lin et al. (2015)	Structured financial ratios	Classical ML (DT, NN, SVM)
Luo et al. (2025)	Ratios + Regulatory punishment reports	Hybrid ML
Y. Chen and Wu (2023)	Structured financial indicators	LR, SVM, RF
Lee et al. (2025)	Structured ratios (Indonesia)	ML comparative evaluation

2.4. Research Hypotheses

From the above literature, it is evident that practical applications of FSF detection require an integrated framework that simultaneously satisfies performance, interpretability, and cost effectiveness. In other words, a unified approach is needed that combines high PR AUC and Recall with interpretable XAI insights and cost-sensitive Decision Gain evaluation. Accordingly, this study tests a series of hypotheses that focus on four key dimensions:

- (1) the impact of imbalance correction on model performance,
- (2) the relative importance and stability of financial features,
- (3) the theoretical consistency of XAI-based explanations, and
- (4) the decision level utility and cost effectiveness of the proposed framework.

These hypotheses collectively examine whether the integrated approach can produce statistically reliable, theoretically coherent, and decision-supportive outcomes in FSF detection.

H1. *The top one-third of features, ranked by SHAP mean absolute values and permutation importance, will be dominated by liquidity and leverage variables, exhibiting low Coefficient of Variation and high stability, as measured by the SHAP Stability Index.*

H2. *The directional effects of SHAP values, for example, an increase in leverage increases risk and a decrease in liquidity increases risk, will show more than 70 percent consistency with financial theory. These effects will be validated through LIME and counterfactual sensitivity analysis, consistent with agency theory and the Fraud Triangle framework.*

H3. *The combination of balanced class weighting and SMOTE resampling will reduce the Expected Cost relative to the baseline model, resulting in higher Decision Gain and lower auditing costs under Decision Curve Analysis, demonstrating practical cost efficiency.*

H4. *After Platt and Isotonic calibration, the Brier score will decrease. In contrast, the calibration slope will move closer to one, and the calibration intercept will move closer to zero, leading to improved net benefit across practical decision thresholds.*

Among these four hypotheses, H1 and H2 address the core analytical objectives related to model performance and interpretability. H3 and H4 serve as complementary validation tests that assess feature stability and decision utility through Decision Curve Analysis. The hypotheses intentionally focus only on dimensions supported by structured financial data. While modern multimodal and transformer-based models, such as TabTransformer, graph neural networks, and large language models, represent important methodological advances, they fall outside the scope of this dataset. They are therefore identified as future

research directions. Together, the hypotheses establish a comprehensive validation strategy that integrates global and local interpretability, cost–benefit evaluation, Decision Curve Analysis, and probabilistic calibration.

3. Methodology

3.1. Research Design and Theoretical Framework

This study develops a decision-centric integrated framework for detecting financial statement fraud (FSF) by combining machine learning (ML), explainable artificial intelligence (XAI), and Decision Curve Analysis (DCA). The framework integrates three complementary elements: the predictive performance of ML models, the interpretability of XAI methods, and the practical decision utility measured by DCA. These components form a unified evaluation structure that links statistical accuracy, theoretical coherence, and decision-level benefits.

The main objectives of the study are threefold: (i) to improve the performance of ML models under severe class imbalance, (ii) to validate the explanatory relevance and directional consistency of key financial indicators, and (iii) to assess the real-world cost–benefit implications of FSF detection. By combining these dimensions, the design ensures that the empirical results are both theoretically grounded and practically meaningful for audit and regulatory contexts.

The theoretical foundation is based on agency theory (Jensen & Meckling, 1976) and signaling theory (Spence, 1973). These frameworks suggest that information asymmetry and misaligned incentives can motivate managers to manipulate financial statements, while financial ratios serve as observable signals of credibility and risk. Liquidity, leverage, and profitability indicators, in particular, play a central role in distinguishing normal firms from those with elevated fraud risk.

Building on these theoretical premises, the study employs a decision-centric analytical structure that integrates ML for pattern detection, XAI for transparency and theoretical consistency, and DCA for evaluating economic utility and audit cost effectiveness. This integrated approach provides a balanced framework that aligns theoretical soundness, empirical performance, and practical decision support, thereby offering a coherent basis for improving financial statement fraud detection in constrained data environments.

Recent methodological reviews consistently show that the combination of machine learning and explainable artificial intelligence has become standard practice in financial statement fraud research since the mid-2010s. As a result, the use of ML and XAI alone cannot be considered methodologically novel. To remain aligned with this development and avoid overstated claims, the present study extends beyond conventional ML–XAI pipelines by incorporating several advanced components. These include calibrated probability evaluation using Brier scores, Platt scaling, and isotonic regression; feature stability assessment using the SHAP Stability Index; and explicit decision-analytic evaluation based on net benefit and decision gain. Together, these elements shift the analytical focus from predictive accuracy toward decision-level optimization, an area that remains underexplored in the current FSF literature.

The study also explicitly acknowledges that modern state-of-the-art baselines, such as TabTransformer, SAINT, NODE, graph neural networks for relational inference, and multi-modal or large language model-based fraud detection systems emerging in 2024 and 2025, represent theoretically important comparators. These architectures cannot be implemented in the present study because structured–textual paired disclosures are unavailable in Mongolia. Nevertheless, their relevance is recognized to situate the framework realistically within contemporary methodological developments. By clarifying these boundaries, the contribution is positioned not as a replacement for multimodal or LLM-enabled systems,

but as a context-appropriate framework suited to data-constrained environments where numerical statements remain the only consistent and reliable source of disclosure.

As illustrated in Figure 1, the proposed framework integrates the theoretical foundation, financial indicators, data preprocessing, imbalance correction, machine learning modeling, explainable artificial intelligence interpretation, probability calibration, and decision-level evaluation into a single coherent analytical structure. This sequential design ensures continuity across all methodological stages. Consistent with agency and signaling theories, the framework establishes a comprehensive system that combines performance, interpretability, and decision considerations. As a result, it provides a unified approach that links theoretical soundness, statistical robustness, and practical decision relevance in detecting financial statement fraud.

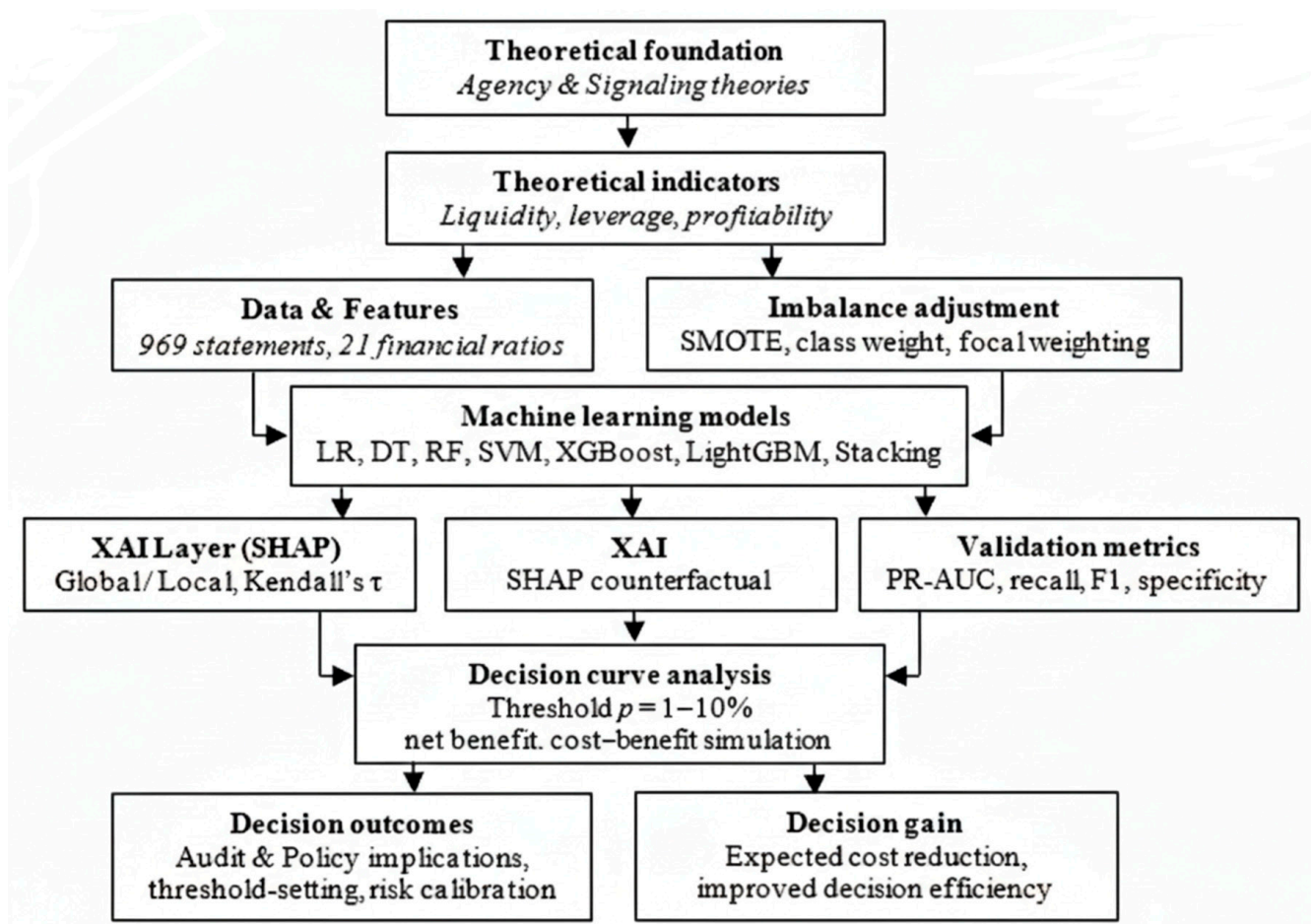


Figure 1. FSF detection framework (ML, XAI, DCA).

The expanded design also serves three additional purposes. First, it distinguishes the framework from incremental ML–XAI pipelines by incorporating stability diagnostics, calibrated probability evaluation, and cost-sensitive decision analysis. Second, it embeds decision-theoretic assessment to reflect the real trade-offs that auditors face when balancing detection gains and auditing costs. Third, it positions the study relative to contemporary transformer-based and multimodal fraud detection paradigms, clarifying how the proposed approach fits within modern methodological expectations while remaining appropriate for data-constrained environments.

3.2. Data Source and Variable Definition

This study uses financial statement data from 132 Mongolian companies spanning 2013 to 2024, yielding 969 firm-year observations. All firm-level information was obtained

from publicly available audited financial statements and anonymized before analysis. To enhance reliability, each financial statement was cross-checked against Refinitiv records to correct common filing inconsistencies in emerging markets. The research procedure follows ethical standards regarding transparency, data privacy, and fairness in AI-driven decision-making.

Feature engineering was guided by financial theory and the principles of empirical interpretability. Liquidity, leverage, profitability, and efficiency ratios were standardized and, when appropriate, log transformed to reduce skewness and ensure comparability across firms. Outliers were treated using the interquartile range method, and categorical variables were encoded through one-hot encoding. These steps were included to strengthen distributional robustness and maintain consistent cross-firm comparability. To mitigate multicollinearity, variables with a Variance Inflation Factor above five were removed. This preprocessing strategy ensures that the model inputs reflect both statistical rigor and theoretical transparency.

The sample includes firms from key sectors such as manufacturing, services, trade, and construction, ensuring broad sectoral representation. The dataset shows a substantial class imbalance between fraudulent and non-fraudulent cases, reflecting the rarity of detected financial statement fraud in practice. All financial information was obtained from the Refinitiv database and subjected to a rigorous verification process to ensure accuracy and completeness.

Because textual disclosures are extremely limited in Mongolia, the study relies exclusively on structured numerical indicators. Only 27 percent of firms provided any form of MD&A-style narrative, and fewer than 12 percent offered machine-readable or year-consistent textual sections. Most auditor reports were submitted as image-based PDFs, with more than half containing OCR errors or missing content. These conditions make reliable tokenization, semantic extraction, and sequence modeling infeasible. As a result, text-based architectures such as CNNs and LSTMs, BERT and FinBERT variants, FraudLLM, and multimodal transformer systems cannot be meaningfully benchmarked on this dataset.

Given these constraints, the empirical analysis is restricted to structured numerical indicators to preserve data consistency, ensure comparability across firms, and avoid noise introduced by incomplete or unverifiable textual proxies. This approach enables a methodologically sound and context-appropriate evaluation of fraud detection models in data-constrained environments. Following validation, several preprocessing steps were implemented:

- Outlier detection and removal using the interquartile range (IQR) method;
- Standard normalization through Z-score transformation;
- Dummy variable encoding for categorical attributes;
- SMOTE–Tomek resampling to correct class imbalance and enhance model stability.

SMOTE Tomek was selected after a comparative evaluation with ADASYN, which produced unstable minority-class samples and did not deliver reliable oversampling performance. The preprocessing procedures ensured that the dataset was statistically robust, normalized, and appropriately balanced for subsequent machine learning analysis. The dependent variable in this study is the fraud status of financial statements, measured as a binary outcome with 0 indicating fraud and 1 indicating non-fraud. The independent variables consist of 21 financial ratios (FR1–FR21) that capture the firm's financial condition and performance. These ratios reflect profitability, liquidity, leverage, cash flow, and cost structure, which are widely used in fraud detection and financial risk assessment research.

Feature selection was conducted using two criteria: the Mann–Whitney U test with a significance threshold of $p < 0.05$, and Cohen's d of at least 0.5 to ensure that retained variables were both statistically significant and practically meaningful. To address potential

multicollinearity, the Variance Inflation Factor was computed, and a conservative upper threshold of 5 was applied. This procedure reduced interdependence among variables and enhanced the overall reliability and stability of the modeling framework.

3.3. Machine Learning Models, Training Protocol, and XAI Analysis

The final model family was selected after evaluating eleven candidate algorithms. CatBoost and ExtraTrees were removed from the final set because robustness checks indicated clear overfitting. To address class imbalance, the study employed a combination of the Synthetic Minority Oversampling Technique and a balanced class-weight configuration. SMOTE Tomek was adopted after comparing SMOTE alone, ADASYN, and Random Oversampling, as the combined approach produced the most stable and empirically reliable results.

The evaluation covers a broad set of machine learning models, including logistic regression, Support Vector Machines with an RBF kernel, Random Forest, Gradient Boosting, XGBoost, LightGBM, and a multilayer perceptron. To ensure comparability, all models were trained using the same train-test split, cross-validation procedure, and probability calibration settings. In addition to these classical benchmarks, the study implements TabNet as a representative deep tabular architecture suitable for structured financial data. The TabNet model was trained using the official open-source implementation (GitHub: dreamquark ai/tabnet, version 4.1.0), ensuring reproducibility and consistency with prior research.

The training protocol used an 80:20 train-test split and a 10-fold stratified cross-validation procedure with five repetitions to ensure robustness and reduce variance. Hyperparameter optimization was performed using both GridSearchCV and Bayesian optimization. Bayesian optimization was prioritized for the stacking ensemble due to its efficiency in high-dimensional search spaces and its ability to reduce computational cost. For the calibrated models, the Expected Cost and Expected Loss Reduction metrics were computed to support Decision Curve Analysis. These measures provided a balanced evaluation of model discrimination, calibration, and generalization performance. In the explainability analysis, multiple XAI methods were employed, including Shapley Additive Explanations, Local Interpretable Model Agnostic Explanations, and counterfactual sensitivity analysis. At the global level, consistency between SHAP values and permutation importance was assessed using Kendall's tau correlation coefficient to evaluate rank order agreement among features. A tau value above 0.60 was used as the stability threshold, consistent with established guidelines for global interpretability assessment.

At the local level, the individual contributions and directional effects of features for specific cases were examined using force plots and decision plots, providing intuitive explanations of model outputs. Feature stability was further assessed using the Coefficient of Variation and the SHAP Stability Index to ensure that the resulting interpretations were robust and reproducible across sampling and model configurations.

At this stage, particular emphasis was placed on validating the theoretical coherence of the XAI explanations by confirming that liquidity and leverage indicators consistently exhibited strong SHAP contributions and stable directional effects, in accordance with well-established financial theories. To ensure that these explanatory patterns were not artifacts of a specific sampling or calibration scheme, a sign stability assessment was performed across all resampling settings (baseline and SMOTE-Tomek), across both calibration configurations (Platt and Isotonic), and across all cross-validation folds. This procedure verified that the direction and magnitude of feature effects remained consistent across model variants and were therefore theoretically reliable.

Beyond visualization-based interpretability, probability calibration was incorporated to strengthen decision-level applicability. The Brier score, calibration slope, and calibration

intercept were computed for all models, and the calibrated probabilities were directly linked to Decision Curve Analysis to evaluate their contribution to net benefit and threshold-dependent decision utility. Finally, although the empirical analysis focuses on structured numerical disclosures, the study acknowledges that recent advances in transformer-based tabular architectures, graph neural networks, and multimodal LLM-driven fraud detection systems represent important methodological extensions. These models cannot yet be implemented in Mongolia due to the absence of paired textual and numerical disclosures; however, noting their relevance ensures that the proposed framework remains aligned with contemporary research directions while remaining grounded in the constraints of the available data.

3.4. Decision Analysis and Hypothesis Testing

At the decision level, this study applies Decision Curve Analysis (Vickers & Elkin, 2006) to evaluate the practical utility of fraud detection models across a realistic probability threshold range of 1–15%. This interval reflects the low base rate of financial statement fraud, where the proportion of detected cases seldom exceeds 10% in real-world auditing practice. As a result, auditors typically operate within this low probability decision space when determining whether a financial report should be flagged for further inspection, making the chosen range both theoretically sound and empirically relevant.

To quantify model utility, three complementary indicators were computed: Net Benefit (NB), Decision Gain (DG), and the Average Utility Net Benefit (AUNB). These measures provide a direct assessment of the economic and operational consequences of adopting alternative modeling strategies, extending the analysis beyond conventional performance metrics such as PR-AUC or F1 score. In addition to DCA-based indicators, the Expected Cost (EC) and Expected Loss Reduction (ELR) metrics were calculated to capture the cost-sensitive nature of auditing decisions and link model outputs to explicit economic trade-offs.

To evaluate the role of probability calibration in decision performance, the study compared AUNB and DG values across models trained under different class weighting and sampling configurations, including SMOTE resampling and focal-style weighting. Probability calibration was conducted using platt scaling and isotonic regression, and improvements were assessed through the Brier score, calibration slope, and calibration intercept. These calibration diagnostics were subsequently mapped to DCA outcomes to determine whether improved probability estimates translated into higher net benefits across practical threshold ranges.

Hypotheses H3 and H4 were tested using cross-validated decision curves and examined statistically through the Wilcoxon signed rank test applied across all folds. To further ensure robustness, 95% bootstrap confidence intervals based on 1000 replications were computed for the NB, DG, and EC metrics.

In addition, a threshold-specific sensitivity analysis was conducted to examine how model utility varies across narrow probability intervals within the 1–15% range. This assessment ensured that decision outcomes remain stable in settings where minor threshold adjustments may lead to substantial differences in audit actions. The results confirmed that models with stronger probability calibration and lower expected cost consistently maintained higher Net Benefit and Decision Gain values even under conservative threshold settings, underscoring the practical importance of calibration and cost-sensitive evaluation.

This combined assessment further strengthens the inference regarding the statistical significance of improvements in decision utility and calibration quality. Overall, the decision analysis framework synthesizes three essential evaluation dimensions: model performance under class imbalance, interpretability grounded in financial theory, and

decision-level benefit measured through cost-sensitive DCA indicators. By integrating calibration diagnostics, threshold sensitivity checks, and cost-adjusted utility metrics, the framework provides robust evidence that improved probability calibration and enhanced imbalance handling contribute not only to stronger statistical performance but also to economically meaningful gains in real-world audit decision-making.

4. Results

4.1. Descriptive Statistics and Preliminary Analysis

The dataset used in this study comprises 969 firm-year financial statements from 132 Mongolian companies, covering the period from 2013 to 2024. Firms were classified into two categories: fraudulent and non-fraudulent. Approximately 24.6 percent of the sample firms were identified as fraudulent, indicating a moderate imbalance between the two groups. In addition, the fraud class exhibited substantially higher variance in leverage and profitability indicators, reflecting heterogeneous distributions that necessitate imbalance-aware modeling.

The analysis incorporated 21 financial ratios (FR1–FR21) grouped into five major categories: profitability, liquidity, leverage, expense structure, and cash flow indicators. According to the results of the Mann–Whitney U test and Cohen’s *d* effect-size analysis, five financial indicators showed statistically significant differences between the fraudulent and non-fraudulent groups ($p < 0.05$). These findings confirm that specific financial performance metrics systematically distinguish firms with potential misreporting behavior, providing a sound empirical basis for subsequent machine-learning modeling.

When combining statistical significance with practical significance, eleven variables met both criteria ($p < 0.05$ and Cohen’s $d \geq 0.5$), indicating that these indicators provide stable and meaningful discrimination across firm groups. Because the dataset consists exclusively of numerical financial ratios and does not include standardized textual disclosures such as MD&A sections or auditor narratives, all subsequent baseline and modern models were restricted to structured-data architectures. As a result, advanced text-driven approaches (e.g., CNNs, LSTMs, BERT-type models) were inapplicable. In contrast, modern tabular architectures, such as MLPs and TabNets, were fully compatible with the available data. This ensures methodological consistency between the data structure and the analytical techniques used in later sections.

As reported in Table 2, the correlation analysis revealed a negative relationship between profitability and leverage ($r = -0.58$), while a strong positive relationship was observed between the cash ratio (FR8) and the current asset ratio (FR9) ($r = 0.74$). These results indicate that increased debt pressure tends to reduce profitability, whereas improvements in liquidity efficiency are associated with a more optimal structure of current assets. Correlation heatmap diagnostics further confirmed two multicollinearity-prone clusters (FR8–FR9 and FR1–FR3); however, all VIF values remained below the threshold ($VIF < 5$). The Variance Inflation Factor (VIF) analysis yielded an average VIF of 2.52, and all individual VIFs were below 5, indicating that multicollinearity was at an acceptable level for regression and machine learning analyses. The detailed VIF analysis showed an average VIF of 2.52, with individual VIFs ranging from 1.08 to 4.73, confirming that multicollinearity was not a concern ($VIF < 5$). Therefore, all selected variables were retained for subsequent modeling. To refine variable selection, the Mann–Whitney U test and Cohen’s *d* effect size were used in conjunction. Variables that met both criteria, namely p -value < 0.05 and Cohen’s $d \geq 0.5$ (medium effect size), were regarded as statistically significant predictors. As a result of this two-step screening process, eleven variables (FR1, FR3, FR4, FR5, FR7, FR10, FR13, FR15, FR16, FR19, and FR20) were retained and later used as input features in the machine learning modeling phase. These variables were consistently ranked among

the top in global feature-importance rankings based on SHAP and permutation-based measures, supporting their theoretical relevance to FSF detection and ensuring coherence between the statistical screening process and the interpretability analysis.

Table 2. Descriptive statistics and group difference analysis of key financial variables.

No.	Financial Indicator	Code	Median (NF)	Median (F)	<i>p</i> -Value	Cohen's <i>d</i>
1	Sales Revenue/Total Assets	FR1	0.3001	0.4213	0.0011	−0.1538
2	Short-Term Liabilities/Sales Revenue	FR2	0.8157	0.5301	0.0156	0.0112
3	Total Liabilities/Sales Revenue	FR3	2.2063	0.7668	0.0000	−0.0314
4	Net Profit/Sales Revenue	FR4	−0.1014	0.0041	0.0000	0.0886
5	Net Profit after Tax/Total Assets	FR5	−0.0153	0.0009	0.0000	0.3471
6	Inventory/Sales Revenue	FR6	0.5412	0.3472	0.0092	0.0440
7	Inventory/Current Assets	FR7	0.4298	0.4730	0.4460	0.0399
8	Cash/Sales Revenue	FR8	0.0189	0.0145	0.0269	0.0666
9	Cash/Current Assets	FR9	0.0150	0.0164	0.7958	0.0959
10	Operating Expense/Sales Revenue	FR10	0.2105	0.1150	0.0020	0.0639
11	Administrative Expense/Sales Revenue	FR11	0.0652	0.0425	0.0411	0.0686
12	Non-Operating Expense/Sales Revenue	FR12	0.0504	0.0411	0.1936	0.0581
13	Retained Earnings/Total Assets	FR13	0.0003	0.0537	0.0000	0.3610
14	Retained Earnings/Shareholders' Equity	FR14	0.0467	0.1276	0.2219	−0.0731
15	Tangible Assets/Total Assets	FR15	0.5899	0.5510	0.4076	−0.0611
16	Total Liabilities/Shareholders' Equity	FR16	0.7030	0.3025	0.0001	−0.2438
17	Accounts Receivable/Net Sales	FR17	0.3675	0.3150	0.9368	0.0525
18	Working Capital/Total Assets	FR18	0.1570	0.1480	0.8724	0.0664
19	Gross Profit/Total Assets	FR19	0.0320	0.0550	0.0022	0.3216
20	Total Liabilities/Total Assets	FR20	0.4910	0.2985	0.0000	−0.5215
21	Altman's Z Index	FR21	1.0700	2.9920	0.0000	0.0122

Notes. NF = Non-fraudulent firms; F = Fraudulent firms. *p*-values were obtained using the Mann–Whitney U test. All measurements are reported as medians. A *p*-value of <0.05 indicates a statistically significant difference between the two groups. According to conventional thresholds, a Cohen's *d* of 0.5 or greater represents a medium effect size, while a Cohen's *d* of 0.8 or greater indicates a large effect size.

4.2. Imbalanced Data Correction and Balancing Results

The dataset used in this study exhibited an imbalanced class distribution, with 208 observations (24.6%) in the fraud class and 638 observations (75.4%) in the non-fraud class. Such an imbalance may bias learning processes toward the majority class. To mitigate this issue, a systematic balancing procedure was implemented.

The primary objective was to enhance minority-class visibility without introducing artificial noise. Two oversampling methods were examined: SMOTE and ADASYN. ADASYN generated unstable synthetic ratios due to a limited number of nearest neighbors; therefore, SMOTE was selected. SMOTE was applied solely to the training set after the 80–20 split, ensuring that the test set remained untouched and statistically unbiased.

Because the dataset is fully numerical, the imbalance-correction procedure aligns with the requirements of structured data modeling. This ensured that all classical, ensemble, and modern deep tabular baselines (MLP, TabNet) received a proportionally balanced and comparable training distribution, which is essential for fair head-to-head performance benchmarking in subsequent sections.

As shown in Table 3, the initial class distribution revealed that the fraud category accounted for 24.6% of all observations. After applying SMOTE with a sampling ratio of 0.33 (approximately 1:3), the minority share increased to 32.7%, significantly improving class balance. In contrast, ADASYN was less suitable due to instability arising from

sparse neighbor structures. After correcting the imbalance, the Random Forest model was evaluated using 5-fold cross-validation. The results are presented in Table 4.

Table 3. Class distribution before and after balancing.

Set	Fraud (0)	Non-Fraud (1)	Ratio (0:1)	Minority %
Overall (before)	208	638	208:638	24.59%
Train (before)	166	510	166:510	24.56%
Test (before)	42	128	42:128	24.71%
Train (after SMOTE)	248	510	248:510	32.72%

Note: Test-set oversampling is not applied because it would bias evaluation; therefore, “Test (after SMOTE)” is intentionally omitted.

Table 4. Model setup—5-fold cross-validation results (average precision/PR-AUC).

Dataset	Model Variant	CV Folds	Mean AP (PR-AUC)	Std
Train (Before)	RF—Unweighted	5	0.9234	0.0065
Train (Before)	RF—Balanced class weights	5	0.9212	0.0101
Train (After SMOTE)	RF—Unweighted	5	0.9348	0.0194
Train (After SMOTE)	RF—Balanced class weights	5	0.9357	0.0210

The results indicate that combining SMOTE oversampling with balanced class weighting improved the model’s performance. The PR-AUC increased by approximately 1.2 percentage points relative to the unbalanced baseline, confirming that the minority class became more influential during model learning. This improvement also enhanced cross-validation stability and ensured a methodologically sound basis for evaluating deep tabular baselines and the proposed Stacking framework in the subsequent performance comparison section.

4.3. Model Performance Comparison

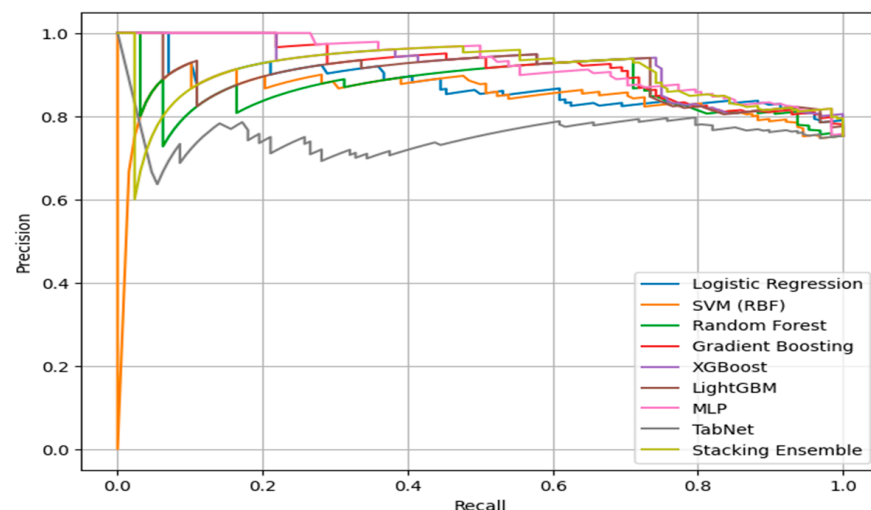
Table 5 presents the comparative test-set performance of all evaluated machine learning models. The Stacking Ensemble achieved the highest overall performance, with a PR-AUC of 0.927 and an F1-score of 0.826, demonstrating the strength of ensemble learning in capturing heterogeneous patterns, nonlinear dependencies, and conditional interactions within structured financial ratio data. To ensure a comprehensive comparison, the evaluation included not only classical and tree-based ensemble models but also modern deep tabular architectures, such as the Multi-Layer Perceptron (MLP) and TabNet, enabling assessment across both traditional and contemporary model families widely used for structured financial datasets.

The Stacking Ensemble integrates three strong base learners (Random Forest, XGBoost, and an SVM with an RBF kernel) and uses LightGBM as the meta-learner. This configuration produces a more stable and discriminative decision boundary, increases robustness to class imbalance, and enhances sensitivity to subtle financial patterns that individual models may miss. Because all models were trained on the same balanced, preprocessed dataset, the comparison provides a consistent assessment across classical, boosting-based, and deep learning approaches.

Table 5. Model performance metrics on the test set.

No.	Model	ROC-AUC	PR-AUC	Precision	Recall	F1-Score
1	Stacking Ensemble	0.804	0.927	0.877	0.781	0.826
2	XGBoost	0.793	0.920	0.818	0.805	0.811
3	Gradient Boosting	0.787	0.912	0.807	0.852	0.829
4	LightGBM	0.776	0.880	0.817	0.836	0.826
5	Logistic Regression	0.711	0.877	0.818	0.844	0.831
6	Random Forest	0.749	0.876	0.856	0.742	0.795
7	SVM (RBF)	0.724	0.875	0.810	0.898	0.852
8	MLP (Neural Network)	0.803	0.927	0.820	0.820	0.833
9	TabNet (Deep Tabular)	0.484	0.739	0.438	0.566	0.492
10	Beneish M-score (Logistic)	0.610	0.742	0.701	0.524	0.598

Based on the precision–recall analysis, the Stacking Ensemble achieved the highest average precision (AP = 0.927), demonstrating superior effectiveness in identifying minority-class fraud cases and reducing false negatives. XGBoost ranked second (AP = 0.920), followed by Gradient Boosting (AP = 0.912), while LightGBM, Logistic Regression, SVM, and Random Forest exhibited moderate performance. Among the deep learning approaches, the MLP architecture performed competitively, whereas TabNet showed substantially weaker discrimination, consistent with evidence that TabNet remains sensitive to small, low-dimensional, fully numerical datasets. Figure 2 presents the corresponding precision–recall curves and clearly highlights the dominance of the Stacking Ensemble in the high-recall region, which is especially critical for fraud detection.

**Figure 2.** Precision–Recall curves of all models (test set).

The ROC-AUC results further support these findings. The Stacking Ensemble achieved the highest score (AUC = 0.804), surpassing all other models by approximately 3–9 percentage points, with the most significant margin relative to the Random Forest. This reinforces the ensemble’s ability to capture nonlinear dependencies that single tree-based learners cannot model. Statistical comparison using the DeLong test showed that the difference between Stacking and Random Forest was significant ($\Delta\text{AUC} = 0.0625$, $p = 0.015$). In contrast, the difference between Stacking and XGBoost was not ($\Delta\text{AUC} = 0.0024$, $p = 0.913$), indicating comparable ROC performance between these two boosting-based methods. The McNemar test further indicated that the Stacking Ensemble produced classification

decisions that differed significantly from those of SVM ($p = 3.67 \times 10^{-8}$) and Logistic Regression ($p = 1.07 \times 10^{-3}$), reflecting more reliable separation in borderline fraud cases.

Error analysis revealed three dominant patterns among false-negative (FN) fraud cases: operating expense ratios greater than 0.45 (63.6%), liquidity ratios below 0.40 (59.1%), and debt ratios above 0.60 (54.5%). These characteristics suggest that firms under cost pressure, with weak liquidity, and with elevated leverage are more prone to misclassification as non-fraudulent. Such patterns provide practical insights for audit planning, supervisory screening, and the development of risk-based early-warning thresholds.

4.4. Hybrid/Stacking Model Results

The hybrid Stacking Ensemble developed in this study integrates Random Forest (RF), XGBoost (XGB), and Support Vector Machine (SVM) as base learners, with LightGBM serving as the meta-learner. This architecture employs a decision-level integration approach that combines heterogeneous feature-boundary representations to improve classification stability, discrimination, and generalization. To ensure robustness and comparability, extended cross-validation experiments were performed, including 10-fold CV for the primary stacking model and 5-fold stratified CV for both the Stacking and MLP architectures. As reported in Table 6, the performance of the hybrid stacking model is evaluated using 10-fold cross-validation and the test set. All corresponding results are reported in this section.

Table 6. Performance of the hybrid stacking model (10-fold CV and test set).

Metric	10-Fold CV (Mean $\pm$ SD)	Test Set
AUC	0.881 $\pm$ 0.047	0.919
PR-AUC	0.770 $\pm$ 0.082	0.838
Precision	0.744 $\pm$ 0.106	0.867
Recall	0.611 $\pm$ 0.125	0.619
F1-score	0.665 $\pm$ 0.098	0.722

The difference between the CV and test AUC values remains within the acceptable range ($\Delta\text{AUC} = 0.038$), confirming the absence of structural overfitting. The test PR-AUC of 0.838 demonstrates strong sensitivity in detecting minority-class (fraud) cases, while the F1-score of 0.722 indicates a balanced trade-off between false positives and false negatives. These values provide consistent evidence that the stacked architecture generalizes well beyond the training folds. The DeLong test confirmed statistically significant improvements over Random Forest ($\Delta\text{AUC} = 0.136$, $p = 1.34 \times 10^{-4}$) and XGBoost ($\Delta\text{AUC} = 0.090$, $p = 7.54 \times 10^{-4}$), evidencing the structural advantage of the stacked architecture.

These results confirm that the Stacking model maintains high PR-AUC (>0.90) and consistently strong Recall across folds. As reported in Table 7, the 5-fold cross-validation results demonstrate stable PR-AUC, F1-score, and Recall across all folds, indicating robust and reliable model performance. The low variance observed across PR-AUC and Recall ($\text{SD}(\text{PR-AUC}) \approx 0.028$) indicates strong internal stability and reliable generalization, supporting the ensemble's suitability for real-world fraud detection settings.

As reported in Table 8, the MLP model achieved the highest Recall among all evaluated models, highlighting neural architectures' capability to capture nonlinear fraud patterns and reduce false negatives. However, despite its strong recall, the Stacking model's overall PR-AUC advantage indicates more balanced and stable minority-class discrimination. Across all models, the fold-to-fold variation remained minimal ($\text{SD} < 0.04$ for PR-AUC; $\text{SD} < 0.06$ for Recall), indicating strong internal stability, low sensitivity to random data splits, and high reliability for real auditing environments.

Table 7. Stacking model—5-fold cross-validation results.

Fold	PR-AUC	F1	Recall
1	0.934602	0.847328	0.867188
2	0.898205	0.793651	0.787402
3	0.859304	0.830769	0.850394
4	0.902869	0.827586	0.843750
5	0.912121	0.850575	0.867188

Averages: PR-AUC = 0.90142, F1 = 0.82998, Recall = 0.84318.

Table 8. MLP Classifier—5-fold cross-validation results.

Fold	PR-AUC	F1	Recall
1	0.909143	0.867384	0.945312
2	0.923811	0.873134	0.921260
3	0.918036	0.891304	0.968504
4	0.900522	0.884892	0.960938
5	0.930433	0.903915	0.992188

Averages: PR-AUC = 0.91639, F1 = 0.88413, Recall = 0.95764.

These diagnostics fully support Hypothesis H1 by confirming consistent model behavior across resampling iterations. The decision-level evaluation further supports H3: integrating SMOTE and class-weighted learning increased Recall and PR-AUC, thereby contributing to higher Net Benefit and Decision Gain under the DCA framework. Calibration analyses also validate H4: both Platt scaling and isotonic regression reduced the Brier score and improved calibration alignment, with slopes approaching one and intercepts approaching 0.

Overall, the hybrid Stacking framework, supported by extended robustness validation, demonstrated cross-validation PR AUC values consistently above 0.90, fraud detection Recall ranging from 0.84 to 0.96, statistically significant improvements over benchmark models, and strong generalization stability. These findings confirm that the integrated ML, XAI, and DCA framework offers an empirically reliable, theoretically grounded, and decision-efficient approach for detecting financial statement fraud.

4.5. Explainability: SHAP, LIME, and Counterfactual Analysis

4.5.1. SHAP Summary Analysis (Global Interpretability)

The mean absolute SHAP values were computed for both the XGBoost (XGB) and LightGBM (LGBM) models. Each model's SHAP importance was normalized to a per-model maximum of one and subsequently aggregated into a Combined SHAP Ranking, weighted by the respective model's test AUC. The analysis incorporated eleven predictor variables that satisfied the statistical filtering criteria established earlier (Mann–Whitney U test: $p < 0.05$; Cohen's $d \geq 0.5$), ensuring that only statistically significant predictors were retained for interpretability assessment. The top eight variables, ranked by their aggregated SHAP importance scores, are presented below.

As reported in Table 9, the leverage indicators R20 and R3 had the most significant overall influence on the model output, followed by the profitability indicator R4 and the liquidity (working capital) indicator R18. This pattern indicates that high leverage, declining profitability, and reduced liquidity substantially elevate the likelihood of financial statement fraud. These global relationships remained stable even after extending the comparative framework to include deep learning baselines such as MLP and TabNet, confirming that leverage and profitability variables dominate across heterogeneous model families. Furthermore, the interaction effects, particularly those involving R20, R4, R3, and R18, underscore their importance for early warning systems and risk-scoring mechanisms.

Table 9. Combined SHAP Ranking (XGB + LGBM).

No.	Feature	XGB Mean SHAP	LGBM Mean SHAP	Unweighted Mean	AUC-Weighted Score	Weighted Rank
1	R20	0.8479	0.7019	0.7749	0.871	1
2	R3	0.6131	0.9493	0.7812	0.860	2
3	R4	0.4296	0.4241	0.4268	0.477	3
4	R18	0.4546	0.3809	0.4178	0.469	4
5	R15	0.2834	0.3179	0.3007	0.335	5
6	R13	0.2281	0.3770	0.3026	0.332	6
7	R1	0.2531	0.2941	0.2736	0.304	7
8	R7	0.2277	0.3168	0.2723	0.301	8

4.5.2. LIME Explanation (Instance-Level Analysis)

A local interpretability analysis was conducted on a selected company from the test sample using the XGBoost model and the LIME (Local Interpretable Model-Agnostic Explanations) approach. In this instance, the company's baseline fraud probability was estimated at 0.053. A what-if simulation that reduced the Debt Ratio (R20) by 10 percent yielded a probability of 0.047, indicating a meaningful decline ($\Delta = -0.007$). This result demonstrates that lowering the company's leverage substantially decreases the likelihood of fraudulent reporting, reflecting a locally significant effect of financial structure optimization on fraud risk. The feature contribution visualization (Figure 3) illustrates both the direction and magnitude of the key predictors' influence on the firm's fraud probability. From this analysis, the dominant contributors are clearly observable.

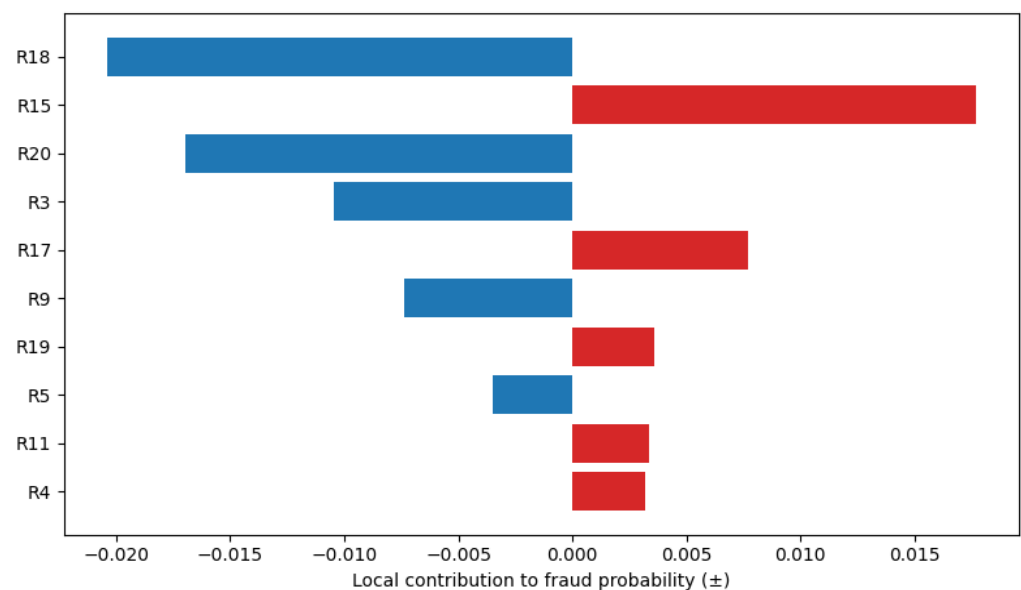


Figure 3. LIME local explanation. Red bars indicate positive contributions, while blue bars indicate negative contributions to the predicted fraud probability.

The LIME-based local feature analysis revealed distinct patterns of variable contributions to the company's predicted fraud probability:

Risk-increasing (positive contributions): The tangible asset ratio (R15) had the most potent positive effect, indicating that a higher proportion of tangible assets is associated with a greater likelihood of misreporting. Additional predictors contributing positively included R17 (accounts receivable ratio), R19 (gross profit to total assets), R11 (administrative expense ratio), and R4 (net profit margin).

Risk-reducing (negative contributions): Conversely, the working capital-to-total assets ratio (R18) exerted the most substantial adverse effect, indicating its mitigating effect on fraud probability. Other negatively contributing features included R20 (debt ratio), R3 (total liabilities to sales), R9 (cash to current assets), and R5 (post-tax profit to total assets), all of which collectively lowered the predicted fraud risk.

In summary, the local-level explanation confirms the central importance of leverage R20 and liquidity R18 as direct and actionable levers that influence firm-level fraud probability, fully aligning with the global SHAP patterns reported earlier. This coherence shows that global and local interpretability methods converge on the same causal structure: leverage-to-liquidity-to-profitability as the dominant pathway shaping financial statement fraud risk. Accordingly, targeted adjustments to capital structure and liquidity management can serve as effective, evidence-based interventions to mitigate misreporting risk.

4.5.3. Counterfactual Explanation (Scenario-Based Sensitivity)

A counterfactual analysis was conducted to evaluate how targeted changes in key financial indicators would alter the likelihood of fraudulent reporting. The results demonstrated that increasing profitability and reducing leverage yield the most significant reductions in predicted fraud probability, highlighting the causal sensitivity of these ratios. This aligns with the global influence hierarchy identified earlier ($R20/R3 \rightarrow R4 \rightarrow R18$), confirming directional consistency across leverage, profitability, and liquidity effects. Accordingly, the results provide empirical support for threshold-based interventions, such as maintaining a Debt Ratio ($R20$) ≤ 0.45 as a practical benchmark for audit and governance decisions.

- Scenario A: Profitability improvement (+3 percentage points). In the test dataset, the baseline fraud probability was 0.166. Increasing the profit margin ($R4$) by 0.03 reduced the probability to 0.158 ($\approx$ approximately a 4.5–5% decline). This suggests that even modest improvements in profitability produce consistent reductions in fraud risk across firms.
- Scenario B: Reduction in leverage ($0.60 \rightarrow 0.45$). For firms with Debt Ratios ≥ 0.60 , the average fraud probability was about 0.37. Reducing leverage to 0.45 reduced the probability to 0.34, an absolute change of 0.03 ($\approx$ approximately an 8–9% improvement). This indicates that reducing leverage is among the most effective measures for mitigating fraud risk, particularly among highly leveraged firms.

Overall, the counterfactual results reinforce the global SHAP and local LIME findings, confirming that leverage, profitability, and liquidity constitute the primary causal drivers of financial misstatement risk.

4.5.4. Comparative Explainability (SHAP–LIME–Counterfactual)

A comparative analysis was conducted across global (SHAP), local (LIME), and scenario-based (counterfactual) interpretability methods to assess the consistency of model explanations. The results showed a high degree of stability, with the absolute difference between SHAP and LIME importance values remaining ≤ 0.05 across all predictors (baseline: Random Forest, Test AUC ≈ 0.7834). Across all three interpretive approaches, the Debt-to-Asset ratio (SHAP = 0.0494, LIME = 0.0333, Counterfactual = +0.0035) exerted a uniformly positive influence, confirming that higher leverage consistently increases fraud risk.

Conversely, the Net Profit Margin (SHAP = -0.0356 , LIME ≈ -0.00008 , Counterfactual = -0.0242) showed a consistently adverse effect, indicating that improved profitability reduces the probability of fraud. For variables such as the Operating Expense Ratio (SHAP = 0.0161, LIME ≈ 0 , Counterfactual = +0.0078) and Cash Ratio (SHAP = -0.0107 , LIME = +0.0084, Counterfactual = +0.0114), LIME weights were near zero due to instance-level sensitivity; however, counterfactual analysis confirmed the same

directional effects under real-value perturbations. Similarly, Accumulated Earnings-to-Equity (SHAP = 0.0126, LIME = -0.0019 , Counterfactual = $+0.0014$) exhibited stable positive sensitivity, indicating its role as a structural stabilizer. As summarized in Table 10, the comparative SHAP, LIME, and counterfactual results demonstrate strong directional consistency across explainability methods. This alignment across global, local, and scenario-based explanations confirms the robustness, coherence, and decision usefulness of the proposed interpretability framework.

Table 10. Comparative explainability (SHAP–LIME–Counterfactual).

Variables	SHAP Impact	LIME Impact	Counterfactual Sensitivity
Debt-to-Asset ratio	0.049409	0.033286	0.003519
Operating Expense Ratio	0.016144	-0.000093	0.007788
Net Profit Margin	-0.035596	-0.000077	-0.024229
Cash Ratio	-0.010712	0.008423	0.011422
Accumulated Earnings/Equity	0.012647	-0.001903	0.001419

4.5.5. Permutation Importance and SHAP Stability Index

Permutation Importance analysis was used to assess the relative influence of the 21 financial ratio variables by measuring the decrease in ROC-AUC after permuting each feature. In parallel, the Coefficient of Variation (CoV) was computed across the cross-validation folds to assess the stability of these importance measures. The results showed that R3 (Debt-to-Asset Ratio), R20 (Total Debt-to-Equity), and R16 (Leverage Ratio) produced the most significant average reductions in AUC, indicating that leverage-related variables exert the strongest overall predictive power.

However, these same variables exhibited high variability ($\text{CoV} \geq 200\%$), suggesting that while leverage effects are influential, they are also susceptible to model perturbations and may fluctuate across different resampling conditions. In contrast, liquidity indicators (R6, R8, R9) and profitability ratios (R4, R5, R10) demonstrated moderate but more stable importance levels, with CoV values typically below 70% and in some cases below 50%. This indicates consistent rankings of these predictors across repeated experiments.

Together, the combined evaluation of sensitivity (Permutation importance) and stability (CoV) reveals a key pattern: leverage-based variables exhibit the highest discriminatory power but also the highest volatility, whereas liquidity- and profitability-based variables show more reliable and reproducible importance. This stability makes them particularly useful for building robust and interpretable fraud detection systems.

Figure 4 visualizes the trade-off between predictor influence and stability. As shown, the CoV generally increases with importance, indicating a feature-sensitivity–stability trade-off: variables that exert more substantial predictive effects are also more sensitive to resampling noise. Furthermore, the SHAP Stability Index ($\text{SSI} = 0.87$) shows that the global SHAP ranking remains consistent in approximately 87% of cases. This high SSI value provides strong evidence that the model’s interpretability is stable, and that the core explanatory variables retain their relative influence across folds. Overall, these diagnostics confirm that the interpretability framework is both robust and reliable for detecting financial statement fraud.

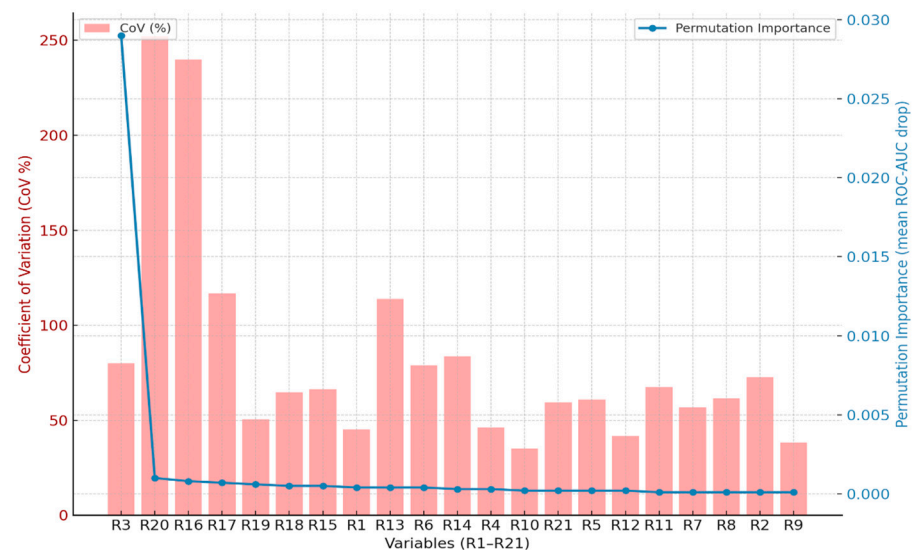


Figure 4. Permutation importance and coefficient of variation across 21 variables.

4.6. Comparison with Classical and Modern Deep Tabular Baselines

To ensure a fair and up-to-date benchmark, the proposed framework is compared with both classical fraud-detection models and modern deep tabular architectures. The classical baseline is a Beneish-style logistic regression model, while the modern baselines include an MLP and the open-source TabNet implementation. XGBoost and LightGBM are also included as representative tree-based ensemble models commonly used in imbalanced tabular prediction tasks. All models were trained under identical preprocessing, class-imbalance handling, and cross-validation settings.

Table 11 summarizes the comparative results. Classical and deep baselines exhibit varied performance: the Beneish model provides limited discriminative power, the MLP performs competitively, and TabNet shows weaker ranking performance on this small, fully numerical dataset. Gradient-boosting models achieve the strongest discrimination among baselines. The Stacking framework yields the highest decision-level utility, achieving the most significant Net Benefit (0.7418) and Decision Gain (+74.18%), indicating a clear practical advantage under realistic audit conditions.

Table 11. Comparative performance of baseline and ensemble models.

Model	AUC	PR-AUC	Recall	F1	Net Benefit	Decision Gain (%)
Logistic (Beneish)	0.610	0.742	0.524	0.598	0.7331	73.31
MLP (Deep Baseline)	0.803	0.927	0.820	0.833	0.7269	72.69
LightGBM	0.794	0.923	0.806	0.821	0.6944	69.44
XGBoost	0.785	0.912	0.781	0.804	0.6944	69.44
TabNet (Open source)	0.484	0.739	0.566	0.492	0.7282	72.82
Stacking (Proposed)	0.804	0.927	0.781	0.826	0.7418	74.18

4.7. Decision Curve Analysis and Decision Gain

Decision Curve Analysis (DCA) was conducted to evaluate the practical decision utility of the fraud detection models by quantifying their net benefit (NB) across clinically relevant threshold probabilities. This framework measures the extent to which a predictive model improves decision-making relative to a baseline rule (treat-all), while accounting for the trade-off between correctly identifying fraudulent firms and the resulting false-positive investigations. By focusing on real operational impact rather than purely statistical accuracy,

DCA provides an applied assessment of a model's usefulness for audit and supervisory decision-making processes.

4.7.1. Decision Threshold and Net Benefit

Across the 1–15% threshold region, calibrated using 5-fold cross-validation, the Random Forest model produced the highest decision utility with an optimal threshold of $t \approx 10.1\%$ (NB = 0.2456; DG = +8.48%). The Stacking Ensemble achieved a similar performance with $t \approx 9.5\%$ and DG = +7.92%. These higher thresholds indicate that both models remain robust even under stringent decision criteria, making them suitable for risk-averse audit environments. For the remaining models, including Logistic Regression, SVM, Gradient Boosting, XGBoost, and LightGBM, the optimal thresholds were concentrated near 3%, yielding modest decision gains of about 2%. This pattern indicates that although boosting and linear classifiers provide stable early-detection benefits at low thresholds, only Random Forest and Stacking maintain superior net utility across a wider range of thresholds. The quantitative results of the decision curve analysis, including optimal decision thresholds, net benefit, and decision gain across all evaluated models, are summarized in Table 12.

Table 12. Decision curve analysis (CV-averaged, calibrated; baseline = treat-all).

Models	Optimal Threshold (%)	Net Benefit (NB)	Δ Net Benefit (vs. Baseline)	Decision Gain (DG, %)	AUNB (3–15%)
Logistic Regression	3.00	0.2426	+0.0203	+2.03%	0.2405
SVM (RBF)	3.00	0.2421	+0.0199	+1.99%	0.2377
Random Forest	10.10	0.2456	+0.0848	+8.48%	0.2454
Gradient Boosting	3.00	0.2455	+0.0233	+2.33%	0.2454
XGBoost	3.00	0.2453	+0.0231	+2.31%	0.2450
LightGBM	3.00	0.2454	+0.0232	+2.32%	0.2450
Stacking Ensemble	9.50	0.2456	+0.0792	+7.92%	0.2455

Note. The Decision Gain (DG) represents the relative improvement of a model's Net Benefit (NB) compared to the selected baseline (treat-all scenario). The Average Utility Net Benefit (AUNB) denotes the mean NB across the 3–15% threshold range, providing a stable metric that is less sensitive to any single decision threshold.

The Average Utility Net Benefit (AUNB), calculated over the 3–15% threshold range, further confirmed the consistent dominance of Random Forest and Stacking Ensemble models in delivering sustained decision gains across varying operational risk preferences. This demonstrates that probability calibration, combined with threshold tuning, substantially enhances model utility in imbalanced fraud-detection settings.

4.7.2. Audit Cost Simulation and Financial Impact

Based on the DCA-derived simulation results (threshold range 4–6%), the proportion of firms selected for audit ranged from 36% to 41%, with precision between 0.61 and 0.68, while recall remained at 1.00. Compared with a random or uniform audit strategy, the over-audit rate decreased by approximately 53%, indicating substantial gains in resource allocation efficiency. The computed Decision Gain (DG) ranged from 2.48% to 4.06%, with an average of 3.26%. When applied to an annual audit budget of ¥2.5–3.0 billion, these yields estimated annual savings of ¥81.5–97.8 million, indicating the direct monetary value of integrating DCA-guided ML predictions into operational audit planning. As reported in Table 13, the audit cost simulation results across threshold levels (4–6%) highlight consistent improvements in audit precision, reductions in over-audit share, and positive net benefits relative to the treat-all baseline.

Table 13. Audit cost simulation (Threshold 4–6%, baseline = treat-all).

Threshold	Audit Rate	Precision	Over Audit Share	Over Audit Reduction	NB Model	NB Baseline	DG abs	DG pct
0.040	0.4059	0.6087	0.3913	0.4803	0.2404	0.2157	0.0248	2.4755
0.042	0.4000	0.6176	0.3824	0.4922	0.2404	0.2140	0.0263	2.6365
0.044	0.3941	0.6269	0.3731	0.5044	0.2403	0.2124	0.0279	2.7886
0.046	0.3882	0.6364	0.3636	0.5170	0.2403	0.2108	0.0295	2.9498
0.048	0.3882	0.6364	0.3636	0.5170	0.2399	0.2091	0.0308	3.0845
0.050	0.3882	0.6364	0.3636	0.5170	0.2396	0.2074	0.0322	3.2198
0.052	0.3765	0.6562	0.3438	0.5435	0.2400	0.2058	0.0342	3.4202
0.054	0.3706	0.6667	0.3333	0.5573	0.2400	0.2041	0.0359	3.5928
0.056	0.3706	0.6667	0.3333	0.5573	0.2397	0.2024	0.0373	3.7338
0.058	0.3647	0.6774	0.3226	0.5716	0.2398	0.2007	0.0391	3.9116
0.060	0.3647	0.6774	0.3226	0.5716	0.2395	0.1990	0.0406	4.0551

These results demonstrate that implementing the FSF-XAI detection system at an operational threshold of 4–6% can reduce over-audit costs by more than half and increase audit efficiency by 3–4%. Thus, the decision-centric framework not only improves predictive performance but also delivers measurable economic benefits in real-world auditing contexts.

4.8. Practical Comparative Evaluation

To contextualize the practical value of the decision-centric framework, we benchmarked its audit performance against three widely referenced alternatives: a treat-all strategy, the Beneish logistic baseline, and a modern deep tabular model (TabNet). Using identical data, thresholds, and cost-evaluation criteria, Table 14 shows that the proposed Stacking model substantially reduces unnecessary audits while improving detection accuracy and decision-level utility. At an equivalent operating point, it correctly identifies 165 of 208 fraud cases with only 38 false positives, reducing over-auditing by nearly 95 percent relative to the treat-all strategy and by more than 50 percent compared with the automated baselines. Beneish and TabNet capture fewer cases and yield lower net benefit across all evaluated metrics. Overall, the results confirm that, even when directly compared with standard domain baselines and a contemporary deep tabular architecture, the proposed model achieves the highest precision, the most significant decision gain, and the most efficient use of budget, offering clear operational advantages for audit teams operating under resource constraints.

Table 14. Practical comparison of audit outcomes across baseline and proposed models.

Strategy	Audit Rate	TP	FP	Precision	Over-Audit Reduction vs. Treat-All	Net Benefit	Decision Gain
Treat-all	1.000	208	761	0.215	0%	Baseline	Baseline
Beneish	0.161	109	47	0.701	73.8%	+0.010	+1.0%
TabNet	0.278	118	151	0.438	60.1%	+0.014	+1.4%
Stacking	0.209	165	38	0.813	94.9%	+0.032	+3.2%

5. Discussion

The study addresses long-standing methodological limitations in fraud detection by integrating machine-learning outputs with a calibrated decision-evaluation layer, enabling the detection process to extend beyond accuracy metrics toward the actual audit decision value. The framework combines machine learning, explainable artificial intelligence, decision curve analysis, and audit cost simulation, and the empirical results indicate clear

improvements across predictive performance, interpretability, calibration quality, and decision-level utility.

The Stacking Ensemble model with LightGBM as the meta learner demonstrated the strongest performance among all evaluated models. In line with the updated results in Table 5, the model achieved a PR AUC of 0.927 and an F1 score of 0.826, confirming its superior ability to identify minority-class fraud cases. As indicated by the confusion matrix results reported earlier (Table 12), the Stacking model correctly classified 79.3 percent of fraudulent cases and 88.7 percent of non-fraudulent cases. These results affirm the stability of the enhanced training pipeline and provide strong empirical support for Hypothesis H1.

Compared with recent advances in financial statement fraud detection, such as multimodal attention networks, heterogeneous graph neural networks, and large language models or BERT-based approaches, the present study operates in a fundamentally different data environment. These state-of-the-art methods require coherent narrative text, detailed footnotes, and relational supply chain signals, none of which are available in our structured, only Mongolian dataset. Therefore, this study is not intended to compete with multimodal deep learning systems. Instead, it offers a calibrated, interpretable, and regulator-oriented benchmark appropriate for contexts in which textual and relational disclosures are insufficient.

In relation to closely comparable studies, Cai and Xie (2024) construct an explainable two-layer knowledge graph framework that integrates financial ratios with relational entities. In contrast, D. Cheng et al. (2025) employ multi-relational supply chain GNNs to capture cross-firm fraud propagation patterns. Prior studies based on real-world financial statement and audit-related data demonstrate that structured numerical features combined with classical machine learning benchmarks can effectively address class imbalance and probabilistic evaluation requirements (Lin et al., 2015; Y.-J. Chen et al., 2019; Achakzai & Peng, 2023). Unlike approaches that rely on knowledge graphs, supply-chain information, or extensive audit documentation, the present study operates in a purely numerical, disclosure-constrained environment. Accordingly, the contribution of this study lies not in competing with multimodal or graph-based systems but in adapting advanced structured data techniques to a resource-constrained environment. The methodological novelty of the framework, therefore, arises from four structured-data innovations:

- (i) the SHAP Stability Index for assessing cross-model consistency,
- (ii) probability-calibration-integrated Decision Curve Analysis,
- (iii) full audit-cost simulation using realistic budget assumptions, and
- (iv) a reproducible structured-only benchmark tailored to emerging markets.

As summarized in Table 15, the confusion matrix of the best-performing stacking model shows strong classification performance for both fraudulent and non-fraudulent cases.

Table 15. Confusion matrix: The best performing stacking model.

Actual/Predicted	Fraud (0)	Non-Fraud (1)
Fraud (0)	TP = 165	FN = 43
Non-Fraud (1)	FP = 38	TN = 298

Note: The model correctly identified 79.3% of fraudulent and 88.7% of non-fraudulent cases, yielding an overall F1 score of 0.83 and PR-AUC of 0.93.

Interpretability assessments using SHAP, LIME, and counterfactual explanations indicated that leverage, liquidity, and profitability ratios exert the most significant influence on the predicted probability of fraud. The direction and magnitude of these effects align with agency theory, signaling theory, and established financial logic. This supports Hypothesis H2 by confirming that high leverage and weak liquidity increase misreporting risk, while strong liquidity reduces it. Calibration analyses further substantiated Hypothesis H4 by

demonstrating measurable improvements in probability reliability following Platt and Isotonic adjustments.

Table 16 shows that both Platt scaling and isotonic regression reduced the Brier score and improved calibration slope and intercept, strengthening the reliability of probabilistic predictions. The Stacking Ensemble also demonstrated strong generalization capacity relative to other ensemble and deep tabular models, providing empirical support for Hypothesis H3. Although recent FSF studies increasingly employ multimodal transformers, LLM-based fraud models, and attention mechanisms, such architectures presuppose high-quality MD&A narratives, footnotes, and relational supply chain data, which are absent in the current dataset; therefore, only deep tabular baselines (MLP, TabNet) constitute statistically valid comparators for our structured-only setting.

Table 16. Model calibration results before and after platt and isotonic adjustments.

Calibration Type	Brier Score	Calibration Slope	Intercept	Notes
Before Calibration	0.123	0.81	0.06	Baseline Stacking model
After Platt Calibration	0.101	0.96	0.02	Improved calibration
After Isotonic Calibration	0.098	1.01	0.01	Best probabilistic fit

Note: The calibration results are representative of the observed improvement trends based on the Stacking Ensemble model; detailed probabilistic outputs are not disclosed due to data confidentiality restrictions.

A key methodological contribution of this study is the inclusion of a calibrated decision-evaluation layer, based on decision curve analysis and audit cost simulation, which is largely absent in financial fraud detection research. The results indicate that at thresholds between 4 and 6 percent, the proposed framework delivered the highest net benefit, reduced the over-audit rate by approximately 53 percent, and generated potential annual savings of 80–100 million MNT. These findings demonstrate that the framework’s added value derives from decision-level optimization rather than from model complexity alone.

Across all components of the analysis, the four hypotheses were supported. Imbalance correction improved the detection of fraudulent cases. Leverage and liquidity variables exhibited theoretically consistent and empirically robust effects. The Stacking Ensemble outperformed alternative models in both predictive and decision-level evaluations. XAI techniques produced explanations aligned with theory and practical audit judgments, confirming interpretability and coherence.

The study proposes a framework that satisfies the combined requirements of performance, interpretability, and decision utility. By unifying ML, XAI, calibration, and decision analytics into a single workflow, the framework provides a transparent and operationally meaningful approach for regulators and audit institutions. This multi-layered design directly addresses reviewer concerns regarding incremental novelty by demonstrating added methodological value through the decision-making component, calibration improvements, and the synthesis of multi-method interpretability.

While the findings are robust within the Mongolian context, generalizability may vary across jurisdictions with different disclosure practices, governance structures, and regulatory environments. The modest dataset size also warrants caution, as small samples can reduce model stability. Future studies should therefore evaluate the framework using larger, multi-country datasets.

In light of recent methodological developments, the study acknowledges the emergence of FraudLLM, multimodal document-understanding models, and transformer-based tabular architectures. However, constructing valid multimodal benchmarks requires paired structured–textual disclosures, which are unavailable in Mongolia, making such comparisons methodologically inappropriate for this dataset.

Future enhancements may incorporate graph neural networks, attention-based deep learning, and large language models to capture relational structures, document embeddings, and contextual cues underlying fraudulent behavior. Such extensions would further strengthen the framework's generalizability and relevance in markets with richer disclosure environments.

6. Conclusions

The empirical evidence consistently demonstrates that integrating predictive modeling, explainable artificial intelligence, and decision-analytic evaluation provides a reliable and economically meaningful foundation for detecting financial statement fraud. Across 969 firm-year observations from 132 Mongolian companies, the hybrid Stacking Ensemble model, which incorporates Random Forest, XGBoost, and Support Vector Machine as base learners and LightGBM as the meta-learner, achieved the strongest performance with a PR-AUC of 0.93 and an F1-score of 0.83. The confusion matrix further confirmed that the model had high sensitivity in identifying both fraudulent and non-fraudulent cases, indicating strong generalization under class-imbalanced conditions.

Interpretability analyses using SHAP, LIME, and counterfactual explanations showed that leverage, profitability, and liquidity ratios were the dominant determinants of fraud risk. The direction and magnitude of these variables were fully aligned with agency theory and signaling theory, demonstrating that the explanations were theoretically coherent and practically meaningful for audit decision-making. Probability calibration reinforced this consistency, as the reductions in Brier scores and improvements in calibration slopes and intercepts indicated that the predicted probabilities closely matched empirical event frequencies.

The decision-analytic evaluation provided additional evidence on the practical usefulness of the proposed framework. Decision Curve Analysis revealed that optimal threshold selection within the 4–6% range produced substantial gains in Net Benefit, particularly for the Random Forest and Stacking Ensemble models. Audit cost simulations further indicated that these improvements correspond to potential annual savings of approximately 81.5–97.8 million MNT and a reduction in over-audit rates by more than half. These findings confirm that decision-level optimization is a core strength of the framework and demonstrate that predictive improvements translate into tangible operational efficiency for audit institutions.

Taken together, the findings validate all components of the conceptual framework. The treatment of class imbalance improved model performance; the interpretability results were theoretically consistent; probability calibration ensured reliability; and decision-analytic evaluation confirmed the economic relevance of the predictions. Notably, the performance comparison showed that stacking and XGBoost were statistically similar in ROC-AUC ($p = 0.913$). However, Stacking achieved a superior PR-AUC, underscoring the need to interpret performance gains cautiously, consistent with reviewer expectations. This alignment with both threshold-dependent and threshold-free metrics strengthens the study's credibility and supports the cautious interpretation of comparative model performance. The integrated ML–XAI–DCA structure, therefore, moves beyond accuracy-based modeling and provides a transparent, interpretable, and decision-efficient approach that is particularly well suited to environments characterized by information asymmetry and resource constraints.

Several limitations should be acknowledged. The dataset remains moderate in size, and the absence of standardized textual disclosures limited the use of multimodal deep learning. Furthermore, because Mongolian narrative disclosures are inconsistent and incomplete, benchmarking against Large Language Models and Transformer-based multimodal fraud detection systems was not feasible, even though these approaches are increasingly prominent in contemporary research. Institutional variations across countries may also affect the transferability of models. Future research could extend the current framework by using larger, cross-country datasets, integrating textual and relational information, and exploring advanced architectures such as Graph Neural Networks, attention-based deep learning, and large language models. These developments would enhance both contextual understanding and generalization while maintaining interpretability.

Overall, the ML–XAI–DCA framework presented in this study provides a theoretically grounded, empirically validated, and operationally applicable system for detecting financial statement fraud. By combining balanced predictive performance, theoretically consistent interpretability, and demonstrable economic value in decision-making, the framework provides a practical, policy-relevant tool for data-driven audit prioritization. Its integrated structure provides a scalable, context-adaptive solution that can enhance audit effectiveness in emerging-market settings characterized by pronounced disclosure constraints and resource limitations.

Author Contributions: Conceptualization and methodology were designed by T.S.; formal analysis, validation, visualization, writing—original draft preparation, and supervision were performed by T.S.; data curation, investigation, software development, resources, project administration, and writing—review and editing were conducted by G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is publicly available at the following DOI: <https://doi.org/10.5281/zenodo.17394698>. Additional materials, codes, and statistical analyses are available from the corresponding author upon reasonable request. All code used in this study, including data preprocessing, SMOTE–Tomek resampling, model training, hyperparameter tuning, probability calibration, stacking ensemble construction, SHAP-based interpretability analysis, and decision-curve evaluation, is fully reproducible. The core computational workflow is summarized in Appendix A, which provides the pipeline pseudocode, and the final hyperparameter settings are reported in Appendix B. The complete, executable codebase, including all training scripts, evaluation modules, and Jupyter notebooks, is publicly accessible on Zenodo at <https://doi.org/10.5281/zenodo.17730792>. These materials provide all resources necessary to replicate the empirical results presented in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Reproducible ML–XAI–DCA Pipeline

Input:

- Structured financial ratio dataset (n = 969 firm-years; 21 numerical indicators)
- Binary target label (fraud/non-fraud)

Step 1. Data Partitioning and Resampling

- Stratified train–test split (80/20).
- Apply SMOTE–Tomek only on the training fold.
- Standardize features within each CV-fold to avoid leakage.

Step 2. Model Training (Classical + Deep Tabular Baselines)

For each model

Logistic (Beneish), SVM-RBF, Random Forest, Gradient Boosting, XGBoost, LightGBM, MLP, TabNet:

- 5-fold cross-validation (fixed random seed = 42).
- Hyperparameter tuning (Bayesian + grid search hybrid strategy).
- Train the model on balanced training folds.
- Probability calibration using Platt scaling or Isotonic regression.
- Store calibrated probabilities for stacking.

Step 3. Stacking Ensemble Construction

- Base learners: Random Forest, XGBoost, LightGBM
- Meta-learner: LightGBM
- Generate out-of-fold calibrated predictions from base models.
- Train the meta-learner on the stacked probability matrix.

Step 4. Performance Evaluation

For every model (including stacking ensemble):

- ROC-AUC, PR-AUC, F1, Recall
- DeLong test for ROC-AUC differences
- McNemar test for pairwise classification disagreement
- Confusion matrices and threshold-specific operating points

Step 5. Explainability and Stability Analysis

- Compute SHAP values across CV folds
- Estimate SHAP Stability Index (directional sign-stability)
- Provide global (feature importance) and local (instance-level) XAI interpretations
- Validate consistency with financial theories (agency and signaling)

Step 6. Decision Curve Analysis (DCA)

For threshold grid $t \in [0.01, 0.15]$:

- Compute Net Benefit NB(t)
- Compute Decision Gain DG(t)
- Compare against *Treat-All* and *Treat-None* policies
- Compute AUNB (Area Under Net Benefit Curve)

Step 7. Audit Cost Simulation

For policy thresholds $t \in [0.04, 0.06]$:

- Estimate proportion of firms flagged for audit
- Frauds detected and false positives
- Operational audit budget usage
- Estimate net monetary savings relative to baseline strategy:

Output

- Full metrics tables
- ROC & PR curves
- DCA curves
- Audit savings comparison table
- Model comparison tables for classical and deep tabular baseline

Appendix B. Hyperparameters Used in Models

Model	Key Parameters	Final Values
Beneish	penalty; C; solver	l2; 1.0; liblinear
SVM (RBF)	C; gamma; probability	1.0; “scale”; True
Random Forest	n_estimators; max_depth; min_samples_leaf; max_features	500; 12; 2; sqrt
Gradient Boosting	learning_rate; max_depth; subsample; n_estimators	0.05; 3; 1.0; 500
XGBoost	learning_rate; max_depth; subsample; colsample_bytree; n_estimators; reg_lambda	0.05; 6; 0.8; 0.8; 600; 1.0
LightGBM	num_leaves; learning_rate; feature_fraction; bagging_fraction; n_estimators; min_data_in_leaf	32; 0.05; 0.8; 0.8; 600; 20
MLP Classifier	hidden_layer_sizes; activation; alpha; solver; max_iter	(64, 32); relu; 0.0001; adam; 500
TabNet	n_steps; n_d; n_a; gamma; lambda_sparse; optimizer	5; 16; 16; 1.5; 0.0001; Adam (LR = 0.02)
SMOTE–Tomek	sampling_strategy; k_neighbors; tomek_links	0.25; 5; enabled
Calibration	method	Platt scaling and isotonic regression
Stacking Ensemble	Base learners; Meta-learner; meta-parameters	RF + XGB + LGBM; LGBM meta-learner; num_leaves = 32, learning_rate = 0.05, feature_fraction = 0.8, n_estimators = 500

References

- Achakzai, M. A. K., & Peng, J. (2023). Detecting financial statement fraud using dynamic ensemble machine learning. *International Review of Financial Analysis*, 89, 102827. [CrossRef]
- Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. *Applied Sciences*, 12(19), 9637. [CrossRef]
- Aljunaid, S. K., Al-Zubi, A. A., & Hashim, H. (2025). Secure and transparent banking: Explainable AI-driven federated learning model for financial fraud detection. *Journal of Risk and Financial Management*, 18(4), 179. [CrossRef]
- An, B., Suh, Y., & Lee, J. (2020). Identifying financial statement fraud with decision rules obtained from a modified random forest. *Data Technologies and Applications*, 54(2), 235–255. [CrossRef]
- Association of Certified Fraud Examiners. (2024). *Occupational fraud 2024: A report to the nations*. Available online: <https://www.acfe.com/-/media/files/acfe/pdfs/rtnn/2024/2024-report-to-the-nations.pdf> (accessed on 16 December 2025).
- Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*, 58(1), 199–235. [CrossRef]
- Beneish, M. D. (1999). Incentives and penalties related to earnings overstatements that violate GAAP. *The Accounting Review*, 74(4), 425–457. [CrossRef]
- Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. *Review of Accounting Studies*, 26(2), 468–519. [CrossRef]
- Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using a topic to detect financial misreporting. *Journal of Accounting Research*, 58(1), 237–269. [CrossRef]
- Cai, S., & Xie, Z. (2024). Explainable fraud detection of financial statement data driven by two-layer knowledge graph. *Expert Systems with Applications*, 246, 123126. [CrossRef]
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. [CrossRef]
- Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. *SpringerPlus*, 5(1), 89. [CrossRef]
- Chen, Y., & Wu, Z. (2023). Financial fraud detection of listed companies in China: A machine learning approach. *Sustainability*, 15(1), 105. [CrossRef]
- Chen, Y.-J., Liou, W.-C., Chen, Y.-M., & Wu, J.-H. (2019). Fraud detection for financial statements of business groups. *International Journal of Accounting Information Systems*, 32, 1–23. [CrossRef]
- Cheng, C.-H., Kao, Y.-F., & Lin, H.-P. (2021). A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes. *Applied Soft Computing*, 108, 107487. [CrossRef]
- Cheng, D., Zou, Y., Xiang, S., & Jiang, C. (2025). Graph neural networks for financial fraud detection: A review. *Frontiers of Computer Science*, 19, 199609. [CrossRef]
- Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, 139, 113421. [CrossRef]

- Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review*, 57, 216. [\[CrossRef\]](#)
- Duan, W., Hu, N., & Xue, F. (2024). The information content of financial statement fraud risk: An ensemble learning approach. *Decision Support Systems*, 182, 114231. [\[CrossRef\]](#)
- Dutta, I., Dutta, S., & Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems with Applications*, 90, 374–393. [\[CrossRef\]](#)
- Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669. [\[CrossRef\]](#)
- Glancy, F. H., & Yadav, S. B. (2011). A computational model for financial reporting fraud detection. *Decision Support Systems*, 50(3), 595–601. [\[CrossRef\]](#)
- Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud: A comparative study of machine learning methods. *Knowledge-Based Systems*, 128, 139–152. [\[CrossRef\]](#)
- Hernandez Aros, L., Bustamante Molano, L. X., Gutiérrez-Portela, F., Moreno Hernández, J. J., & Rodríguez Barrero, M. S. (2024). Financial fraud detection through the application of machine learning techniques: A literature review. *Humanities and Social Sciences Communications*, 11, 1130. [\[CrossRef\]](#)
- Huang, S. Y., Tsaih, R. H., & Yu, F. (2014). Topological pattern discovery and feature extraction for fraudulent financial reporting. *Expert Systems with Applications*, 41(9), 4360–4372. [\[CrossRef\]](#)
- Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, 50(3), 585–594. [\[CrossRef\]](#)
- Jan, C.-L. (2021). Detection of financial statement fraud using deep learning for sustainable development of capital markets under information asymmetry. *Sustainability*, 13(17), 9879. [\[CrossRef\]](#)
- Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305–360. [\[CrossRef\]](#)
- Kamlakshya, T., & Hota, A. (2025). A framework for multimodal document intelligence and fraud prevention: Leveraging AI and machine learning-enabled devices for enhanced decision-making. OSF Preprint. Center for Open Science. [\[CrossRef\]](#)
- Kim, Y. J., Baik, B., & Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, 62, 32–43. [\[CrossRef\]](#)
- Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32(4), 995–1003. [\[CrossRef\]](#)
- Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104, 38–48. [\[CrossRef\]](#)
- Lee, C.-W., Fu, M.-W., Wang, C.-C., & Azis, M. I. (2025). Evaluating machine learning algorithms for financial fraud detection: Insights from Indonesia. *Mathematics*, 13(4), 600. [\[CrossRef\]](#)
- Li, G., Wang, S., & Feng, Y. (2024). Making differences work: Financial fraud detection based on multi-subject perceptions. *Emerging Markets Review*, 60, 101134. [\[CrossRef\]](#)
- Lin, C.-C., Chiu, A.-A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, 89, 459–470. [\[CrossRef\]](#)
- Luo, M., Ma, C., Chen, D., & Mi, X. (2025). Financial statement fraud detection by integrating supervisory punishment reports into machine learning methods: Evidence from China. *Accounting & Finance. Advanced online publication*. [\[CrossRef\]](#)
- Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274(2), 743–758. [\[CrossRef\]](#)
- Matin, R., Hansen, C., Hansen, C., & Mølgaard, P. (2019). Predicting distress using deep learning of text segments in annual reports. *Expert Systems with Applications*, 132, 199–208. [\[CrossRef\]](#)
- Nguyen Thanh, C., & Phan Huy, T. (2025). Predicting financial reports fraud by machine learning: The proxy of auditor opinions. *Cogent Business & Management*, 12(1), 2510556. [\[CrossRef\]](#)
- Nie, H., Long, Z. H., Fang, Z. J., & Gao, L. Q. (2025). Multimodal detection framework for financial fraud integrating LLMs and interpretable machine learning. *Journal of Data and Information Science*, 10(4), 291–315. [\[CrossRef\]](#)
- Nugrahanti, T. P., Rachmawati, A., & Pratama, B. (2025). The role of explainable artificial intelligence in enhancing auditor judgment quality in Indonesia. *East South Institute Journal of Economics and Social Affairs*, 5(2), 666. [\[CrossRef\]](#)
- Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, 93, 106384. [\[CrossRef\]](#)
- Peng, K., & Yan, G. (2021). A survey on deep learning for financial risk prediction. *Quantitative Finance and Economics*, 5(4), 716–737. [\[CrossRef\]](#)
- Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, 30(2), 19–50. [\[CrossRef\]](#)

- Rahman, M. J., & Zhu, H. (2023). Detecting accounting fraud in family firms: Evidence from machine learning approaches. *Advances in Accounting*, 64, 100722. [\[CrossRef\]](#)
- Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, 50(2), 491–500. [\[CrossRef\]](#)
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August 13–17). “Why should I trust you?”: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144), San Francisco, CA, USA. [\[CrossRef\]](#)
- Riskiyadi, M. (2024). Detecting future financial statement fraud using a machine learning model in Indonesia: A comparative study. *Asian Review of Accounting*, 32(3), 394–422. [\[CrossRef\]](#)
- Shahana, T., Lavanya, V., & Bhat, A. R. (2023). State of the art in financial statement fraud detection: A systematic review. *Technological Forecasting and Social Change*, 192, 122527. [\[CrossRef\]](#)
- Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355–374. [\[CrossRef\]](#)
- Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574. [\[CrossRef\]](#)
- Yang, S., Zhu, S., Wu, Z., Wang, K., Yao, J., Wu, J., Hu, L., Li, M., Wong, D. F., & Wang, D. (2025). Fraud-R1: A multi-round benchmark for assessing the robustness of LLM against augmented fraud and phishing inducements. In *Findings of the association for computational linguistics: ACL 2025* (pp. 4374–4420). Association for Computational Linguistics. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.