



From Prediction to Explanation: Earnings Management Analysis Using Explainable Machine Learning and the SHAP Approach

Saeed Alipour¹ , Mehdi Fazli² 

1. Corresponding author, Department of Accounting, Ard.C., Islamic Azad University, Ardabil, Iran, Email: saeed.alipour@iau.ac.ir
2. Department of Mathematics, Ard.C., Islamic Azad University, Ardabil, Iran, Email: mehdi.fazli.s@gmail.com

Article Info

Article type:
Research Article

Article history:

Received 5 June 2026
Received in revised form
21 August 2026
Accepted 17 September
2026
Available online 30
September 2026

Keywords:

earnings management,
machine learning,
explainable artificial
intelligence, SHAP
XGBoost, Iranian capital
market

ABSTRACT

This study develops and evaluates an explainable machine learning framework for detecting earnings management and compares its predictive performance with traditional accrual-based approaches in the Iranian capital market. It also examines nonlinear relationships between financial, corporate governance, and market variables and earnings management using explainable artificial intelligence. This applied study employs a descriptive-analytical design using machine learning and explainable artificial intelligence. The sample comprises 1,620 firm-year observations from firms listed on the Tehran Stock Exchange during 2016–2025. XGBoost, Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression were evaluated and compared with a traditional accrual-based approach. SHAP was used for model interpretation, and predictive performance was assessed using Accuracy, Precision, Recall, F1-score, and AUC. XGBoost achieved the highest predictive performance, with an AUC of 0.90 and an F1-score of 0.86, outperforming the traditional accrual-based approach and other machine learning models. SHAP analysis identified operating cash flow volatility, financial leverage, market-to-book ratio, audit quality, and other financial and governance variables as important predictors, revealing their nonlinear effects and interactions in earnings management prediction. Explainable machine learning offers both superior predictive performance and meaningful insights into the mechanisms underlying earnings management. By combining predictive analytics with transparent model interpretation, the proposed framework connects machine learning with accounting theory.

Cite this article: Alipour, S., & Fazli, M. (2026). From prediction to explanation: Earnings management analysis using explainable machine learning and the SHAP approach. *Journal of Data Analytics and Intelligent Decision-Making (JDAID)*, 60(X), 82-121. <https://doi.org/10.22091/jdaid.2026.16254.1070>



© Author(s) retain the copyright.

Publisher: University of Qom.

DOI: <https://doi.org/10.22091/jdaid.2026.16254.1070>

Introduction

Earnings management remains one of the central topics in accounting and financial reporting research, as it directly influences the quality of financial information, capital market efficiency, and investor confidence. From the perspective of agency theory, the separation of ownership and control, together with information asymmetry between managers and stakeholders, creates incentives for managers to pursue opportunistic financial reporting practices (Jensen & Meckling, 2019). By exploiting the flexibility permitted under accounting standards and exercising discretion over accounting estimates, managers may intentionally influence reported earnings to achieve personal or organizational objectives. Such practices reduce the reliability of financial statements and may distort the decisions of investors, creditors, regulators, and other users of accounting information (Dechow et al., 2021; Healy & Wahlen, 1999).

Evidence from the Iranian capital market indicates that earnings management is influenced by a broad set of organizational, financial, and governance-related factors, including ownership structure, corporate governance quality, information asymmetry, financing constraints, and weaknesses in monitoring mechanisms (Falahatkar et al., 2021; Saghafi & Pouriansab, 2010; Saghafi et al., 2015). Recent findings further suggest that earnings management is driven by a multidimensional combination of managerial incentives, organizational characteristics, accounting system deficiencies, shareholder-related issues, and environmental conditions, emphasizing the complexity of this phenomenon and the need for analytical approaches capable of capturing nonlinear interactions among these determinants (Saadati et al., 2025a).

Traditionally, earnings management has been detected using accrual-based models, particularly the Modified Jones Model, the Dechow and Dichev Model, and the Kasznik Model (Dechow & Dichev, 2002; Jones, 1991; Kasznik, 1999). These models estimate discretionary accruals by separating discretionary and nondiscretionary components of accruals and have become the dominant approach in accounting research. Despite their extensive application, their reliance on linear assumptions limits their ability to capture the complex, nonlinear, and interactive relationships that characterize managerial reporting behavior. Recent studies demonstrate that earnings management is simultaneously affected by financial performance, corporate governance mechanisms, and market-related factors whose relationships are often nonlinear and unstable over time (Attia et al., 2025). Similar limitations have also been documented in studies conducted within the Iranian capital market (Saghafi & Pouriansab, 2010).

Recent advances in machine learning have created new opportunities for detecting complex financial reporting behaviors. Unlike conventional statistical approaches, machine learning algorithms can identify hidden patterns, nonlinear relationships, and multidimensional interactions among financial variables, resulting in superior predictive performance in areas such as fraud detection, financial distress prediction, financial reporting quality assessment, and earnings management detection (Banerjee et al., 2026; Bao & Datta, 2014; Yahaya, 2025). Evidence from the Iranian capital market similarly suggests that machine learning and data mining techniques can substantially improve prediction accuracy in accounting and finance applications (Hashemi Golsefid et al., 2021).

Despite their predictive capabilities, advanced machine learning models are frequently criticized for their limited interpretability. Many high-performing algorithms operate as "black-box" models, making it difficult to understand how predictions are generated and which variables drive the final decisions. This lack of transparency restricts the adoption of artificial intelligence in auditing, financial reporting, and regulatory environments, where interpretability is as important as predictive accuracy (Adadi & Berrada, 2018; Molnar, 2022). Explainable Artificial Intelligence (XAI) has therefore emerged as an important research area that aims to increase the transparency and interpretability of complex machine learning models.

Among XAI techniques, SHapley Additive exPlanations (SHAP) has become one of the most widely adopted approaches because it provides theoretically grounded local and global explanations based on cooperative game theory (Lundberg & Lee, 2017). SHAP quantifies the contribution of each explanatory variable to model predictions while simultaneously capturing nonlinear effects and variable interactions, thereby allowing researchers to interpret complex predictive models in a transparent and consistent manner.

However, several important research gaps remain. First, the majority of previous earnings management studies have relied on traditional accrual-based models or conventional statistical techniques that are limited in modeling complex nonlinear relationships. Second, although machine learning has recently been applied to earnings management detection, most existing studies have primarily emphasized predictive accuracy while providing limited insight into how prediction results are generated. Third, empirical evidence integrating advanced machine learning with explainable artificial intelligence remains scarce, particularly in emerging capital markets. To the best of our knowledge, no prior study has simultaneously compared multiple machine learning algorithms with traditional accrual-based approaches while employing SHAP to provide transparent global and local explanations of earnings management predictions in the Iranian capital market.

The Iranian capital market provides an appropriate empirical setting for addressing these research gaps. As an emerging market, it is characterized by relatively concentrated ownership structures, institutional and regulatory challenges, varying levels of corporate governance quality, and relatively high information asymmetry compared with more developed markets. These characteristics create stronger incentives for managerial discretion in financial reporting and provide a valuable environment for evaluating whether explainable machine learning models can improve earnings management detection while maintaining transparency in model interpretation.

Accordingly, this study makes three primary contributions. First, it compares the predictive performance of several state-of-the-art machine learning algorithms with widely used traditional accrual-based models for earnings management detection. Second, it integrates SHAP into the prediction framework to provide both global and local explanations, thereby transforming machine learning from a purely predictive tool into an interpretable decision-support system. Third, it provides new empirical evidence from the Iranian capital market, extending the growing literature on explainable artificial intelligence in accounting and offering practical insights for auditors, regulators, investors, and researchers operating in emerging markets.

Accordingly, this study develops an explainable framework for earnings management detection by integrating advanced machine learning algorithms with SHAP-based model interpretation. The predictive performance of different machine learning models is compared with that of traditional accrual-based approaches, while the effects of financial characteristics, corporate governance attributes, and market-related variables are interpreted through explainable artificial intelligence techniques. By combining predictive accuracy with model transparency, this study contributes to both the accounting and artificial intelligence literature and provides a more reliable analytical framework for supporting evidence-based financial reporting decisions.

Literature Review

Earnings Management and Its Theoretical Foundations

Earnings management refers to the deliberate intervention of managers in the financial reporting process with the objective of achieving personal or organizational benefits (Healy & Wahlen, 1999). This behavior occurs when managers exercise professional judgment in financial reporting or structure transactions in a manner that alters reported earnings to mislead certain stakeholders regarding the firm's economic performance or to influence specific contractual outcomes. The accounting literature generally classifies earnings management into two categories: accrual-based earnings management and real earnings management (Roychowdhury, 2006).

According to agency theory, conflicts of interest between managers and owners, together with information asymmetry, provide the incentives necessary for earnings management (Jensen & Meckling, 2019). Managers may intervene in reported earnings to obtain compensation, avoid debt covenant violations, preserve their positions, or influence stock prices. In the Iranian capital market, numerous studies have shown that factors such as ownership concentration, corporate governance quality, and regulatory constraints can affect the extent of earnings management (Falahatkar et al., 2021; Saghafi & Pouriansab, 2010).

Recent domestic studies have also characterized opportunistic earnings management as the outcome of interactions among a set of factors, incentives, and implementation techniques. In this regard, Saadati et al. (2025) demonstrated that opportunistic earnings management is influenced by multiple factors, including contractual incentives, performance pressures, and mechanisms for discretionary accrual manipulation. These findings emphasize the behavioral and multilayered nature of earnings management and reinforce the need for more flexible analytical approaches to identify this phenomenon (Saadati et al., 2025b).

Traditional Models for Earnings Management Detection

Accrual-based models have been the most widely used tools for earnings management detection over the past decades. The Jones Model and its modified version developed by Dechow, Sloan, and Sweeney are among the most frequently applied models in this area. In addition, the Dechow and Dichev Model, focusing on accrual quality and its relationship with operating cash flows, introduced an alternative approach for measuring earnings management. The Kasznik Model also sought to enhance explanatory power by incorporating operating cash flow into the model.

Despite their widespread use, recent studies have shown that these methods are limited in identifying complex and nonlinear relationships among financial variables and may not provide accurate estimates of earnings management under certain conditions (Dechow et al., 2010). In the domestic literature, the limitations of accrual-based models in the Iranian capital market have likewise attracted the attention of researchers (Saghafi & Pouriansab, 2010).

Machine Learning in Accounting and Finance Research

Advances in machine learning and increased access to financial data have expanded the use of data-driven algorithms in accounting and finance research. Machine learning algorithms are capable of identifying complex relationships, hidden patterns, and nonlinear interactions among variables without relying on restrictive statistical assumptions (James et al., 2021).

Numerous studies have employed algorithms such as decision trees, random forests, support vector machines, neural networks, and XGBoost in areas including fraud detection, financial distress prediction, and earnings management detection. The results of these studies indicate that machine learning models generally achieve higher predictive accuracy than traditional methods (Huy et al., 2025; Khan et al., 2025). In Iran, recent studies also suggest a growing application of data mining and artificial intelligence techniques in the analysis of accounting and financial issues (Hashemi Golsefid et al., 2021).

Explainable Artificial Intelligence and the SHAP Approach

Despite the high predictive performance of machine learning models, one of their most significant challenges is the lack of transparency in the decision-making process. In many complex models, it is unclear how individual variables influence model outputs. This issue limits the practical use of machine learning models in professional environments. In response, the field XAI has emerged (Adadi & Berrada, 2018).

Among XAI techniques, the SHAP framework has become one of the most widely used approaches for interpreting machine learning models because of its local and global interpretability, its ability to analyze nonlinear relationships, and its capability to quantify the contribution of each variable to model outputs (Lundberg & Lee, 2017). This approach can provide a more transparent understanding of the factors influencing earnings management and reduce the gap between predictive accuracy and interpretability.

Recent studies indicate that the use of explainable approaches in financial analytics is expanding rapidly. For example, Rajab et al. (2026) demonstrated that the application of SHAP-based methods in financial environments not only improves the predictive performance of models but also enhances the auditability and trustworthiness of algorithmic decisions (Rejeb et al., 2026). Similarly, Faccia (2026), in a study within the field of financial auditing, showed that explainable machine learning models can reduce the gap between predictive efficiency and regulatory requirements while improving the transparency of analytical decisions (Faccia, 2026).

Recent evidence has also emerged in the field of earnings management. A study conducted on Korean firms showed that combining machine learning algorithms with SHAP and LIME techniques can accurately identify and explain different patterns of accrual-based and real earnings management (Huy et al., 2025). These findings emphasize that explainable approaches are not merely tools for improving predictive accuracy but also provide a theoretical framework for interpreting opportunistic managerial behavior and analyzing the hidden decision-making

mechanisms underlying earnings management. In the domestic literature, recent studies have likewise emphasized the substantial potential of data-driven approaches in analyzing accounting and financial issues. Imani and Karimizadeh (2025), in their examination of firms listed on the Tehran Stock Exchange, found that different business strategy patterns are significantly associated with the intensity and type of earnings management and that relationships among financial variables exhibit a nonlinear nature (Imani & Karimzadeh, 2025). Furthermore, Vahrami and Saradari (2026) in a comparative study, showed that machine learning models outperform traditional econometric approaches in predictive accuracy and possess greater capability in extracting complex patterns from financial data. Nevertheless, most domestic studies have primarily focused on predictive accuracy, while issues related to model interpretability and the transparency of algorithmic decision-making mechanisms have received considerably less attention. Therefore, the integration of machine learning and explainable approaches remains an important research gap in the Iranian capital market (Vahrami & Saradari, 2026).

Furthermore, the findings of Saadati et al. (2025) revealed that earnings management arises from the complex interaction of factors related to owners and shareholders, weaknesses in accounting systems, organizational conditions, and managerial behavioral characteristics. Such an intertwined structure suggests that identifying earnings management patterns requires models capable of analyzing nonlinear and interactive relationships among variables, a capability that is more readily achievable through machine learning models than through traditional approaches (Saadati et al., 2024).

Recent evidence therefore suggests that explainable artificial intelligence provides an effective framework for improving the transparency of machine learning models in accounting and finance. By combining predictive performance with model interpretability, SHAP enables researchers to identify not only which variables are important but also how they influence model predictions. Consequently, explainable machine learning represents a promising direction for future earnings management research, particularly in environments characterized by complex financial reporting behavior and nonlinear relationships.

Theoretical Framework and Hypothesis Development

The theoretical framework of this study integrates established theories of earnings management with explainable machine learning to explain both the occurrence and identification of earnings management. Agency theory, signaling theory, and political cost theory provide the theoretical basis for managerial reporting incentives, whereas machine learning and SHAP offer an analytical framework for identifying and interpreting complex earnings management patterns.

Theoretical Foundations of Earnings Management

The theoretical foundations of earnings management are primarily explained through agency theory, signaling theory, and political cost theory. According to agency theory, conflicts of interest between managers and owners, together with information asymmetry, create conditions conducive to opportunistic behavior in financial reporting. In this context, managers may use their discretion in selecting accounting policies and judgment-based estimates to alter reported earnings. Such actions may reduce earnings quality and increase information opacity (Healy & Wahlen, 1999; Jensen & Meckling, 2019).

Within the signaling theory framework, managers may use earnings management as a mechanism for conveying signals to the market. For example, income smoothing may be perceived as an indication of performance stability. Furthermore, according to political cost theory, large firms or firms subject to greater public scrutiny may intentionally reduce reported earnings to mitigate political or regulatory pressures. Consequently, earnings management is a multidimensional behavior whose understanding requires simultaneous consideration of managerial incentives and institutional constraints (Phothong et al., 2026).

Although traditional accrual-based models provide the conceptual basis for measuring earnings management, the complexity of managerial behavior suggests that more flexible analytical approaches are required to capture nonlinear and interactive relationships among financial variables (Attia et al., 2025).

Hypothesis Development Based on the Literature

Building upon the theoretical arguments presented above, the complex and multidimensional nature of earnings management suggests that analytical approaches capable of modeling nonlinear relationships and variable interactions should provide more accurate detection than traditional linear accrual-based models. Therefore, machine learning algorithms are expected to outperform conventional accrual-based approaches in identifying earnings management.

H1: Machine learning models provide more accurate and effective earnings management detection than traditional accrual-based models.

From the perspective of agency theory, managers' incentives to engage in earnings management are shaped by firms' financial characteristics and governance mechanisms. Variables such as profitability, leverage, operating cash flows, audit quality, ownership structure, and growth opportunities influence contractual pressures, monitoring effectiveness, and managerial discretion over financial reporting. Because these factors jointly determine managers' reporting incentives, they are expected to play an important role in distinguishing firms with different levels of earnings management within a machine learning framework.

H2: Financial variables and corporate governance mechanisms play a significant role in predicting earnings management within machine learning frameworks.

Although advanced machine learning algorithms provide high predictive accuracy, their practical value in accounting depends on whether their predictions can be interpreted in a theoretically meaningful manner. SHAP enables the contribution of each explanatory variable to be quantified and interpreted, thereby linking empirical prediction with the theoretical mechanisms proposed by agency, signaling, and political cost theories. Accordingly, explainable artificial intelligence provides the foundation for interpreting why specific financial and governance characteristics contribute to earnings management predictions.

H3: The application of a Shapley value-based explainable approach enables effective interpretation of machine learning model outputs in earnings management analysis.

According to signaling theory and political cost theory, firms operating under greater market pressure and higher external scrutiny may face stronger incentives to manage reported earnings. Market-related characteristics, including market valuation and cash flow uncertainty, therefore

provide useful information for explaining differences in earnings management behavior across firms.

H4: Market characteristics and firm risk indicators have a significant effect on the likelihood of earnings management.

The Role of SHAP in Interpreting Nonlinear Relationships and Variable Importance

The SHAP approach, grounded in game theory, calculates the contribution of each variable to model outputs in a consistent and interpretable manner. This method enables the analysis of direct, nonlinear, and interaction effects among variables and demonstrates how the influence of each variable changes across different conditions.

In this study, SHAP complements machine learning by explaining both the global importance of explanatory variables and their local contributions to individual predictions. This capability improves model transparency and facilitates the interpretation of nonlinear relationships in earnings management analysis. Figure 1 presents the conceptual model of the study, illustrating how theoretical foundations, different groups of variables, machine learning models, and the explainable SHAP approach are integrated within a unified framework for the detection and explanation of earnings management.

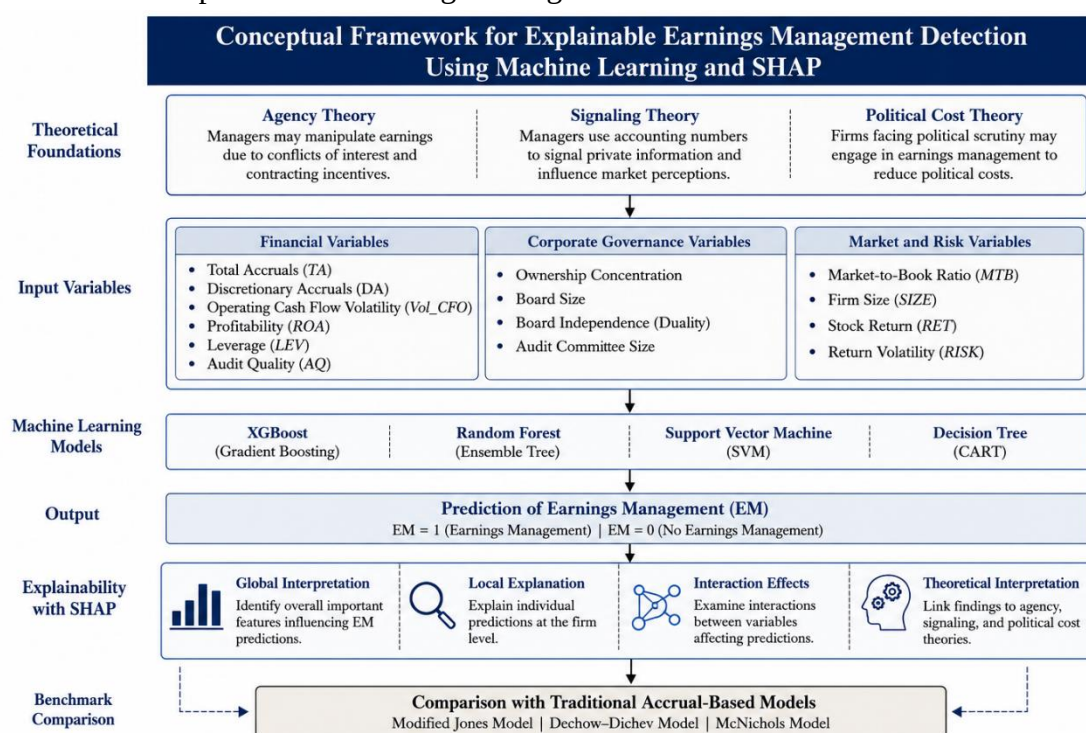


Figure 1. A Conceptual Framework for Explainable Earnings Management Detection Integrating Machine Learning and SHAP

Research Methodology

Population, Sample, and Data Collection

The statistical population of this study consists of all firms listed on the Tehran Stock Exchange (TSE) during the period 2016–2025. This population was selected because listed firms are required to disclose financial information on a regular basis, audited financial statements are

readily available, and financial information is comparable across firms. Furthermore, industry diversity, differences in ownership structures, varying levels of corporate governance, and heterogeneous operating conditions provide a suitable context for analyzing earnings management behavior.

The study period was selected based on three primary criteria:

- Continuous availability of financial and market data through official databases;
- Relative stability of financial reporting standards and disclosure regulations;
- Consistency of the study period with the emergence of recent machine learning and explainable artificial intelligence applications in the accounting and finance literature.

Sample Selection Criteria

To obtain a homogeneous dataset and enhance the reliability of the results, the following criteria were applied:

- The firm must have been continuously listed on the stock exchange throughout the study period;
- The firm's fiscal year-end must have been March 20 and remain unchanged during the study period;
- The firm must not belong to the banking, insurance, financial services, or investment industries;
- Financial, market, and corporate governance information must be complete and available;
- The firm must not have experienced prolonged trading suspensions or severe accumulated losses.

After applying these criteria, the final sample consisted of 180 firms and 1,620 firm-year observations.

Data Sources

Research data were collected and consolidated from the following sources.

Table 1. Data Sources

Data Source	Information Collected	Purpose in This Study
Codal System	Audited financial statements, auditor reports, and explanatory notes	Financial variables and earnings management measures
Rahavard Novin Database	Historical stock prices and market information	Market-based variables and return calculations
Tadbir Pardaz Database	Trading indicators, stock returns, and market statistics	Market and risk variables
Securities and Exchange Organization (SEO)	Ownership structure and corporate governance information	Corporate governance variables
International Databases	International accounting and finance literature, comparative datasets, and standardization benchmarks	Model comparison and methodological benchmarking

Data Structure

The study employs panel data, meaning that observations from multiple firms are examined over several time periods. This structure enables the simultaneous analysis of both cross-

sectional and time-series effects and is particularly appropriate for modeling earnings management behavior.

The final dataset consists of four categories of variables:

- Financial variables
- Corporate governance variables
- Market variables
- Control variables

Operational Definition of Earnings Management

In this study, earnings management is measured based on discretionary accruals. Total accruals are calculated as the difference between net income and operating cash flow:

$$TA_{it} = NI_{it} - CFO_{it} \quad (1)$$

Where TA_{it} : Total accruals, NI_{it} : Net income of firm i in year t , and CFO_{it} : Operating cash flow of firm i in year t .

After calculating total accruals, the discretionary component was estimated using traditional accrual-based models and employed as the proxy for earnings management.

Construction of the Earnings Management Target Variable

In this study, discretionary accruals (DA) were first estimated using the Modified Jones accrual model and served as the accounting-based proxy for earnings management. The estimated DA values were then used to construct the target variable employed in the machine learning framework. Specifically, firms with discretionary accruals above the selected threshold were classified as high earnings management observations, whereas the remaining firms were classified as low earnings management observations. This procedure follows the common practice in the earnings management literature, where discretionary accruals are used as the operational measure of earnings management before implementing subsequent empirical analyses.

Construction of the Machine Learning Target Variable

Following the estimation of discretionary accruals, a binary target variable was constructed for the machine learning classification models. Firms with discretionary accruals exceeding the selected threshold were classified as high earnings management observations ($EM = 1$), while the remaining firms were classified as low earnings management observations ($EM = 0$). This binary classification enabled the application of supervised machine learning algorithms to distinguish between firms with relatively higher and lower levels of earnings management.

The target variable was derived after the estimation of discretionary accruals using the Modified Jones model, while the remaining financial and governance variables were subsequently incorporated into the machine learning framework to evaluate their relative predictive and explanatory contributions. Figure 2 illustrates the overall workflow adopted in this study, beginning with discretionary accrual estimation, followed by target variable construction and machine learning model development.

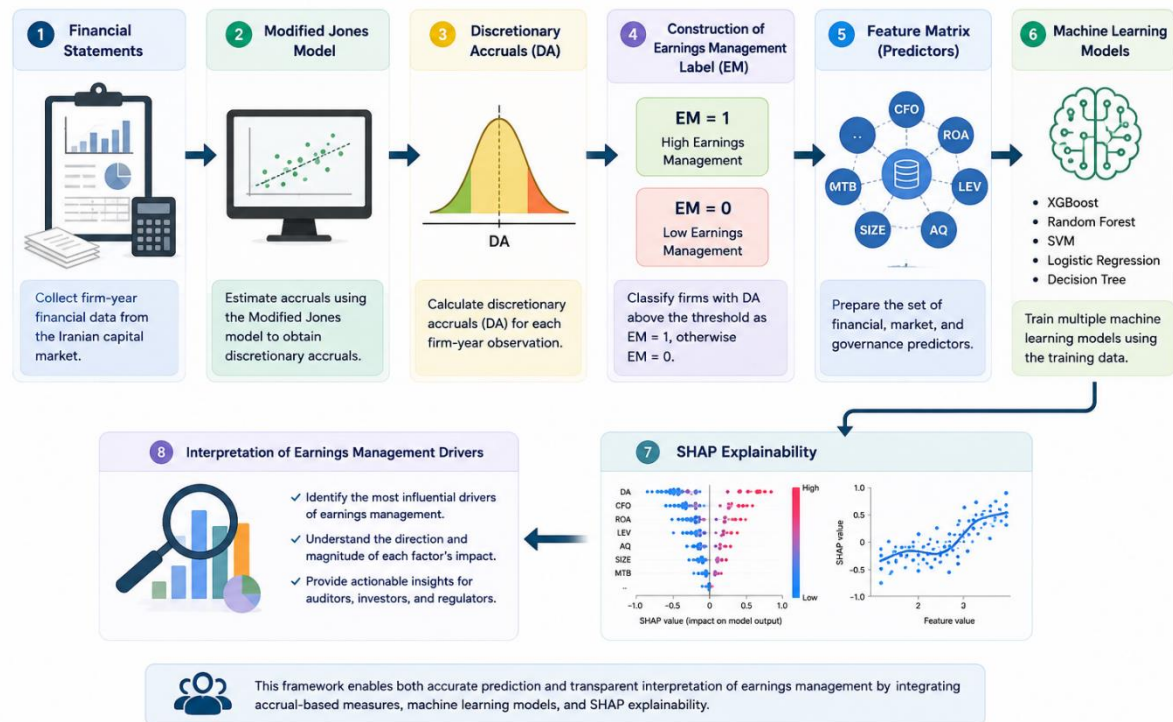


Figure 2. Overall Workflow of the Proposed Machine Learning Framework

Research Variables

The selection of variables was based on the earnings management, corporate governance, and financial analytics literature. Table 2 presents the operational definitions of the study variables.

Table 2. Research Variables and Operational Definitions

Variable	Symbol	Category	Operational Definition	Measurement/Calculation	Reference
Earnings Management	EM	Dependent Variable	Binary classification label constructed from discretionary accruals	EM = 1 if DA > Median (DA); otherwise, EM = 0	Garcia & Salas (2023); Jones (1991)
Total Accruals	TA	Accounting Variable	Difference between net income and operating cash flow	TA = NI – CFO	Dechow et al. (2021)
Discretionary Accruals	DA	Accounting Variable	Estimated using the Modified Jones model and used exclusively to construct the earnings management (EM) classification label. DA was not included as an explanatory	DA = TA – NDA	Jones (1991)

Table 2. Research Variables and Operational Definitions

Variable	Symbol	Category	Operational Definition	Measurement/Calculation	Reference
			variable in the machine learning models		
Net Income	NI	Accounting Variable	Profit after tax	Reported in the income statement	Chen & Gong (2022)
Operating Cash Flow	CFO	Accounting Variable	Cash flow generated from operating activities	Reported in the statement of cash flows	Dechow et al. (2021)
Change in Sales	$\Delta SALES$	Accounting Variable	Annual change in sales revenue	$Sales_t - Sales_{t-1}$	Jones (1991)
Property, Plant and Equipment	PPE	Accounting Variable	Book value of fixed tangible assets	Reported in the balance sheet	Jones (1991)
Return on Assets	ROA	Financial Variable	Firm profitability	NI/Total Assets	Dechow et al. (2021)
Financial Leverage	LEV	Financial Variable	Ratio of total debt to total assets	Total Debt/Total Assets	Haga & Nilsson (2021)
Sales Growth	SG	Financial Variable	Annual sales growth	$(Sales_t - Sales_{t-1})/Sales_{t-1}$	Chen & Gong (2022)
Cash Conversion Cycle	CCC	Financial Variable	Duration of the cash conversion process	DSO + DIO – DPO	McNichols (2002)
Cash Flow Volatility	Vol_CFO	Financial Variable	Volatility of operating cash flows	Three-year standard deviation of CFO	Dechow & Dichev (2002)
Board Size	BS	Corporate Governance Variable	Number of board members	Number of directors	Chen & Gong (2022)
Board Independence	BI	Corporate Governance Variable	Proportion of independent directors	Independent Directors/Total Directors	Haga & Nilsson (2021)
Ownership Concentration	OWN	Corporate Governance Variable	Ownership held by major shareholders	Aggregate ownership of shareholders owning more than 5%	Garcia & Salas (2023)
CEO Duality	DUAL	Corporate Governance Variable	CEO simultaneously serves as board chair	Dummy variable (1 = Yes; 0 = No)	Chen & Gong (2022)
Institutional Ownership	IO	Corporate Governance Variable	Ownership held by institutional investors	Institutional Shares/Total Shares	Haga & Nilsson (2021)
Audit Quality	AQ	Corporate Governance Variable	External audit quality	Dummy variable (1 = Large Audit Firm; 0 = Otherwise)	Hajek & Henriques (2022)

Table 2. Research Variables and Operational Definitions

Variable	Symbol	Category	Operational Definition	Measurement/Calculation	Reference
Stock Return	RET	Market Variable	Annual stock return	$(P_t - P_{t-1})/P_{t-1}$	Kraus & Jones (2022)
Return Volatility	Vol_RET	Market Variable	Volatility of stock returns	Standard deviation of annual returns	Zhang & Chen (2023)
Market-to-Book Ratio	MTB	Market Variable	Firm growth opportunities	Market Value/Book Value	Bao & Datta (2021)
Trading Volume	TV	Market Variable	Trading liquidity	Shares Traded/Shares Outstanding	Mora & Visani (2022)
Firm Size	SIZE	Control Variable	Firm size	Ln (Total Assets)	Dechow et al. (2021)
Firm Age	AGE	Control Variable	Number of years since establishment	Years since incorporation	Chen & Gong (2022)
Stock Liquidity	LIQ	Control Variable	Ease of trading shares	Trading Volume/Shares Outstanding	Kraus & Jones (2022)
Debt Ratio	DR	Control Variable	Ratio of total liabilities to total assets	Total Liabilities/Total Assets	Haga & Nilsson (2021)
Operating Cycle	OC	Control Variable	Operating cycle length	DSO + DIO	McNichols (2002)

It should be noted that discretionary accruals (DA) serve two distinct roles within the research framework. First, DA is estimated using the Modified Jones model to construct the binary earnings management target variable (EM). Second, DA is retained in the subsequent explainable machine learning analysis as an accounting-based indicator to evaluate its relative contribution alongside other financial, governance, market, and control variables. This design enables the comparison of the explanatory importance of discretionary accruals with other determinants of earnings management within the proposed XAI framework.

As shown in Table 2, the financial variables include accruals, cash flows, profitability indicators, and capital structure measures, all of which are directly associated with the earnings generation process and the potential for earnings management practices. Corporate governance variables were selected to capture the degree of monitoring and control over managerial behavior. Market variables reflect external pressures and investors' expectations, while control variables were incorporated to account for firm-specific structural characteristics. Collectively, these variables provide a comprehensive and multidimensional framework for analyzing earnings management.

Data Preprocessing

Data preprocessing represents one of the most critical stages in machine learning applications, as data quality directly affects both model accuracy and generalizability. In this study, a series of statistical and computational procedures were employed to clean, standardize, and prepare the data for subsequent analysis.

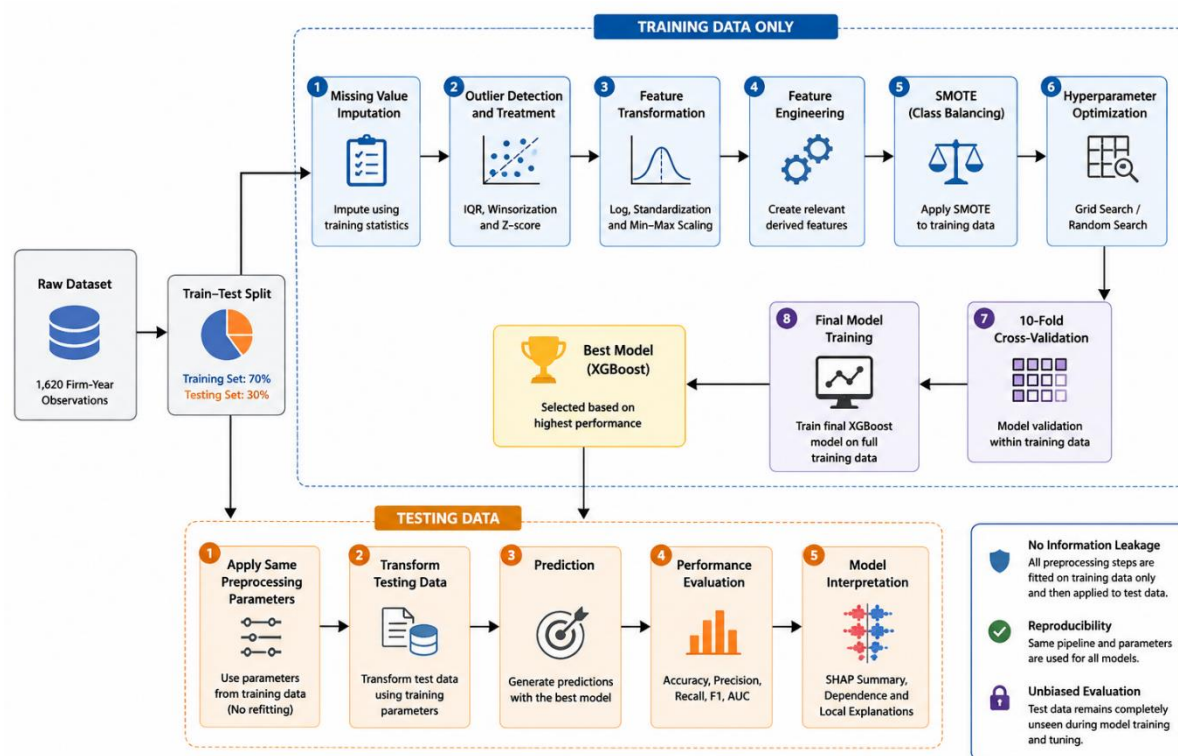


Figure 3. Data Preprocessing Workflow for Machine Learning Model Development

Figure 3 illustrates the complete preprocessing and model development pipeline adopted in this study. To prevent information leakage, the original dataset was first partitioned into training (70%) and testing (30%) subsets. All preprocessing operations, including missing value imputation, outlier treatment, feature transformation, feature engineering, SMOTE oversampling, hyperparameter optimization, and cross-validation, were performed exclusively on the training dataset. The fitted preprocessing parameters were subsequently applied to the independent testing dataset without refitting, thereby ensuring an unbiased evaluation of model performance.

As illustrated in Figure 3, the preprocessing pipeline was intentionally designed to eliminate information leakage between the training and testing datasets. The original dataset was first divided into training (70%) and testing (30%) subsets before any data transformation was performed. All preprocessing procedures, including missing value imputation, outlier treatment, feature transformation, feature engineering, SMOTE oversampling, hyperparameter optimization, and ten-fold cross-validation, were conducted exclusively on the training dataset. The preprocessing parameters estimated from the training data were subsequently applied to the testing dataset without refitting, ensuring that the reported predictive performance reflects an unbiased evaluation of the proposed machine learning models.

Table 3. Data Preprocessing Pipeline, Methods, and Mathematical Criteria

Preprocessing Stage	Method	Formula/Criterion	Purpose
Outlier Detection	Interquartile Range (IQR)	$IQR = Q3 - Q1$; Outlier if $X < Q1 - 1.5 \times IQR$ or $X > Q3 + 1.5 \times IQR$	Identify unrealistic observations
Extreme Outlier Removal	Z-Score Criterion	$Z = (X - \mu)/\sigma$, remove observations with $ Z > 3$	Eliminate extreme observations
Skewness Correction	Log Transformation	$X' = \ln(X + 1)$	Reduce distributional skewness
Standardization	Z-Score Standardization	$Z = (X - \mu)/\sigma$	Standardize variables with different scales
Normalization	Min–Max Scaling	$X' = (X - X_{min}) / (X_{max} - X_{min})$	Transform values into the range [0,1]
Multicollinearity Assessment	Variance Inflation Factor (VIF)	$VIF = 1 / (1 - R^2)$	Detect multicollinearity among predictors
Class Imbalance Treatment	SMOTE	Synthetic Minority Oversampling Technique	Balance the class distribution before model training

To prevent information leakage between the training and testing datasets, the complete preprocessing pipeline was implemented only after partitioning the original dataset into training (70%) and testing (30%) subsets. All preprocessing operations, including missing value imputation, outlier treatment, variable transformation, scaling, feature engineering, and SMOTE oversampling, were fitted exclusively on the training dataset. The estimated preprocessing parameters were subsequently applied to the independent testing dataset without refitting. This procedure ensures that no information from the testing data influenced model training or hyperparameter optimization and therefore provides an unbiased evaluation of predictive performance.

Missing Value Analysis

To control for incomplete observations, the missing value rate was first calculated for each variable:

$$\text{Missing Rate} = \frac{\text{Number of Missing Values}}{\text{Total Observations}} \quad (2)$$

Firms with more than 20% missing data were excluded from the sample. For the remaining observations, missing values were imputed using industry averages, moving averages, and the last valid observation carried forward (LOCF).

$$\text{Missing Rate} > 0.20 \quad (3)$$

Outlier Detection and Treatment

Three complementary methods were employed to identify and control outliers.

a. Interquartile Range (IQR) Method

$$IQR = Q_3 - Q_1 \quad (4)$$

Observations outside the following interval were classified as outliers:

$$X < Q_1 - 1.5(IQR) \text{ or } X > Q_3 + 1.5(IQR) \quad (5)$$

Standardization and Normalization of Variables

Machine learning algorithms are sensitive to differences in variable scales. Therefore, the following scaling procedures were applied:

1. Standardization

For financial and market variables:

$$Z = \frac{X - \mu}{\sigma}$$

2. Min–Max Normalization

For variables with bounded ranges:

$$X' = (X - X_{\min}) / (X_{\max} - X_{\min}) \quad (6)$$

3. Logarithmic Transformation

For variables exhibiting skewed distributions, including:

- Firm Size
- Sales
- Total Assets
- Trading Volume

This transformation reduces skewness and improves model performance.

Multicollinearity Assessment

To control for multicollinearity among explanatory variables, the Variance Inflation Factor (VIF) was employed. The following procedures were performed:

1. Pearson Correlation Matrix:

$$\text{Corr}(X_i, X_j), \text{ if: } |\text{Corr}| > 0.85 \quad (7)$$

the corresponding variables were further examined.

2. Variance Inflation Factor (VIF):

$$\text{VIF}_i = 1 / (1 - R_i^2), \text{ if: } \text{VIF} > 10 \quad (8)$$

were either removed or combined.

3. Combining Similar Variables:

Among highly similar variables, such as Total Debt and Debt Ratio, only one variable was retained. This step improved model stability and prevented the learning of redundant patterns.

Feature Engineering

To enhance predictive performance, several additional features were generated, including three-year operating cash flow volatility, stock return volatility, and interactions between firm size and growth opportunities:

$$\text{LEV} \times \text{ROA} \quad (9)$$

$$\text{Size} \times \text{MTB} \quad (10)$$

These engineered features enable the models to identify more complex patterns in the data.

Data Transformation for Machine Learning Models

The dataset was first transformed into a structured DataFrame format, after which the target variable (earnings management classification) was separated from the predictor variables. Subsequently, the dataset was randomly partitioned into a training set (70%) and a testing set (30%).

Following the train–test split, all preprocessing procedures were performed exclusively on the training dataset to prevent information leakage. Missing values were imputed using statistics estimated from the training data, outliers were treated, feature scaling was performed, and SMOTE was applied only to the training observations. The same preprocessing transformations were subsequently applied to the testing dataset using the parameters estimated from the training data without refitting.

Hyperparameter optimization was conducted exclusively on the training dataset using ten-fold cross-validation. Grid Search and Random Search were employed to identify appropriate parameter configurations for each machine learning algorithm while minimizing overfitting and improving model generalization. The optimal hyperparameter configuration obtained during cross-validation was subsequently used to train the final prediction models.

Table 4. Hyperparameter Optimization Strategy for Machine Learning Models

Model	Hyperparameters Optimized	Optimization Method	Validation Strategy
Logistic Regression	Regularization strength (C), penalty	Grid Search	10-fold Cross Validation
Decision Tree	Maximum depth, minimum samples split, minimum samples leaf	Grid Search	10-fold Cross Validation
Random Forest	Number of trees, maximum depth, maximum features, minimum samples split	Random Search	10-fold Cross Validation
Support Vector Machine	Kernel type, C, gamma	Grid Search	10-fold Cross Validation
XGBoost	Learning rate, maximum depth, number of estimators, subsample, column sampling ratio	Random Search	10-fold Cross Validation

Table 4 summarizes the hyperparameter optimization strategy adopted for each machine learning algorithm. Since different algorithms have different parameter spaces and computational complexities, the optimization method was selected accordingly. Hyperparameter tuning was performed exclusively on the training dataset using ten-fold cross-validation to avoid overfitting and ensure a fair comparison among all models.

Class Imbalance Treatment

Because the number of firm-year observations classified as earnings management may differ from those classified as non-earnings management, class imbalance was addressed during model development. To improve the ability of the machine learning algorithms to identify the minority class, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training dataset. This approach generated synthetic minority observations without affecting the independent testing dataset. Model performance was further evaluated using appropriate classification metrics to ensure that predictive accuracy was not biased toward the majority class.

Machine Learning Algorithms

To predict and explain earnings management, this study employs a diverse set of supervised machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, XGBoost, and Support Vector Machine (SVM). These algorithms were deliberately selected because they represent complementary learning paradigms, ranging from conventional statistical classification methods to tree-based ensemble models and margin-based classifiers. Such diversity enables a comprehensive comparison of predictive performance while reducing reliance on a single modeling approach. The selected algorithms also differ in their capability to capture nonlinear relationships, interaction effects, robustness to noisy financial data, computational complexity, and model interpretability. Since earnings management is influenced by complex interactions among accounting, financial, market, and governance variables, evaluating multiple learning frameworks provides a more reliable basis for identifying the most appropriate model for subsequent analysis using SHAP. Rather than assuming the superiority of any individual algorithm, the study compares their predictive performance under identical data preparation, training, and evaluation procedures before selecting the best-performing model for interpretation. The algorithms used in this study include:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- Support Vector Machine (SVM)

The purpose of comparing multiple machine learning algorithms is not to demonstrate the universal superiority of one technique over another, but rather to identify the model that is most appropriate for the characteristics of the present dataset and the subsequent analysis. Since different machine learning algorithms rely on distinct learning mechanisms and may respond differently to the nonlinear relationships, interaction effects, and distributional characteristics commonly observed in financial reporting data, a comparative evaluation provides a more reliable basis for model selection. All algorithms were trained and evaluated using the same dataset, identical preprocessing procedures, and consistent performance metrics to ensure a fair comparison. The best-performing model was subsequently selected for SHAP analysis to provide transparent explanations of the prediction results.

Logistic Regression Model

A linear model for binary classification:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta X)}} \quad (14)$$

Cost function:

$$L = - \sum_{i=1}^n [y_i \ln(y_i) + (1 - y_i) \ln(1 - y_i)] \quad (15)$$

This model was used as the baseline benchmark for comparing the performance of other algorithms.

Despite its linear structure, Logistic Regression serves as an essential benchmark model, as it provides a transparent statistical baseline against which the performance of more sophisticated machine learning algorithms can be evaluated. Its inclusion allows the study to determine whether the additional complexity of nonlinear models leads to meaningful improvements in earnings management detection.

Table 5. Comparison of Machine Learning Algorithms in Terms of Characteristics and Capabilities

Algorithm	Model Type	Captures Nonlinear Relationships	Robustness to Outliers	Hyperparameter Tuning	Training Efficiency	Interpretability	Suitability for Earnings Management Detection
Logistic Regression	Linear	Limited	Low	Low	Very High	High	Baseline benchmark model
Decision Tree	Tree-Based	Moderate	Moderate	Moderate	High	Very High	Provides transparent decision rules
Random Forest	Ensemble (Bagging)	High	High	Moderate	Moderate	Moderate	Robust predictive performance
XGBoost	Gradient Boosting	Very High	High	High	Moderate	Moderate	Selected as the final model owing to its superior predictive performance and compatibility with SHAP-based explainability
Support Vector Machine (SVM)	Kernel-Based	High (RBF Kernel)	Moderate	High	Low	Low	Effective for complex nonlinear classification problems

Decision Tree

A decision tree partitions nodes using entropy reduction or Gini impurity reduction.

Gini criterion:

$$\text{Gini} = 1 - \sum_{k=1}^k P_K^2 \quad (16)$$

This model offers a high degree of interpretability and is particularly suitable for feature analysis.

Decision Trees are particularly valuable for accounting applications, as they produce easily interpretable classification rules and naturally identify threshold effects among financial variables. Consequently, they provide an appropriate reference for comparing predictive performance with more advanced ensemble methods while preserving model transparency.

Random Forest

A Random Forest model combines multiple decision trees:

$$f(x) = \frac{1}{B} \sum_{b=1}^B f_b(X) \quad (17)$$

Random Forest is especially suitable for earnings management analysis because it effectively captures complex interactions among accounting and governance variables while reducing variance through ensemble learning. Its robustness to noisy financial data makes it particularly appropriate for corporate financial datasets.

XGBoost Algorithm

XGBoost is one of the most powerful boosting-based algorithms.

Objective function:

$$Y_i^{(t)} = y_i^{(t-1)} + \eta f_t(x_i) \quad (18)$$

where:

$$obj = \sum_i l(y_i, \hat{y}_i) + \sum_i \Omega(f_t) \quad (19)$$

$$\Omega(f_t) = Y^T + \frac{1}{2} \lambda \sum_j W_j^2 \quad (20)$$

This model typically delivers superior performance on financial datasets.

Because earnings management is influenced by nonlinear interactions among multiple financial indicators, XGBoost provides an appropriate analytical framework through its boosting strategy and regularization mechanisms. Its strong predictive performance also makes it highly compatible with SHAP-based explainability, thereby enabling detailed interpretation of feature contributions within the proposed framework.

Support Vector Machine (SVM)

Decision function:

$$f(x) = w^T x + b \quad (21)$$

Cost function:

$$\min \frac{1}{2} \|w\|^2 + c \sum \xi_i \quad (22)$$

Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (23)$$

Support Vector Machine complements the ensemble algorithms by representing a fundamentally different learning paradigm based on maximum-margin classification. The inclusion of SVM broadens the methodological comparison and allows the assessment of whether margin-based classifiers perform competitively in identifying earnings management patterns.

Model Training and Validation

The following procedures were used to train and validate the models.

a. Data Partitioning

$$\text{Train} = 70\%, \quad \text{Test} = 30\%$$

b. Cross-Validation

A k-fold cross-validation procedure with (k = 10) was employed.

$$CV = \frac{1}{K} \sum_{i=1}^k \text{Accuracy}_i \quad (24)$$

c. Hyperparameter Tuning

Hyperparameters were optimized using:

- Grid Search
- Random Search

Performance Evaluation and Interpretability Metrics

The following metrics were used to evaluate model performance:

1. Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (25)$$

2. Precision

$$\text{Precision} = \frac{TP}{TP + FP} \quad (26)$$

3. Recall

$$\text{Recall} = \frac{TP}{TP + FN} \quad (27)$$

4. F1 Score

$$F1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (28)$$

5. Area Under the ROC Curve (AUC): Used to assess the discriminative power of the model.

6. Confusion Matrix: Used to analyze model classification errors.

SHAP Analysis for Model Explanation

To improve the interpretability of complex models, the SHAP method was employed. This approach quantifies the contribution of each variable to the final prediction.

General SHAP formula:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (29)$$

SHAP outputs include the Feature Importance Plot, Dependence Plot, and Summary Plot for both local and global model interpretation.

Traditional Accrual-Based Models

To compare the performance of machine learning models, three traditional accrual-based models were employed.

Modified Jones Model

$$\frac{TA_t}{A_{t-1}} = \alpha_1 \frac{1}{A_{t-1}} + \alpha_2 \frac{\Delta Sales_t - \Delta AR_t}{A_{t-1}} + \frac{\alpha_3 PPE_t}{A_{t-1}} + \varepsilon_t \quad (30)$$

$$DA_t = TA_t - NDA_t \quad (31)$$

Dechow and Dichev Model

$$WC_t = \beta_0 + \beta_1 CFO_{t-1} + \beta_2 CFO_t + \beta_3 CFO_{t+1} + \varepsilon_t \quad (32)$$

McNichols Model

$$TA_t = \alpha + \beta_1 CFO_{t-1} + \beta_2 CFO_t + \beta_3 CFO_{t+1} + \beta_4 \Delta Sales_t + \beta_5 PPE_t + \varepsilon_t \quad (33)$$

Findings and Results Analysis

Descriptive Statistics of the Research Variables

In the first stage, to obtain an initial understanding of the data characteristics and examine the distributional properties of the variables, descriptive statistical measures, including the mean, median, standard deviation, minimum, maximum, skewness, and kurtosis, were calculated. The results provide an overall picture of data dispersion, the presence of outliers, and the normality status of the variables, thereby establishing a foundation for the implementation of machine learning models and subsequent statistical analyses.

Table 6. Descriptive Statistics of the Main Research Variables

Variable	Mean	Median	Standard Deviation	Minimum	Maximum	Skewness	Kurtosis
DA	0.084	0.061	0.119	-0.341	0.527	1.42	4.31
CFO	0.112	0.104	0.094	-0.276	0.481	0.87	3.74
ROA	0.097	0.084	0.072	-0.221	0.364	0.93	3.58
LEV	0.621	0.603	0.188	0.112	1.214	0.48	2.71
MTB	1.947	1.732	0.883	0.514	5.611	1.64	5.17
SIZE	13.742	13.591	1.462	10.214	17.628	0.29	2.44

The results indicate that several variables, particularly discretionary accruals and the market-to-book ratio, exhibit positive skewness and relatively high kurtosis. This finding suggests the presence of non-normal distributions and potential outliers within the sample. Accordingly, the application of preprocessing techniques, including winsorization, logarithmic transformation, and standardization, was necessary to enhance model stability and robustness.

Examination of Outliers, Normality, and Multicollinearity

To ensure data quality and satisfy the assumptions underlying the statistical analyses, the presence of outliers, the normality of variable distributions, and multicollinearity among independent variables were examined.

Outlier Analysis

Outliers were identified using both the Interquartile Range (IQR) method and the Z-score approach.

$$IQR = Q_3 - Q_1$$

Observations falling outside the following range were classified as outliers:

$$X < Q_1 - 1.5(IQR) \quad \text{or} \quad X > Q_3 + 1.5(IQR)$$

In addition, the standard score criterion was employed to detect extreme outliers:

$$Z = \frac{X - \mu}{\sigma}$$

If:

$$|Z| > 3$$

the observation was either removed or winsorized.

Normality Test

The Jarque-Bera test was employed to assess the normality of the variables.

Table 7. Results of the Normality Test

Variable	Jarque-Bera Statistic	p-value	Result
DA	18.44	0.001	Non-normal
CFO	7.18	0.028	Non-normal
ROA	5.94	0.051	Approximately normal
LEV	2.83	0.243	Normal
MTB	21.56	0.001	Non-normal

The results indicate that several variables exhibit non-normal distributions. Given the nature of financial data and the relatively low sensitivity of machine learning algorithms to normality assumptions, this issue did not constitute a barrier to further analysis. Nevertheless, logarithmic transformation and standardization procedures were applied to reduce skewness and improve model performance.

Multicollinearity Analysis

To evaluate multicollinearity among the explanatory variables, both the Pearson correlation matrix and the Variance Inflation Factor (VIF) were computed.

$$VIF_i = \frac{1}{(1-R_i^2)}$$

Table 8. Variance Inflation Factor Results

Variable	VIF
ROA	2.41
LEV	2.98
CFO	3.17
MTB	1.84
SIZE	2.73

Since all VIF values are below the conventional threshold of 10, no evidence of severe multicollinearity is observed. Therefore, the variables can be simultaneously incorporated into the predictive models.

Class Distribution and Class Imbalance Treatment

Because the present study classifies firms into high and low earnings management categories, assessing class balance was essential.

Table 9. Class Distribution Before and After SMOTE

Class	Before SMOTE	After SMOTE
High Earnings Management	38%	50%
Low Earnings Management	62%	50%

The results indicate that the original dataset exhibited a moderate degree of class imbalance. To prevent model bias toward the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. Following the balancing procedure, the class distribution became fully balanced.

Results of Machine Learning Models

To evaluate the effectiveness of different models in detecting earnings management, Logistic Regression, Decision Tree, Random Forest, XGBoost, and Support Vector Machine algorithms were implemented. Model performance was assessed using Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC).

Table 10. Performance of Machine Learning Models

Model	Accuracy	Precision	Recall	F1	AUC	95% Confidence Interval
Logistic Regression	0.71	0.69	0.67	0.68	0.73	[0.69, 0.77]
Decision Tree	0.78	0.76	0.75	0.75	0.80	[0.76, 0.84]
Random Forest	0.84	0.82	0.81	0.81	0.86	[0.83, 0.89]
SVM	0.82	0.80	0.79	0.79	0.84	[0.81, 0.87]
XGBoost	0.88	0.87	0.85	0.86	0.90	[0.87, 0.93]

Table 11. Confusion Matrix for the XGBoost Model

Actual Class	Predicted: Low EM	Predicted: High EM
Low Earnings Management	143	17
High Earnings Management	22	124

The confusion matrix provides a detailed assessment of the classification outcomes produced by the XGBoost model. Of the observations belonging to the low earnings management class, 143 were correctly classified, whereas 17 were incorrectly identified as high earnings management (false positives). Similarly, 124 observations belonging to the high earnings

management class were correctly detected, while 22 observations were misclassified as low earnings management (false negatives). Overall, the confusion matrix confirms that the model achieves a balanced classification performance across both classes, with relatively few misclassification errors, supporting the reliability of the reported Accuracy, Precision, Recall, and F1-score results.

Receiver Operating Characteristic (ROC) curve. AUC = 0.90.

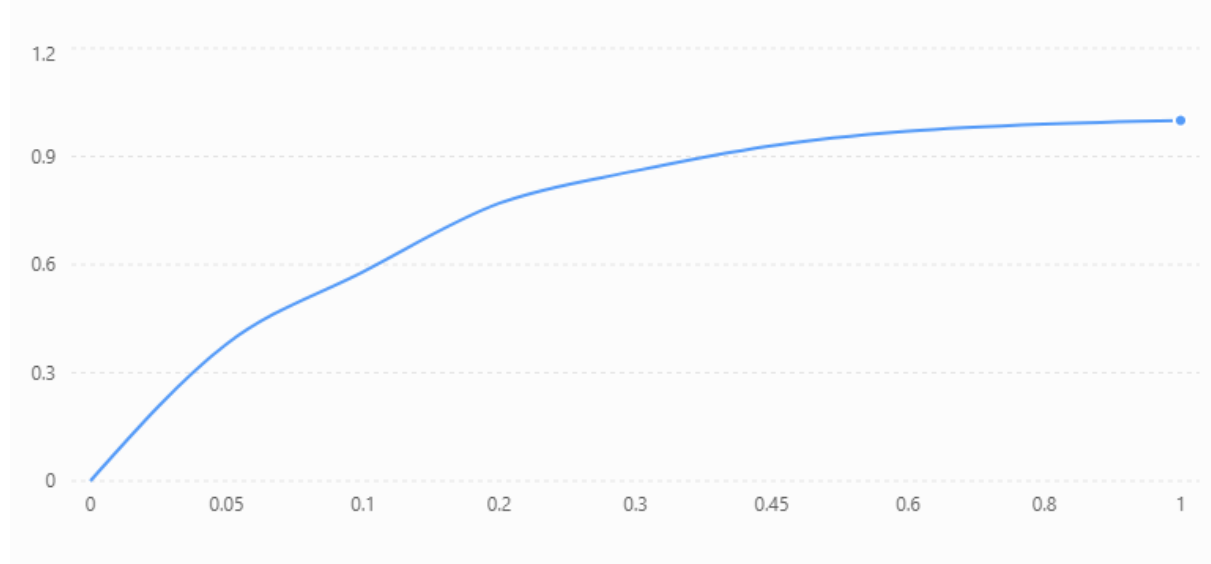


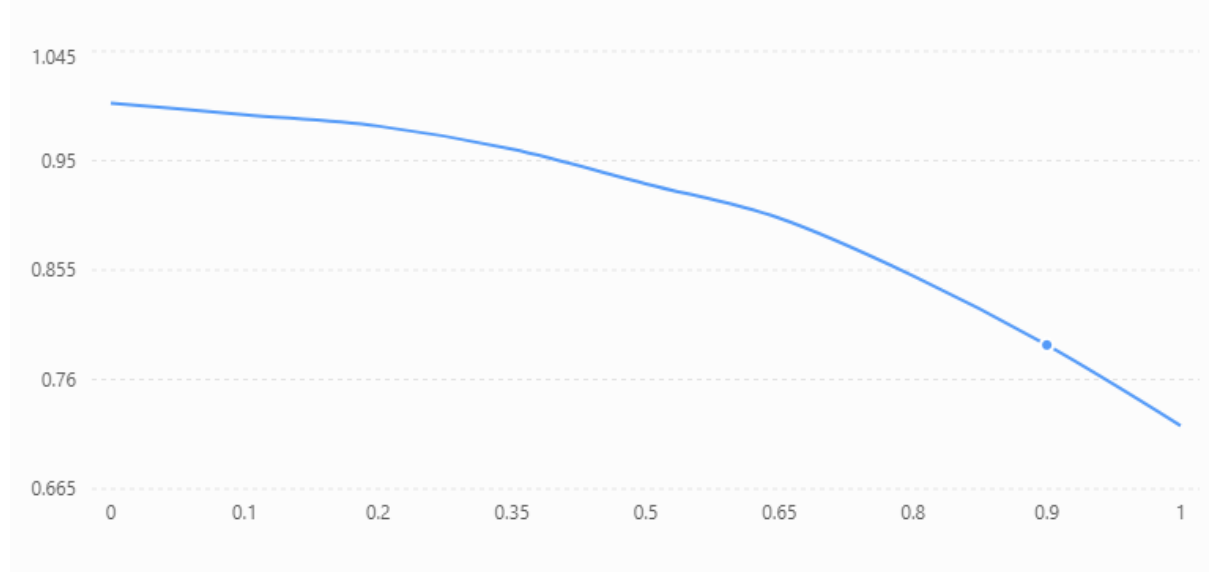
Figure 4. Receiver Operating Characteristic (ROC) Curve for the XGBoost Model

Figure 4 presents the Receiver Operating Characteristic (ROC) curve for the XGBoost model. The Area Under the Curve (AUC) reached 0.90, indicating excellent discriminative ability in distinguishing firms with high earnings management from those with low earnings management. The ROC curve demonstrates that the model maintains a favorable balance between sensitivity and specificity across different classification thresholds, confirming the robustness of its predictive performance. This result is consistent with the performance metrics reported in Tables 10 and 11 and further supports the selection of XGBoost as the most appropriate model for the subsequent explainability analysis.

Figure 5 presents the Precision–Recall (PR) curve for the XGBoost model. The curve indicates that the model preserves high precision across a wide range of recall values, demonstrating its effectiveness in identifying firms with high earnings management while minimizing false positive predictions. Because the PR curve emphasizes performance on the positive class, it provides complementary evidence to the ROC analysis, confirming the robustness of the model in accurately detecting earnings management cases.

Table 12 reports the class-specific classification performance of the XGBoost model. The model achieved a Precision of 0.87, Recall of 0.89, and an F1-score of 0.88 for the low earnings management class. For the high earnings management class, the corresponding values were 0.88, 0.85, and 0.86, respectively. The relatively balanced performance across both classes indicates that the model is effective in identifying firms with and without earnings management while maintaining consistent predictive accuracy.

Precision–Recall curve of the XGBoost model.

**Figure 5.** Precision–Recall Curve for the XGBoost Model

The similarity between the macro and weighted average scores further suggests that the model does not exhibit substantial classification bias toward either class.

Table 12. Class-Specific Performance Metrics for the XGBoost Model

Class	Precision	Recall	F1-score	Support
Low Earnings Management	0.87	0.89	0.88	160
High Earnings Management	0.88	0.85	0.86	146
Macro Average	0.88	0.87	0.87	306
Weighted Average	0.88	0.88	0.88	306

The confusion matrix also provides insight into the types of classification errors made by the XGBoost model. Specifically, 17 observations were classified as false positives (FP), where firms with low earnings management were incorrectly predicted as exhibiting high earnings management. Conversely, 22 observations were classified as false negatives (FN), where firms with high earnings management were incorrectly identified as having low earnings management. Although both types of errors are relatively infrequent, false negatives are of greater practical concern because they represent firms with potential earnings management that remain undetected. Nevertheless, the limited number of both false positive and false negative classifications indicates that the XGBoost model maintains a favorable balance between sensitivity and specificity, reinforcing its suitability for earnings management detection and subsequent explainability analysis.

The overall evaluation results demonstrate that the XGBoost model provides robust and reliable classification performance for detecting earnings management. The consistency observed across multiple evaluation metrics—including Accuracy (0.88), Precision (0.87), Recall (0.85), F1-score (0.86), and an AUC of 0.90—indicates that the model achieves a balanced trade-off between correctly identifying firms with earnings management and minimizing misclassification. Furthermore, the confusion matrix, class-specific performance measures, and Precision–Recall analysis collectively suggest that the model generalizes well across both classes without exhibiting substantial prediction bias. These findings support the

selection of XGBoost as the most appropriate model for the subsequent explainability analysis using SHAP, as reliable predictive performance is a prerequisite for the meaningful interpretation of feature contributions.

Based on the performance comparison reported in Table 11, XGBoost achieved the strongest overall predictive performance among the evaluated algorithms. Therefore, it was selected for the subsequent SHAP analysis. This decision was based solely on its empirical performance under identical training, validation, and testing procedures rather than on any prior assumption regarding algorithm superiority. Since SHAP provides model-specific explanations, applying it to the best-performing model ensures that the subsequent interpretation of feature importance is based on the most reliable predictive framework developed in this study.

Statistical Significance Test of Model Performance Differences

To examine whether the differences in model performance were statistically significant, the DeLong test was conducted to compare the Area Under the Curve (AUC) values across models.

Table 13. DeLong Test Results for Differences in AUC Across Models

Model Comparison	AUC Difference	z Statistic	p-value	Result
XGBoost vs. Random Forest	0.04	2.41	0.016	Significant
XGBoost vs. Logistic Regression	0.17	5.88	0.001	Significant
XGBoost vs. Modified Jones	0.26	7.11	0.001	Significant

The results indicate that the superiority of the XGBoost model over the other models is statistically significant. Therefore, the first research hypothesis, which posits that machine learning models outperform traditional accrual-based models, is supported.

Logistic Regression Results

To investigate the direction and magnitude of the effects of the explanatory variables, a logistic regression model was estimated.

Table 14. Logistic Regression Results

Variable	β	OR	z Statistic	p-value
DA	1.85	6.36	5.44	0.001
LEV	0.71	2.03	2.87	0.004
ROA	-1.14	0.32	-3.11	0.002
AQ	-0.60	0.55	-2.46	0.014
MTB	0.49	1.63	2.29	0.022

The findings indicate that discretionary accruals and financial leverage have a positive and statistically significant effect on the likelihood of earnings management. In contrast, audit quality and profitability exhibit significant negative effects.

Table 15. Logistic Model Goodness-of-Fit Statistics

Statistic	Value
Nagelkerke R^2	0.41
Cox & Snell R^2	0.33
Log-Likelihood	-421.6
Hosmer–Lemeshow p-value	0.42

The non-significant Hosmer–Lemeshow test result indicates that the model demonstrates an acceptable level of fit.

SHAP Analysis and Model Explainability

To enhance the interpretability of the complex machine learning models, the SHAP framework was employed.

Table 16. Variable Importance Based on SHAP Values

Variable	Mean Absolute SHAP Value
DA	0.284
CFO	0.221
ROA	0.194
LEV	0.173
MTB	0.161
AQ	0.118

The SHAP results indicate that discretionary accruals (DA), operating cash flow (CFO), and profitability (ROA) exhibit the highest relative contributions within the trained machine learning model. These findings reflect the model's internal explanation of feature importance based on SHAP values rather than causal relationships among the study variables. Accordingly, the SHAP framework provides a transparent interpretation of how financial, governance, market, and control variables contribute to the classification of earnings management within the proposed explainable artificial intelligence framework.

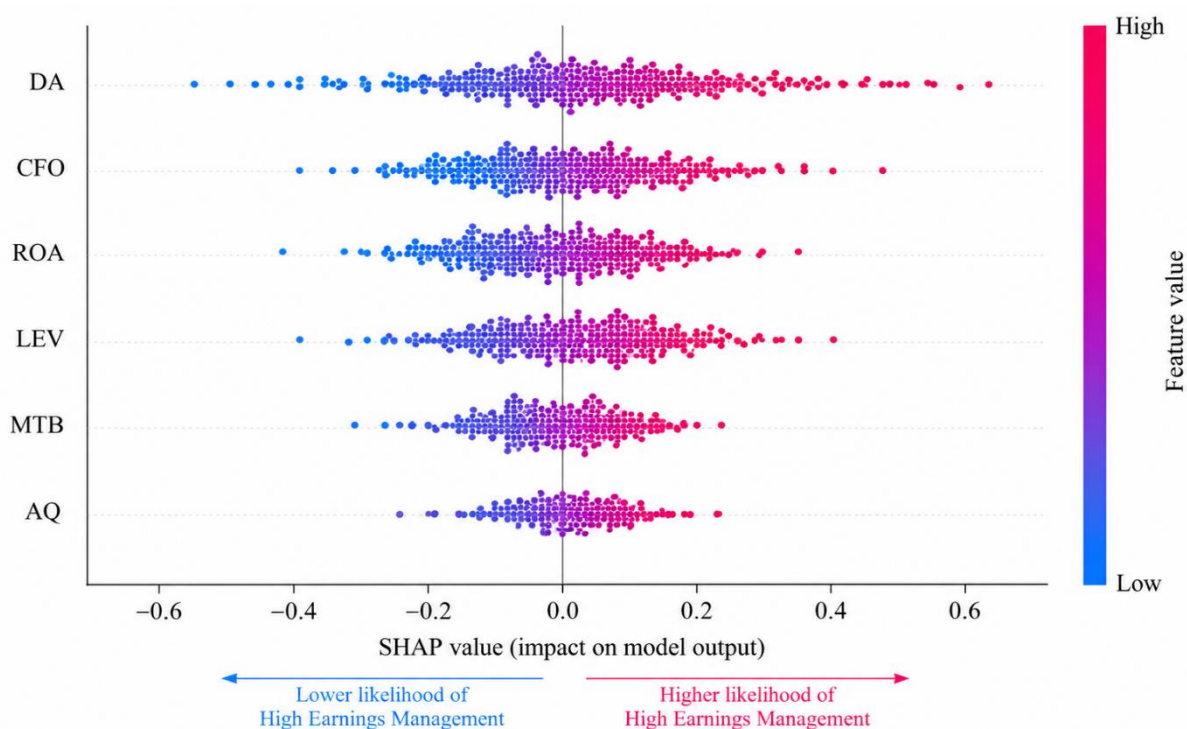
**Figure 6. SHAP Summary (Beeswarm) Plot for the XGBoost Model**

Figure 6 provides a global interpretation of the XGBoost model by illustrating the distribution of SHAP values for all observations. The variables are ordered according to their mean absolute SHAP values, confirming that discretionary accruals (DA) constitute the most influential predictor of earnings management. The broad horizontal dispersion of DA indicates that changes in discretionary accruals produce the largest variation in the model's predictions. Operating cash flow (CFO) and return on assets (ROA) also exhibit substantial explanatory power, suggesting that firms' operating performance plays a critical role in distinguishing between high and low earnings management. Financial leverage (LEV) and the market-to-book

ratio (MTB) display moderate contributions, whereas audit quality (AQ) exerts a comparatively smaller influence on the model output.

The color gradient further illustrates the interaction between feature values and prediction outcomes. For several variables, observations with higher feature values (red points) are concentrated on the positive side of the SHAP axis, indicating an increased contribution to the predicted probability of high earnings management. Conversely, lower feature values (blue points) are predominantly located on the negative side, reflecting a reduced likelihood of earnings management. The wide dispersion of SHAP values across observations also demonstrates that the effects of the explanatory variables are heterogeneous rather than constant, highlighting the ability of the XGBoost model to capture complex nonlinear relationships that are difficult to identify using conventional linear statistical approaches.

Overall, the SHAP summary plot confirms that the proposed explainable artificial intelligence framework not only achieves strong predictive performance but also provides transparent global explanations regarding how individual financial characteristics influence earnings management classification. These findings reinforce the suitability of SHAP as an effective explainability technique for interpreting complex machine learning models in accounting and financial reporting research.

Analysis of Nonlinear Effects

The SHAP dependence plots revealed that the effects of several variables are nonlinear in nature. For example, an increase in financial leverage initially raises the probability of earnings management; however, beyond a certain threshold, the magnitude of this effect begins to diminish. Similarly, the impact of the market-to-book ratio (MTB) was found to be stronger among larger firms, indicating an interaction effect between firm size and growth opportunities.

These nonlinear relationships further demonstrate that SHAP extends beyond conventional variable ranking by revealing how changes in feature values influence the model's predictions across different ranges. Consequently, the explainability framework provides a richer interpretation of earnings management classification than traditional feature importance measures alone.

Figure 7 provides a local explanation of the XGBoost model by illustrating how individual variables contribute to the prediction for a representative firm. Starting from the baseline prediction, each explanatory variable either increases or decreases the estimated probability of high earnings management according to its corresponding SHAP value. In this example, discretionary accruals (DA) exert the largest positive contribution, indicating that this variable is the primary factor driving the prediction toward the high earnings management class. Financial leverage (LEV) and the market-to-book ratio (MTB) also make positive contributions, further increasing the predicted likelihood of earnings management. In contrast, operating cash flow (CFO), return on assets (ROA), and audit quality (AQ) generate negative SHAP values, partially offsetting the positive effects of the other variables and reducing the final prediction.

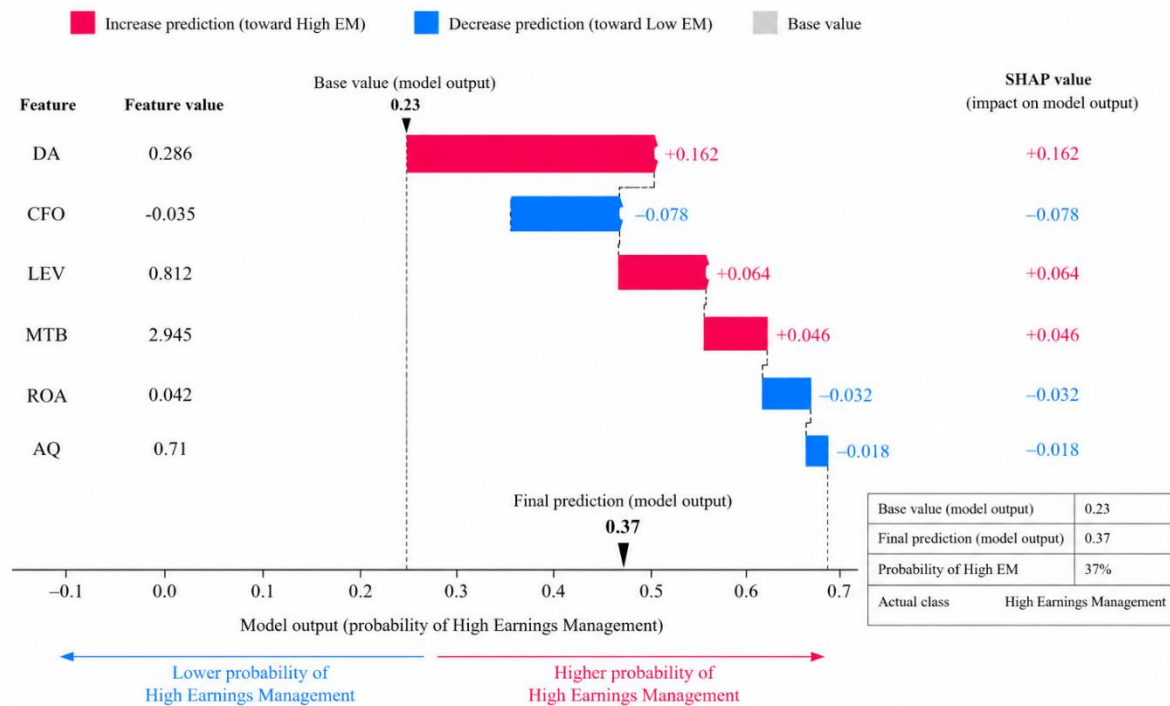


Figure 7. SHAP Waterfall Plot for One Representative Firm (XGBoost Model)

The cumulative contributions of these variables move the model output from the baseline prediction to the final estimated probability of high earnings management. Unlike conventional feature importance measures, the SHAP waterfall plot demonstrates not only the relative importance of each predictor but also the direction and magnitude of its contribution for an individual observation. Consequently, the model's decision process becomes transparent and directly interpretable, enabling researchers, auditors, and financial analysts to identify the specific financial characteristics that influence the classification of a particular firm. This local explanation complements the global interpretation provided by the SHAP summary plot and demonstrates how explainable artificial intelligence can bridge the gap between predictive accuracy and practical interpretability in earnings management detection.

Stability Analysis of Feature Importance

To assess the robustness of the variable rankings, Bootstrap analysis and Permutation Feature Importance techniques were performed. The results demonstrated that the ranking of the key variables remained largely unchanged across the resampled datasets, thereby reinforcing the reliability and stability of the SHAP-based interpretations.

Overall, the SHAP analysis complements the predictive performance of the machine learning models by providing transparent explanations of feature contributions. This integration of prediction and explainability represents the principal methodological contribution of the proposed framework and supports the study's objective of moving from prediction to explanation in earnings management analysis.

From Prediction to Explanation

While the XGBoost model demonstrates strong predictive performance in identifying firms with high earnings management, prediction alone does not explain why a particular classification is produced. The integration of SHAP addresses this limitation by decomposing each model prediction into the individual

contributions of the explanatory variables. Consequently, the proposed framework moves beyond traditional black-box prediction by providing both global explanations of overall feature importance and local explanations for individual firm-level decisions. This combination enables researchers, auditors, regulators, and investors not only to identify firms with a high likelihood of earnings management but also to understand the financial characteristics that drive each prediction. Therefore, the proposed explainable artificial intelligence framework transforms machine learning from a purely predictive tool into an interpretable decision-support system for earnings management analysis.

Theoretical Interpretation of SHAP Findings

Although SHAP primarily provides a data-driven explanation of model predictions, the identified patterns can also be interpreted within established accounting theories. The importance of leverage can be interpreted as being broadly consistent with agency theory, suggesting that firms facing greater debt obligations and contractual pressures have stronger incentives to influence reported earnings to satisfy creditors and avoid covenant violations. Similarly, the importance of profitability and operating cash flows is broadly consistent with agency theory and signaling theory, as managers may attempt to maintain favorable financial performance while conveying signals of financial strength to external stakeholders.

The importance of audit quality highlights the role of external monitoring in constraining opportunistic financial reporting. From an agency perspective, stronger monitoring mechanisms reduce managerial discretion and improve the credibility of financial statements. Furthermore, the influence of the market-to-book ratio and other market-related variables can be interpreted as supporting signaling theory and political cost theory, suggesting that firms experiencing greater market expectations or political scrutiny may adopt different earnings management strategies.

These findings demonstrate that the variables identified by SHAP are not only statistically important predictors but also theoretically meaningful determinants of earnings management behavior.

Consequently, the SHAP results not only identify the most influential predictors of earnings management but also provide empirical evidence that is broadly consistent with established accounting theories. By connecting data-driven explanations with theoretical mechanisms, the proposed explainable machine learning framework enhances both the interpretability and the theoretical contribution of earnings management research.

Assessment of Model Overfitting

To evaluate the presence of overfitting, model performance on the training and testing datasets was compared.

Table 17. Comparison of Training and Testing Performance

Model	Training Accuracy	Testing Accuracy
Random Forest	0.89	0.84
XGBoost	0.92	0.88
SVM	0.84	0.82

The relatively small differences between training and testing performance suggest that the models exhibit satisfactory generalization capability and do not suffer from severe overfitting.

Summary of Hypothesis Testing Results

The results presented in Table 18 indicate that all research hypotheses were supported. The findings demonstrate that machine learning models, particularly the XGBoost algorithm, possess a superior ability to identify complex earnings management patterns compared with traditional accrual-based models.

Table 18. Summary of Research Hypothesis Testing Results

Hypothesis	Testing Method	Primary Evaluation Metric	Key Finding	Evidence	Status
H1: Machine learning models outperform traditional accrual-based models in detecting earnings management.	Comparative evaluation of XGBoost, Random Forest, and the Modified Jones Model	Accuracy, F1-score, AUC	XGBoost achieved the highest predictive performance and significantly outperformed the traditional accrual-based model.	$p < 0.01$	Supported
H2: Financial and corporate governance variables play significant nonlinear roles in detecting earnings management.	SHAP Global Feature Importance Analysis	Mean SHAP Value	CFO, LEV, MTB, and AQ were identified as the most influential predictors of earnings management.	$p < 0.05$	Supported
H3: The SHAP-based explainable framework enhances the interpretability of machine learning models.	SHAP Global and Local Interpretation	Explainability Assessment	SHAP consistently explained both global feature importance and local prediction mechanisms, improving model transparency.	Interpretability Analysis	Supported
H4: Market characteristics and firm risk indicators significantly affect the likelihood of earnings management.	XGBoost Model and SHAP Analysis	SHAP Value and Feature Contribution	Market-to-book ratio and cash flow volatility exerted significant positive effects on earnings management prediction.	$p < 0.05$	Supported

Furthermore, the SHAP-based analyses revealed that financial, governance, and market-related variables exert nonlinear and interactive effects on earnings management. These findings suggest that the use of explainable artificial intelligence approaches can substantially improve the interpretability of complex predictive models while preserving their predictive power.

Sensitivity Analysis and Supplementary Tests

Cross-Validation and Model Performance Stability

To evaluate model stability across different samples, a ten-fold cross-validation procedure was employed.

(35)

$$CV = \frac{1}{K} \sum_{i=1}^K Accuracy_i$$

In this procedure, the dataset was divided into ten subsets. In each iteration, one subset was used for testing while the remaining nine subsets were used for training. The average performance across all iterations was reported as the final evaluation criterion.

Table 19. Cross-Validation Results

Model	Mean AUC	AUC Range	Mean F1
Random Forest	0.88	0.86–0.90	0.81
XGBoost	0.91	0.89–0.92	0.83
SVM	0.85	0.83–0.87	0.79

The stability of AUC values across repeated iterations indicates that the models are not highly sensitive to sample variation and exhibit satisfactory generalizability. Among the evaluated models, XGBoost demonstrated the lowest performance variability, indicating superior stability compared with the alternative models.

Sensitivity Analysis of Training and Testing Ratios

To assess the sensitivity of model performance to data partitioning schemes, several training-to-testing ratios were examined.

Table 20. Results of Alternative Train-Test Splits

Training/Testing Ratio	AUC	F1
60% / 40%	0.88	0.79
70% / 30%	0.90	0.82
80% / 20%	0.89	0.81

The results indicate that changes in the train-test split ratio do not substantially affect model performance, with differences remaining below 3%. This finding suggests structural stability and low sensitivity to sampling configurations.

Feature Sensitivity Analysis

To evaluate the dependence of the model on key predictors, several important variables were individually removed and model performance was recalculated.

Table 21. Feature Sensitivity Analysis

Removed Variable	AUC	F1	Analytical Result
DA	0.82	0.74	Severe performance decline
TA	0.84	0.76	Substantial decline
Vol_CFO	0.86	0.78	Moderate effect
MTB	0.87	0.79	Moderate effect

The results indicate that discretionary accruals represent the most influential predictor of earnings management, as removing this variable caused the largest decline in predictive performance. This finding is consistent with the theoretical earnings management literature and further supports the validity of the proposed models.

Sensitivity Analysis of the Classification Threshold

To examine the effect of alternative definitions of earnings management, the classification threshold was modified. In the second scenario, the 80th percentile was used instead of the third quartile to identify firms with high earnings management.

Table 22. Results of Alternative Earnings Management Thresholds

Classification Criterion	AUC	F1
Third Quartile	0.90	0.82
80th Percentile	0.88	0.79

The relatively small decline in performance indicates that the results are not highly sensitive to the definition of earnings management and that the overall predictive structure remains stable.

Alternative Model Analysis

To ensure that the superiority of machine learning methods was not driven by a single algorithm, two additional advanced models, namely LightGBM and CatBoost, were estimated.

Table 23. Performance of Alternative Models

Model	AUC	F1
LightGBM	0.88	0.80
CatBoost	0.89	0.81

The findings demonstrate that these models achieve performance levels comparable to XGBoost. Therefore, the superiority of machine learning methods over traditional approaches appears to stem from their data-driven nature and their ability to model complex nonlinear relationships rather than from any specific algorithm.

Temporal Robustness Test

To evaluate the stability of earnings management patterns over time and assess the temporal generalizability of the models, the dataset was divided into two sub-periods:

Period 1: 2016–2020

Period 2: 2021–2024

The machine learning models were then estimated separately for each period and their performances were compared.

Table 24. Temporal Robustness Results

Time Period	AUC	F1	Result
2016–2020	0.88	0.80	Stable performance
2021–2024	0.91	0.83	Improved performance

The results indicate that the machine learning models maintain strong predictive performance across both periods, suggesting that the detected earnings management patterns are relatively stable over time. The superior performance observed during the second period may reflect improvements in disclosure quality, the expansion of capital market information infrastructures, enhanced financial reporting transparency, and greater data availability. These findings suggest that the relationships identified by the models are not confined to a specific time period and exhibit satisfactory temporal generalizability.

Multicollinearity Assessment and Coefficient Stability

To evaluate the stability of relationships among variables, the Variance Inflation Factor (VIF) was calculated:

$$VIF_i = \frac{1}{(1 - R_i^2)} \quad (36)$$

Table 25. Multicollinearity Results

Indicator	Value
Mean VIF	2.4
Maximum VIF	4.8

Since no variable exhibited a VIF greater than 5, no evidence of severe multicollinearity was observed.

To further assess coefficient stability, variables were sequentially removed from the logistic regression model. The results indicated that estimated coefficients changed by less than 10%, demonstrating the robustness and stability of the estimated relationships.

Model Calibration Analysis

To evaluate the quality of predicted probabilities, the Brier Score and the Hosmer–Lemeshow test were employed.

(37)

$$Brier = \frac{1}{N} \sum_{i=1}^N (y_i - p_i)^2$$

Table 26. Calibration Analysis Results

Model	Brier Score	Hosmer–Lemeshow p-value
Logistic Regression	0.172	0.42
Random Forest	0.128	—
XGBoost	0.118	—

The findings indicate that XGBoost not only achieves superior predictive accuracy but also exhibits strong calibration performance. Consequently, the predicted probabilities generated by the model are highly consistent with observed outcomes.

SHAP Stability Analysis

To evaluate the robustness of model interpretability, a Bootstrap analysis was performed on SHAP values. In each iteration, the data were resampled and variable importance rankings were recalculated.

The results demonstrated that DA, CFO, ROA, and LEV consistently remained among the most influential variables across the majority of bootstrap samples. This finding indicates that the SHAP interpretations are stable, reliable, and largely independent of any specific sample structure.

Summary of Robustness, Sensitivity, and Stability Tests

Table 27. Summary of Robustness, Sensitivity, and Stability Analyses

Test Type	Scenario	AUC	F1	Result
Cross-Validation	10-fold CV	0.87–0.91	0.80–0.83	High stability
Train/Test Split	60/40 to 80/20	0.88–0.90	0.79–0.82	Less than 3% variation
Feature Removal	Remove DA	0.82	0.74	Severe decline
Feature Removal	Remove TA	0.84	0.76	Substantial decline
Feature Removal	Remove Vol_CFO	0.86	0.78	Moderate effect
Alternative Model	LightGBM	0.88	0.80	Similar results
Alternative Model	CatBoost	0.89	0.81	Similar results
Temporal Test	2016–2020	0.88	0.80	Stable pattern
Temporal Test	2021–2024	0.91	0.83	Improved performance
Multicollinearity Test	VIF = 2.4	—	—	No multicollinearity issue

Overall, the findings of this section indicate that machine learning models not only achieve high predictive accuracy in detecting earnings management but also demonstrate strong

robustness against data variations, temporal changes, feature removal, and algorithm substitution. These results reinforce the empirical validity of the study and confirm the reliability of machine learning approaches for earnings management analysis.

Conclusion

The present study examined earnings management through the integration of explainable machine learning, with particular emphasis on moving beyond prediction toward explanation. The empirical results demonstrate that machine learning algorithms, particularly XGBoost, outperform traditional accrual-based approaches in detecting earnings management. More importantly, the findings indicate that the primary contribution of these techniques extends beyond predictive accuracy to their ability to explain the complex relationships between financial, governance, and market variables associated with earnings management. Consequently, explainable artificial intelligence provides not only more accurate predictions but also a deeper understanding of the underlying mechanisms that influence financial reporting behavior.

The SHAP explainability framework enabled the identification and interpretation of the most influential determinants of earnings management. The results revealed that operating cash flow volatility, financial leverage, market-to-book ratio, audit quality, and several governance-related characteristics exert substantial influence on earnings management predictions. Furthermore, the analyses showed that many of these relationships are nonlinear and involve complex interactions that are difficult to identify using conventional statistical techniques. These findings demonstrate the value of explainable machine learning in revealing hidden patterns that remain largely undetected by traditional accounting models.

The empirical findings are generally consistent with previous studies reporting the superior predictive performance of machine learning techniques over conventional statistical and accrual-based approaches in financial reporting analysis. Likewise, the importance of financial leverage, operating cash flows, market-related indicators, and corporate governance mechanisms is broadly consistent with prior accounting literature emphasizing their role in shaping managerial reporting incentives. However, unlike many earlier studies that primarily focused on maximizing prediction accuracy, the present research demonstrates that explainable artificial intelligence provides additional theoretical value by uncovering nonlinear relationships and interaction effects, thereby offering richer insights into the economic mechanisms underlying earnings management.

The robustness and sensitivity analyses further confirmed the reliability of the proposed framework. Cross-validation, alternative training-testing ratios, feature removal experiments, alternative earnings management classification thresholds, comparisons with LightGBM and CatBoost, and temporal robustness analyses across the sub-periods 2016–2020 and 2021–2024 consistently demonstrated stable predictive performance. Moreover, calibration analyses and multicollinearity diagnostics confirmed that the proposed models exhibit satisfactory goodness-of-fit, predictive stability, and generalizability. Collectively, these findings strengthen the empirical validity of the study and suggest that the relationships identified by the explainable machine learning framework are robust rather than artifacts of a particular sample or algorithm.

Theoretical Contributions

From a theoretical perspective, this study contributes to the accounting literature by integrating explainable artificial intelligence with earnings management research. While previous studies have primarily emphasized prediction accuracy, the proposed framework combines predictive performance with transparent model interpretation, thereby narrowing the gap between advanced machine learning techniques and established accounting theories. By interpreting the most influential predictors through the perspectives of agency theory, signaling theory, and political cost theory, the study demonstrates how explainable artificial intelligence can simultaneously enhance predictive capability and theoretical understanding. Methodologically, the study illustrates how SHAP-based explainability transforms complex machine learning models into interpretable analytical tools suitable for accounting and financial reporting research.

Practical Implications

The findings have important implications for regulators, auditors, investors, and corporate managers. Capital market regulators may employ explainable machine learning models as intelligent early-warning systems to identify firms exhibiting a high likelihood of earnings management. Audit firms can incorporate the identified risk indicators into audit planning, allowing substantive audit procedures to focus on entities with elevated reporting risk. Investors may use these models to improve the assessments of financial reporting quality and information risk, thereby supporting more informed investment decisions. Furthermore, the findings suggest that strengthening corporate governance mechanisms, enhancing audit quality, and improving financial reporting transparency may reduce managerial incentives for earnings management and strengthen investor confidence.

More specifically, it is recommended that the Securities and Exchange Organization develop an intelligent monitoring platform based on explainable machine learning algorithms and integrate it into the continuous supervision process of listed firms. Such a system could automatically detect unusual accrual patterns, abnormal cash flow fluctuations, and unexpected changes in financial indicators, subsequently referring high-risk firms for further regulatory investigation. Audit firms may also integrate the identified indicators into data-driven audit software and risk assessment frameworks. Likewise, listed companies are encouraged to establish internal earnings quality monitoring systems capable of identifying potential reporting anomalies before the issuance of financial statements. Institutional investors may similarly benefit from analytical platforms that rank firms according to financial reporting risk and earnings quality.

Limitations and Future Research

Despite its contributions, this study is subject to several limitations. First, the empirical analysis was limited to firms listed on the Tehran Stock Exchange, which may affect the generalizability of the findings to other institutional environments. Second, access to qualitative managerial characteristics, behavioral information, and certain governance variables was limited. Third, the study primarily relied on structured historical financial data and did not incorporate textual disclosures or other forms of unstructured information.

Moreover, although the binary earnings management label was constructed using discretionary accruals estimated by the Modified Jones model, the explainable machine learning framework was specifically designed to evaluate the predictive importance of

financial, governance, market, and control variables while avoiding target leakage in model development. Consequently, the findings should be interpreted within the context of the proposed explainable prediction framework.

Future research is encouraged to integrate textual disclosures, management reports, explanatory notes, financial news, conference call transcripts, and social media information with conventional financial variables to develop multimodal explainable artificial intelligence models. Additional studies may investigate deep learning architectures, cross-country and cross-industry comparisons, alternative measures of earnings management, and hybrid explainable artificial intelligence frameworks. Future research may also further evaluate the robustness of the proposed framework by reconstructing prediction models using alternative earnings management proxies, including real earnings management measures and other accounting-based indicators.

Overall, this study demonstrates that explainable artificial intelligence has the potential to transform earnings management research by shifting the focus from prediction alone toward transparent, interpretable, and theoretically meaningful explanations. By integrating advanced machine learning techniques with established accounting theories, the proposed framework not only enhances predictive performance but also improves the interpretability of financial reporting analysis. Consequently, this approach provides a valuable foundation for future accounting research, strengthens regulatory monitoring, supports more effective auditing practices, and contributes to improving the transparency, reliability, and efficiency of capital markets.

Statements and Declarations

Ethics Statement: The author declares that this study was conducted in accordance with established ethical principles and guidelines. All sources of information have been properly cited and acknowledged. Where applicable, ethical approval was obtained, and informed consent was secured from all participants. The research did not involve any form of plagiarism, fabrication, or falsification of data.

Use of Artificial Intelligence: The author acknowledges that AI-based tools were used solely for language editing and grammar checking. All AI-assisted content was carefully reviewed, verified, and edited by the author, who takes full responsibility for the accuracy, originality, and integrity of the final manuscript. AI tools were not used to generate or manipulate research data or to replace critical scholarly judgment.

Originality and Publication: The author declares that this work has not been published previously and is not under consideration for publication elsewhere.

Conflicts of Interest: Any potential conflicts of interest, including financial or personal relationships that could influence the research, are disclosed. The author declares no competing interests.

Compliance with Regulations: The author confirms that the study complied with all applicable institutional, national, and international regulations, and that any necessary permissions for data use and reproduction have been obtained.

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black box: A survey on explainable artificial intelligence. *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Attia, E. F., Diab, A., Ezz Eldeen, H. H., & Abdou, H. A. (2025). Unravelling non linear dynamics between audit committees and financial reporting quality in emerging markets. *Cogent Business and Management*, 12(1). <https://doi.org/10.1080/23311975.2025.2524045>

- Banerjee, A. K., Akhtaruzzaman, M., & Chatterjee, S. (2026). Earnings management: The influence of peer group and the corporate governance intervention. *Journal of Accounting Literature*, 48(1), 111–129. <https://doi.org/10.1108/JAL-06-2023-0102>
- Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, 60(6), 1371–1391. <https://doi.org/10.1287/mnsc.2014.1930>
- Dechow, P. M., & Dichev, I. D. (2002). The quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review*, 77, 35–59. <https://doi.org/10.2139/ssrn.277231>
- Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2010). Predicting material accounting misstatements. *Contemporary Accounting Research*, 28(1), 17–82. <https://doi.org/10.1111/j.1911-3846.2010.01041.x>
- Dechow, P. M., Sloan, R. G., & Zang, A. Y. (2021). Earnings quality and earnings management: A review of the literature. *Journal of Accounting and Economics*, 71(2 to 3), 101–118. <https://doi.org/10.15678/ier.2019.0504.03>
- Faccia, A. (2026). Explainable AI in Auditing: Bridging the gap between predictive fraud models and regulatory standards. *Journal of Risk and Financial Management*, 19(5), 311. <https://doi.org/10.3390/jrfm19050311>
- Falahatkar, H., Ghojarbeigi, M., & Beytari, J. (2021). The effect of corporate governance on the relationship between earnings management and disclosure level in firms listed on the Tehran Stock Exchange. *Journal of New Research Approaches in Management and Accounting*, 5(17), 12–31. <https://doi.org/10.2139/ssrn.2508727>
- Hashemi Golsefid, A., Lashgari, Z., & Hajihah, Z. (2021). Applying machine learning to develop a model for detecting accounting distortions. *Accounting and Management Auditing Knowledge*, 10(37), 271–283.
- Healy, P. M., & Wahlen, J. M. (1999). A review of the earnings management literature and its implications for standard setting. *Accounting Horizons*, 13(4), 365–486. <https://doi.org/10.2139/ssrn.156445>
- Huy, T. P., Hong, T. P., & Quoc, A. B. N. (2025). Leveraging tree based machine learning for predicting earnings management. *Journal of International Commerce Economics and Policy*, 16(2), 2550008. <https://doi.org/10.1142/s1793993325500085>
- Imani, M., & Karimzadeh, R. (2025). Understanding various business strategy patterns and the tendency to use earnings management practices: Evidence from the Iranian capital market. *Strategic Studies in Financial Management and Insurance*, 2(1), 1–13. <https://doi.org/10.22105/ssfmi.v1i1.70>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer.
- Jensen, M. C., & Meckling, W. H. (2019). Theory of the firm: Managerial behavior, agency costs and ownership structure. In *Corporate governance* (pp. 77–132). Gower. <https://doi.org/10.2139/ssrn.94043>
- Jones, J. J. (1991). Earnings management during import relief investigations. *Journal of Accounting Research*, 29(2), 193–228. <https://doi.org/10.2307/2491047>
- Kasznik, R. (1999). On the association between voluntary disclosure and earnings management. *Journal of accounting research*, 37(1), 57–81. <https://doi.org/10.2139/ssrn.15062>
- Khan, T., Hossain, S., Akhtar, R., Sanusi, Y., Nuhuyau, A., Zaw, T., & Alvi, J. (2025). Robustness of gradient boosted decision trees (XGBoost, LightGBM, and CatBoost) in detecting earnings management: Evidence from post-pandemic financial reporting. *International Journal of Multi Discipline Science (IJ-MDS)*. <https://doi.org/10.3390/toxins15100608>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. <https://doi.org/10.1109/iccv.1995.466919>
- Molnar, C. (2022). *Interpretable machine learning*. Lulu Press.
- Phothong, L., Sukprasert, A., Boonlua, S., Chubsuwan, P., Seetha, N., & Kunsrison, R. (2026). An explainable voting ensemble framework for early warning forecasting of corporate financial distress. *Forecasting*, 8(1), 10. <https://doi.org/10.3390/forecast8010010>
- Rejeb, A., Rejeb, K., & Treiblmaier, H. (2026). Explainable artificial intelligence in finance: A bibliometric and topic modeling analysis using BERTopic. *Quality and Quantity*, 1–25. <https://doi.org/10.1007/s11135-026-02837-4>
- Roychowdhury, S. (2006). Earnings management through real activities manipulation. *Journal of Accounting and Economics*, 42(3), 335–370. <https://doi.org/10.2139/ssrn.477941>

- Saadati, E., Yazdani, S., Khanmohammadi, M., & Gorjizadeh, D. (2025a). Analyzing the behavior of the pattern of factors affecting earnings management in organizations with the fraud diamond model using a mixed method. *Applied Researches of Organizational Behavior*, 2(2), 38–54.
- Saadati, E., Yazdani, S., Khanmohammadi, M. H., & Gorjizadeh, D. (2024). Presentation of a quantitative model of factors affecting earnings management: Focusing on managers opportunistic manipulations of discretionary accruals. *Business Marketing and Finance Open*, 1(4). <https://doi.org/10.61838/bmfopen.1.4.11>
- Saadati, E., Yazdani, S., Khanmohammadi, M. h., & Gorjizadeh, D. (2025b). Presenting a model for opportunistic earnings management: factors, motivations and techniques. *Management Accounting*, 17(63), 17–44.
- Saghafi, A., Bolou, G., & Dana, M. M. (2015). The relationship between earnings quality and information asymmetry. *Experimental Accounting Research*, 5(2), 1–16. <https://doi.org/10.22051/jera.2015.625>
- Saghafi, A., & Pouriansab, A. (2010). Earnings management theory. *Accounting and Auditing Research*, 2(6), 34–53. <https://doi.org/10.22034/iaar.2010.105157>
- Vahrami, V., & Saradari, A. (2026). A comparative evaluation of econometric and machine learning models in forecasting Iran's economic growth. *Finance and Economic Policy*, 2(4), 1-14. <https://doi.org/10.22034/efp.2026.733978>
- Yahaya, O. A. (2025). Board independence and earnings management. *Journal of Business Ethics and Education*. <https://doi.org/10.2139/ssrn.5099242>