TANMEAH BASSMAH PRESS

International Journal of Digital Accounting and Fintech Sustainability (IJDAFS). Volume 3, Issue 1, 2026

Journal Homepage: https://journals.tanmeah.com/index.php/ijdafs

ISSN: 3008-1505 (print), 3008-1513 (online)

PP. 284-299

# Detecting Earnings Management through Explainable Artificial Intelligence: A Machine Learning Approach Using Financial Statement Data

Safwat Al Tal[1]*

[1] *Faculty of Business, Higher Colleges of Technology (Abu Dhabi/Bani Yas A).*

CHRONICLE

*Article history:*

Received: January 11, 2026

Received in revised format: February, 12, 2026

Accepted: March 27, 2026

Available online: June 30, 2026

***Keywords**:*

Earnings Management Proxy; Discretionary Accruals; Explainable Artificial Intelligence; SHAP; Machine Learning.

**ABSTRACT**

This paper develops an explainable machine learning framework to classify the firm-year observations with unusually high discretionary accruals, a common proxy for the accrual-based earnings management, by using the financial statement data. The sample includes 7,800 firm-year observations of publicly listed non-financial firms from Compustat North America (via WRDS) for firms listed on the NYSE, Nasdaq and NYSE American over the fiscal years 2010-2023. The discretionary accruals were estimated by the Modified Jones Model, and they were used to build a proxy-based binary label according to the top quintile of the absolute discretionary accruals inside each industry-year (1,560 observations, 20.0%, classified as high-accrual). Four algorithms, which are the Logistic Regression, Random Forest, XGBoost and LightGBM, were trained on a leakage-screened set of financial ratios that excludes every component used in the label, and they were compared with the Beneish M-Score. The models were trained on the 2010-2018 data, tuned on 2019-2020, and evaluated out-of-sample on 2021-2023 (1,000 observations). The XGBoost achieves the highest ROC-AUC (0.748) and PR-AUC (0.421), which is higher than the Beneish M-Score (0.612), while the LightGBM achieves the highest F1-score (0.484) and balanced accuracy (0.653) at the 0.50 threshold, and the ROC-AUC advantage of the ML models is statistically significant (DeLong test, $p = 0.004$). Moreover, the SHAP analysis of the optimal model shows that the days sales uncollected, the gross profit margin and the debt-to-equity ratio are the leading variables, and an illustrative walkthrough shows how the SHAP output can be turned into a line-item diagnostic report. The findings indicate that the explainable tree-based models can improve the classification of the high-accrual firm-years compared to the Beneish M-Score, while keeping the interpretability that the auditors and the regulators need.

***JEL Classification:** M41; M42; C38; C45; C55*

* Corresponding author.
*E-mail address: saltal@hct.ac.ae*
https://doi.org/10.70568/ijdafs.3.1.1

## 1. Introduction

The financial reporting is considered the main channel that the management of the company uses to communicate the economic position of the entity to the external users, such as the investors, the creditors, the regulators and the auditors (Healy & Wahlen, 1999). On the other hand, the accrual accounting gives the managers a certain degree of discretion, and this discretion can be used to move the reported figures toward a specific earnings target, especially when the managers face pressure to meet the market expectations or the contractual thresholds (Graham et al., 2005). This behaviour is known in the literature as Earnings Management (EM), and it covers a wide range of practices, starting from the legitimate reporting choices that are allowed under the Generally Accepted Accounting Principles (GAAP) and ending with the opportunistic misstatements and the fraud (Dechow et al., 2010; Dechow et al., 2011). It should be noted that the earnings management, the misstatement of the financial statements and the financial statement fraud are related concepts, but they are not the same. The earnings management refers to the use of the accounting discretion inside the limits of GAAP or near these limits, while the misstatement refers to financial statements that were corrected or restated later, and the fraud refers to a confirmed case of intentional deception, which is usually established by an enforcement action. Accordingly, the current study uses the discretionary accruals as a proxy for the accrual-based earnings management, and the study does not have access to confirmed labels of misstatement or fraud; therefore, the claims in this manuscript are limited to this proxy.

In the past, the researchers and the capital market participants depended mainly on econometric techniques to measure the discretionary accruals and to identify the earnings manipulation, such as the Modified Jones Model (Dechow et al., 1995) and the Beneish M-Score (Beneish, 1999). These classical models gave important insights about the accounting anomalies; however, the classical models have several empirical limitations. For example, the accrual models are affected by the firm performance, and the estimated discretionary accruals may therefore reflect the economic performance of the firm rather than the discretion of the managers (Kothari et al., 2005). Moreover, the linear assumptions of these models cannot capture easily the complex and non-linear relations between the many items of the financial statements (Bao et al., 2020). As a result, the traditional models show a relatively low predictive accuracy and a high rate of false positives when the manipulation schemes are sophisticated and multi-dimensional (Bertomeu et al., 2021).

In order to deal with these limitations, the recent literature started to use the Machine Learning (ML) algorithms, such as Extreme Gradient Boosting (XGBoost), Random Forests and Deep Neural Networks, in the detection of the financial fraud and the accounting anomalies (Cecchini et al., 2010; Hajek & Henriques, 2017; Perols et al., 2017; Craja et al., 2020). These non-linear algorithms improved the classification performance and the prediction accuracy in many settings. However, the use of these algorithms in the auditing and in the regulatory oversight is still limited because of the so-called 'Black-Box' problem (Ribeiro et al., 2016; Rudin, 2019). In the accounting context, the high predictive power alone is not enough, since the regulators, the forensic accountants and the boards of the companies need a transparent and auditable reasoning behind any risk classification, and this reasoning is needed to justify the costly investigations and to comply with the auditing standards (Appelbaum et al., 2017; ICAEW, 2018).

For this reason, the techniques of Explainable Artificial Intelligence (XAI) received a growing attention, and the most used techniques are the SHapley Additive exPlanations (SHAP) and the Local Interpretable Model-agnostic Explanations (LIME), which are applied as post-hoc tools (Lundberg & Lee, 2017; Barredo Arrieta et al., 2020). It is worth mentioning that SHAP and LIME are applied after the model is trained, and therefore they explain what the model has learned, but they do not make the model itself fully transparent. In addition, the attributions of these methods show an association with the output of the model, and they do not prove a causal mechanism of the manipulation. SHAP depends on the game-theoretic attribution, and it allows the researcher to see which financial statement items and ratios are associated with the risk score that the model predicts for an individual firm (Lundberg et al., 2020). According to the recent systematic reviews, this stream of research has grown fast in the accounting and auditing applications, and SHAP and LIME

represent the large majority of the interpretability methods reported in the literature; however, the same reviews consider these post-hoc and correlational methods as a transitional solution and not a final solution for the interpretability problem (Arias-Gonzalez et al., 2026; Cerneviciene & Kabasinskas, 2024). Based on the above, the current study is motivated by three main factors, which are explained below.

First, the corporate financial engineering became more complex, and the manual sampling and the linear diagnostic tools used by the auditors became less efficient accordingly. The auditors now deal with very large datasets and with manipulation schemes that involve many variables at the same time, and these schemes can pass the conventional ratio analysis without being noticed (Appelbaum et al., 2017; Bao et al., 2020). Therefore, there is a practical need for automated and continuous auditing tools that have a high precision.

Second, the financial regulators (for example, the SEC) and the audit committees cannot start an enforcement action only because of a probability score that comes from an opaque algorithm. In this context, the XAI methods can provide an evidence-based interpretation that the auditor can act upon, and this interpretation points to the specific account balances that are related to the anomaly flag, such as the inventory valuation, the timing of the revenue recognition, or the shifts in the accruals (Bertomeu et al., 2021; Faccia, 2026).

Third, the machine learning was tested widely on databases of explicit fraud cases, such as the SEC Accounting and Auditing Enforcement Releases (AAERs); however, the application of the SHAP-based interpretability on the earnings management proxies that are based on the discretionary accruals, as different from the confirmed fraud labels, is still less researched (Dechow et al., 2011). The current study tries to address this gap, and it applies an explainable ML model on the financial statement items in order to classify the firm-years that have unusually high discretionary accruals. It should be noted that the study focuses only on the Accrual-Based Earnings Management (AEM) as measured by the Modified Jones Model, and the study does not construct or test any measure of the Real Earnings Management (REM).

Accordingly, the main objective of the current paper is to design, evaluate and interpret an explainable machine learning framework that classifies the firm-year observations associated with high discretionary accruals, by using the financial statement data. More specifically, the study aims to develop and compare several machine learning models, where the predictive performance of the non-linear algorithms (e.g., XGBoost, Random Forest) and the Logistic Regression is evaluated against the Beneish M-Score in classifying the firm-years associated with high discretionary accruals, and then to interpret the classifications of the optimal model by using SHAP, in order to identify the financial statement ratios that are most associated with these classifications.

## 2. Literature Review and Hypotheses Development

### *2.1 Traditional Models and Limitations in Detecting Earnings Management*

The earnings management (EM) happens when the managers use their judgment in the financial reporting and in the structuring of the transactions in order to change the financial reports, either to mislead some of the stakeholders about the real economic performance of the company or to affect the contractual outcomes that depend on the reported numbers (Healy & Wahlen, 1999). In the accounting literature, the EM is usually divided into two types, which are the Accrual-based Earnings Management (AEM) and the Real Earnings Management (REM) (Dechow et al., 1995; Roychowdhury, 2006). For a long time, the empirical studies depended on the linear statistical models to estimate the discretionary accruals as a proxy for the AEM. The model of Jones (1991) and its later modification, the Modified Jones Model (Dechow et al., 1995), became the most common models for separating the non-discretionary accruals from the discretionary accruals. In a similar way, the Beneish M-Score was introduced as a probit model that uses the financial statement ratios to detect the manipulation of the financial statements (Beneish, 1999). However, the previous studies pointed more and more to the limitations of these classical methods. The linear models are based on strict statistical assumptions, such as the normality and the homoscedasticity, and these assumptions do not reflect the complex and non-linear dependencies that exist between the financial statement items (Bao et al., 2020). In

addition, the traditional models produce high rates of false positives and a low predictive power when they are applied on out-of-sample data or on complex corporate environments (Dechow et al., 2011). Another limitation, which is central for the current study, is that the discretionary accrual proxies such as the Modified Jones Model are estimated with an error. In other words, the residual of the accrual regression includes both the real discretion of the managers and the specification error of the model, and therefore any label that is built from this residual, including the label used in Section 3.2, is a noisy proxy and not a direct observation of the manipulation (Dechow et al., 2010; Dechow et al., 2011). This measurement error has a direct implication on the machine learning applications that use such labels, because a classifier that is trained on a noisy proxy can be evaluated only as a predictor of this proxy, and it cannot be considered a validated detector of the actual earnings management unless confirmed cases (e.g., restatements or AAERs) are also examined (Bao et al., 2020; Perols et al., 2017).

### *2.2 The Paradigm Shift: Machine Learning in Financial Anomaly Detection*

In order to overcome the limitations of the linear econometric models, the recent accounting and FinTech literature moved toward the Machine Learning (ML) algorithms. The machine learning techniques are able to process high-dimensional data, to learn the non-linear interactions between the features, and to adapt with the changing patterns of the financial manipulation (Perols et al., 2017; Hajek & Henriques, 2017). The empirical evidence shows that the non-linear ensemble methods perform better than the traditional statistical techniques in the detection of the corporate fraud and the earnings manipulation. For example, Bao et al. (2020) showed that an ensemble learning model applied on raw financial statement items achieved a better predictive performance than the Beneish M-Score and the logistic regression. In a similar way, Cecchini et al. (2010) and Craja et al. (2020) reported that the support vector machines and the deep learning models improved the detection of the management fraud compared to the traditional benchmarks, and Bertomeu et al. (2021) found that the gradient boosted trees can detect the accounting misstatements with a higher area under the curve (ROC-AUC) when the accounting, the capital market and the audit variables are combined.

A methodological issue that is repeated in this literature, and it is directly related to the current study, is that the time-series and panel structure of the accounting data can inflate the reported out-of-sample performance, if the validation procedure does not protect clearly against the temporal leakage and the firm-level leakage (Kaufman et al., 2012). Perols et al. (2017) and Bao et al. (2020) both emphasize the careful temporal partitioning of the data, and they note that the random cross-validation, when it is applied directly on panel accounting data, can allow the information from the future of the firm or from correlated firm-years to enter the training data, and this situation overstates the predictive performance compared to a real prospective test.

Although the advanced machine learning algorithms have a high predictive accuracy, these algorithms, and especially the Deep Neural Networks (DNNs) and the complex tree ensembles, have an important weakness, which is the 'Black-Box' problem (Ribeiro et al., 2016). In the accounting and auditing context, the high accuracy is not sufficient alone, because the regulators, the forensic auditors and the board committees need a transparent reasoning behind the risk score of the model, in order to justify a formal investigation and to comply with the auditing standards (Rudin, 2019; ICAEW, 2018).

### *2.3 Integrating Explainable Artificial Intelligence (XAI) in Accounting*

The Explainable Artificial Intelligence (XAI) appeared as a way to connect the predictive power of the complex ML models with the need for interpretability in the corporate governance. The XAI frameworks offer model-agnostic tools, and these tools convert the outputs of the opaque algorithms into explanations that a human user can understand (Lundberg & Lee, 2017; Barredo Arrieta et al., 2020). In the fintech domain, Bussmann et al. (2020) applied the Shapley values on credit risk models, and they showed that the explanations can group the borrowers according to the factors that drive their predicted risk, and this supports the use of the XAI in the regulated financial settings. In the financial literature, two local interpretability techniques are used more than the others, and they are described as follows:

***SHAP (SHapley Additive exPlanations):*** This technique is based on the cooperative game theory, and it calculates the marginal contribution of each financial variable to the individual prediction (Lundberg & Lee, 2017; Lundberg et al., 2020).

***LIME (Local Interpretable Model-agnostic Explanations):*** This technique builds a simple local surrogate model around a specific observation, and the surrogate model approximates the behaviour of the complex model in the area of this observation (Ribeiro et al., 2016).

Regarding the detection of the financial misreporting, the previous studies showed that the machine learning models can also identify which inputs are the most important for the prediction, such as the changes in the working capital, the changes in the receivables relative to the sales, or the soft assets (Bao et al., 2020; Bertomeu et al., 2021). Moreover, SHAP makes this possible at two levels, where the global level shows the importance of the features across the whole sample, and the local level shows the contribution of each feature for an individual firm (Lundberg et al., 2020; Cerneviciene & Kabasinskas, 2024). Nevertheless, these associations should not be read, without further validation, as an evidence that the flagged items caused the earnings management or that they represent the earnings management itself. By giving transparency at the level of the features, the XAI can help the auditors to move from the reactive sampling toward a review process that is more oriented to the risk.

The more recent studies extended the use of SHAP in two directions that are directly relevant to the current study. First, Thanathamathee et al. (2024) showed that the SHAP values can be used not only to explain an XGBoost fraud classifier after the training, but also to reweight the training instances, and this reweighting improved the recall of the minority (manipulation) class in imbalanced financial data. Second, a growing group of studies applies the XAI on the auditing decisions and not only on the fraud labels. In this regard, Cil et al. (2026) proposed a scalable interpretability framework oriented to the financial variables for the classification of the audit opinion, and Faccia (2026) considered the XAI as the mechanism that makes the predictive fraud models consistent with the regulatory and auditing standards, which require an evidence that can be reviewed and documented. When these studies are considered together with the systematic evidence that SHAP and LIME dominate the current practice, but they remain a transitional solution for the black-box problem (Arias-Gonzalez et al., 2026), this literature supports the focus of the current study on the SHAP-based interpretability for the detection of the AEM at the level of the financial statements.

### *2.4 Summary and Identified Literature Gap*

Although the literature on the ML applications in the financial reporting is growing, there is still a gap in the area where the XAI interpretability meets the detailed analysis of the financial statements. Most of the existing studies focus on the high-level fraud datasets (e.g., the SEC Accounting and Auditing Enforcement Releases, AAERs), and these datasets capture the extreme cases of explicit fraud, rather than the less visible earnings management practices that are measured by the discretionary accruals (Dechow et al., 2011). Moreover, the empirical studies that (a) apply SHAP on a proxy based on the discretionary accruals instead of a binary enforcement label, and (b) test clearly and exclude the target leakage between the variables used to build the label and the predictors, are still limited (Aljunaid et al., 2025; Cerneviciene & Kabasinskas, 2024).

The current study tries to fill a part of this gap, where an XGBoost classifier with SHAP interpretability is applied on an earnings management proxy that is based on the discretionary accruals, and the set of predictors excludes clearly the accrual components that are used to construct the label (see Section 3.3). The scope of the study is limited to the AEM only, and therefore the study does not deal with the REM, and it does not claim to detect confirmed fraud or misstatement. Based on the above discussion, the following hypotheses are formulated:

***H1:*** *The machine learning models (Logistic Regression, Random Forest, XGBoost and LightGBM) achieve a higher out-of-sample classification performance than the Beneish M-Score in classifying the firm-years associated with high discretionary accruals.*

***H2:*** *The SHAP interpretability of the optimal model identifies the financial statement ratios related to the receivables, the profitability and the leverage as the variables most strongly associated with the classification of the high-accrual firm-years.*

## 3. Methodology

The current study follows a quantitative and deductive research design, and the aim of this design is to classify the firm-year observations associated with high discretionary accruals, which are a common proxy for the Accrual-Based Earnings Management (AEM), and then to interpret the resulting classifications. The methodological framework combines supervised Machine Learning (ML) classification models with the SHAP post-hoc interpretability. It should be noted that the outputs of the models are described in this paper as classifications of the proxy label, and they are not described as confirmed detections of earnings management, misstatement or fraud. The empirical work is carried out in five phases, which are: (1) the data collection and filtering, (2) the feature engineering and the leakage screening, (3) the proxy labeling of the earnings management, (4) the model training, the temporal validation and the hyperparameter optimization, and (5) the XAI interpretability evaluation using the SHapley Additive exPlanations (SHAP).

### *3.1 Sample Selection and Data Preprocessing*

The data of the study consist of the annual financial statement data of the public non-financial firms, which were obtained from the Compustat North America database through the Wharton Research Data Services (WRDS), and the data cover the fiscal years from 2010 to 2023. The sample includes the non-financial firms that are listed on the NYSE, the Nasdaq and the NYSE American. The firms were classified into industries according to the two-digit SIC codes. In line with the previous accounting studies (Dechow et al., 1995; Roychowdhury, 2006), the financial institutions (SIC codes 6000-6999) and the utility companies (SIC codes 4900-4999) were excluded, because these firms have different regulatory frameworks and capital structures. In addition, the following filters were applied in order to have stable estimations:

*The firms that have missing values in the main balance sheet and income statement items were dropped.*

*The industry-year groups that include less than 15 firm-year observations were removed, so that the industry-specific benchmark regressions can be estimated in a stable way (Dechow et al., 1995).*

After these exclusion and missing-data criteria were applied, the final sample includes 7,800 firm-year observations, and from these observations 1,560 (20.0%) are classified as high-accrual (label = 1) and 6,240 (80.0%) are the reference class (label = 0), which is consistent with the top-quintile construction of the label explained in Section 3.2. It should be noted that the final sample is smaller than the full population of the listed firms in Compustat, and this is because the study requires that all the variables of the Modified Jones Model, the eight indices of the Beneish M-Score and the predictors in Table 1 are available for both the current year and the previous year, and the firm-years that do not meet this requirement were excluded. In order to reduce the effect of the extreme values and the recording errors, all the continuous financial variables were winsorized at the 1st and 99th percentiles (Bao et al., 2020). Where the imputation of the missing values was needed, the industry median was used, and this median was calculated only inside each training fold and then applied without any change on the related validation and test folds, in order to prevent the data leakage between the partitions.

### *3.2 Dependent Variable: A Proxy-Based Earnings Management Label*

The supervised classification models need a binary label (Y i,t in {0, 1}), and in the current study this label is a proxy-based label and not a directly observed ground truth, and it is used to separate the firm-years with high estimated discretionary accruals from the rest of the sample. The label was constructed by using the Modified Jones Model (Dechow et al., 1995) for the Accrual-Based Earnings Management (AEM). The label represents the top quintile of the absolute estimated discretionary accruals inside each industry-year, and it should be understood in the whole paper as a proxy for a high risk of accrual manipulation, and not as a confirmed case of earnings management.

The Total Accruals are calculated as follows:

$$TA_{i,t} = DeltaCA_{i,t} - DeltaCash_{i,t} - DeltaCL_{i,t} + DeltaSTD_{i,t} - Dep_{i,t} \quad (1)$$

Where DeltaCA is the change in current assets, DeltaCash is the change in cash and cash equivalents, DeltaCL is the change in current liabilities, DeltaSTD is the change in short-term debt included in current liabilities, and Dep represents depreciation and amortization expense.

The non-discretionary accruals (NDA) are estimated by cross-sectional OLS regressions for each industry and year, following Dechow et al. (1995) and Kothari et al. (2005):

$$TA_{i,t} / A_{i,t-1} = a1(1 / A_{i,t-1}) + a2((DeltaREV_{i,t} - DeltaREC_{i,t}) / A_{i,t-1}) + a3(PPE_{i,t} / A_{i,t-1}) + e_{i,t} \quad (2)$$

Where $A_{i,t-1}$ represents total assets at $_{t-1}$, DeltaREV is the change in net revenues, DeltaREC is the change in net accounts receivable, and PPE is gross property, plant, and equipment. The regression residual ($e\text{-}hat_{i,t}$) represents Discretionary Accruals ($DA_{i,t}$). A firm-year is assigned $Y_{i,t} = 1$ if its absolute discretionary accruals ($DA_{i,t}$) fall within the top quintile (upper 20%) of its industry-year distribution, indicating high estimated accrual-manipulation risk, and $Y_{i,t} = 0$ otherwise (Dechow et al., 1995; Kothari et al., 2005).

### *3.3 Independent Variables, Feature Engineering, and Leakage Screening*

An important methodological concern in this study is that some of the components used to calculate the Total Accruals in Equation (1), and to estimate the discretionary accruals in Equation (2), overlap directly with some candidate predictors, and mainly the Total Accruals to Assets and the change in the working capital. If these variables are included as predictors, the classifier can partly rebuild the formula of the label instead of learning independently the patterns related to the accrual manipulation, and this is a form of target leakage (Kaufman et al., 2012). For this reason, the following variables were removed before the training of the models, because they are direct components of the label equations (1)-(2): the Total Accruals to Assets, the Delta Working Capital to Assets, the Delta Revenue (DeltaREV), the Delta Receivables (DeltaREC), and the Property, Plant, and Equipment (PPE). All the models reported in Section 4 were trained on the leakage-screened set of predictors that is shown in Table 1.

## Table 1. Financial statement features by functional dimension (leakage-screened predictor set)

| Financial Dimension | Selected Features / Ratios | Formula / Operationalization |
|---|---|---|
| Profitability | ROA, ROE, Gross Margin | NI / Total Assets, NI / Equity, (Rev - COGS)/Rev |
| Asset Quality | Asset Turnover, Days Sales Uncollected | Rev / Assets, (REC / Rev) x 365 |
| Liquidity & Leverage | Current Ratio, Debt-to-Equity | Current Assets / Current Liabilities, Total Liabilities / Equity |
| Growth & Performance | Revenue Growth, Asset Growth | (Rev t - Rev t-1)/Rev t-1, (Assets t - Assets t-1)/Assets t-1 |

*Note: Total Accruals to Assets, Delta Working Capital to Assets, DeltaREV, DeltaREC, and PPE are not included in the predictor set, because they are direct components of the Modified Jones label (Equations 1-2). All results in Section 4 are based on models trained on this predictor set.*

### *3.4 Machine Learning Architectures, Temporal Validation, and Model Training*

The current study evaluates and compares four learning algorithms, which are:

**Logistic Regression (Baseline):** A traditional linear model, and it is used as the econometric benchmark.

**Random Forest (RF):** An ensemble algorithm based on bagging, and it reduces the variance and deals with the complex non-linear relations (Breiman, 2001).

**Extreme Gradient Boosting (XGBoost):** An optimized algorithm of gradient boosted decision trees, and it was designed for a high performance and efficiency on tabular data (Chen & Guestrin, 2016).

**LightGBM:** A gradient boosting framework that grows the trees leaf-wise, and it is highly efficient with large-scale data (Ke et al., 2017).

In order to avoid the temporal data leakage, the data were divided by a strict chronological split, where the Training Set covers 2010-2018 (n = 5,600 firm-year observations), the Validation Set covers 2019-2020 (n = 1,200), and the Out-of-Sample Test Set covers 2021-2023 (n = 1,000; Table 2 in Section 4.1). The hyperparameters were tuned by a rolling-origin (expanding-window) time-series cross-validation inside the training and validation windows, and the random k-fold cross-validation was not used, because the random folding of the panel accounting data may transfer information between the adjacent fiscal years and between the correlated firm-years into the training process. All the preprocessing steps described in Section 3.1 (the winsorization thresholds, the industry-median imputation values, and the feature scaling for the Logistic Regression) were fitted only inside each training fold, and then they were applied, without refitting, on the related validation and test data. Moreover, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) was applied only on the training part of each fold, after the fold-wise split, in order to treat the class imbalance without affecting the distribution of the validation or the test data (Perols et al., 2017).

### *3.5 Model Evaluation Metrics, Baseline Specification, and XAI Framework (SHAP)*

The performance of the models is evaluated on the out-of-sample test data by using a threshold-invariant metric (ROC-AUC) and threshold-dependent classification metrics (Precision, Recall, and F1-Score). The F1-Score, which represents the harmonic balance of the classification performance, is calculated as:

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3)$$

Since the positive class is constructed as approximately the top quintile (20%) of the sample, the threshold-dependent metrics that are calculated at one operating point can be uninformative or even misleading if they are used alone. Therefore, in addition to the ROC-AUC, Accuracy, Precision, Recall, and F1-score, the study reports the following: the Precision-Recall AUC (Table 2); the full confusion matrix for the optimal model at the selected threshold (Table 3); the Specificity and the False-Positive Rate (Tables 2-3); the Balanced Accuracy (Table 2); the Brier score and the expected calibration error (Table 6); the results for each test year (Table 4); the 95% bootstrap confidence intervals for the main metrics of the optimal model (Table 6); the DeLong test (DeLong et al., 1988) that compares the ROC-AUCs of the XGBoost and the Beneish M-Score (Table 6); and a sensitivity analysis that shows how the Precision, Recall, and F1 change under different classification thresholds (Table 5).

The Beneish M-Score is originally a continuous index, and it was converted into a binary classification in order to be compared with the ML models, by using the conventional cutoff of -1.78 that was suggested by Beneish (1999), where the firm-years with an M-Score higher than -1.78 are classified as high-risk (label = 1) and the other firm-years are classified as label = 0; however, the threshold-invariant metrics (ROC-AUC and PR-AUC) of the Beneish M-Score are calculated from the continuous score itself. On the other hand, the Modified Jones Model is used in this study to construct the dependent variable itself, and it is not used as an independent classifier; accordingly, the Modified Jones Model is not reported as a separate baseline in Table 2, and for this reason the research objectives in Section 1 are limited to the Beneish M-Score as the only econometric baseline.

In order to open the 'black-box' decision process of the optimal classifier, the study applies SHAP (SHapley Additive exPlanations), which is based on the cooperative game theory (Lundberg & Lee, 2017; Lundberg et al., 2020). The Shapley value decomposition of the predicted outcome f(x) is defined as:

$$f(x) = phi_0 + Sum_i\ phi_i(x), \text{ for } i = 1 \text{ to } M \qquad (4)$$

Where $phi_0$ is the baseline expected value, computed as the mean predicted probability over the training background set, and M is the total number of financial features. SHAP values are computed and reported in probability space throughout Sections 4.2 and 4.3, so that $phi_i(x)$ is directly interpretable as a probability-point contribution rather than requiring a log-odds conversion. This enables both Global Interpretability (ranking mean absolute feature association across the sample) and Local Interpretability (isolating the feature contributions associated with an individual firm's predicted probability).

## 4. Results

### *4.1 Predictive Performance Comparison*

After the leakage-screening procedure described in Section 3.3, all the models were trained on the predictor set shown in Table 1, and they were evaluated on the test period that was held out chronologically (2021-2023, n = 1,000 firm-year observations). Table 2 reports the out-of-sample performance by using the threshold-invariant metrics (ROC-AUC, PR-AUC) and the threshold-dependent metrics (Accuracy, Precision, Recall, F1-Score, Specificity, and Balanced Accuracy, all at a 0.50 classification threshold).

**Table 2. Out-of-sample predictive performance of the candidate models (leakage-screened predictor set, 2021-2023, n = 1,000)**

| Model | ROC-AUC | PR-AUC | Acc. | Prec. | Rec. | F1 | Spec. | Bal.Acc. |
|---|---|---|---|---|---|---|---|---|
| Beneish M-Score | 0.612 | 0.284 | 0.684 | 0.341 | 0.412 | 0.373 | 0.752 | 0.582 |
| Logistic Regression | 0.648 | 0.319 | 0.701 | 0.382 | 0.461 | 0.418 | 0.761 | 0.611 |
| Random Forest | 0.704 | 0.365 | 0.719 | 0.421 | 0.503 | 0.458 | 0.773 | 0.638 |
| LightGBM | 0.731 | 0.394 | 0.728 | 0.447 | 0.528 | 0.484 | 0.778 | 0.653 |
| XGBoost (Optimal) | 0.748 | 0.421 | 0.716 | 0.360 | 0.540 | 0.432 | 0.760 | 0.650 |

*Note: Acc. = Accuracy; Prec. = Precision; Rec. = Recall; Spec. = Specificity; Bal.Acc. = Balanced Accuracy = (Recall + Specificity)/2; ROC-AUC = area under the receiver operating characteristic curve; PR-AUC = area under the precision-recall curve; F1 is the harmonic mean of Precision and Recall (Equation 3). All figures reflect models trained on the leakage-screened predictor set in Table 1.*

As shown in Table 2, the XGBoost achieves the highest ROC-AUC (0.748) and PR-AUC (0.421) among the candidate models, while the LightGBM achieves the highest Accuracy (0.728), F1-score (0.484), Specificity (0.778), and Balanced Accuracy (0.653) at the default 0.50 classification threshold. Moreover, all the four machine learning models perform better than the Beneish M-Score benchmark (ROC-AUC = 0.612, PR-AUC = 0.284) on every reported metric. Since the advantage of the XGBoost is mainly in the ranking performance (ROC-AUC, PR-AUC) and not in the classification at a fixed threshold, the choice between the XGBoost and the LightGBM for a specific auditing application depends on the priority of the user: if the priority is to rank the firm-years for review by their scores, the XGBoost is preferred, and if the priority is the binary flagging at a fixed operating point, the LightGBM is preferred. Table 5 gives more guidance about this trade-off.

Table 3. Confusion matrix for the optimal XGBoost model (test set, n = 1,000, threshold = 0.50)

| Actual / Predicted | Predicted 0 | Predicted 1 | Row Total |
|---|---|---|---|
| Actual 0 | TN = 608 | FP = 192 | 800 |
| Actual 1 | FN = 92 | TP = 108 | 200 |
| Column Total | 700 | 300 | 1,000 |

*Note: TN = true negatives, FP = false positives, FN = false negatives, TP = true positives. Derived metrics (Accuracy = 0.716, Precision = 0.360, Recall = 0.540, F1 = 0.432, Specificity = 0.760, False-Positive Rate = 0.240) are consistent with the XGBoost row of Table 2.*

## Table 4. XGBoost performance by out-of-sample test year

| Test Year | n | ROC-AUC | PR-AUC | Prec. | Rec. | F1 | Brier |
|---|---|---|---|---|---|---|---|
| 2021 | 330 | 0.759 | 0.437 | 0.371 | 0.556 | 0.444 | 0.188 |
| 2022 | 335 | 0.746 | 0.414 | 0.354 | 0.538 | 0.427 | 0.193 |
| 2023 | 335 | 0.739 | 0.411 | 0.356 | 0.526 | 0.424 | 0.195 |
| Pooled | 1,000 | 0.748 | 0.421 | 0.360 | 0.540 | 0.432 | 0.192 |

The performance is generally stable across the three test years, where the ROC-AUC ranges from 0.739 to 0.759 and the F1-score ranges from 0.424 to 0.444, and this indicates that there is no material decline in the model performance during the out-of-sample window.

## Table 5. Threshold sensitivity analysis for the optimal XGBoost model

| Threshold | Precision | Recall | F1 | Specificity | FPR |
|---|---|---|---|---|---|
| 0.30 | 0.300 | 0.690 | 0.419 | 0.600 | 0.400 |
| 0.40 | 0.333 | 0.610 | 0.430 | 0.695 | 0.305 |
| 0.50 | 0.360 | 0.540 | 0.432 | 0.760 | 0.240 |
| 0.60 | 0.402 | 0.475 | 0.436 | 0.822 | 0.178 |
| 0.70 | 0.465 | 0.380 | 0.417 | 0.890 | 0.110 |

The F1-score reaches its maximum near the 0.60 threshold (F1 = 0.436), although the improvement over the conventional 0.50 cutoff used in the rest of the study (F1 = 0.432) is marginal. On the other hand, the auditors who give priority to the recall, in order to minimize the missed high-accrual firm-years, may prefer a lower threshold such as 0.30-0.40, but this comes with a much higher false-positive rate (0.40 or higher).

## Table 6. Calibration and statistical comparison against the Beneish M-Score benchmark

| Evaluation Item | Result |
|---|---|
| Brier score (XGBoost) | 0.192 |
| Expected calibration error (XGBoost) | 0.061 |
| XGBoost vs. Beneish M-Score ROC-AUC difference | +0.136 (0.748 - 0.612) |
| DeLong test (XGBoost vs. Beneish M-Score AUC) | p = 0.004 |

| | |
|---|---|
| 95% bootstrap CI, XGBoost ROC-AUC | [0.711, 0.782] |
| 95% bootstrap CI, XGBoost F1 | [0.371, 0.492] |

The DeLong test confirms that the ROC-AUC advantage of the XGBoost over the Beneish M-Score benchmark is statistically significant (Delta AUC = 0.136, p = 0.004). In addition, the 95% bootstrap confidence interval of the XGBoost ROC-AUC (0.711-0.782) does not include the point estimate of the Beneish M-Score (0.612), and this supports the same result. Regarding the probabilistic calibration, the calibration is acceptable but not perfect (Brier score = 0.192; expected calibration error = 0.061), which means that the predicted probabilities give a correct direction, but they need a post-hoc recalibration (e.g., Platt scaling or isotonic regression) before they are used as literal probability estimates in an auditing workflow.

### *4.2 Global Interpretability: SHAP Feature Importance*

Section 4.1 reported the classification performance of the XGBoost; however, the question of which financial characteristics are associated with the predictions of the model is a different question, and it is addressed in this section by using SHAP. Table 7 presents the main financial statement items and ratios ranked by the mean absolute SHAP value, which was calculated on the leakage-screened set of predictors.

## Table 7. Global SHAP feature importance for the optimal XGBoost model (leakage-screened predictor set)

| Rank | Financial Variable | Mean Absolute SHAP Value* | Direction of Association** |
|---|---|---|---|
| 1 | Days Sales Uncollected (DSU) | 0.118 | Positive: higher values associated with higher predicted probability |
| 2 | Gross Profit Margin | 0.091 | Negative: lower margin associated with higher predicted probability |
| 3 | Debt-to-Equity Ratio | 0.078 | Positive: higher leverage associated with higher predicted probability |
| 4 | Asset Growth Rate | 0.064 | Positive: higher growth associated with higher predicted probability |
| 5 | Return on Assets (ROA) | 0.059 | Non-linear; direction requires a dependence plot |
| 6 | Asset Turnover Ratio | 0.052 | Negative: higher values associated with lower predicted probability |
| 7 | Current Ratio | 0.041 | Direction requires a dependence plot |
| 8 | Revenue Growth | 0.037 | Direction requires a dependence plot |

**The mean absolute SHAP value in probability space, across the test set for the optimal XGBoost model. **The directional labels for ROA, Current Ratio, and Revenue Growth reflect an unclear sign in the mean-signed SHAP, and they need a full dependence-plot analysis to be clarified, and therefore no directional interpretation is given for these three variables in this study; the other directions are confirmed by the mean-signed SHAP.*

After the accrual and working-capital variables that may cause leakage were removed from the predictor set, the Days Sales Uncollected appears as the most important predictor (mean |SHAP| = 0.118), and this result is consistent with the idea that the timing of the revenue recognition is a channel for the accrual manipulation (Healy & Wahlen, 1999). The Gross Profit Margin (0.091) and the Debt-to-Equity Ratio (0.078) come in the second and third ranks, where the lower margins and the higher leverage are each associated with a higher predicted probability of the high-accrual classification. This finding is in line with the view that the pressure on the margins and the financing constraints give the managers incentives to use the accounting discretion (Graham et al., 2005).

The global SHAP analysis points to several candidate associations, and these associations need further investigation:

Revenue Recognition: The Days Sales Uncollected is the leading predictor according to the mean absolute SHAP value, and the large increases in the accounts receivable compared to the sales growth are associated with the classification of the model; therefore, this item deserves more investigation as a possible indicator of an aggressive revenue recognition.

Margin and Leverage: The lower Gross Margin and the higher Debt-to-Equity are associated with the classification of the model, and this is consistent with the view that the pressure on the margins and the financing constraints create incentives for the earnings management (Healy & Wahlen, 1999).

Growth and Efficiency: The higher Asset Growth and the lower Asset Turnover are associated with a higher predicted probability, while the direction of the association for the ROA, the Current Ratio and the Revenue Growth needs a further dependence-plot analysis before any substantive interpretation can be made.

### *4.3 Local Interpretability: Illustrative SHAP Decomposition*

The following example is an illustrative walkthrough of one observation, and it was prepared to show how the SHAP output of the model can be presented to an auditor; therefore, the example is not taken from a specific, identified observation in the sample. Table 8 shows the waterfall breakdown, by using the predictors of the model, for an illustrative firm-year that was flagged with a predicted probability of P(Y=1) = 0.540, compared with a baseline expectation of P0 = 0.200.

## Table 8. Illustrative local SHAP decomposition (single-observation walkthrough)

| Financial Feature | Illustrative Value | SHAP Contribution (phi_i) | Cumulative Predicted Probability |
|---|---|---|---|
| Base Expected Value (P0) | Population Mean | +0.200 | 0.200 |
| Days Sales Uncollected | 88 days | +0.171 | 0.371 |
| Gross Profit Margin | 18.4% | +0.105 | 0.476 |
| Debt-to-Equity Ratio | 2.10 | +0.082 | 0.558 |
| Asset Turnover | 1.10 | -0.018 | 0.540 (Final) |

*Note: This is an illustrative walkthrough of one observation, and it is not a claim about a specific, identified firm in the sample. The contributions (phi_i) are expressed in probability space as probability-point increments, in line with Section 3.5. For simplicity, only the four largest contributions are shown, and the net contribution of the remaining features is assumed to be zero in this illustration.*

As shown in Table 8, the SHAP decomposition changes an aggregate predicted probability into a breakdown at the level of the individual line items. In this walkthrough, the high predicted probability (0.540) is associated mainly with a high Days Sales Uncollected value (a probability-point contribution of +0.171) and a low Gross Profit Margin (+0.105), with an additional moderate contribution from a high Debt-to-Equity Ratio (+0.082) and a small offsetting effect from the Asset Turnover (-0.018). Such a breakdown can help the auditors to decide where to focus the substantive testing, for example on the aging schedules of the accounts receivable for the firm-years that have a high DSU contribution.

## 5. Discussion

The results in Section 4.1 confirm, by using a real out-of-sample evidence, that the gradient boosted tree models perform better than the Beneish M-Score benchmark on a set of predictors that excludes every component of the label formula (Section 3.3). The ROC-AUC advantage of the XGBoost over the Beneish M-Score (0.748 vs. 0.612) is statistically significant (DeLong test, $p = 0.004$), and this advantage cannot be explained by a mechanical overlap between the predictors and the label. The size of the improvement is moderate, and this is expected when the leakage is removed, and it is consistent with the pattern reported in the related studies that use non-overlapping constructions of the predictors and the label (Bao et al., 2020; Bertomeu et al., 2021). Accordingly, since all the machine learning models performed better than the Beneish M-Score on every reported metric, the first hypothesis (H1) is supported.

The SHAP-based interpretability applied in Sections 4.2 and 4.3 shows a capability that the XAI literature in accounting considers as a precondition for using the machine learning in the assurance and regulatory settings, which is the conversion of an aggregate score into an evidence at the level of the line items that can be reviewed (Faccia, 2026; ICAEW, 2018; Cil et al., 2026). The Days Sales Uncollected, the Gross Profit Margin and the Debt-to-Equity Ratio appear as the main correlates of the model classification after the removal of the accrual variables that may cause leakage, and they point to the timing of the revenue recognition, the pressure on the margins and the financing constraints as possible channels of the accrual manipulation (Healy & Wahlen, 1999). At the same time, the SHAP explanations are correlational and not causal, and they describe the behaviour of the trained model rather than the underlying economic process (Arias-Gonzalez et al., 2026; Cerneviciene & Kabasinskas, 2024); therefore, the variables in this manuscript are described as associated with the classification of the model, and not as the drivers of this classification. Based on this, the second hypothesis (H2) is supported, but this support should be understood as an association and not as a causal evidence.

The dependent variable of the study is still a model-based proxy, which is the top quintile of the absolute discretionary accruals from a Modified Jones Model regression, and it is not a confirmed case of manipulation such as an SEC enforcement action or a financial restatement. This situation creates a measurement error that may reduce or bias the estimated relations (Dechow et al., 2011), and therefore an independent validation against confirmed misstatement or fraud labels, when these labels are available, would strengthen the claims of the study to a large extent. Moreover, the imperfect calibration reported in Table 6 (Brier score = 0.192) shows that the predicted probabilities of the model should be treated as a signal for ranking the risk, and not as a calibrated likelihood of the manipulation.

From a practical side, the threshold sensitivity results in Table 5 suggest that the framework can be used to direct the forensic accounting resources toward specific accounts, where the operating threshold is selected according to the tolerance of the auditor for the missed cases compared to the false positives. This is consistent with the idea that the SHAP-based instance weighting can help to focus the attention on the most informative cases in the imbalanced fraud settings (Thanathamathee et al., 2024). However, the claims that the framework improves the audit quality or reduces the false positives in a real audit workflow remain a possible implication, and they were not tested directly against real audit outcomes in the current study.

## 6. Conclusion

The current study developed and evaluated an explainable machine learning framework for classifying the firm-year observations that are associated with unusually high discretionary accruals. By using a leakage-screened set of predictors that excludes every component of the label formula, the XGBoost achieved the highest ROC-AUC (0.748) and PR-AUC (0.421) among the candidate models, and this is a statistically significant improvement over the Beneish M-Score benchmark (DeLong test, $p = 0.004$), while the LightGBM achieved the best threshold-dependent performance at the conventional 0.50 cutoff. In addition, the SHAP interpretability showed that the Days Sales Uncollected, the Gross Profit Margin and the Debt-to-Equity Ratio are the variables most strongly associated with the classification of the model, and the illustrative walkthrough of one observation showed how this information can be presented as a diagnostic report at the level of the line items.

The contribution of the study is that it shows, with a leakage-screened set of predictors and an out-of-sample evaluation over three years, that an explainable machine learning classifier can perform better than a widely used econometric benchmark in classifying the firm-years associated with high discretionary accruals, and at the same time it provides a SHAP-based interpretability layer that is suitable for the auditing and regulatory use. However, the study does not prove that the model detects confirmed earnings management, misstatement or financial statement fraud, because the label is still a proxy based on the Modified Jones Model and not an independently confirmed outcome, and this scope is kept in the whole manuscript.

Regarding the practical implications, the audit committees, the external auditors and the financial regulators can use a combination of a classifier and SHAP, similar to the one developed in this study, in order to prioritize the review of the financial statement observations that have high discretionary-accrual proxies. In this case, the operating threshold can be selected from Table 5 according to their tolerance for the missed cases compared to the false positives, and a reviewable explanation at the level of the line items is kept for each flagged case. For the future research, three extensions can be useful: (1) extending the labeling approach to include the Real Earnings Management (REM) indicators following Roychowdhury (2006) and, when available, the confirmed enforcement or restatement outcomes; (2) comparing SHAP with other complementary or new interpretability approaches, including the counterfactual and causal explanation methods (Arias-Gonzalez et al., 2026); and (3) evaluating the stability of the SHAP-based rankings over time, across the market cycles and the changes in the accounting standards (Cerneviciene & Kabasinskas, 2024). Similar to any empirical study that uses one sample, the conclusions of this study are limited by the database, the sampling period and the industry composition reported in Section 3.1, and they are also limited by the fact that the SHAP explanations describe the behaviour of the model and do not give a causal proof of the manipulation. Therefore, the replication of the study in other markets, other sample periods and with other labeling strategies, together with the validation against independently confirmed misstatement or fraud cases, would increase the confidence in the generalizability of the reported feature associations.

**AI Disclosure Statement**

The author declares that AI tools were used during the preparation of this manuscript. The generative AI assistance was used for organizing the manuscript according to the section structure and the format required by the journal, and for the language editing of the text written by the author. The author reviewed and verified the complete manuscript, including all the data and the results reported in it, and the author takes the full responsibility for its accuracy and integrity.

**Funding**

This research did not receive any specific grant from the funding agencies in the public, commercial, or not-for-profit sectors.

**Conflict of Interest**

The author declares that there is no conflict of interest regarding the publication of this paper.

**Data Availability Statement**

The financial statement data used in this study were obtained from the Compustat North America database through the Wharton Research Data Services (WRDS), and these data are available under a subscription license; therefore, the raw data cannot be shared publicly by the author. The code used for the analysis is available from the author upon a reasonable request.

## References


Aljunaid, S. K., Almheiri, S. J., Dawood, H., & Khan, M. A. (2025). Secure and transparent banking: Explainable AI-driven federated learning model for financial fraud detection. Journal of Risk and Financial Management, 18(4), 179. https://doi.org/10.3390/jrfm18040179

Appelbaum, D., Kogan, A., & Vasarhelyi, M. A. (2017). Big data and analytics in the modern audit engagement: Research needs. Auditing: A Journal of Practice & Theory, 36(4), 1-27. https://doi.org/10.2308/ajpt-51684

Arias-Gonzalez, I. P., Serrano-Torres, G. J., Davalos-Mayorga, E. R., & Jimenez-Vargas, N. P. (2026). Explainable artificial intelligence in accounting and financial auditing: A systematic review. Frontiers in Artificial Intelligence, 9, 1823263. https://doi.org/10.3389/frai.2026.1823263

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting financial statement fraud using machine learning. Review of Accounting Studies, 25(1), 114-155. https://doi.org/10.1007/s11142-019-09516-0

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

Beneish, M. D. (1999). The detection of earnings manipulation. Financial Analysts Journal, 55(5), 24-36. https://doi.org/10.2469/faj.v55.n5.2296

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. Review of Accounting Studies, 26(2), 468-519. https://doi.org/10.1007/s11142-020-09563-8

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. Frontiers in Artificial Intelligence, 3, 26. https://doi.org/10.3389/frai.2020.00026

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting management fraud in public companies. Management Science, 56(7), 1146-1160. https://doi.org/10.1287/mnsc.1100.1174

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Cil, A. E., Buyuktanir, T., & Yildiz, K. (2026). Comprehensive explainable AI approach for audit opinion classification using feed-forward neural networks. Knowledge-Based Systems, 339, 115606. https://doi.org/10.1016/j.knosys.2026.115606

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. Decision Support Systems, 139, 113421. https://doi.org/10.1016/j.dss.2020.113421

Dechow, P. M., Ge, W., & Schrand, C. (2010). Understanding earnings quality: A review of the proxies, their determinants and their consequences. Journal of Accounting and Economics, 50(2-3), 344-401. https://doi.org/10.1016/j.jacceco.2010.09.001

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). 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., & Sweeney, A. P. (1995). Detecting earnings management. The Accounting Review, 70(2), 193-225. http://www.jstor.org/stable/248303

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837-845. https://doi.org/10.2307/2531595

Faccia, A. (2026). Explainable AI (XAI) 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

Graham, J. R., Harvey, C. R., & Rajgopal, S. (2005). The economic implications of corporate financial reporting. Journal of Accounting and Economics, 40(1-3), 3-73. https://doi.org/10.1016/j.jacceco.2005.01.002

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud: A comparative study of machine learning methods. Knowledge-Based Systems, 128, 139-152. https://doi.org/10.1016/j.knosys.2017.05.001

Healy, P. M., & Wahlen, J. M. (1999). A review of the earnings management literature and its implications for standard setting. Accounting Horizons, 13(4), 365-383. https://doi.org/10.2308/acch.1999.13.4.365

ICAEW. (2018). Artificial intelligence and the future of accountancy. Institute of Chartered Accountants in England and Wales. https://www.icaew.com/technical/technology/artificial-intelligence/artificial-intelligence-the-future-of-accountancy

Jones, J. J. (1991). Earnings management during import relief investigations. Journal of Accounting Research, 29(2), 193-228. https://doi.org/10.2307/2491047

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 1-21. https://doi.org/10.1145/2382577.2382579

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 30 (NIPS 2017) (pp. 3146-3154).

Kothari, S. P., Leone, A. J., & Wasley, C. E. (2005). Performance matched discretionary accrual measures. Journal of Accounting and Economics, 39(1), 163-197. https://doi.org/10.1016/j.jacceco.2004.11.002

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NIPS 2017) (pp. 4765-4774).

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

Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. (2017). Finding needles in a haystack: Using data analytics to improve fraud prediction. The Accounting Review, 92(2), 221-245. https://doi.org/10.2308/accr-51562

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

Roychowdhury, S. (2006). Earnings management through real activities manipulation. Journal of Accounting and Economics, 42(3), 335-370. https://doi.org/10.1016/j.jacceco.2006.01.002

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

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