

Article

Enhancing Corporate Transparency: AI-Based Detection of Financial Misstatements in Korean Firms Using NearMiss Sampling and Explainable Models

Woosung Kim  and Sooin Kim *Department of Business Administration, Konkuk University, Seoul 05029, Republic of Korea;
kimws@konkuk.ac.kr

* Correspondence: sooinkim@konkuk.ac.kr

Abstract

Corporate transparency is vital for sustainable governance. However, detecting financial misstatements remains challenging due to their rarity and resulting class imbalance. Using financial statement data from Korean firms, this study develops an integrated AI framework that evaluates the joint effects of sampling strategy, model choice, and interpretability. Across multiple imbalance ratios, NearMiss undersampling consistently outperforms random undersampling—particularly in recall and F1-score—showing that careful data balancing can yield greater improvements than algorithmic complexity alone. To ensure interpretability rests on reliable predictions, we apply Shapley Additive Explanations (SHAP) and Permutation Feature Importance (PFI) only to high-performing models. Logistic regression emphasizes globally influential operating and financing accounts, whereas Random Forest identifies context-dependent patterns such as ownership structure and discretionary spending. Even with a reduced feature set identified by explainable AI, models maintain robust detection performance under low imbalance, highlighting the practical value of interpretability in building simpler and more transparent systems. By combining predictive accuracy with transparency, this study contributes to trustworthy misstatement detection tools that reinforce investor confidence, strengthen responsible corporate governance, and reduce information asymmetry. In doing so, it advances the United Nations Sustainable Development Goal 16 (Peace, Justice, and Strong Institutions) by supporting fair, accountable, and sustainable economic systems.

Keywords: financial misstatement detection; AI-based framework; NearMiss sampling; explainable AI; class imbalance; corporate transparency; sustainable corporate governance; sustainable development (SDG 16)



Academic Editor: Hyunchul Ahn

Received: 14 August 2025

Revised: 21 September 2025

Accepted: 25 September 2025

Published: 9 October 2025

Citation: Kim, W.; Kim, S. Enhancing Corporate Transparency: AI-Based Detection of Financial Misstatements in Korean Firms Using NearMiss Sampling and Explainable Models. *Sustainability* **2025**, *17*, 8933. <https://doi.org/10.3390/su17198933>

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) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Auditing has long been regarded as a cornerstone of capital market integrity, providing assurance that financial statements are prepared in accordance with accounting standards. However, audits face structural limitations stemming from managerial discretion, reliance on cooperation from the audited firm, and the constraints of risk-based, sample-based testing. These limitations often prevent auditors and regulators from detecting intentional misconduct, collusion, or subtle reporting irregularities in a timely manner. Consequently, a proactive assessment of the likelihood of financial misstatements has become a critical supplement to traditional audit and supervisory procedures.

Financial misstatements—whether arising from intentional fraud or unintentional error—distort firms’ true financial condition and undermine market credibility. In particular, material misstatements carry serious consequences for investors, regulators, and the broader economy. Ref. [1] estimated that corporate misreporting destroyed approximately 1.6% of the equity value of large U.S. firms annually, amounting to \$830 billion in 2021. It is therefore essential to distinguish between material and immaterial errors: while immaterial errors may be corrected without major impact, material misstatements can erode investor trust, compromise regulatory oversight, and destabilize sustainable economic systems. Because such cases undermine not only financial accuracy but also broader stakeholder confidence, they are directly linked to the issue of corporate transparency.

Corporate transparency plays a pivotal role in sustaining investor confidence, reducing information asymmetry, and strengthening accountability. Ref. [1] further reported that only one-third of corporate fraud cases were detected, and that prior to the Sarbanes–Oxley Act, as many as 41% of large U.S. firms were materially misreporting their statements, with 10% engaged in securities fraud. These undetected cases imposed an estimated annual cost of \$254 billion on investors, illustrating how failures in detection directly undermine transparency. Transparent reporting is not only a matter of regulatory compliance but also a foundation for sustainable corporate governance and the United Nations Sustainable Development Goal 16 (Peace, Justice, and Strong Institutions). Given the limitations of traditional audit and supervisory mechanisms, there is a growing need for data-driven approaches capable of identifying hidden irregularities before they escalate into systemic risks.

Against this backdrop, advances in artificial intelligence (AI) and machine learning offer powerful tools for detecting anomalies and uncovering non-linear patterns in financial data that traditional methods often overlook. However, their effectiveness has been constrained by the extreme class imbalance between misstated and non-misreported cases, which biases classifiers toward the majority class and reduces recall for rare but critical events. Prior research has emphasized algorithm selection but has given limited attention to how sampling strategies and class ratios jointly influence both predictive performance and interpretability.

This study seeks to fill this gap by developing an integrated AI-based framework that systematically evaluates the interaction between class imbalance, undersampling methods (NearMiss vs. Random), and explainable AI (XAI) techniques. The framework investigates how different undersampling strategies affect performance across imbalance ratios, how interpretability differs between linear and non-linear models under these conditions, and whether XAI can help simplify detection systems while preserving predictive power.

Building on these questions, the study makes several contributions. First, it provides one of the few systematic comparisons of undersampling strategies in financial misstatement detection, showing that careful data balancing can yield greater performance gains than algorithmic complexity alone. Second, it integrates predictive modeling with interpretability by applying Shapley Additive Explanations (SHAP) and Permutation Feature Importance (PFI) only to high-performing models, thereby generating reliable and transparent explanations. This analysis underscores that interpretability is model-dependent: linear models emphasize variables with consistent and proportional effects on misstatement risk, whereas tree-based models capture conditional and non-linear influences, where a variable’s importance may vary depending on firm-specific thresholds or contextual interactions. Finally, the study demonstrates that even with a reduced feature set identified through XAI, models preserve strong detection capability under low imbalance, offering practical guidance for building parsimonious yet effective detection systems. By combining predictive accuracy with transparency, this framework contributes to trustworthy mis-

statement detection tools that reinforce investor confidence, support sustainable corporate governance, and advance SDG 16.

The structure of this paper is as follows. Section 2 provides a literature review on financial misstatement detection and the methodological approaches commonly used in this field. Section 3 introduces the machine learning algorithms and XAI techniques employed for classification and interpretation. Section 4 describes the financial statement dataset used in this study and outlines the key variables and descriptive statistics. Section 5 presents and discusses the empirical results, including model performance and interpretability analyses. Finally, in Section 6, the key findings are summarized and their implications for practical application and future research are discussed.

2. Literature Review

2.1. Foundations and Classical Approaches

Reliable financial information is fundamental for the sound functioning of capital markets. External audits and regulatory oversight serve as key mechanisms for maintaining the credibility of financial reporting. Audits provide reasonable assurance that financial statements are prepared in accordance with accounting standards. However, due to inherent limitations, they cannot guarantee complete accuracy. According to IAS 200, the limitations of auditing stem from three characteristics of financial reporting and audit procedures. First, financial reporting involves management's judgment and estimation, which can lead to ambiguity or discretion in presentation. Second, the audit process depends on cooperation from the audited firm. As a result, auditors face substantial difficulty in detecting violations of accounting standards, especially when management engages in intentional misstatements or fraud, particularly through collusion or document falsification. Third, auditors are constrained by time and resource considerations, which necessitate risk-based planning and sample-based testing. However, sample-based testing may fail to detect misstatements if relevant transactions are excluded [2]. Similarly, supervisory authorities also operate under structural constraints when monitoring the accuracy of financial statements. For instance, in Korea, the Financial Supervisory Service (FSS) inspects firms selectively, based on risk assessments or external referrals, rather than through comprehensive reviews of all listed companies. These limitations underscore the importance of assessing the ex-ante likelihood of financial misstatements for enhancing the effectiveness of audit and regulatory enforcement mechanisms. In response, a substantial body of accounting literature has sought to identify key indicators and determinants of financial misstatements [3–7].

Although recent reviews emphasized growing methodological diversity [8,9], early misstatement detection relied on classical approaches. In particular, traditional accounting research often employed indirect proxies—such as discretionary accruals—as indicators of potential misreporting [10–12]. While these measures provided useful signals, they captured only limited aspects of managerial discretion and were often insufficient for distinguishing between benign reporting choices and intentional manipulation. Over time, the literature gradually moved beyond these proxies and began to develop predictive models grounded in statistical techniques, marking the first step toward more systematic and data-driven approaches.

Notable early models include the M-score proposed by [13], which combined eight financial ratios to detect earnings manipulation with explicit attention to type I and type II errors. Ref. [14] advanced this line of work with the F-score, a scaled logistic regression model that integrated financial, non-financial, and market-based variables, demonstrating superior performance relative to accrual-based measures when tested on U.S. Securities and Exchange Commission (SEC)'s Accounting and Auditing Enforcement Releases (AAERs). Ref. [15] further contributed by developing both Z-score and non-Z-score models to identify

misstated statements. The model achieves an accuracy of 84.21% when the Z-Score is included, whereas excluding the Z-Score improves the accuracy to 86.84%. Collectively, these approaches established an early foundation for misstatement detection research, offering useful benchmarks while also revealing limitations in capturing the complex, non-linear relationships and rare-event distributions that characterize financial reporting data.

2.2. Machine Learning and Recent Advances

The adoption of machine learning and advanced statistical methods has further expanded the scope of financial misstatement detection. The literature has gradually shifted toward predictive modeling using statistical and machine learning approaches [16,17], moving beyond traditional approaches such as accrual-based measures, financial ratios, and early statistical models (e.g., Logistic regression). A range of methods were applied: Ref. [18] employed Support Vector Machines (SVMs) with a financial kernel to improve detection over Logistic regression; Ref. [19] compared multiple classifiers and found Logistic regression and SVMs to perform competitively; Ref. [20] applied evolutionary algorithms such as genetic algorithms and Markovian Learning Estimation of Distribution Algorithm (MARLEDA) to develop fuzzy rule-based classifiers that outperformed traditional models; Ref. [21] demonstrated that decision trees (C5.0) achieved the highest accuracy in a hybrid framework; and ref. [22] showed that Classification and Regression Tree (CART) models were effective in distinguishing misstated statements in Chinese firms. More recent studies highlighted the advantages of ensemble methods: Ref. [23] employed Gradient Boosted Regression Trees (GBRT) to detect misstatements and extracted interpretable rules using the In-Trees algorithm, while ref. [24] found Random Forest to be particularly effective, with the debt-to-equity ratio emerging as a key predictor. Ref. [25] confirmed that neural networks and classification trees were valuable for detecting financial distress and fraud, and ref. [26] benchmarked seven widely used detection models—including the M-score, F-score, and machine learning variants—using SEC enforcement data.

Despite these advances, one of the most significant challenges in misstatement detection is the severe class imbalance: misstated cases are extremely rare relative to the large number of accurately reported ones. This imbalance biases standard classifiers toward the majority class, producing deceptively high accuracy but low recall for true misstatements [27]. To address this issue, studies have incorporated specialized resampling methods. Ref. [28] applied the MetaCost framework to distinguish intentional and unintentional errors under asymmetric misclassification costs. Ref. [29] developed a fraud detection model using RUSBoost algorithm in MATLAB, which integrates random undersampling with boosting, and demonstrated its superiority over traditional benchmarks such as [14,18]. Ref. [30] confirmed the effectiveness of RUSBoost in identifying material misstatements among Korean firms, while ref. [31] proposed a Modified Random Forest (MRF) approach that combined sub-dataset modeling with selective ensemble learning to produce both strong predictive performance and interpretable decision rules. Although these approaches are effective, they remain largely dependent on random undersampling techniques. In contrast, distance-based methods such as NearMiss [32] prioritize majority-class cases located near the decision boundary, offering theoretical advantages for generalization. However, while such methods have been applied in transactional fraud contexts [33,34], to the best of our knowledge they have not yet been systematically employed in the domain of financial misstatement detection.

Beyond addressing class imbalance, interpretability has become a critical concern in financial misstatement detection, as stakeholders require not only accurate but also transparent decision-support systems. XAI techniques such as SHAP and PFI provide valuable tools for identifying which input variables most influence predictions and for

clarifying how their effects vary across models [35,36]. However, prior studies cautioned that applying XAI to underperforming classifiers could yield misleading or uninformative explanations. For this reason, interpretability should be integrated only after sufficient detection performance is achieved. Once applied, XAI reveals that different algorithms emphasize different types of signals: linear models highlight variables with consistent and proportional effects on misstatement risk, while tree-based models capture conditional and non-linear patterns, where a variable's influence may depend on firm-specific thresholds or contextual interactions. This distinction underscores the importance of considering both model performance and interpretability when evaluating detection systems, as each contributes uniquely to sustainable governance and reliable decision-making.

Taken together, the literature demonstrates important progress but also leaves key gaps. Most prior work has relied on random undersampling, with little systematic evaluation of distance-based methods such as NearMiss in financial misstatement contexts. Moreover, while XAI has been proposed as a promising tool, its integration has been limited, and little is known about how interpretability varies across models and sampling strategies. By explicitly addressing these gaps, the present study provides one of the first comprehensive assessments that combines sampling strategies, machine learning models, and XAI to enhance both predictive performance and interpretability.

2.3. Research Gap and Contribution

Although prior studies have made important progress in developing models for financial misstatement detection, several critical limitations remain. First, most existing research has focused on algorithm selection or model complexity, while giving limited attention to how sampling strategies and class imbalance ratios jointly affect predictive outcomes. This is a notable omission, as financial misstatement cases are extremely rare and imbalanced distributions can strongly bias classifier performance. Second, while undersampling methods such as RUSBoost have been widely employed, the reliance on random undersampling restricts generalizability and risks discarding informative cases. Distance-based approaches such as NearMiss, which preserve boundary-relevant examples, remain largely unexplored in the financial reporting context despite their theoretical advantages. Third, although XAI techniques such as SHAP and PFI have been recognized as useful for interpreting model behavior, prior studies have not systematically examined how interpretability varies across different algorithms and sampling conditions. In particular, the relationship between linear and non-linear models in terms of the stability and context-dependence of feature importance remains under-investigated. Finally, much of the prior literature has relied on a limited set of theoretically defined financial ratios or pre-processed indicators, potentially overlooking the richness of raw accounting variables.

This study contributes to the literature by addressing these limitations in several ways. It systematically evaluates the interaction between class imbalance ratios and undersampling strategies, offering one of the few comprehensive comparisons between random and distance-based methods in financial misstatement detection. It further integrates predictive performance with interpretability by applying SHAP and PFI only to high-performing models, thereby ensuring that explanations are both reliable and transparent. By comparing linear and tree-based models, the study highlights how interpretability profiles differ across algorithms, revealing whether variables exert consistent global effects or conditional, context-dependent influences. Finally, by employing a comprehensive set of 90 raw accounting variables directly drawn from financial statements, the study departs from prior reliance on narrow, ratio-driven indicators and demonstrates the practical value of XAI for guiding feature selection and model simplification. Together, these contributions advance the development of trustworthy, transparent, and empirically grounded detection systems

that support investor confidence, evidence-based regulatory oversight, and sustainable corporate governance.

While many prior studies frame detection in terms of fraud, this study adopts the broader and more neutral term financial misstatement to encompass both intentional and unintentional irregularities. This choice avoids assuming that all irregularities are intentional and allows us to capture both errors and deliberate misconduct within a single detection framework.

3. Methodology

This section presents the machine learning algorithms and XAI techniques employed in this study for financial misstatement classification and post-hoc model interpretation. To capture a broad spectrum of modeling behaviors, we consider both linear and non-linear classifiers. Logistic regression is included as a representative linear model, valued for its interpretability and extensive use in prior financial misstatement detection research. In this study, Logistic regression serves as a complementary model that allows comparison between linear and non-linear structures. Its transparent coefficients provide a clear reference for interpreting feature effects. When contrasted with tree-based models, these estimates highlight how different algorithmic assumptions lead to varying assessments of variable importance. To account for complex interactions among financial indicators, we further employ a range of tree-based models—including decision trees and ensemble methods such as Random Forest, XGBoost, Gradient Boosting, CatBoost, and AdaBoost. While individual trees are inherently interpretable, ensemble variants improve predictive performance at the cost of transparency. To address this challenge, we complement these models with two XAI techniques—SHAP and PFI—which clarify how input features influence predictions, thereby balancing predictive power with interpretability in auditing and regulatory contexts. All algorithms and experiments were implemented in Python 3.13.7, using scikit-learn (logistic regression, decision tree, random forest, gradient boosting, AdaBoost, and permutation feature importance), imbalanced-learn (NearMiss undersampling), SHAP (explainable AI), as well as the XGBoost and CatBoost libraries.

3.1. Logistic Regression

Logistic regression is a linear algorithm widely used for binary classification. In financial misstatement detection research, it has been applied to identify misstatements using financial ratios such as liquidity, leverage, and profitability [15,19]. Its main strength lies in interpretability: each coefficient directly shows how a variable affects the likelihood of misstatement, which is particularly valuable in auditing and regulatory contexts where transparency is essential.

However, prior studies also point to clear limitations. Logistic regression assumes a linear relationship between predictors and the log-odds of the outcome. This restricts its ability to capture complex interactions among financial indicators, which can reduce predictive accuracy in practice. For this reason, more recent studies have increasingly relied on non-linear and ensemble approaches.

In our study, logistic regression is retained not only for its interpretability but also to provide a complementary perspective to tree-based models. By applying SHAP and PFI, we are able to compare how feature importance differs between linear and non-linear structures. This comparison sheds light on how modeling assumptions influence the interpretation of financial indicators in misstatement detection, thus addressing a gap left by earlier studies that often evaluated models in isolation.

3.2. Decision Tree

Decision trees are non-parametric, tree-structured algorithms widely used for classification tasks, including financial misstatement detection. They recursively partition the feature space into regions that are increasingly homogeneous with respect to the target class. This process, known as recursive binary splitting, selects an input feature and a threshold at each node to divide the data into two subsets. The algorithm evaluates candidate splits using an impurity measure such as the Gini index or entropy. For example, the Gini impurity quantifies the likelihood of misclassification within a node, and the algorithm seeks the split that maximizes the reduction in impurity across child nodes. This process continues until a stopping criterion is met, such as a maximum depth or a minimum number of samples per node. To classify a new observation, the model traverses the tree from root to leaf, applying a sequence of binary decision rules. The final prediction corresponds to the majority class among the training samples in the reached terminal node.

Decision trees offer key advantages in misstatement detection, as they capture complex, non-linear interactions among financial variables while maintaining an intuitive and interpretable structure. This transparency is particularly valuable in auditing and regulatory settings, where the explainability of model decisions is essential [19,22,25].

3.3. Random Forest

Random Forest is an ensemble learning algorithm that constructs many decision trees and combines their predictions to improve classification accuracy and generalization. RF addresses the limitations of single decision trees by constructing a large number of trees and aggregating their predictions, thereby reducing variance and improving model stability. Each tree is trained on a bootstrap sample drawn with replacement from the original dataset. At each node, a random subset of predictors is considered for splitting, and the optimal split is selected based on a chosen impurity measure such as the Gini index. This dual-randomization—across both data samples and feature subsets—produces decorrelated trees, which enhances the ensemble's ability to generalize to unseen data. Final predictions are obtained through majority voting across all trees in the forest.

In the context of financial misstatement detection, Random Forest offers several practical advantages [24,31]. It effectively captures non-linear relationships and higher-order interactions among financial indicators that may signal misreporting. Moreover, Random Forest is resilient to missing values, enabling reliable predictions even when certain disclosures are incomplete—a frequent challenge in real-world financial data. Importantly, Random Forest includes built-in mechanisms for evaluating feature importance, which align well with post hoc interpretability methods such as SHAP and PFI. This synergy supports transparent and explainable model outputs, which is essential for deployment in auditing and regulatory contexts.

3.4. Gradient Boosting Algorithm

Gradient Boosting is an ensemble learning technique that constructs a strong predictive model by iteratively adding decision trees, each of which aims to minimize the residual errors of the existing ensemble [37]. Unlike Random Forest, which employs a bagging strategy by training multiple trees in parallel on bootstrap samples to reduce variance, Gradient Boosting follows a sequential boosting paradigm: each subsequent tree is trained to correct the prediction errors made by its predecessors. While bagging improves generalization by averaging uncorrelated predictors, boosting primarily reduces bias by focusing learning on difficult-to-classify instances.

For financial misstatement detection, Gradient Boosting offers distinct advantages. It can capture complex, non-linear interactions among financial indicators and is partic-

ularly effective in handling class imbalance by concentrating learning on minority cases. These properties make it well-suited to financial misstatement detection, where misstated observations are rare but carry high significance. For this reason, recent studies have employed boosting algorithms and demonstrated their strong performance in detecting accounting fraud [23,29]. Despite these strengths, boosting models can be complex and prone to overfitting if not properly regularized. This limitation has motivated the development of extensions such as XGBoost, CatBoost, and AdaBoost, which introduce additional mechanisms to improve scalability, robustness, and handling of specific data challenges. These extensions are discussed in the following subsections.

3.5. XGBoost Algorithm

XGBoost (eXtreme Gradient Boosting) extends the standard Gradient Boosting framework through several innovations that improve both regularization and computational efficiency. A key enhancement is the incorporation of L1 (Lasso) and L2 (Ridge) regularization terms into the objective function, which penalize model complexity and encourage sparsity in leaf weights. These mechanisms help mitigate overfitting, particularly in high-dimensional datasets. XGBoost also introduces a novel approach to handling missing values and sparse inputs. Rather than relying on imputation, the algorithm learns the optimal default direction for each split during training, allowing it to incorporate missing values seamlessly without degrading predictive accuracy.

Beyond these improvements, XGBoost offers distinct advantages for financial misstatement detection. Regularization helps prevent overfitting when the number of financial indicators is large relative to misstated cases. Its ability to handle missing values natively makes it robust to incomplete corporate disclosures, a frequent challenge in real-world data. Moreover, its computational scalability enables efficient training on large firm-level datasets. Empirical studies confirm these strengths, showing that XGBoost consistently delivers high predictive accuracy in detecting accounting misstatements [38].

3.6. CatBoost Algorithm

CatBoost extends the Gradient Boosting framework with specialized mechanisms designed to handle categorical variables effectively and reduce overfitting [39]. A major innovation is the use of ordered target statistics for encoding categorical features. Instead of relying on one-hot encoding or arbitrary numerical mappings, CatBoost computes statistical encodings based on target distribution estimates while minimizing the risk of target leakage. Another distinguishing feature is ordered boosting, which modifies the way residuals are calculated during training. By computing residuals only on preceding observations within random permutations of the data, CatBoost reduces the likelihood of overfitting. In addition, CatBoost builds symmetric trees—splitting at the same feature across all branches at a given depth—which improves both stability and computational efficiency. In general, these features make CatBoost particularly advantageous in domains where categorical variables are prevalent, as it can capture informative patterns without requiring extensive preprocessing.

Although CatBoost has demonstrated strong performance in related contexts such as financial transaction fraud and credit card fraud detection [40–42], it has not previously been applied to the detection of financial misstatements. Our study therefore explores its effectiveness in this novel domain and benchmarks its performance against other boosting algorithms.

3.7. AdaBoost Algorithm

AdaBoost departs from Gradient Boosting in both its mathematical foundation and its optimization strategy (Freund & Schapire, 1997 [43]). While Gradient Boosting incre-

mentally fits new decision trees to the negative gradient of a specified loss function using gradient-based optimization, AdaBoost employs an adaptive reweighting mechanism. Specifically, at each iteration, the algorithm increases the weights of misclassified instances, thereby directing subsequent learners to concentrate on the most challenging cases in the training set. Each decision tree (often a shallow “stump”) contributes to the final ensemble based on its classification accuracy, with more accurate learners receiving higher weights in the model’s aggregated output. This approach enables AdaBoost to enhance performance even with relatively simple base learners. However, such emphasis on hard-to-classify examples can also make the algorithm sensitive to noisy data and outliers, which may disproportionately influence the learning process. Despite this limitation, AdaBoost remains an effective and interpretable method for classification tasks, particularly when model simplicity is prioritized.

In the context of financial misstatement detection, AdaBoost has rarely been applied in prior research, with only limited examples such as [44], who explored data mining techniques to identify financial restatements. Given this gap, we include AdaBoost in our study to evaluate its potential in this domain and to benchmark its performance against more widely used boosting algorithms.

3.8. Explainable AI Techniques: SHAP and PFI

As the complexity of machine learning models increases, their interpretability has become a critical concern, particularly in domains where transparency is essential for decision-making. Black-box models such as Random Forests and boosting algorithms often deliver high predictive performance, yet their opacity hinders stakeholder trust and limits adoption in sensitive fields such as financial misstatement detection. To address this issue, explainable artificial intelligence (XAI) techniques have emerged to provide interpretable insights into model behavior.

Among these, SHAP offers a principled method for attributing model predictions to individual features. Based on Shapley values from cooperative game theory [45], SHAP decomposes a model’s prediction into additive contributions from each feature by averaging their marginal impact across all possible feature combinations. This ensures consistency and local accuracy, making SHAP particularly well-suited for understanding feature effects in non-linear models. In contrast, PFI provides a model-agnostic approach to quantifying the global relevance of each feature. By randomly permuting the values of a variable and measuring the resulting drop in model performance, PFI evaluates how much the model relies on that feature to make accurate predictions. A larger decline indicates greater importance, whereas a negligible change suggests a limited role. However, PFI assumes feature independence and may yield biased results when predictors are strongly correlated.

SHAP and PFI thus serve complementary roles in model interpretation: SHAP focuses on instance-level explanations, while PFI highlights overall feature relevance across the dataset. Both techniques were employed in this study to better understand the behavior of tree-based models in financial misstatement detection, where recent studies have also demonstrated their usefulness [35,36].

3.9. Nearmiss Undersampling

NearMiss is a distance-based undersampling technique designed to address class imbalance in binary classification problems. Unlike random undersampling, which removes majority class instances indiscriminately, NearMiss instead retains majority examples that are closest to the minority class in the feature space. By focusing on these boundary-near instances, the method enhances the classifier’s ability to learn from the most informative and ambiguous cases. This, in turn, improves detection performance for the minority class.

Originally proposed by [32], NearMiss aims to preserve class boundaries while mitigating the imbalance ratio. By concentrating on samples near decision margins, it increases model sensitivity to underrepresented patterns without discarding valuable information from the majority class. This characteristic has proven useful in domains such as medical diagnosis, credit card fraud, and fault detection, where false negatives are especially costly [33,46].

Despite these theoretical strengths, NearMiss has received little attention in financial misstatement detection. To our knowledge, no prior study has systematically applied this technique in the accounting domain. By introducing NearMiss to this context, our study addresses this gap and provides a new methodological perspective on handling class imbalance in misstatement detection tasks.

4. Data

Our initial sample comprises all publicly listed non-financial firms on the Korea Stock Exchange (KSE) and the Korea Securities Dealers Automated Quotations (KOSDAQ) from 2009 to 2023. The sample period begins in 2009, the earliest year for which inspection-related enforcement data are available, and ends in 2023, one year prior to the most recent inspection cycle. We hand-collected data on firms identified as violating accounting standards through inspections conducted by the Financial Supervisory Service (FSS) between 2011 and 2024. Because these inspections are carried out on financial statements that have already been disclosed, they apply to earlier reporting periods. Firm-year observations with missing total asset values are excluded to ensure data accuracy, while other missing items are set to zero. After applying these criteria, the final dataset consists of 29,134 firm-year observations.

Under the Act on External Audit of Stock Companies, the Securities and Futures Commission (SFC) oversees the financial statements submitted by firms subject to this legislation. The SFC may delegate authority to the Governor of the FSS to inspect a company's accounting records, financial documents, business operations, and overall financial condition, including those of its affiliates. When the FSS identifies accounting misstatements, the firm may be subject to enforcement sanctions, such as monetary penalties, auditor designation by the SFC, restrictions on equity issuance, dismissal of management, or referral to the Prosecutors' Office. In addition, the FSS publicly discloses firms found to have violated accounting standards. These disclosures specify the year of inspection, the fiscal years of the statements under review, the fiscal year(s) during which the violations occurred, and the nature of the violations.

In this study, we define financial misstatements as cases in which firms were found by the FSS to have violated accounting standards, based on official inspection outcomes. While this definition ensures objectivity and consistency with regulatory findings, it may not capture all instances of misreporting. For example, some firms may not have been inspected by the FSS, and others may have voluntarily restated their financials without regulatory sanction. Nonetheless, the FSS enforcement data provide a well-established and authoritative benchmark that has been widely adopted in prior studies on financial misstatement detection. Based on this classification, we construct a binary dependent variable indicating whether a firm-year observation involves a financial misstatement.

Table 1 presents the number of public firms, the number of accounting misstatements detected by the FSS, and the proportion of misstatements relative to all public firms in our sample, based on financial data from Dataguide Pro and enforcement disclosures from the FSS. In each year, fewer than 4 percent of public firms were identified as committing financial misstatement. In total, there are 598 such cases, representing approximately 2 percent of all firm-year observations. Because the FSS conducts inspections several

years after financial statements are disclosed, the frequency of detected misstatements is lower in recent years. This lag suggests that undetected accounting violations may exist in those periods.

Table 1. Yearly Distribution of Observations and Detected Misstatements.

Year	Number of Observations (A)	Number of Misstatements (B)	Misstatement Rate (B/A)
2009	1651	42	2.54%
2010	1653	59	3.57%
2011	1688	55	3.26%
2012	1694	65	3.84%
2013	1714	60	3.50%
2014	1759	68	3.87%
2015	1844	69	3.74%
2016	1895	62	3.27%
2017	1967	47	2.39%
2018	2037	37	1.82%
2019	2115	23	1.09%
2020	2179	7	0.32%
2021	2254	3	0.13%
2022	2304	1	0.04%
2023	2380	0	0.00%
Total	29,134	598	2.05%

Notes: Based on Dataguide Pro and FSS enforcement disclosures, 2011–2024.

In this study, we train our models using a broad set of raw accounting variables. While many prior studies rely on financial ratios derived from theoretical frameworks, we do not restrict our inputs to any predetermined set. Instead, we utilize line-item level accounts that are directly observable in firms’ financial statements, including the statement of financial position, income statement, and cash flow statement. This approach allows for greater transparency and replicability, as it depends only on readily available data. A full list of variables used in this study is presented in the Appendix A.

5. Results and Discussion

5.1. Effects of Undersampling and Class Imbalance on Model Performance

This section presents the empirical results of our classification experiments on financial misstatement detection, focusing on how different undersampling strategies and class imbalance ratios affect model performance. In particular, we compare two undersampling methods—NearMiss and random undersampling—across multiple class distributions to evaluate their effectiveness in training financial misstatement detection models. To complement this performance analysis, the next subsection provides an interpretability assessment using XAI techniques, applied only to those models and class ratio settings that demonstrated sufficiently high performance.

5.1.1. Experimental Design and Performance Metrics

The experiments were conducted using stratified 5-fold cross-validation to ensure robustness and generalizability. In this approach, the dataset was partitioned into five folds such that the proportion of misstated and non-misstatement cases was preserved within each fold. This stratification ensures that the minority class (misstated cases) is adequately represented during both training and validation phases. For each fold, the ratio of non-misstatement to misstated cases was systematically varied, and the selected undersampling method was applied to the training set. Seven commonly used classification algorithms were evaluated: Logistic regression, Decision Tree, Random Forest, Gradient Boosting,

XGBoost, CatBoost, and AdaBoost. Model hyperparameters were optimized using grid search within the training folds.

Model performance was evaluated using four standard classification metrics: accuracy, precision, recall, and F1-score. Accuracy measures the overall proportion of correctly classified instances among all predictions. Precision quantifies the proportion of correctly predicted misstated cases among all cases predicted as misstated, indicating how often the model's positive predictions are correct. Recall, which is particularly important in the context of financial misstatement detection, measures the proportion of actual misstated cases that were correctly identified by the model. In addition to the recall for the misstatement class, we also report the macro-average recall, which reflects the model's ability to detect both classes by computing the unweighted average recall across the misstatement and non-misstatement classes. The F1-score, defined as the harmonic mean of precision and recall, provides a balanced assessment of the model's capacity to identify financial misstatements while limiting false positives.

The formal definitions of these metrics are presented below.

$$Accuracy (A) = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision (P) = \frac{TP}{TP + FP} \quad (2)$$

$$Recall (R) = \frac{TP}{TP + FN} \quad (3)$$

$$F1-score (F) = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

In these definitions, TP (True Positive) denotes the number of misstated cases correctly identified as misstated, while TN (True Negative) represents the number of non-misstatement cases accurately classified as non-misstatement. FP (False Positive) refers to the instances in which non-misstatement cases were incorrectly labeled as misstated, and FN (False Negative) indicates the misstated cases that the model failed to detect. This study adopts the convention of designating misstated cases as the positive class, consistent with established practices in financial misstatement detection, where the primary objective is to maximize the identification of rare but critical positive instances. Accordingly, the metrics of precision, recall, and F1-score are all calculated relative to the model's ability to correctly identify financial misstatements. This focus reflects the practical priority in financial misstatement detection of minimizing false negatives, which carry significant regulatory and financial risks.

Specifically, accuracy measures the overall proportion of correct predictions among all cases, precision quantifies the proportion of cases predicted as misstated that are truly misstated (thus reflecting the reliability of positive predictions), recall captures the model's sensitivity to detecting actual misstatement cases, and the F1-score provides a harmonic mean that balances precision and recall. Together, these metrics enable a comprehensive evaluation of classification performance, highlighting both the model's general accuracy and its effectiveness in identifying financial misstatements.

5.1.2. Empirical Results and Interpretation

Table 2 presents the model performance across different undersampling strategies and class ratio settings characterized by a low proportion of majority class instances. A, F, P, and R denote accuracy, F1-score, precision, and recall, respectively. LR refers to Logistic regression; Tree indicates decision tree; RF denotes Random Forest; and GB represents the Gradient Boosting Algorithm.

Table 2. Performance Metrics under Lower Class Imbalance.

Class Ratio	Algorithm	Random Sampling				NearMiss			
		A	F	P	R	A	F	P	R
500:598	LR	0.673	0.718	0.677	0.764	0.842	0.859	0.836	0.885
	Tree	0.672	0.694	0.706	0.682	0.903	0.911	0.911	0.911
	RF	0.782	0.814	0.761	0.876	0.947	0.951	0.956	0.946
	XGBoost	0.783	0.810	0.775	0.848	0.939	0.943	0.955	0.931
	GB	0.755	0.786	0.750	0.824	0.940	0.944	0.954	0.935
	CatBoost	0.776	0.805	0.765	0.849	0.942	0.946	0.954	0.938
	AdaBoost	0.665	0.691	0.694	0.689	0.903	0.911	0.916	0.906
600:598	LR	0.648	0.674	0.626	0.731	0.841	0.847	0.818	0.878
	Tree	0.656	0.656	0.655	0.657	0.905	0.904	0.907	0.901
	RF	0.783	0.788	0.768	0.809	0.949	0.948	0.964	0.933
	XGBoost	0.779	0.784	0.764	0.806	0.936	0.935	0.944	0.926
	GB	0.768	0.773	0.756	0.791	0.942	0.940	0.960	0.921
	CatBoost	0.794	0.801	0.772	0.833	0.946	0.945	0.963	0.926
	AdaBoost	0.668	0.674	0.661	0.687	0.917	0.917	0.924	0.910
1000:598	LR	0.655	0.537	0.540	0.535	0.820	0.775	0.729	0.828
	Tree	0.695	0.586	0.596	0.575	0.918	0.890	0.894	0.886
	RF	0.806	0.718	0.788	0.659	0.944	0.922	0.964	0.885
	XGBoost	0.815	0.750	0.759	0.741	0.952	0.934	0.961	0.908
	GB	0.801	0.724	0.752	0.699	0.946	0.925	0.957	0.895
	CatBoost	0.825	0.760	0.782	0.739	0.951	0.933	0.964	0.903
	AdaBoost	0.715	0.630	0.612	0.649	0.913	0.883	0.888	0.878
1500:598	LR	0.718	0.440	0.508	0.388	0.825	0.714	0.667	0.768
	Tree	0.745	0.545	0.555	0.535	0.915	0.847	0.870	0.826
	RF	0.827	0.627	0.813	0.510	0.951	0.909	0.971	0.853
	XGBoost	0.846	0.703	0.780	0.640	0.945	0.898	0.946	0.855
	GB	0.814	0.627	0.732	0.548	0.945	0.898	0.946	0.855
	CatBoost	0.838	0.690	0.762	0.630	0.951	0.909	0.963	0.861
	AdaBoost	0.733	0.548	0.529	0.569	0.900	0.825	0.828	0.821

Notes: Based on Dataguide Pro and FSS enforcement disclosures, 2011–2024.

Several important observations can be drawn from these results, offering insights into algorithmic behavior, sampling strategies, and the trade-offs among performance metrics. Ensemble models based on decision trees—particularly Random Forest, XGBoost, and CatBoost—consistently outperformed simpler classifiers such as Logistic regression and single decision trees. This finding aligns with prior studies demonstrating the advantages of tree-based ensembles in financial misstatement detection and other high-dimensional tasks. Notably, even single decision trees often achieved higher recall and F1-scores than Logistic regression. This difference reflects their underlying assumptions: Logistic regression imposes additive and proportional effects of each feature on the log-odds, whereas tree-based methods capture threshold effects and interactions, enabling features with weak average influence to become highly predictive in specific contexts.

When examining the effect of class ratio, recall improved markedly when datasets were close to balanced (e.g., 500:598 or 600:598). As imbalance increased, recall declined

across models, while accuracy often rose because the majority class was easier to predict. This occurs because models exposed to a larger volume of normal cases become biased toward labeling observations as non-misstatement cases. Consequently, a model may appear strong in accuracy but still miss many misstated cases, underscoring the importance of recall in audit applications.

The configuration of training data—combining class ratio and sampling method—had a stronger impact on performance than algorithm choice itself. NearMiss generally outperformed random undersampling across accuracy, precision, recall, and F1-score. Random undersampling, especially under high imbalance, often reduced recall and F1-score because majority cases were removed without regard to their similarity to misstatement cases. By contrast, NearMiss enriched the training set with borderline examples, helping models learn more discriminative patterns. For practitioners, this shows that appropriate resampling is essential, as relying solely on raw data may limit the ability to capture the defining characteristics of financial misstatements.

When focusing specifically on recall, sampling design often outweighed algorithm sophistication. Under NearMiss with near-balanced ratios, even simple classifiers achieved higher recall than advanced ensembles trained on imbalanced data. The strongest recall and F1-scores were obtained by Random Forest with NearMiss under low imbalance, demonstrating the benefit of combining balanced sampling with ensemble learning. Notably, however, Random Forest’s recall deteriorated more quickly than XGBoost’s as imbalance increased, likely because bootstrap aggregation introduces variability less suited to skewed data, whereas XGBoost’s gradient-based boosting is more resilient in such settings.

Table 3 reports performance metrics under higher imbalance, and Figure 1 provides a graphical comparison across class ratios and sampling methods. Beyond the general trend of rising accuracy and falling recall, a distinct pattern emerges in precision. Simpler models such as logistic regression, single decision trees, and ensembles built from weak learners like AdaBoost lost precision as imbalance increased, mirroring the decline in recall. In contrast, more complex ensembles—particularly Random Forest, CatBoost, and XGBoost—displayed the opposite tendency. Under random sampling, their precision rose sharply with increasing imbalance, whereas under NearMiss sampling, precision remained relatively stable or increased only slightly even as recall continued to decline.

Table 3. Performance Metrics under Higher Class Imbalance.

Class Ratio	Algorithm	Random Sampling				NearMiss			
		A	F	P	R	A	F	P	R
2000 :598	LR	0.752	0.384	0.449	0.336	0.829	0.656	0.611	0.707
	Tree	0.769	0.506	0.499	0.513	0.915	0.815	0.819	0.811
	RF	0.840	0.533	0.812	0.396	0.954	0.891	0.974	0.821
	XGBoost	0.862	0.653	0.776	0.564	0.952	0.890	0.947	0.839
	GB	0.838	0.578	0.724	0.482	0.953	0.893	0.949	0.843
	CatBoost	0.865	0.653	0.796	0.554	0.952	0.888	0.959	0.826
	AdaBoost	0.769	0.518	0.498	0.540	0.922	0.833	0.825	0.841

Table 3. Cont.

Class Ratio	Algorithm	Random Sampling				NearMiss			
		A	F	P	R	A	F	P	R
5000:598	LR	0.868	0.199	0.282	0.154	0.882	0.455	0.449	0.460
	Tree	0.863	0.352	0.355	0.349	0.935	0.687	0.705	0.671
	RF	0.907	0.232	0.952	0.132	0.962	0.793	0.966	0.672
	XGBoost	0.926	0.519	0.842	0.375	0.962	0.797	0.935	0.694
	GB	0.911	0.388	0.731	0.264	0.959	0.785	0.906	0.692
	CatBoost	0.918	0.477	0.756	0.348	0.961	0.789	0.932	0.684
	AdaBoost	0.856	0.359	0.342	0.378	0.935	0.695	0.693	0.697
10,000:598	LR	0.924	0.122	0.173	0.094	0.923	0.294	0.303	0.286
	Tree	0.920	0.260	0.271	0.251	0.951	0.570	0.566	0.575
	RF	0.947	0.125	0.976	0.067	0.971	0.662	0.965	0.503
	XGBoost	0.956	0.400	0.857	0.261	0.972	0.679	0.943	0.530
	GB	0.947	0.225	0.626	0.137	0.968	0.652	0.852	0.528
	CatBoost	0.953	0.349	0.788	0.224	0.970	0.663	0.931	0.515
	AdaBoost	0.914	0.288	0.270	0.309	0.949	0.560	0.546	0.574
15,000:598	LR	0.949	0.107	0.159	0.080	0.948	0.236	0.273	0.207
	Tree	0.946	0.272	0.281	0.264	0.959	0.452	0.464	0.440
	RF	0.963	0.089	0.983	0.047	0.975	0.527	0.964	0.363
	XGBoost	0.969	0.345	0.920	0.212	0.977	0.574	0.943	0.413
	GB	0.963	0.153	0.550	0.087	0.973	0.519	0.804	0.383
	CatBoost	0.967	0.300	0.782	0.186	0.976	0.566	0.935	0.406
	AdaBoost	0.935	0.236	0.256	0.261	0.955	0.444	0.425	0.465
20,000:598	LR	0.960	0.095	0.138	0.072	0.958	0.175	0.205	0.152
	Tree	0.957	0.243	0.250	0.237	0.964	0.363	0.375	0.353
	RF	0.972	0.055	1.000	0.028	0.979	0.438	0.950	0.284
	XGBoost	0.976	0.310	0.903	0.187	0.980	0.501	0.940	0.341
	GB	0.971	0.129	0.512	0.074	0.977	0.423	0.741	0.296
	CatBoost	0.975	0.292	0.835	0.177	0.980	0.480	0.932	0.323
	AdaBoost	0.952	0.245	0.225	0.269	0.961	0.361	0.342	0.383

Notes: Based on Dataguide Pro and FSS enforcement disclosures, 2011–2024.

This divergent behavior reflects how ensemble models adapt their decision thresholds in the presence of extreme imbalance. As misstated cases become increasingly scarce, these models adopt a more conservative labeling strategy, assigning positive predictions only when the signals of misreporting are strongest. This approach boosts the proportion of correctly classified misstated cases among positive predictions, thereby elevating precision. However, it also causes a substantial loss in recall, as many actual misstated cases are left undetected.

Overall, the results highlight several key insights. First, decision tree-based models consistently outperformed Logistic regression, consistent with prior studies and reflecting that hierarchical and non-linear relationships in financial data are better captured by tree structures than by linear assumptions. Second, recall improved most under near-balanced class ratios, confirming that reducing imbalance is essential for detecting rare misstatement cases that accuracy alone may obscure. This effect was further enhanced by

NearMiss undersampling, which preserves boundary-relevant examples and enables more discriminative learning. While much prior research has focused on identifying superior algorithms, our findings demonstrate that the configuration of training data—particularly class balance and the choice of undersampling method—can be even more decisive for effective misstatement detection.

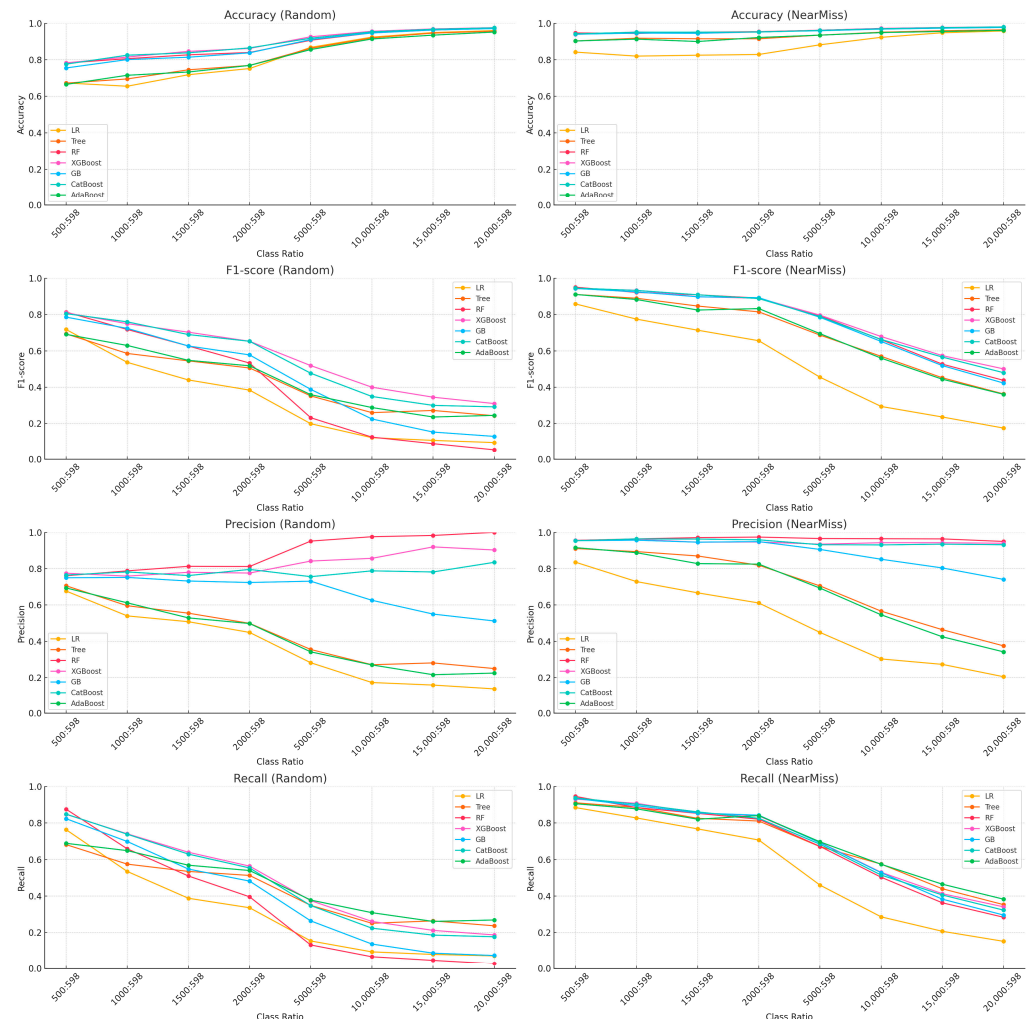


Figure 1. Performance comparison of models by class ratio and sampling method. Notes: Based on Data Guide Pro and FSS enforcement disclosures, 2011–2024.

5.2. Understanding Variable Influence Through XAI: Linear vs. Random Forest

The preceding section focused on evaluating model performance under different class ratios and undersampling strategies, identifying combinations that optimized accuracy, precision, recall, and F1-score. While these metrics are critical, they do not explain how or why a model arrives at a particular prediction. In high-stakes applications like misstatement detection, interpretability is equally important, as it enables practitioners to understand the underlying logic of the model, identify the most influential variables, and gain domain-specific insights. It also facilitates meaningful comparison between linear models (e.g., Logistic regression) and more complex, non-linear models (e.g., Random Forest), revealing not only which variables matter, but also how different modeling approaches attribute importance differently. In this context, XAI plays a key role by offering systematic methods to quantify and visualize how each input variable contributes to its prediction.

5.2.1. Experimental Design for Interpretability Analysis

To explore how interpretability varies across model types, we compare two representative algorithms: Logistic regression and Random Forest. These models reflect the two major methodological paradigms used in this study—linear models and tree-based ensembles. Logistic regression assumes a linear relationship between predictors and the log-odds of the outcome, with each feature exerting a proportional and additive influence. In contrast, decision tree-based models like Random Forest rely on hierarchical splitting rules, where the importance of a feature may vary depending on the values of other features. For instance, a variable that plays a minor role for small firms may become highly significant for large firms due to threshold effects embedded in the tree structure. These contrasting assumptions suggest that the two models may attribute importance to different predictors or interpret the same predictors differently. To ensure a fair and meaningful analysis, we applied XAI techniques only to configurations where overall model performance was satisfactory. In particular, we focused on low-imbalance scenarios (500:598 and 600:598) using the NearMiss sampling strategy, which consistently yielded high performance. Among them, Random Forest achieved the strongest overall results, making it the most appropriate non-linear benchmark. Accordingly, we selected Random Forest to represent tree-based ensemble methods, and Logistic regression to represent linear models, enabling a focused comparison of how interpretability manifests across fundamentally different modeling approaches.

To interpret how individual features influence predictions, we apply two complementary XAI methods: SHAP and PFI. SHAP is grounded in cooperative game theory and calculates the marginal contribution of each feature to a specific prediction. The SHAP value for feature j is defined as:

$$\phi_j(f, x) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f_{S \cup \{j\}}(x) - f_S(x)], \quad (5)$$

where $f_S(x)$ is the expected model output using only features in subset S , and N is the full set of features. SHAP evaluates how much feature j shifts the model's prediction away from the baseline (average prediction), by averaging its marginal contribution across all possible feature combinations. A positive SHAP value means the feature increases the likelihood of misstatement, whereas a negative value lowers it. This makes SHAP particularly useful for understanding localized, instance-level explanations in non-linear models such as Random Forests.

In contrast, PFI provides a global perspective by measuring how model performance changes when the values of a specific feature are randomly permuted. Because detecting misstated firms is more important, we define the importance of feature j as its drop in recall:

$$I_j = R(f, X) - R(f, X_{\pi(j)}), \quad (6)$$

where $R(f, X)$ is the recall on the original data, and $R(f, X_{\pi(j)})$ is the recall after permuting feature j . A higher I_j indicates that the feature plays a more critical role in identifying misstatements.

Together, SHAP and PFI offer a complementary perspective: SHAP provides localized, additive explanations for individual predictions, while recall-based PFI captures global feature relevance in terms of detection sensitivity. We report the top features from both, focusing on consistent predictors and model differences.

5.2.2. Empirical Results and Feature Importance Interpretation

Tables 4–7 summarize the top 25 features with the highest importance in predicting financial misstatements, as evaluated using PFI and SHAP, across different class imbalance

settings and model specifications (Logistic regression and Random Forest). Consistent with the previous experiments, we applied 5-fold cross-validation, and the reported values represent the average SHAP and PFI scores across folds.

Table 4. Feature Importance in the Logistic Regression Model (500:598 case).

Logistic Regression Model				
Rank		PFI		SHAP
1	EXPSOTH	0.214572	COGS	224.9572806
2	RDTOT_NCS	0.191985	SALES	167.3553631
3	EXPSRAW	0.178871	DIVOUT	40.59469412
4	RDTOT	0.176138	EXPSTOT	33.7529652
5	SALES	0.175956	BORROWINGS_INC	24.81593337
6	INTEXPS	0.149727	INTEXPS	24.10728485
7	EXPSEMP	0.14408	EMPEXPS	23.05113003
8	EXPSTOT	0.106557	EXPSLOGI	21.50173439
9	RDEXP_NCS	0.0981785	LTBOND	21.43660984
10	BORROWINGS_INC	0.0928962	TA	21.18560568
11	DIVOUT	0.084153	BORROWINGS_DEC	17.88946911
12	EXPSCOGSOTH	0.0797814	OI	15.73371231
13	COGS	0.0741348	SALESPAYABLE	15.68294781
14	SALESPAYABLE	0.0661202	CURFINDEBT	11.3551659
15	TA	0.0637523	AOCI	11.04812898
16	TAXEXPENSE	0.0628415	EXPSOTH	10.93668593
17	LTBOND	0.0599271	TAXEXPENSE	10.39631399
18	EXPSGOOD	0.0571949	LTFININS	9.931261421
19	CTNDOI	0.0562842	INVENTORY	8.750782913
20	RDEXP	0.0539162	DEFTAXASSET	8.591080109
21	RETEARNING	0.0522769	EXPSRAW	8.571741124
22	OI	0.0486339	SELLEXPS	8.103085643
23	NI	0.045173	EXPSRENT	7.989591611
24	EXPSDEP	0.0440801	CASH	7.521913646
25	BORROWINGS_DEC	0.0437158	OCI	7.228263512

Notes: Based on Dataguide Pro and FSS enforcement disclosures, 2011–2024.

Table 5. Feature Importance in the Random Forest Model (500:598 case).

Random Forest Model				
Rank		PFI		SHAP
1	INTEXPS	0.0229508	INTEXPS	0.199420242
2	COMSTOCK	0.00491803	SALES	0.079550986
3	EXPSGOOD	0.00255009	EXPSTOT	0.055302776
4	RETEARNING	0.00200364	COMSTOCK	0.036331639
5	CLS	0.00182149	CLS	0.027605796
6	SHTLOAN	0.00163934	RDTOT	0.01647187

Table 5. *Cont.*

Random Forest Model				
Rank		PFI		SHAP
7	RECEIVABLES	0.00145719	LTBOND	0.006915916
8	INVENTORY	0.00145719	BOND_INC	0.006796722
9	AMTEXPS	0.00127505	LTBOND_INC	0.006556615
10	EXPSOTH	0.00127505	RDEXP	0.005646649
11	NI	0.00127505	NI	0.005459114
12	RENTEXPS	0.0010929	EXPSOTH	0.005421135
13	SALESPAYABLE	0.0010929	RETEARNING	0.005313095
14	TA_DEC	0.0010929	SHTBOND	0.005244374
15	CASH	0.000910747	SALESPAYABLE	0.004962876
16	PPE	0.000910747	BORROWINGS_DEC	0.004868177
17	DEFTAXASSET	0.000910747	CASH	0.004706878
18	RDEXPS	0.000910747	DIVOUT	0.004585846
19	DISCTNDOI	0.000910747	EMPEXPS	0.004481484
20	TA_INC	0.000910747	TA_INC	0.004139376
21	OCI	0.000728597	TA	0.003939147
22	EXPSBAD	0.000728597	EXPSLOGI	0.003838548
23	EPS_B	0.000728597	OI	0.003732307
24	OI	0.000546448	DEFTAXASSET	0.003614583
25	BORROWINGS_DEC	0.000546448	LTBORROWINGS	0.003327861

Notes: Based on Dataguide Pro and FSS enforcement disclosures, 2011–2024.

Table 6. Feature Importance in the Logistic Regression Model (600:598 case).

Logistic Regression Model				
Rank		PFI		SHAP
1	SALES	0.282304	COGS	61.53445135
2	EXPSOTH	0.234558	SALES	45.59872648
3	RDTOT	0.2202	EXPSLOGI	25.53142992
4	EXPSRAW	0.197329	BORROWINGS_INC	19.68857208
5	RDTOT_NCS	0.174958	DIVOUT	17.69124581
6	EXPSEMP	0.153255	OI	14.14559552
7	INTEXPS	0.143072	EMPEXPS	12.70261301
8	BORROWINGS_INC	0.107179	INTEXPS	12.1180114
9	DIVOUT	0.0943239	EXPSTOT	9.989317447
10	EXPSTOT	0.0864775	LTBOND	8.602903724
11	SALESPAYABLE	0.0809683	LTFININS	7.631990299
12	EXPSCOGSOTH	0.0801336	TA	7.398501825
13	TA	0.0776294	EXPSOTH	6.496375107
14	RECEIVABLES	0.0742905	SALESPAYABLE	6.446622027
15	RDEXP_NCS	0.0729549	TAXEXPENSE	6.040313269

Table 6. Cont.

Logistic Regression Model				
Rank	PFI		SHAP	
16	COGS	0.0716194	INVENTORY	5.892306269
17	RDEXP	0.0687813	TAXPAYABLE	5.211915496
18	TAXEXPENSE	0.0647746	PPE	4.601470208
19	EXPSGOOD	0.0604341	RECEIVABLES	4.512144215
20	LTBOND	0.0590985	RDTOT	3.650243276
21	OI	0.0587646	RETEARNING	3.568175264
22	RETEARNING	0.0552588	EXPSRENT	3.497344791
23	DEFTAXASSET	0.0520868	COMSTOCK	3.370334989
24	EXPSLOGI	0.0504174	RDEXPS	3.344085502
25	SHTLOAN	0.0398998	BORROWINGS_DEC	3.066701029

Notes: Based on Dataguide Pro and FSS enforcement disclosures, 2011–2024.

Table 7. Feature Importance in the Random Forest Model (600:598 case).

Random Forest Model				
Rank	PFI		SHAP	
1	INTEXPS	0.0247078	INTEXPS	0.217342819
2	COMSTOCK	0.00467446	SALES	0.065933972
3	INVENTORY	0.00283806	EXPSTOT	0.042612209
4	RDTOT	0.00217028	COMSTOCK	0.039590058
5	AMTEXPS	0.00217028	CLS	0.024625831
6	EXPSGOOD	0.00217028	RDTOT	0.01095623
7	EXPSOTH	0.00200334	EXPSOTH	0.01047819
8	CLS	0.00166945	EMPEXPS	0.010419105
9	RECEIVABLES	0.0015025	CASH	0.007371457
10	CASH	0.0015025	RETEARNING	0.007330844
11	PPE	0.00133556	TA_INC	0.006274468
12	RETEARNING	0.00133556	SHTLOAN	0.006131024
13	TA	0.00116861	TA_DEC	0.00598646
14	SALESPAYABLE	0.00116861	LTBOND_INC	0.005816972
15	TA_INC	0.00116861	SALESPAYABLE	0.005810318
16	SHTLOAN	0.00100167	TA	0.005521536
17	EXPSTOT	0.00100167	NI	0.00517267
18	RDEXP_NCS	0.000834725	SHTBOND	0.005113623
19	SHTBORROWINGS	0.000834725	BOND_INC	0.004639136
20	COGS	0.000834725	DIVOUT	0.004391363
21	DISCTNDOI	0.000834725	BORROWINGS_DEC	0.004366225
22	EPS_B	0.000834725	INTANGIBLES	0.004243866
23	EXPSBAD	0.000834725	RDEXPS	0.003807693
24	TA_DEC	0.000834725	LTBORROWINGS	0.003581214
25	RDTOT_NCS	0.00066778	RDTOT_NCS	0.00348957

Notes: Based on Dataguide Pro and FSS enforcement disclosures, 2011–2024.

In the Logistic regression models (Tables 4 and 6), both PFI and SHAP consistently highlight several core variables, including sales (SALES), cost of goods sold (COGS), and raw material costs (EXPSRAW)—all fundamental elements of a firm’s core operating activities. Financing-related variables such as borrowings increase/decrease (BORROWINGS_INC, BORROWINGS_DEC), dividend payments (DIVOUT), long-term bonds (LTBOND), and interest expenses (INTEXPS) also rank highly, indicating that external financing, especially debt obligations, is closely linked to financial misstatement risk. Additional overlapping features include trade payables (SALESPAYABLE), tangible assets (TA), operating income (OI), and income tax expense (TAXEXPENSE), reinforcing the role of both operational performance and capital structure in misstatement prediction.

PFI further identifies several R&D-related variables (e.g., RDTOT, RDEXP) as essential, as their permutation significantly degrades model performance. These accounts often involve managerial discretion and are difficult for outsiders to interpret, making them prone to manipulation. Other notable features include net income (NI), retained earnings (RETEARNING), operating income from continuing operations (CTNDOI), and various cost accounts such as personnel expenses (EXPSEMP), other costs of goods sold (EXPSCOGSOTH), and sales of goods (EXPSGOOD). Together, these highlight the cost structure as a key area of potential irregularities.

SHAP, by contrast, emphasizes liquidity and working-capital indicators such as cash (CASH), inventory (INVENTORY), deferred tax assets (DEFTAXASSET), and long-term financial instruments (LTFININS). These reflect a firm’s ability to deploy resources and withstand financial strain. SHAP also points to current financial liabilities (CURFIND-EBT), detailed selling and administrative expenses (e.g., EMPEXPS, SELLEXPS), and non-operating items such as other comprehensive income (OCI) and accumulated OCI (AOCI). These suggest that SHAP captures nuanced, instance-level patterns that may not heavily affect average error but still reveal irregularities.

The dominance of variables such as sales, cost of goods sold, cost of raw materials, and borrowings increase/decrease in the Logistic regression models can be partly attributed to the model’s linear nature. Logistic regression captures features with strong, global, and monotonic relationships to the target variable. Accordingly, line-item accounts that scale with firm size or exhibit straightforward manipulation patterns—such as revenue inflation, cost suppression, or aggressive debt financing—are more likely to be detected. Moreover, the prominence of variables within the operating-cost-income structure (e.g., Sales → COGS → OI) aligns with common earnings manipulation strategies, which frequently involve adjusting top-line or expense accounts in relatively transparent and linear ways.

In the Random Forest models (Tables 5 and 7), interest expenses (INTEXPS), common stock capital (COMSTOCK), and ownership of the largest shareholder (CLS) consistently rank within the top five features in both SHAP and PFI, underscoring their robustness as predictive indicators. SHAP also highlights several financing-related variables, including short-term bonds (SHTBOND), bond issuances (BOND_INC, LTBOND, LTBOND_INC), and long-term borrowings (LTBORROWINGS), showing the model’s sensitivity to firms’ capital structure. These findings align with prior accounting literature that emphasizes the role of leverage and debt covenants in incentivizing earnings management.

Additional important features include profitability indicators (e.g., net income, operating income), operational accounts (e.g., cash, trade payables, deferred tax assets, other expenses), and asset-related variables (e.g., capital expenditures, tangible assets, asset disposals). PFI also identifies inventory (INVENTORY), trade receivables (RECEIVABLES), and property, plant, and equipment (PPE) as key predictors, along with disclosure-related items such as basic earnings-per-share (EPS_B) and operating income from discontinued

operations (DISCTNDI). This suggests that PFI is particularly sensitive to structural variations in financial statements, often prioritizing asset-related categories subject to accounting discretion or judgment. In contrast, SHAP places greater emphasis on sales (SALES), R&D expenditures (RDTOT, RDEXP), and other discretionary or complex accounting items, reflecting its strength in capturing non-linear interactions and context-dependent influences.

Taken together, the two models capture well-established motivations and mechanisms underlying financial misstatements. As highlighted in previous research [10,12], earnings manipulation is frequently driven by the need to meet debt covenants or to compensate for deteriorating cash flow positions. It often involves irregular shifts in cost structures and reliance on opaque accounts such as capitalized R&D expenditures. While both models reflect these dynamics, the importance rankings differ due to their underlying algorithmic structures.

Logistic regression, as a linear model, estimates the marginal effect of each feature under the assumption of *ceteris paribus* variation. Consequently, it tends to assign importance more evenly across a broad set of operating, investing, and financing variables. Random Forest, by contrast, relies on recursive partitioning based on variable thresholds, which heightens sensitivity to structural discontinuities and extreme values. For example, external borrowings may increase the likelihood of earnings management due to covenant pressure, while also signaling stronger creditor oversight. Our results suggest that the Random Forest model captures this duality by assigning relatively higher importance to capital structure-related variables.

The differing rankings between SHAP and PFI further show that these interpretability techniques reveal complementary aspects of model behavior. PFI highlights variables with broad, global effects by measuring performance changes under random permutation, while SHAP captures localized, instance-specific contributions and interaction effects. From a practical perspective, this suggests that detection systems benefit from a wide set of financial variables, particularly those linked to cash flow, financing, and discretionary expenses. It also underscores the importance of model choice in shaping interpretability outcomes—linear models may suit rule-based audit settings, whereas tree-based models offer richer insights into non-linear patterns relevant for forensic or exploratory analysis.

In sum, the interpretability analysis shows that linear and tree-based models, as well as SHAP and PFI, capture complementary dimensions of financial misstatements. These differences highlight the need to consider both model type and interpretability method when designing detection systems. For practitioners, the key implication is that effective systems should combine diverse financial variables with methods that balance transparency and the ability to capture complex reporting behaviors.

Table 8 reports the classification performance of models trained using only the top 25 features identified by SHAP and PFI. Despite relying on a reduced set of top-ranked variables, the models exhibit performance generally comparable to those using the full set of input features. This result highlights the practical value of XAI techniques in identifying a parsimonious yet effective subset of predictors for financial misstatement detection.

Table 8. Model Performance with Top 25 Features under Low Class Imbalance.

Class Ratio	Algorithm	NearMiss			
		A	F	P	R
500:598	Logistic regression	0.849	0.868	0.827	0.913
	Random Forest	0.934	0.947	0.951	0.943

Table 8. *Cont.*

Class Ratio	Algorithm	NearMiss			
		A	F	P	R
600:598	Logistic regression	0.834	0.841	0.804	0.883
	Random Forest	0.945	0.944	0.957	0.931
1000:598	Logistic regression	0.806	0.762	0.704	0.831
	Random Forest	0.947	0.926	0.966	0.890

Notes: Based on Dataguide Pro and FSS enforcement disclosures, 2011–2024.

6. Conclusions

This study provides a comprehensive evaluation of financial misstatement detection using machine learning, focusing on how class imbalance, undersampling strategies, and model interpretability jointly influence detection outcomes. Systematic experiments across multiple class ratios show that the structure of training data plays a decisive role in model performance. NearMiss sampling consistently outperforms random undersampling, particularly in recall and F1-score—critical for identifying rare misstatement cases—demonstrating that strategic data balancing can produce greater improvements than increased model complexity alone.

Beyond predictive accuracy, this study highlights the importance of interpretability in high-stakes domains such as financial reporting. SHAP and PFI analyses reveal that Logistic regression emphasizes globally influential operating and financing variables, whereas Random Forest identifies context-dependent patterns, such as ownership structure and discretionary spending. These complementary insights illustrate how model architecture shapes interpretive outcomes. Moreover, even with a reduced set of key features, models maintain competitive performance under low-imbalance conditions, underscoring the value of explainable AI in guiding feature selection and model simplification.

By aligning balanced sampling with interpretable modeling, this study contributes to the development of transparent, efficient, and accountable detection systems that reinforce investor confidence, support responsible corporate governance, and promote the integrity of capital markets. Such systems not only enhance the effectiveness of audits and regulatory oversight but also contribute to the long-term sustainability of economic systems by reducing information asymmetry and fostering fair competition.

Future research could explore hybrid sampling approaches that combine undersampling and oversampling to strengthen robustness under extreme imbalance. Incorporating unstructured data—such as notes to the financial statements, audit opinions, and market disclosures—may enrich predictive signals. Methodologically, expanding interpretability analysis to advanced architectures (e.g., neural networks [47,48], transformers [49]) and leveraging time-series or sequential models could further extend the scope of intelligent, trustworthy systems for detecting financial irregularities. Future research could also incorporate formal statistical inference techniques to complement the machine learning-based evaluation. Such approaches would enable researchers to test the robustness of performance differences and link the findings more directly to theory.

Author Contributions: Conceptualization, S.K. and W.K.; methodology, W.K.; software, W.K.; validation, S.K.; formal analysis, W.K.; investigation, S.K.; resources, S.K.; data curation, S.K. and W.K.; writing—original draft preparation, W.K.; writing—review and editing, S.K.; visualization, W.K.; supervision, S.K.; funding acquisition, W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by Konkuk University in 2023.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Financial statement data used in this study were obtained from the Dataguide Pro database provided by FnGuide Inc., Seoul, Republic of Korea. Due to licensing restrictions, the data cannot be shared publicly. Researchers may access the data through a paid subscription service.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

This appendix provides the full list of raw accounting variables used as input features in our analysis. The variables are extracted directly from firms' financial statements—including the statement of financial position, income statement, and cash flow statement—without relying on predefined financial ratios.

Table A1. List of Variables and Descriptions.

Variable Name	Description
RDTOT	R&D Expenditures—Consolidated
RDTOT_NCS	R&D Expenditures
RDCAP	R&D Assets—Consolidated
RDCAP_NCS	R&D Assets
RDEXP	R&D Expenses—Consolidated
RDEXP_NCS	R&D Expenses
CLS	Ownership of Largest Shareholders—Common Stocks
PLS	Ownership of Largest Shareholders—Preferred Stocks
INVENTORY	Inventory
SHTLOAN	Short-term Loans Receivable
RECEIVABLES	Trade Receivables
SHTCONTAT	Short-term Contract Assets
SHTCO2AT	Short-term Emission Rights
CASH	Cash and Cash Equivalents
TA	Tangible Assets
PPE	Property, Plant and Equipment
INTANGIBLES	Intangible Assets
GOODWILL	Goodwill
RDASSET	Development costs recorded as intangible assets
IP	Industrial Property Rights
LTFININS	Long-term Financial Instruments
LTAFVPL	FVTPL Financial Assets
LTFVOCI	FVOCI Financial Assets
LTACSEC	Held-to-Maturity Securities
LTACFIN	Amortized Cost Financial Assets
LTLOAN	Long-term Loans Receivable
DEFTAXASSET	Deferred Tax Assets
LTCONTAT	Long-term Contract Assets
LTCO2AT	Long-term Emission Rights
SHTBOND	Short-term Bond Payable
SHTBORROWINGS	Short-term Borrowings
CURLTDEBT	Current portion of long-term borrowings
CURFINDEBT	Current Financial Liabilities
SHTLEASE	Short-term Lease Liabilities
SALESPAYABLE	Trade Payables
DBCURDT	Current Portion of Defined Benefit Liabilities
TAXPAYABLE	Current Tax Payable
SHTCO2DT	Short-term Emission Liabilities
LTBOND	Long-term Bond Payable

Table A1. Cont.

Variable Name	Description
LTBORROWINGS	Long-term Borrowings
LTLEASE	Long-term Lease Liabilities
DBLT	Non-current Defined Benefit Liabilities
LT taxpayable	Long-term Income Tax Payable
LTCO2DT	Long-term Emission Liabilities
COMSTOCK	Common Stock Capital
PRESTOCK	Preferred Stock Capital
AOCI	Accumulated Other Comprehensive Income
RETEARNING	Retained Earnings
SALES	Sales Revenue
COGS	Cost of Sales
EMPEXPS	Personnel Expenses
DEPEXPS	Depreciation Expense
AMTEXPS	Amortization of Intangible Assets
RDEXPS	R&D Expenses
ADVEXPS	Advertising Expenses
SELLEXPS	Selling Expenses
RENTXPS	Rental Expenses
LEASEEXPS	Lease Expenses
OI	Income (Loss) from Operations
INTEXPS	Interest Expense
TAXEXPENSE	Income Tax Expense
CTNDOI	Net Profit (Loss) from Continuing Operations
DISCTNDOI	Net Profit (Loss) from Discontinued Operations
NI	Net Income (Loss)
OCI	Other Comprehensive Income
CEPS_B	Basic EPS from Continuing Operations
EPS_B	Basic Earnings per Share
CEPS_D	Diluted EPS from Continuing Operations
EPS_D	Diluted Earnings per Share
EXPSTOT	Expense Classified by Nature
EXPSMAT	Change in Finished Goods and Work-in-Progress
EXPSSVC	Capitalized Services Rendered
EXPSRAW	Raw Materials and Supplies Used
EXPSGOOD	Sales of Goods
EXPSCOGSOTH	Other Costs
EXPSEMP	Employee Benefits Expense
EXPSDEP	Depreciation, Amortization and Impairment Losses
EXPSTAX	Taxes and Public Charges
EXPSBAD	Bad Debt Expense
EXPSLOGI	Logistic Expenses
EXPSPROMO	Advertising and Sales Promotion Expenses
EXPSRENT	Rental and Lease Expenses
EXPSRND	Ordinary R&D Expenses
EXPSOTH	Other Expenses
OCF	Cash Flows from Operating Activities
TA_DEC	Decrease in Tangible Assets
PPE_DEC	Decrease in Property, Plant and Equipment
TA_INC	Increase in Tangible Assets
PPE_INC	Increase in Property, Plant and Equipment
RDASSET_INC	Increase in Development Costs
CASHTAX	Income Taxes Paid
BOND_INC	Increase in Bonds Payable
LTBOND_INC	Increase in Long-term Bonds
BORROWINGS_INC	Increase in Borrowings
BOND_DEC	Decrease in Bonds Payable
LTBOND_DEC	Decrease in Long-term Bonds
BORROWINGS_DEC	Decrease in Borrowings
DIVOUT	Dividends Paid
AR (Target variable)	Detected Financial Misstatements

References

1. Dyck, A.; Morse, A.; Zingales, L. How pervasive is corporate fraud? *Rev. Account. Stud.* **2024**, *29*, 736–769. [\[CrossRef\]](#)
2. Lokanan, M.; Sharma, S. The use of machine learning algorithms to predict financial statement fraud. *Br. Account. Rev.* **2024**, *56*, 101441, reprinted in *Br. Account. Rev.* **2025**, *57*, 101222. [\[CrossRef\]](#)
3. Beasley, M.S. An empirical analysis of the relation between the board of director composition and financial statement fraud. *Account. Rev.* **1996**, *71*, 443–465.
4. Bell, T.B.; Carcello, J.V. A decision aid for assessing the likelihood of fraudulent financial reporting. *Audit. J. Pract. Theory* **2000**, *19*, 169–184. [\[CrossRef\]](#)
5. Hogan, C.E.; Rezaee, Z.; Riley, R.A.; Velury, U.K. Financial statement fraud: Insights from the academic literature. *Audit. J. Pract. Theory* **2008**, *27*, 231–252. [\[CrossRef\]](#)
6. Trompeter, G.M.; Carpenter, T.D.; Desai, N.; Jones, K.L.; Riley, R.A. A synthesis of fraud-related research. *Audit. J. Pract. Theory* **2013**, *32* (Suppl. S1), 287–321. [\[CrossRef\]](#)
7. Ruhnke, K.; Schmidt, M. Misstatements in financial statements: The relationship between inherent and control risk factors and audit adjustments. *Audit. J. Pract. Theory* **2014**, *33*, 247–269. [\[CrossRef\]](#)
8. Yu, S.J.; Rha, J.S. Research trends in accounting fraud using network analysis. *Sustainability* **2021**, *13*, 5579. [\[CrossRef\]](#)
9. Ramos Montesdeoca, M.; Sanchez Medina, A.J.; Blazquez Santana, F. Research topics in accounting fraud in the 21st century: A state of the art. *Sustainability* **2019**, *11*, 1570. [\[CrossRef\]](#)
10. Healy, P.M. The effect of bonus schemes on accounting decisions. *J. Account. Econ.* **1985**, *7*, 85–107. [\[CrossRef\]](#)
11. Jones, J.J. Earnings management during import relief investigations. *J. Account. Res.* **1991**, *29*, 193–228. [\[CrossRef\]](#)
12. Dechow, P.M.; Sloan, R.G.; Sweeney, A.P. Detecting earnings management. *Account. Rev.* **1995**, *70*, 193–225.
13. Beneish, M.D. The detection of earnings manipulation. *Financ. Anal. J.* **1999**, *55*, 24–36. [\[CrossRef\]](#)
14. Dechow, P.M.; Ge, W.; Larson, C.R.; Sloan, R.G. Predicting material accounting misstatements. *Contemp. Account. Res.* **2011**, *28*, 17–82. [\[CrossRef\]](#)
15. Spathis, C.T. Detecting false financial statements using published data: Some evidence from Greece. *Manag. Audit. J.* **2002**, *17*, 179–191. [\[CrossRef\]](#)
16. Qin, R. Identification of accounting fraud based on support vector machine and logistic regression model. *Complexity* **2021**, *2021*, 5597060. [\[CrossRef\]](#)
17. Ramzan, S.; Lokanan, M. The application of machine learning to study fraud in the accounting literature. *J. Account. Lit.* **2024**, *47*, 570–596. [\[CrossRef\]](#)
18. Cecchini, M.; Aytug, H.; Koehler, G.J.; Pathak, P. Detecting management fraud in public companies. *Manag. Sci.* **2010**, *56*, 1146–1160. [\[CrossRef\]](#)
19. Perols, J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Audit. J. Pract. Theory* **2011**, *30*, 19–50. [\[CrossRef\]](#)
20. Alden, M.E.; Ciconte, W.A.; Steiger, J. Detection of financial statement fraud using evolutionary algorithms. *J. Emerg. Technol. Account.* **2012**, *9*, 71–94. [\[CrossRef\]](#)
21. Chen, S.; Goo, Y.J.J.; Shen, Z.D. A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements. *Sci. World J.* **2014**, *2014*, 968712. [\[CrossRef\]](#)
22. Bai, B.; Yen, J.; Yang, X. False financial statements: Characteristics of China's listed companies and CART detecting approach. *Int. J. Inf. Technol. Decis. Mak.* **2008**, *7*, 339–359. [\[CrossRef\]](#)
23. Bertomeu, J.; Taylor, D.J.; Xue, Y. Using machine learning to detect misstatements. *Rev. Account. Stud.* **2021**, *26*, 468–519. [\[CrossRef\]](#)
24. Liu, C.; Chan, Y.; Kazmi, S.H.A.; Fu, H. Financial fraud detection model: Based on random forest. *Int. J. Econ. Financ.* **2015**, *7*, 178. [\[CrossRef\]](#)
25. Liou, F.M. Fraudulent financial reporting detection and business failure prediction models: A comparison. *Manag. Audit. J.* **2008**, *23*, 650–662. [\[CrossRef\]](#)
26. Beneish, M.D.; Vorst, P. The cost of fraud prediction errors. *Account. Rev.* **2022**, *97*, 91–121. [\[CrossRef\]](#)
27. Van Vlasselaer, V.; Bravo, C.; Caelen, O.; Eliassi-Rad, T.; Akoglu, L.; Snoeck, M.; Baesens, B. APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decis. Support Syst.* **2015**, *75*, 38–48. [\[CrossRef\]](#)
28. Kim, Y.J.; Park, H.J.; Lee, J.H. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Syst. Appl.* **2016**, *62*, 32–43. [\[CrossRef\]](#)
29. Bao, Y.; Ke, B.; Li, B.; Yu, Y. Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *J. Account. Res.* **2020**, *58*, 199–235. [\[CrossRef\]](#)
30. Na, H.J.; Jung, T. An explorative study to detect accounting fraud using a machine learning approach. *Korean Account. Rev.* **2022**, *47*, 177–205. (In Korean) [\[CrossRef\]](#)

31. An, B.; Suh, Y. Identifying financial statement fraud with decision rules obtained from Modified Random Forest. *Data Technol. Appl.* **2020**, *54*, 235–255. [\[CrossRef\]](#)
32. Mani, I.; Zhang, I. kNN approach to unbalanced data distributions: A case study involving information extraction. In Proceedings of the Workshop on Learning from Imbalanced Datasets, Washington, DC, USA, 21 August 2003; ICML: San Diego, CA, USA; Volume 126, pp. 1–7.
33. Mqadi, N.M.; Naicker, N.; Adeliyi, T. Solving misclassification of the credit card imbalance problem using near miss. *Math. Probl. Eng.* **2021**, *2021*, 7194728. [\[CrossRef\]](#)
34. Zhu, H.; Zhou, M.; Liu, G.; Xie, Y.; Liu, S.; Guo, C. NUS: Noisy-sample-removed undersampling scheme for imbalanced classification and application to credit card fraud detection. *IEEE Trans. Comput. Soc. Syst.* **2023**, *11*, 1793–1804. [\[CrossRef\]](#)
35. Lin, K.; Gao, Y. Model interpretability of financial fraud detection by group SHAP. *Expert Syst. Appl.* **2022**, *210*, 118354. [\[CrossRef\]](#)
36. Zhang, C.A.; Cho, S.; Vasarhelyi, M. Explainable artificial intelligence (XAI) in auditing. *Int. J. Account. Inf. Syst.* **2022**, *46*, 100572. [\[CrossRef\]](#)
37. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [\[CrossRef\]](#)
38. Ali, A.A.; Khedr, A.M.; El-Bannany, M.; Kanakkayil, S. A powerful predicting model for financial statement fraud based on optimized XGBoost ensemble learning technique. *Appl. Sci.* **2023**, *13*, 2272. [\[CrossRef\]](#)
39. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Brooklyn, NY, USA, 2018; Volume 31.
40. Nguyen, N.; Duong, T.; Chau, T.; Nguyen, V.H.; Trinh, T.; Tran, D.; Ho, T. A proposed model for card fraud detection based on Catboost and deep neural network. *IEEE Access* **2022**, *10*, 96852–96861. [\[CrossRef\]](#)
41. Chen, Y.; Han, X. CatBoost for fraud detection in financial transactions. In Proceedings of the 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 15–17 January 2021; pp. 176–179.
42. Theodorakopoulos, L.; Theodoropoulou, A.; Tsimakis, A.; Halkiopoulou, C. Big data-driven distributed machine learning for scalable credit card fraud detection using PySpark, XGBoost, and CatBoost. *Electronics* **2025**, *14*, 1754. [\[CrossRef\]](#)
43. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [\[CrossRef\]](#)
44. Dutta, I. Data Mining Techniques to Identify Financial Restatements. Ph.D. Thesis, University of Ottawa, Ottawa, ON, Canada, 2018.
45. Shapley, L.S. Stochastic games. *Proc. Natl. Acad. Sci. USA* **1953**, *39*, 1095–1100. [\[CrossRef\]](#)
46. Goyal, S. Handling class-imbalance with KNN (neighbourhood) under-sampling for software defect prediction. *Artif. Intell. Rev.* **2022**, *55*, 2023–2064. [\[CrossRef\]](#)
47. Du, X. Optimized convolutional neural network for intelligent financial statement anomaly detection. *J. Comput. Technol. Softw.* **2024**, *3*.
48. Nemati, Z.; Mohammadi, A.; Bayat, A.; Mirzaei, A. Predicting fraud in financial statements using supervised methods: An analytical comparison. *Int. J. Nonlinear Anal. Appl.* **2024**, *15*, 259–272.
49. Tang, Y.; Liu, Z. A distributed knowledge distillation framework for financial fraud detection based on transformer. *IEEE Access* **2024**, *12*, 62899–62911. [\[CrossRef\]](#)

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.