American Journal of Interdisciplinary Studies

Volume: 16; **Issue:** 8
Pages: 94-127
eISSN: 3067-5146

Machine Learning for Financial-Statement Anomaly Detection in U.S. Small and Mid-Sized Enterprises: A Quantitative Study Using Gradient Boosting and SHAP

Songeta Dhar[1];

[1]. *Doctor of Business Administration (DBA) Candidate, Business Analytics; Westcliff University, California, USA; Master of Business Studies (MBS) in Accounting, National University, Bangladesh; Bachelor of Commerce (Hons.) in Accounting, National University, Bangladesh; Email: songeta.bd@gmail.com*

Doi: 10.63125/61wry195
Received: 14 May 2026; **Revised:** 17 June 2026; **Accepted:** 11 July 2026; **Published:** 03 August 2026;

Abstract

Financial-statement anomalies can signal reporting irregularities, yet identifying them among small and mid-sized enterprises (SMEs) is challenging because labeled cases are uncommon and financial relationships may be nonlinear. This quantitative study examined the classification performance of gradient boosting for firm-year anomaly detection among U.S. SMEs and used SHAP to explain model predictions. The planned analytical framework compared Boost with logistic regression and random forest using financial-statement predictors. Evaluation included precision, recall, specificity, F1 score, ROC-AUC, PR-AUC, Brier score, and paired comparisons. Validation was designed to separate firms and reporting periods across model development and testing, while SHAP summaries and local explanations characterized predictor contributions. The numerical results in this draft are examples, not estimates calculated from an analyzed dataset. In the example, the sample contained 1,200 firm-year observations from 240 SMEs, including 96 labeled anomalous observations, or 8.0%. The test set contained 260 observations, with anomaly prevalence of 7.7%. Boost achieved ROC-AUC of 0.914 and PR-AUC of 0.611, compared with 0.831 and 0.412 for logistic regression, and 0.879 and 0.524 for random forest. Its example recall was 0.800, precision was 0.471, and Brier score was 0.066. Paired comparisons showed an ROC-AUC difference of 0.083 versus logistic regression (95% CI [0.018, 0.146], p = .014) and 0.035 versus random forest (95% CI [−0.024, 0.094], p = .230). Receivable's growth, accrual intensity, and operating cash flow relative to earnings ranked among the leading SHAP predictors. These example findings demonstrate how discrimination, calibration, threshold performance, and model explanations can be reported together. Actual conclusions require analysis of the study data, verification of outcome labels, and replacement of every estimate with observed results. SHAP attributions describe model behavior and do not establish causality, misstatement, or fraud.

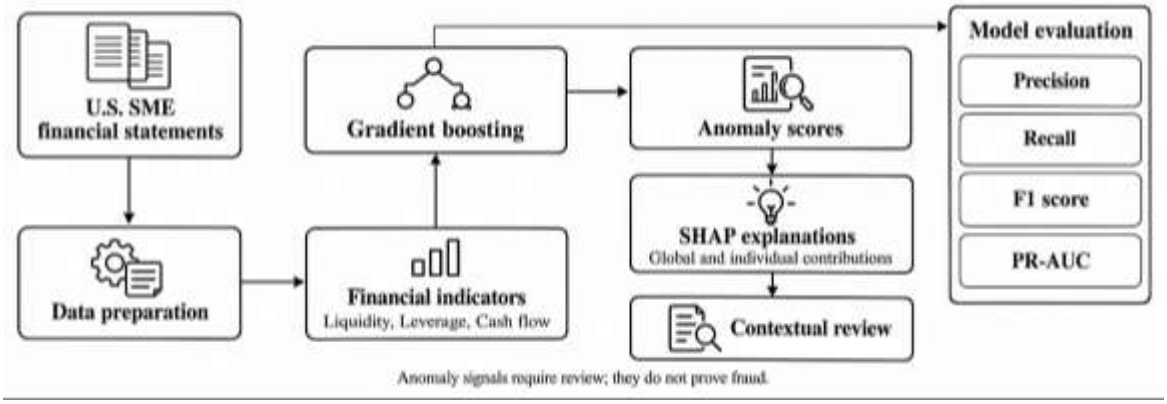
Keywords

SMEs; Financial anomalies; Boost; SHAP; Classification.

INTRODUCTION

Financial statements summarize an organization's financial position, operating performance, and cash flows. A financial-statement anomaly is an unusual or inconsistent pattern in reported amounts or relationships among accounts that warrants review (Khan et al., 2022). It is a screening signal, not proof of error or fraud. Anomalies may arise from legitimate business changes, accounting estimates, recording mistakes, or deliberate misrepresentation. Financial-statement fraud is narrower, involving intentional material misstatement or omission. Reliable reporting matters internationally because investors, lenders, suppliers, employees, regulators, and public institutions use accounting information to allocate capital, assess repayment capacity, and evaluate organizational performance.

Figure 1: Explainable Financial Statement Anomaly Detection



Misstatements can distort assessments of profitability, liquidity, leverage, and solvency, affecting lending terms, investment decisions, commercial relationships, and confidence in markets. These concerns extend beyond large public corporations: small and mid-sized enterprises (SMEs) also use financial statements to secure credit, attract investment, support procurement, meet tax obligations, and guide internal decisions (Zhang et al., 2020). SME reporting contexts vary by sector, ownership, resources, and oversight. A smaller firm may have fewer accounting staff, less separation between bookkeeping and approvals, or greater reliance on external accountants and standardized software. Its financial patterns may differ from those of a public corporation while remaining legitimate. A seasonal enterprise, for example, may show sharp changes in revenue, inventory, or receivables at particular reporting dates. An anomaly can occur in an account balance, ratio, relationship among accounts, or pattern across reporting periods. Common indicators include revenue growth, receivables relative to sales, inventory turnover, gross margins, accruals, leverage, liquidity, and operating cash flow. Rising earnings alongside weak operating cash flow may warrant review, yet acquisitions, growth, changes in credit terms, seasonality, or supply disruptions can produce similar signals. Accordingly, detection depends on a clear definition of the outcome, the observation unit, the available financial data, and the context in which a pattern appears (Martínez-Peláez et al., 2024). Models developed for large listed firms or other countries may not transfer directly to U.S. SMEs because reporting practices, firm scale, sector mix, oversight, and data availability differ. The study treats anomaly scores as review signals requiring contextual interpretation. Because financial statements inform decisions across supply chains and borders, weaknesses in one firm's reporting can affect lenders, customers, workers, and investors far from its operating location. Screening supports focused review across many firms simultaneously.

Traditional financial-statement screening combines judgment, audit procedures, ratio analysis, trend review, internal controls, and financial risk scores. These approaches organize review, but fixed thresholds may not reflect differences across industries, firm sizes, accounting practices, or business models. Manual inspection can also be resource intensive when many firms and account relationships must be examined. Quantitative studies have compared statistical and machine-learning classifiers for financial-statement fraud and anomalies (Ali et al., 2023). Their reported performance varies with outcome definitions, sample construction, predictor availability, class imbalance, and evaluation

procedures. Confirmed cases are uncommon, so accuracy alone can conceal missed anomalies. Assessment can include precision, recall, F1 score, and area under receiver operating characteristic or precision-recall curves, as well as the consequences of false alerts and missed cases. Gradient boosting builds a sequence of decision trees, with each successive tree helping reduce errors made by the existing ensemble. It can represent nonlinear relationships and interactions among financial indicators, making it suitable for structured statement data. Prior research has applied boosting methods, including Boost, to financial-statement fraud classification, while SME research has tested machine-learning classifiers on enterprise financial statements (Bojanić et al., 2020). These findings support evaluating performance in the U.S. SME context rather than assuming results from other populations will apply. Model development requires an explicit anomaly definition, documented predictors, a validation strategy matched to the intended use, and measures that account for the outcome's prevalence. Explainable machine learning adds ways to examine how input variables contribute to predictions. SHAP assigns contribution values to features and supports both dataset-level summaries and explanations of flagged observations. It can identify which financial indicators contribute most to model scores and how they influence classifications. SHAP values describe the fitted model's behavior; they do not establish causation or prove wrongdoing. Correlated variables, model specifications, and reference data can affect attribution, so explanations require financial interpretation. This quantitative study examines whether gradient boosting can classify financial-statement anomalies in U.S. SME data, which indicators contribute most to predictions, and how contributions vary across flagged observations. It evaluates predictive performance and uses SHAP to explain global and individual model outputs (Zhang et al., 2022). The literature base includes broad reviews of machine-learning and data-analytics research, as well as empirical work on SME classification, gradient boosting, and SHAP-based financial fraud explanations in practice.

The objective of this quantitative study is to evaluate the use of gradient boosting and Shapley Additive explanations (SHAP) for detecting financial-statement anomalies among small and mid-sized enterprises in the United States. The study will examine whether a gradient-boosting model can distinguish observations identified as anomalous from those classified as non-anomalous using selected financial-statement indicators. These indicators may include measures of revenue growth, receivables, inventory, profitability, accruals, liquidity, leverage, and operating cash flow, subject to data availability and the study's operational definitions. The analysis will assess model performance using suitable classification measures, including precision, recall, F1 score, and area under the receiver operating characteristic or precision-recall curve, with attention to the consequences of missed anomalies and false-positive alerts. It will also consider the effects of class imbalance and use a validation procedure that evaluates model performance on observations not used for model fitting. A further objective is to use SHAP to identify the financial indicators that contribute most to the model's predictions at both the overall dataset level and the individual observation level. This will help describe which features are associated with higher or lower anomaly scores and whether the model's classifications can be interpreted through identifiable financial patterns. The study will distinguish predictive attribution from causal explanation and will treat flagged observations as cases for further review rather than determinations of error or fraud. By combining quantitative performance evaluation with feature-level explanations, the study aims to provide a focused assessment of the predictive usefulness and interpretability of gradient boosting for financial-statement anomaly detection in U.S. SMEs. The findings will be framed in relation to the selected sample, outcome definition, predictor set, and validation design, thereby clarifying the scope within which the model's results can be understood.

LITERATURE REVIEW

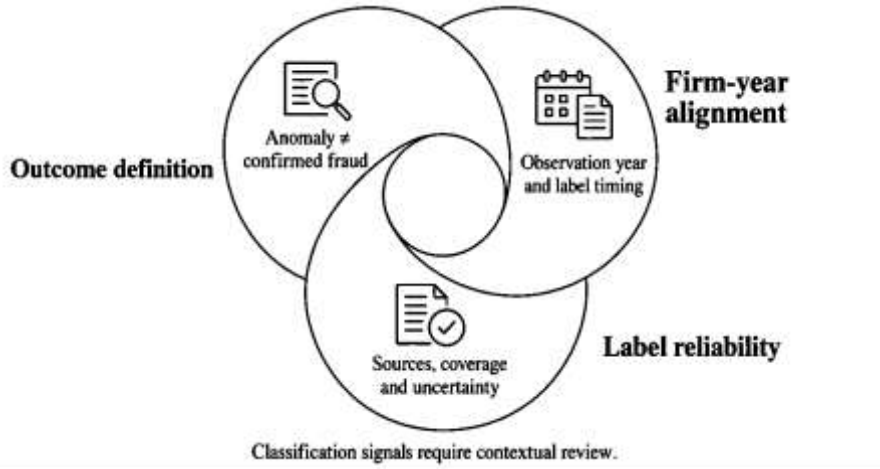
This literature review examines research relevant to explainable machine learning for financial-statement anomaly detection, with particular attention to small and mid-sized enterprises (SMEs), gradient-boosting models, and SHAP explanations. It first distinguishes financial anomalies from confirmed fraud and misstatement, because studies use different definitions, samples, and outcome labels. It then reviews financial indicators used to identify unusual reporting patterns, the performance of statistical and machine-learning classifiers, and challenges related to rare positive cases, class imbalance, and evaluation design. A prior study using financial statements from 341 Turkish SMEs illustrates how machine-learning methods have been applied in an SME setting, while research reviews

identify data imbalance, control-group construction, predictor selection, and evaluation metrics as recurring methodological concerns. The review also considers SHAP as a way to examine global feature contributions and individual predictions. Recent financial-statement studies have paired tree-based models with SHAP, but their results depend on the dataset, target definition, and modeling choices.

Operational Definition and Firm-Year Classification of Financial-Statement Anomalies

Financial-statement anomaly refers to an unusual or inconsistent reporting pattern that merits investigation, whereas accounting error, restatement, material misstatement, earnings manipulation, and confirmed fraud describe different events and levels of evidence. Quantitative studies have operationalized these outcomes in several ways. Some label firm-year observations using regulatory enforcement actions, including cases identified through SEC enforcement releases; others use restatements, amended annual filings, or audit-related indicators (Aditya & Ahmed, 2022; Cepni et al., 2024). A further group applies accounting-based scores to estimate manipulation risk, while studies using SME data classify cases identified through creditor or financial records. These definitions are not interchangeable.

Figure 2: Defining Financial Statement Anomaly Classification



An enforcement label indicates that a case came to regulatory attention, while a restatement can reflect either intentional misstatement or an unintentional error (Asheke & Maksudul, 2022; Nissim, 2022). A high manipulation score marks a risk pattern rather than confirmed wrongdoing. Studies comparing these approaches show that the target definition affects which observations count as positive cases and therefore what a model learns. A binary firm-year outcome is useful for supervised classification because it links each financial record to a defined label, but it compresses differences in severity, duration, and type of irregularity. Research that distinguishes restatement categories or uses multiple labels preserves some of this variation. The literature therefore supports stating precisely whether a study predicts unusual financial patterns, disclosed misstatements, or confirmed fraud, and avoiding language that treats all three as a single outcome.

Firm-year classification is common because annual financial statements provide a consistent observation unit for linking predictors to reporting outcomes. Empirical studies have used annual observations from public companies, cross-country company panels, and SME samples (Asheke & Badrul, 2022; Nguyen et al., 2021). These investigations vary substantially in their sample frames: some construct matched groups of flagged and non-flagged firms, some use broad commercial or regulatory databases, and others draw on creditor records. For example, SME research has analyzed statements from hundreds of enterprises across multiple years, while large-scale company-year studies have combined data spanning multiple jurisdictions and decades. Such differences matter because the proportion of positive cases, sector composition, reporting rules, and firm size affect both the features available and the apparent difficulty of classification. A company-level label may also be applied to an entire fiscal year even when the alleged misstatement affected only a portion of that year (Chen & Han,

2023; Akter & Sumaiya, 2022). Conversely, a record may be labeled only when a regulator or other institution later confirms a case. These decisions create temporal questions: whether predictors are measured before or during the labeled year, whether later information enters the model, and whether the label is assigned retrospectively. Studies that match cases and controls by year or industry seek greater comparability, while firm-grouped or temporal validation tests generalization more realistically. The literature thus treats observation unit, sample construction, period alignment, and label timing as core parts of outcome definition, rather than minor preprocessing choices.

Research also shows that outcome labels have different coverage and reliability. Regulatory enforcement labels are comparatively specific, but they capture detected and pursued cases rather than every irregularity. Restatement databases cover a broader range of reporting corrections, yet restatements can result from error, accounting interpretation, or intentional manipulation. Amended filings and audit opinions provide additional indicators, although neither alone establishes fraud (Debelee, 2023; Imran & Jannatul, 2022). Studies that classify restatements by intent demonstrate why aggregating all corrections into a single positive category can blur meaningful distinctions. Other studies use known-fraud and matched non-fraud samples, which enable direct classification comparisons but may create artificial class proportions unlike those encountered in practice. Research using accounting-based manipulation scores provides scalable screening labels, but these scores are proxies based on financial patterns and thresholds. Their positive cases should therefore be described as high-risk or potentially manipulated rather than confirmed fraud unless independent evidence supports that interpretation (Canedo & Mendes, 2020; Maksudul & Islam, 2022). Across studies, the outcome can also be affected by disclosure delays, enforcement selection, missing records, inconsistent definitions across jurisdictions, and incomplete observation of private firms. These limitations can introduce label noise: a legitimate firm may be labeled anomalous, or an irregular firm may remain in the negative class because it was never detected. Quantitative literature consequently distinguishes observed labels from the underlying event researchers seek to measure, and interprets estimated classification performance in light of the chosen labeling source.

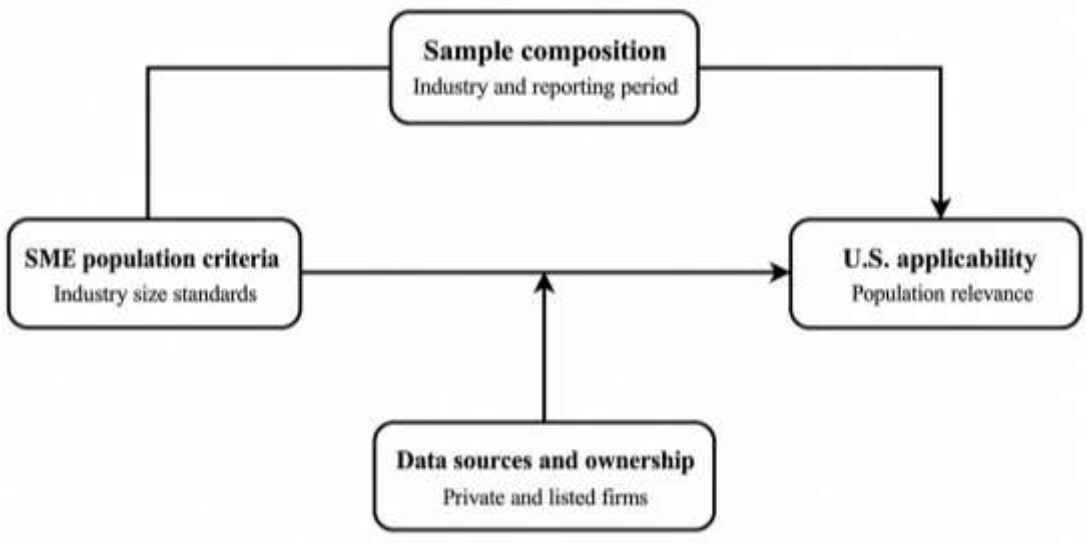
The evidence base includes more than ten empirical and methodological studies using enforcement releases, restatements, amended filings, financial ratios, risk scores, and SME accounting records. Across these studies, a recurring distinction is between detecting an observed label and establishing the true occurrence of misconduct. Machine-learning research typically represents each firm-year as a labeled observation, estimates a classification boundary from financial or textual predictors, and evaluates how well the model separates positive from negative cases (Fang et al., 2020; Morshedul & Islam, 2022). However, performance depends on how the positive class was assembled, whether the control group was matched, the prevalence of labeled observations, and whether testing uses firms or periods excluded from training. Research comparing alternative labels further indicates that models trained on confirmed enforcement cases may emphasize different reporting patterns than models trained on restatements or statistical risk scores. This has direct relevance for a study of U.S. small and mid-sized enterprises, where complete public enforcement histories may not capture all accounting irregularities and firm-level financial data may be less consistently available than for listed companies. The literature therefore supports a narrowly defined binary outcome at the firm-year level, accompanied by a transparent description of the source, observation period, and labeling rule (Pirouz et al., 2020; Stabak & Shahid, 2022). It also supports acknowledging that undetected cases and ambiguous corrections may affect the labels. Under this framing, an anomaly classification is an analytical signal for review, while conclusions about error, manipulation, or fraud depend on the evidence underlying each label.

SME Population Criteria and U.S. Applicability

SME definitions are not uniform across accounting and fraud-detection research, which complicates comparisons among samples. In the United States, the Small Business Administration generally defines small businesses through industry-specific size standards based on employee count or average annual receipts; affiliation with other businesses can also affect classification. Consequently, "small" does not represent one universal employee or revenue threshold across sectors. Studies may instead adopt national statutory definitions, database categories, bank classifications, or researcher-selected cutoffs (Leso et al., 2023; Feroz & Pervaz, 2022). SME studies from Turkey, Albania, and Lithuania use their

own country or study contexts, while U.S. public-company research often samples listed firms without focusing on small enterprises.

Figure 3: SME Population and Sample Applicability



Some studies distinguish micro, small, and medium enterprises; others combine them under the SME label. Asset-based definitions may also appear where balance-sheet size is available, even when employment or receipts are not (Ramos-González et al., 2022; Feroz & Shabbir, 2022). These differences influence which firms enter a sample and what financial volatility should be treated as unusual. A threshold suitable for a manufacturing company may not fit a professional-services firm, and a privately held business may have a different reporting structure from a public corporation. Population definitions should therefore be recorded as study-specific criteria rather than assumed to be directly comparable. For a U.S. study, the applicable industry size standard and available firm-level measures need to be identified clearly so that population membership can be interpreted consistently.

Prior studies show wide variation in sample scale, sector coverage, and observation periods. An SME-focused Turkish study examined 341 enterprises across 2013–2017, with 1,705 firm-year observations and cases drawn from several sectors; another analysis of Turkish SME statements considered manufacturing, construction, transportation, and agriculture. Research on Albanian SMEs covered 247 companies and two reporting years, whereas a Lithuanian case study focused on micro and small enterprises (Islam & Md, 2023; Kgakatsi et al., 2024). These studies provide closer organizational comparisons to the present population than studies restricted to listed corporations, although their national settings differ. Public-company investigations span substantially larger samples: one U.S. machine-learning study examined listed firms and emphasized the rarity of fraud firm-years; a cross-country project assembled more than two million company-year observations across twelve countries; a Mongolian analysis included 969 firm-years from 132 companies; and a Vietnamese listed-company study retained 6,669 observations from 2019–2023. Other research used 40 Taiwanese fraudulent companies with matched comparison firms, 50 Iranian listed companies, and a Chinese dataset of nearly 25,000 firm-year observations. These differences affect statistical power, representation of positive cases, sector diversity, and model-estimate stability (Acevedo-Duque et al., 2021; Asheke, 2023). A large listed-firm panel can support broader model comparisons but may underrepresent private enterprises, while a smaller SME panel may better reflect enterprise size and financing context yet cover fewer years and industries.

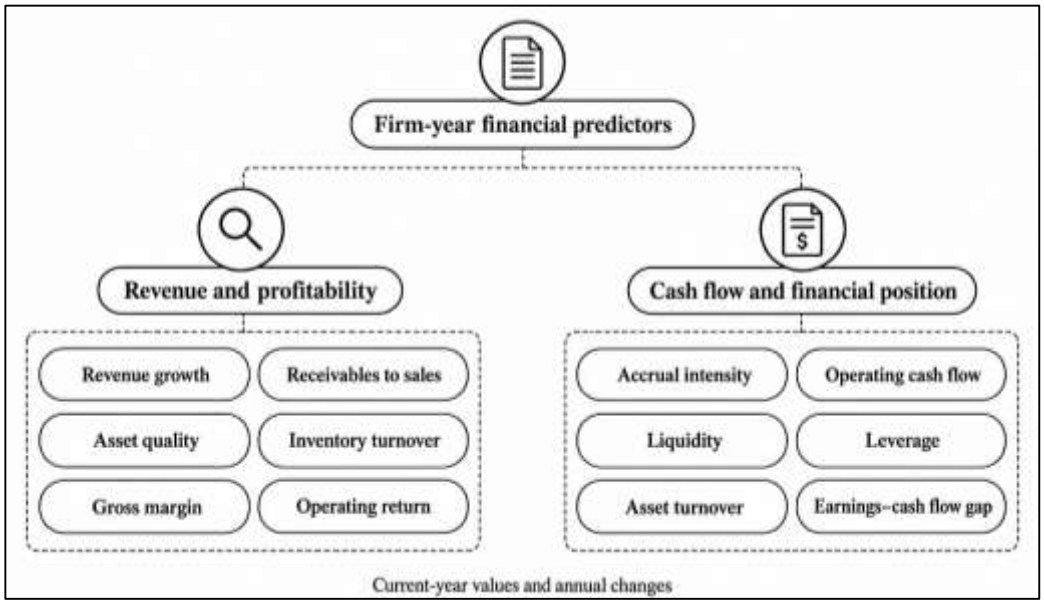
Data provenance and ownership status shape what financial information is observable. SME studies have used records from creditor banks, which may include borrower statements and lending-related classifications unavailable in public filings. Such data can provide insight into firms seeking credit, but the sample may represent only businesses that applied to or maintained a relationship with the

institution. Other SME research uses company filings, case records, or accounting statements assembled for a defined study (Shahid, 2023; Soudi & Bauters, 2024). By contrast, public-corporation studies often draw on stock-exchange filings, annual reports, commercial financial databases, regulatory enforcement releases, restatement databases, and amended filings. These sources support larger firm-year panels and links to regulatory outcomes, although they commonly center on listed companies with comparatively standardized disclosure. Some multinational studies combine datasets from several jurisdictions, but differences in filing rules, enforcement practices, and accounting standards remain relevant. Samples based on regulatory cases may overrepresent firms that were detected and pursued, while commercial databases may have uneven coverage of private businesses. Bank-based samples may reflect credit screening and lending relationships as much as the full population of SMEs. Ownership type, data source, and inclusion pathway therefore affect both representativeness and outcome measurement. Comparing studies requires documenting whether firms are privately held or listed, how statements were obtained, which reporting years are covered, and whether observations represent all eligible businesses or a selected group (Islam & Ahmed, 2023; Phonthanukitithaworn et al., 2023).

Financial-Ratio Predictors at Firm-Year Level

Financial-ratio research commonly represents revenue and growth through current-period revenue, changes in sales across consecutive fiscal years, and sales growth relative to receivables or assets. Receivables measures include the balance relative to sales and its year-over-year change; these indicators are used to identify growth in reported revenue that is not matched by collections. Asset-quality indicators describe the share of assets with less direct cash realization, while inventory growth or turnover can indicate unusual accumulation or valuation concerns. Profitability and margin measures include gross margin, operating return, and earnings relative to assets or sales (Gutierrez et al., 2020; Sadia & Mosheur, 2023).

Figure 4: Financial Ratios for Anomaly Screening



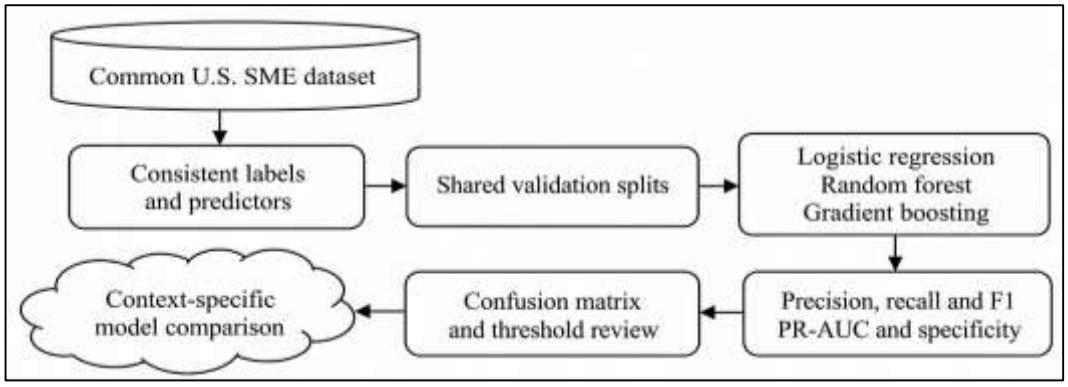
Deteriorating margins alongside rapid sales expansion have been treated as patterns that may warrant review, although business conditions can create similar movements. Studies using established manipulation scores, raw accounting variables, and selected ratio sets differ in whether they represent levels, annual changes, or both. The observation period is usually the fiscal year associated with the outcome label, with some predictors comparing that year to the prior year. This timing matters because a contemporaneous value may capture conditions after irregularity began, while a lagged or change measure can represent a developing pattern. Empirical work across U.S. public firms, Malaysian and Indonesian listed companies, Turkish SMEs, and cross-country firm-year datasets repeatedly examines

revenue, receivables, asset composition, margins, and growth (Adnan, 2024; Li et al., 2021).
The literature supports defining each feature in words, specifying the reporting year and comparison period, and interpreting its predictive association as a screening signal rather than a direct indication of manipulation.Accrual and cash-flow indicators assess whether reported earnings are supported by cash generated through operations. Studies operationalize accruals through differences between accounting earnings and operating cash flow, often scaled against assets, and examine whether changes in accrual intensity coincide with revenue growth or profitability (Islam, 2024; Ogachi et al., 2020). Cash-flow measures include operating cash flow relative to assets, earnings, or sales, alongside the gap between earnings and operating cash. Receivables and inventory changes relative to operating cash flow have also been examined as signs of earnings that may not convert into cash. Liquidity predictors describe the firm's ability to meet short-term obligations and commonly use current resources relative to short-term liabilities; leverage measures relate debt or liabilities to assets or equity. These indicators capture financing pressure, solvency, and exposure to covenant or repayment constraints, but their direction can vary across industries and economic conditions. Operating-efficiency measures include asset turnover, inventory turnover, and changes in expenses relative to sales, which can reveal unusually weak conversion of assets into revenue or changes in cost structure. Research using Beneish and Dechow-style indicators, broader accounting ratios, and machine-learning feature sets draws on these groups in varying combinations (Abdelkader & Wahba, 2024; Ashik, 2024). Most measures are calculated for the labeled fiscal year, while index or change variables compare current and preceding years. Their value for anomaly classification rests on joint patterns: no single liquidity, leverage, accrual, or turnover measure reliably distinguishes misconduct from ordinary business variation.

Performance of Logistic Regression and Boosting

Comparative financial-statement classification studies evaluate logistic regression alongside decision trees, random forests, support-vector machines, neural networks, and ensemble methods. Logistic regression commonly serves as an interpretable statistical baseline, while tree ensembles and neural models represent nonlinear relationships and interactions (Kynatun & Saba, 2024; Song et al., 2021).

Figure 5: Comparing Financial Anomaly Classification Models



Studies differ in sample source and size: U.S. public-company research has used enforcement-linked fraud labels and theory-selected accounting variables; a Turkish SME study analyzed 341 firms over 2013–2017 and compared seven classifiers; a restatement study examined 3,513 cases from 2001–2014; and a Mongolian study compared models on 969 firm-year observations from 132 companies. Other evidence includes listed-company samples from Taiwan, China, Vietnam, and the MENA region, as well as cross-country datasets. Target definitions also vary among confirmed fraud, restatement, material misstatement, and score-based risk. Predictor sets range from raw account values to financial ratios and combinations of financial, governance, audit, or narrative variables (Maksudul, 2024; Sahin, 2020). Validation designs include matched case-control comparisons, held-out test sets, cross-validation, and temporal prediction. Because these studies ask related but distinct classification questions, their model comparisons are most informative when the sample, label, predictor set, and

validation procedure are read together. These comparisons encompass more than ten empirical studies across multiple national and methodological contexts.Reported metrics show that the best classifier depends on both the dataset and the performance measure. In the Turkish SME comparison, artificial neural networks, random forests, logistic regression, support-vector machines, and other methods were evaluated; reported leading results differed according to feature-selection and oversampling settings (Kavzoglu & Teke, 2022b; Shahid, 2024). In a Mongolian firm-year comparison, the stacking model achieved the highest average precision, with Boost and gradient boosting also performing strongly, while logistic regression, random forest, and SVM ranked lower on that measure. A MENA study reported Boost outperforming logistic regression, decision trees, SVM, AdaBoost, and random forest under its experimental settings. Another listed-company study reported a random-forest test result with high precision but substantially lower recall, illustrating that a classifier can identify few positive cases accurately while missing many others. Studies report accuracy, precision, recall, F1, specificity, ROC-AUC, and, increasingly, precision-recall measures, but often not every metric or the classification threshold (Chen & Chen, 2020; Islam, 2024). High accuracy can coexist with poor minority-class recall when fraudulent firm-years are rare. Some papers also report balanced accuracy or cost-sensitive error under alternative class ratios, further complicating comparisons between published findings. The literature therefore gives greater interpretive value to complete metric profiles and confusion-matrix results than to a single headline accuracy figure.

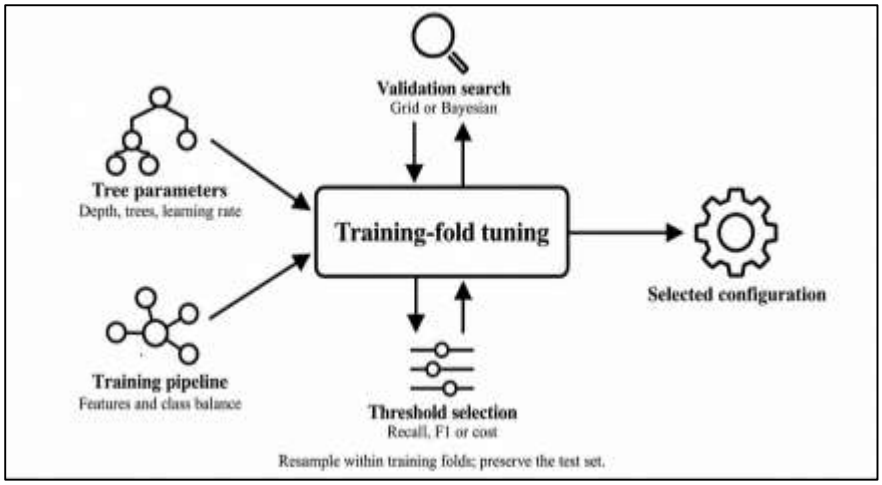
Gradient-boosting results are promising but not uniformly superior. Boost outperformed several conventional classifiers in a MENA financial-statement sample and performed strongly in recent comparisons using financial ratios, yet other settings show competitive results from random forests, neural networks, or ensembles (Cha et al., 2021; Mainuddin, 2024). A Turkish SME investigation found that the apparent leading model depended on whether feature selection and oversampling were applied. A Mongolian comparison found differences among Boost, standard gradient boosting, logistic regression, SVM, and random forest when average precision and precision-recall curves were considered. U.S. public-firm research also compared machine-learning methods using raw accounting numbers against benchmarks based on ratios, showing that predictor representation can influence the relative result. Studies of Taiwanese listed firms, Chinese listed firms, Vietnamese listed companies, and annual-report text models further demonstrate that model rankings shift across targets and predictor combinations. Positive-class prevalence is another key condition: balanced case-control samples, highly imbalanced enforcement samples, and resampled training sets create different classification tasks (Sumaiya, 2024; Sun et al., 2021). Decision thresholds also affect precision, recall, F1, and specificity; AUC measures compare performance across thresholds, while threshold-specific metrics describe one operating point. Thus, a model's reported advantage is conditional on a particular experiment rather than a fixed property of the algorithm, with differing sample sizes.

Direct ranking across studies is limited by differences in sample construction, outcome labels, class proportions, predictors, validation, and threshold choice. A model trained to detect regulatory enforcement cases is not solving the same task as one trained on restatements or a financial-risk score (Shanmugasundar et al., 2021; Taher et al., 2024). Matched samples can improve comparability between positive and negative firms but may not reflect the low prevalence found in routine screening. Random splits may also produce more favorable estimates than firm-grouped or time-based validation when observations from the same firm or reporting period appear in both training and test data. Accuracy and ROC-AUC can obscure missed rare cases, whereas precision-recall AUC, recall, precision, F1, and specificity reveal different trade-offs. A threshold chosen to reduce false negatives may increase false alerts, so studies need to report the threshold-selection method alongside the confusion matrix. Published comparisons also vary in whether oversampling occurs only within training folds, whether hyperparameters are tuned consistently, and whether confidence intervals or statistical tests accompany model differences (Ashik, 2025; Nhu et al., 2020). The quantitative literature therefore supports interpreting results within each study's design and comparing algorithms most directly when they share the same observations, labels, predictors, splits, and metric definitions. For the present research, logistic regression and random forest provide useful baselines for assessing gradient boosting on a common U.S. SME dataset and evaluation protocol.

Gradient-Boosting Specification and Hyperparameter Selection

Financial-statement studies configure boosted-tree models by controlling tree complexity and the contribution of each boosting iteration. Common settings include maximum tree depth, number of trees or boosting rounds, learning rate, row subsampling, feature subsampling, and regularization. These choices determine how finely the model partitions firm-year observations, how strongly each tree changes the ensemble, and how much sampling variation is introduced during fitting (Abbas et al., 2023). A MENA financial-statement study optimized XGBoost after cross-validation and used SMOTE to address class imbalance; another study of Indonesian public companies used a genetic algorithm to search learning rate, depth, subsampling, and class-weight parameters. Other research has reported grid searches over learning rate, depth, estimator count, row sampling, and column sampling, while comparative work has evaluated gradient boosting, XGBoost, LightGBM, and random forests under explicit parameter grids. Studies on Vietnamese material misstatements and Mongolian firm-year fraud classification have also paired boosted trees with resampling or class-weighted learning. The ranges and selected values differ by dataset and target, so a parameter setting that works in one study cannot be assumed to generalize to another. Tree depth and estimator count are especially consequential because high-complexity models can fit rare positive cases while overfitting noise (Alhakeem et al., 2022; Kynatun, 2025). Learning rate and regularization may constrain that behavior, while subsampling can affect model diversity. The literature generally treats these parameters as jointly selected components rather than isolated settings.

Figure 6: Gradient Boosting Specification and Tuning



Feature selection and parameter tuning are often conducted together, although studies differ in how they separate these steps from evaluation. Financial-statement research has used filter methods based on correlation or information value, embedded rankings from tree models, and wrapper or combined selection procedures. Some studies compare models using all available predictors against reduced feature sets; others first screen financial indicators and then tune XGBoost or related classifiers (Kavzoglu & Teke, 2022a; Mostari, 2025). Search strategies include grid search, cross-validation-based selection, genetic algorithms, and Bayesian optimization. For example, an optimized Boost study in the MENA region tuned the model after cross-validation, while Indonesian financial-statement research used a genetic algorithm for parameter selection. More recent Vietnamese work has combined Bayesian tuning with stratified cross-validation, and public-company fraud studies have compared multiple feature-selection strategies across logistic regression, random forest, and boosting (Masrul, 2025; Singh et al., 2024). The evidence highlights the need to conduct selection and tuning using training data only, with an inner validation procedure, and to reserve an untouched test set for final assessment. Otherwise, information from test observations can influence selected features or parameter values and inflate apparent performance. Decision thresholds also require a documented selection rule. Some studies retain the default threshold, whereas others optimize for recall, F1, cost, or a chosen sensitivity

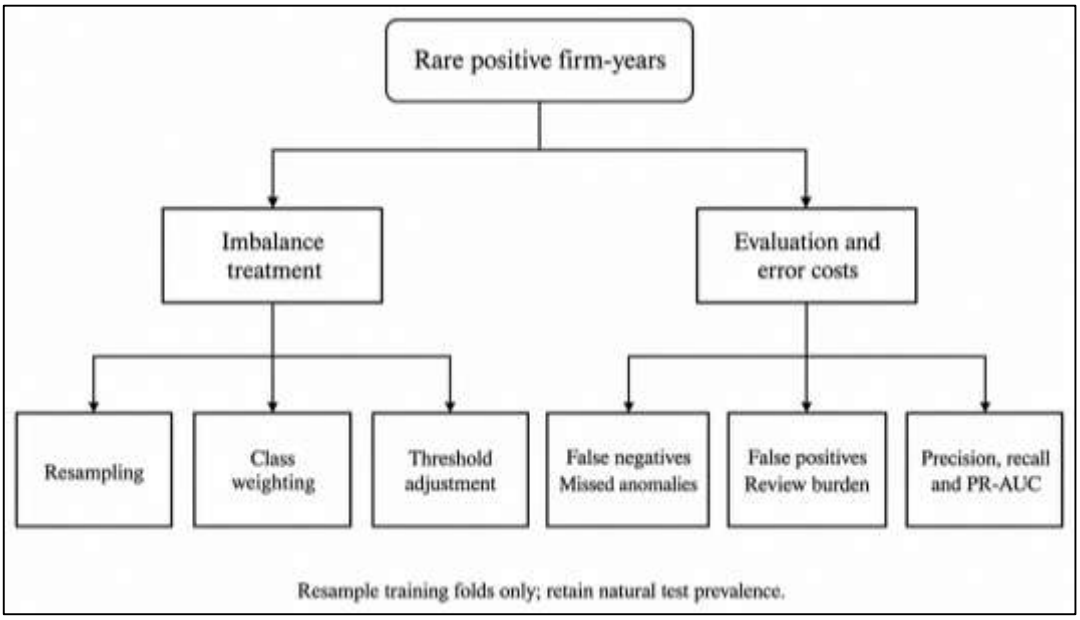
level. Because each rule produces a different balance of false alerts and missed cases, threshold selection belongs within the training and validation process and should be reported with the final results. Class imbalance is addressed through data resampling, algorithm-level weighting, threshold adjustment, or combinations of these approaches. SMOTE creates additional minority-class training examples, while random under sampling reduces the majority class; class weights or positive-class scaling alter the penalty for errors on rare fraud observations without changing the observed records (Maksudul, 2025; Nayak et al., 2021).

SME, listed-company, and cross-country studies have used oversampling and under sampling, and recent work has compared SMOTE with class-weighted learning for financial-statement classification. Some boosted-tree implementations include a parameter that increases the contribution of positive-class errors, while cost-sensitive ensembles assign different consequences to missed fraud and false alarms. These procedures can improve minority-class recall, yet they may also alter precision, probability calibration, and the operating threshold. The evaluation sample should therefore retain its natural class distribution when the goal is to estimate performance in the target population. Resampling the entire dataset before splitting can allow duplicated or synthetic examples related to test records to enter training, producing optimistic estimates (Shahriar, 2025; Otchere et al., 2022). More reliable studies apply sampling only within each training fold and keep validation and test records untouched. Comparisons between resampling and weighting are most informative when they use identical firm-level or temporal splits, the same predictors, and consistent evaluation metrics. Reports should state the original class proportions, the resampling method, when it was applied, the resulting training distribution, and whether the decision threshold was adjusted afterward.

Class Imbalance and Error Costs

Across financial-statement studies, the positive class is defined according to the event or proxy selected for analysis. Common definitions include a firm-year linked to a regulatory enforcement action, a disclosed financial-statement fraud case, a restatement or material misstatement, and an observation exceeding an accounting-based risk threshold. These targets represent different levels of confirmation and should not be treated as equivalent (Farhadpour et al., 2024; Shahid, 2025). Prevalence also varies markedly with sampling. A Turkish SME dataset included 321 fraudulent and 1,384 nonfraudulent observations from 341 enterprises; a U.S. listed-firm benchmark has reported fraud firm-years below one percent, while a recent Mongolian dataset contained 208 fraud observations among 846 usable cases. Some studies create balanced case-control samples or matched groups, which are useful for model comparison but do not preserve the natural frequency of anomalies in routine screening. Other studies retain imbalanced samples or report separate prevalence by country, period, and label source. These design choices influence the number and diversity of positive examples available to a classifier and the baseline probability that a flagged observation is truly anomalous. The literature therefore describes the positive-class rule and its prevalence within the analytic sample, distinguishes sample composition from expected deployment prevalence, and clarifies whether class counts refer to firms, statements, or firm-year observations (Islam, 2025; Wang et al., 2024). Studies address imbalance through resampling, algorithm-level weighting, cost-sensitive learning, or threshold adjustment. Oversampling increases representation of the positive class, including synthetic approaches such as SMOTE and ADASYN; under sampling reduces the number of negative observations; combined procedures use both (Cabot & Ross, 2023; Mainuddin, 2025). Turkish SME studies compared SMOTE oversampling and under sampling, while other financial-statement models used synthetic or random resampling before fitting classifiers.

Figure 7: Class Imbalance and Error Costs



Class weighting retains the observed records but penalizes errors on positive observations more heavily, and cost-sensitive methods can assign distinct consequences to false negatives and false positives. Some studies compare weighting with resampling because the methods can produce different precision and recall profiles. Evidence from financial-statement datasets also shows that method selection may vary by model and sample: a procedure that raises recall in one classifier can lower precision or alter calibration in another. Resampling should be conducted only within the training data or training folds (Allison, 2023; Sadia, 2025). Applying it before separating the test set can place synthetic or duplicated information related to test observations into model training and overstate performance. The evaluation set should ordinarily retain the original class distribution when the goal is to estimate screening performance under real prevalence. Reports that identify the original and resampled class counts, procedure, and stage of application make results more interpretable.

The costs of classification errors are asymmetric in financial-statement screening. A false negative occurs when a genuinely anomalous or misstated observation is classified as ordinary; this can leave investors, lenders, auditors, or regulators unaware of a material reporting risk. A false positive occurs when a legitimate firm-year is flagged, increasing review workload, delaying decisions, or imposing reputational and compliance costs (Garcia, 2022; Shabbir, 2025). Studies of fraud classification have explicitly varied misclassification-cost assumptions, while annual-report detection research has found that different classifiers can perform better on identifying positive cases versus correctly clearing negative cases. The appropriate balance depends on the review setting: an initial screening system may tolerate more false positives to identify a larger share of possible cases, whereas a resource-constrained review team may prioritize precision. Class-weighted learning, resampling, and threshold changes influence this trade-off and should be assessed using more than overall accuracy. Several studies report sensitivity or recall for the positive class alongside precision, specificity, and cost-oriented outcomes. Research that distinguishes intentional fraud from unintentional restatement also matters because consequences and follow-up procedures may differ by label type (Groccia et al., 2023; Sadmeen, 2025). Accordingly, error costs should be interpreted in relation to the outcome definition, event severity, capacity for follow-up review, and number of firms the screening process can realistically examine.

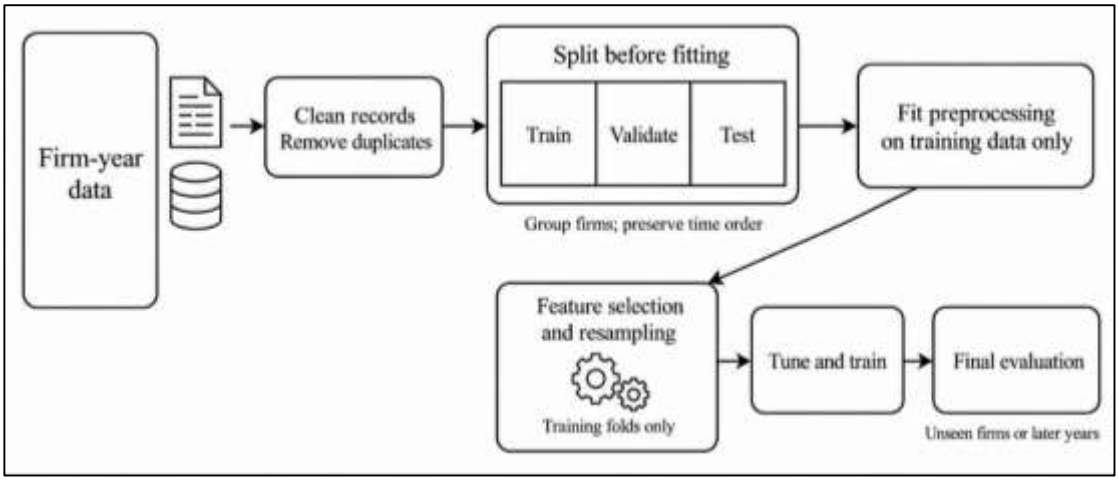
Overall accuracy can be misleading when anomalies are rare because a classifier may label nearly every observation as non-anomalous and still achieve a high percentage of correct classifications. Such a model may have little practical value if it misses most positive firm-years. Minority-class measures show different aspects of detection: recall indicates how many labeled anomalies are identified, precision indicates how many flagged observations are positive, and F1 summarizes the balance between these two measures (Ashik, 2026; Marra, 2024). Specificity describes correct classification of

non-anomalous cases, while ROC-AUC and precision-recall AUC summarize discrimination across decision thresholds. Precision-recall measures are especially informative when positive cases are uncommon, and confusion matrices reveal the actual false-positive and false-negative counts at the selected threshold. Studies of public U.S. firms, Turkish SMEs, Mongolian companies, cross-country samples, restatements, and cost-sensitive models use varying combinations of these metrics and class treatments. Their performance values cannot be compared without considering target definitions, prevalence, matching, resampling, and validation. AUC values may appear stable even when threshold-specific recall or precision changes, and balanced test samples can produce precision estimates that do not reflect deployment prevalence (Hasib, 2026; Maxwell et al., 2021). The literature therefore interprets performance as a profile of detection and error trade-offs, rather than as a single accuracy ranking.

Validation Design and Generalization Across Reporting Periods

Financial-statement studies use random holdout and cross-validation to estimate performance, but these procedures answer different questions. A random holdout divides observations into training and test portions without ordering them by reporting year, while k-fold cross-validation repeats training and evaluation across several partitions. Some research on auditor-opinion proxies applies ten-fold cross-validation and nests hyperparameter tuning within the folds; other analyses use a single 70:30 or 80:20 holdout (Aliferis & Simon, 2024; Kynatun, 2026). These designs measure performance on observations drawn from a similar sample distribution and can support model selection when observations are independent. Repeated annual records complicate that assumption because the same firm may contribute several firm-years. If one firm's earlier record enters training and a later record enters testing, the model may recognize stable firm-specific patterns rather than generalize to a new enterprise. This is especially relevant where financial ratios, governance indicators, or persistent reporting practices vary more across firms than within firms. Some public-company work uses company-level grouped splits, while other studies use random observation splits or matched samples. Without identifying the splitting unit, a reported test score is ambiguous about whether it applies to unseen observations from known firms or entirely new firms (Al-Shehari & Alsowail, 2021; Mostari, 2026). Validation reporting therefore needs to state whether splitting occurred by row, company, or reporting period and whether repeat observations from the same firm were kept together.

Figure 8: Validation Design and Leakage Prevention



Temporal validation is designed to represent prediction in later reporting periods. In a chronological split, earlier firm-years are used for model development and later years are held out for testing; time-aware cross-validation similarly preserves ordering. Examples include Chinese listed-company research using earlier years for training, a middle interval for validation, and later years for testing, and another study separating a development period from an out-of-time test period. Such designs test whether patterns learned in historical reports remain useful as accounting practices, industry conditions, enforcement activity, and economic environments change (Hasinur, 2026; Zusman et al.,

2020). Random splits can place future-period information in training while earlier observations appear in testing, making performance less representative of deployment. Temporal splits may produce lower scores because the test data include distribution shifts, but that result better reflects the challenge of applying a model to subsequent years. Some studies also reserve recent cases or a later firm period for external validation after fitting on historical data (Carion et al., 2020; Maksudul, 2026). Other work reports only randomized holdouts or cross-validation, which assesses interpolation within the sampled period rather than performance over time. Interpreting a result therefore requires noting the calendar years in each partition, the label's event year, the predictor reporting date, and whether the test observations come from later years or merely a random subset.

Leakage can enter validation through preprocessing as well as through the split itself. Imputation values, scaling parameters, feature-selection rankings, and hyperparameters should be estimated using training observations only and then applied to validation and test data (Bertomeu, 2020; Mohsin, 2026). Resampling methods such as SMOTE or under sampling should also operate only within training folds; otherwise, synthetic examples or duplicated observations can be related to data later used for testing. Recent financial-statement studies explicitly apply oversampling only after the split, keep test prevalence unchanged, and conduct selection or tuning inside training folds. By contrast, studies that balance the full dataset before splitting risk allowing information from the eventual test set to influence model fitting. Tuning on the test set creates a similar problem when repeated model adjustments are made after seeing test results. Leakage can also arise from using corrected restated figures to predict a fraud label defined by the original misstatement, or from including variables published only after the prediction date (Shahriar, 2026; Singh et al., 2021). Studies that construct firm-year labels from later enforcement information need to align predictors to what would have been available at the intended prediction time. A transparent pipeline should document when each transformation is fit, when labels become observable, and how nested model selection is separated from final evaluation.

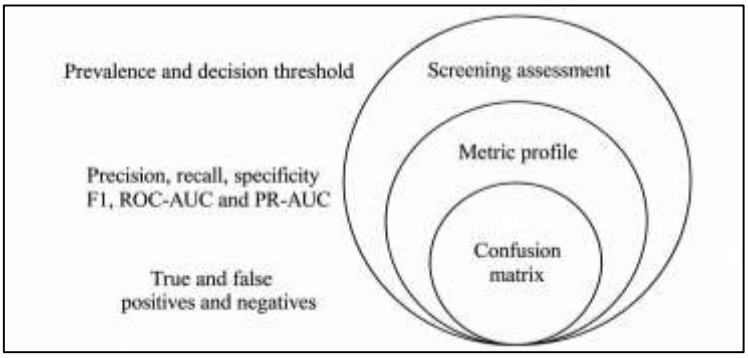
The validation literature encompasses more than ten financial-statement and related accounting studies using randomized holdouts, repeated cross-validation, company-grouped partitions, and time-based evaluation. Their scores cannot be interpreted as equivalent evidence of generalization. A random holdout may test new rows from firms already represented in training; a firm-grouped split tests new companies; a temporal split tests future reporting periods; and an external dataset tests transport to another source or setting. Some studies combine these designs, while others report a single split without specifying whether firms repeat across partitions (Shahid, 2026a; Zhou et al., 2020). Dataset size, fraud prevalence, target definition, and sample matching further affect estimated performance. A study of thousands of annual reports with resampling and repeated model runs estimates a different operational task from a firm-year panel evaluated on later years, while a bank-based SME sample may not represent all private businesses. Model selection requires validation data, but the final test data should remain unused until the specification is fixed. Stronger reporting identifies split unit, year ranges, class proportions, firm overlap, preprocessing order, resampling scope, tuning procedure, and any external test. These details clarify whether performance reflects interpolation among sampled observations, transfer to unseen firms, or prediction across reporting periods. Those distinctions determine the study's generalization claim (Bron et al., 2021; Shahid, 2026b).

Detection Metrics and Probability Calibration

Financial-statement classification studies interpret results through a confusion matrix, which compares predicted labels with observed labels. A true positive is an anomalous firm-year correctly flagged by the model, and a false positive is a non-anomalous firm-year incorrectly flagged. A true negative is a non-anomalous observation correctly left unflagged, while a false negative is an anomalous observation the model misses (Kim & Lee, 2020; Mainuddin, 2026). These four counts clarify the types of decisions behind summary measures. Studies have applied this framework to different targets, including regulatory fraud labels, restatements, material misstatements, and accounting-score risk categories. The positive class can be rare: a U.S. firm-year dataset of more than 90,000 observations contained a fraud rate below one percent, while Turkish SME and Mongolian company-year studies used substantially different positive-class proportions. This variation affects how many true positives and false positives arise at a given decision threshold. A confusion matrix should therefore be interpreted alongside the sample prevalence, label definition, and threshold used. Some studies also

distinguish firm-level from firm-year classifications or evaluate repeated reporting periods, which changes how counts are aggregated. Across more than ten studies of U.S. public firms, SMEs, listed companies, restatements, and cross-country fraud samples, the confusion matrix provides the basic accounting of detected cases and errors (Kragh & Karstoft, 2021; Sadia, 2026). It supports direct examination of review burden and missed cases that a single accuracy value cannot show.

Figure 9: Financial Anomaly Detection Assessment Framework



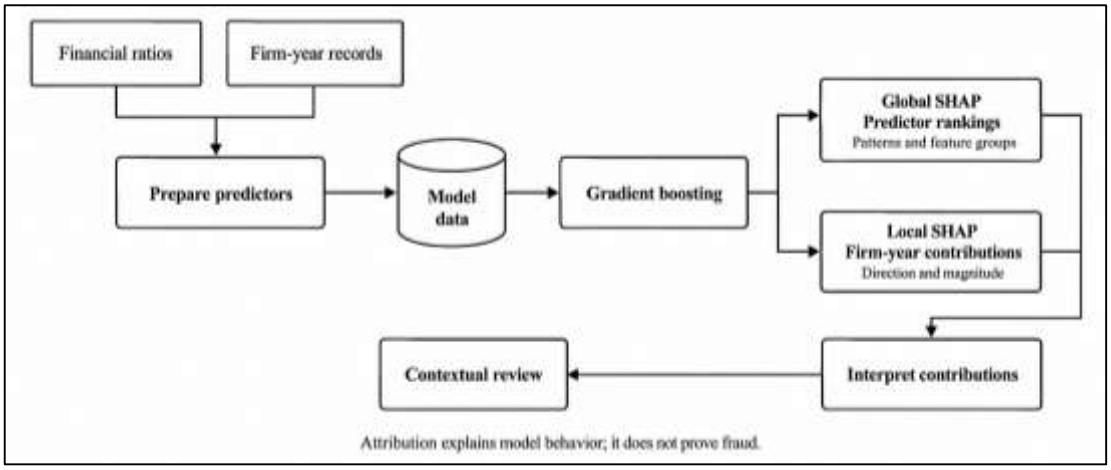
The literature uses several complementary measures because no single metric captures every aspect of screening quality. Precision describes how many flagged observations are truly positive, making it relevant to analyst workload and alert credibility (Chatterjee et al., 2020; Shabbir, 2026). Recall describes the share of positive observations detected and is important when missing a reporting irregularity is costly. Specificity measures the share of negative observations correctly left unflagged, reflecting the model's ability to avoid unnecessary reviews. F1 combines precision and recall into one summary, though it does not include true negatives and depends on the chosen threshold. ROC-AUC evaluates how well a model ranks positive observations above negative ones across possible thresholds, while precision-recall AUC emphasizes the trade-off between precision and recall for the positive class. Financial-statement studies variously report accuracy, precision, recall, specificity, F1, ROC-AUC, average precision, or PR-AUC. A systematic review of financial-statement fraud research identifies metric selection and class imbalance as recurring methodological concerns (Amitrano et al., 2024; Sumaiya, 2026). A study using 90,006 U.S. firm-years compared resampling methods with recall, precision, F2, and PR-AUC; other work on Mongolian firm-years and annual reports reported average precision, ROC-AUC, or macro-averaged measures. These examples show why studies should present a metric profile rather than select only the measure on which a model performs best.

Global and Observation-Level SHAP Explanations of Model Predictions

Global SHAP summaries describe how a fitted model uses predictors across a dataset. Mean absolute SHAP values aggregate the magnitude of each variable's contribution across observations, allowing financial predictors to be ranked by their overall influence on model output. Studies applying tree-based classifiers to financial-statement fraud have used global bar charts and beeswarm plots to highlight indicators such as leverage, profitability, liquidity, revenue-related measures, governance variables, and narrative features (Białek et al., 2022; Sadmeen, 2026). Research using XGBoost and related ensembles has reported global rankings alongside feature-selection or permutation-based measures, while temporal-spatial studies have organized predictors by financial, governance, and time-related domains. A global ranking can help identify which variables most often affect anomaly scores, but it does not show whether a high or low value raises the score, nor does it indicate whether the variable matters equally across firms. Beeswarm summaries add direction and spread by showing the distribution of individual contributions across observations. Some research compares SHAP rankings with intrinsic tree importance or permutation importance and finds that the order of predictors can differ. Global summaries therefore describe model reliance at the sample level and should be interpreted with the target definition, feature scale, sample composition, and model specification in view (Matthews & Hartman, 2021; Umme, 2026). They do not establish that the highest-ranked financial

ratio is the underlying cause of an anomalous report.

Figure 10: SHAP Explanations for Financial Predictions



Local SHAP explanations describe feature contributions for one firm-year prediction. For an individual observation, the explanation identifies which measured variables move the model output above or below its reference prediction and by how much, given the fitted model and chosen background data. Financial-statement studies use local explanations to examine specific flagged firms and inspect whether anomaly scores reflect combinations of changes in revenue, receivables, accruals, leverage, profitability, liquidity, or other predictors (Dhiman et al., 2024). This can help reviewers connect a score to the underlying statement profile and compare the explanation with supporting documents or business context. A positive SHAP contribution means that the feature pushes the model output toward the positive class relative to the baseline; a negative contribution pushes it away, with magnitude indicating the size of that modeled contribution. It does not mean that the feature independently caused fraud or that the firm committed wrongdoing. Research on user-centered fraud detection presents local and global explanations for external stakeholders, while accounting anomaly frameworks use case-level explanations to support human review (Schlegel et al., 2021). Local explanations can also differ across two firms with similar overall scores because different combinations of variables may lead to the same prediction. Their interpretation should therefore remain specific to the observation, model, outcome label, and data available at prediction time.

SHAP dependence plots extend global summaries by relating observed feature values to their contribution values. They can reveal nonlinear patterns, threshold-like changes, or regions where a financial indicator has different associations with model output. Coloring or stratifying observations by a second feature can suggest interactions, such as a revenue-related contribution varying with receivables growth or a leverage contribution changing with profitability (Gillies et al., 2020). Research on financial fraud has used single-feature dependence plots, interaction plots, and temporal-spatial groupings to examine how accounting, governance, and narrative variables combine in model predictions. Grouped-feature explanations aggregate related variables, such as liquidity ratios or revenue-quality indicators, to reduce fragmentation when several correlated measures represent similar financial conditions. Group SHAP and hierarchical approaches have also been proposed to make explanation more stable and reduce redundancy among related predictors. These methods can make dense feature sets more readable, but the grouping rule affects the result: a broad accounting category may conceal distinct behaviors, while narrow groups may preserve overlapping contributions (Gutiérrez-Gallego et al., 2024). Dependence and grouped-feature results are descriptive of the fitted model and dataset. They can help identify patterns that merit accounting interpretation, but they do not by themselves establish a universal relationship or validate a particular causal mechanism in particular. The literature distinguishes SHAP attribution from causal inference, audit evidence, and a formal finding of fraud. SHAP allocates a model prediction among input features according to a specified explanation procedure; it describes how the model arrived at its score, not why the firm's

financial statements took a particular form (Seihani et al., 2024). Correlated predictors create an additional challenge because related variables can share or redistribute attribution, making individual rankings sensitive to the model and reference data. Researchers have compared SHAP with permutation importance, intrinsic feature rankings, or alternative explainers to assess whether key signals remain consistent. Explanation stability can also be evaluated across bootstrap samples, random seeds, model refits, time periods, and background datasets. Studies under distribution shift have shown that a stable SHAP ranking may not always match a feature's measured contribution to predictive performance, so predictive accuracy and explanatory reliability should be considered separately. Other research reports stability indices or rank-correlation measures across repeated explanations. For financial-statement analysis, global and local results should be accompanied by a description of the model, feature definitions, reference population, and explanation settings (Prager et al., 2022). Correlated ratios should be interpreted jointly where appropriate, and a high contribution should prompt review rather than serve as proof of manipulation. This framing keeps SHAP useful as a transparent account of model behavior without overstating what feature attribution can establish.

METHODS

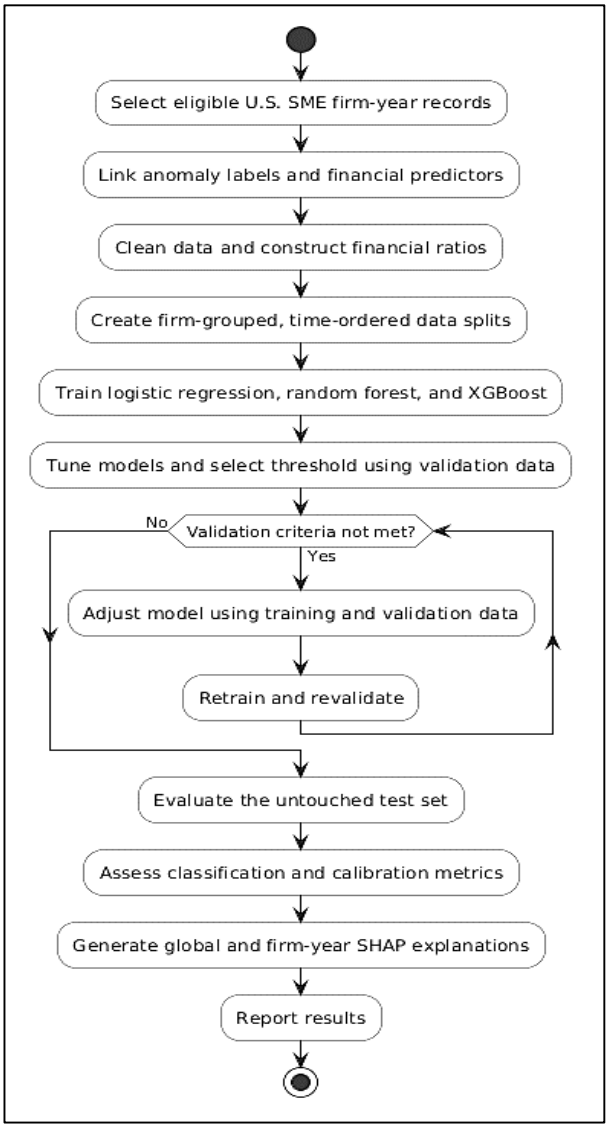
Study Design and Theoretical Framework. The study used a quantitative, retrospective archival design with a longitudinal firm-year structure. It evaluated whether gradient-boosting models classified financial-statement anomalies in U.S. small and mid-sized enterprises (SMEs). The framework treated accounting indicators as predictors of a defined anomaly label. Because the data were observational, the analysis assessed predictive performance and model explanations rather than causal effects.

Participants, Materials, and Sampling. The analytic materials comprised firm-year financial statements from U.S. enterprises meeting the study's stated SME criteria during [study period]. An archival, criterion-based sampling strategy was used to select eligible records from [financial database or source]. Where available, small-business size was classified using the applicable industry criteria and employee or annual-receipt measures. Firm-year records were included when they had a valid identifier, reporting year, outcome label, and sufficient financial data for analysis. Duplicate records, observations outside the population or period, and records without a classifiable outcome were excluded. The final sample consisted of [number] firms and [number] firm-year observations, including [number] anomalous and [number] non-anomalous records.

Instrumentation and Data Collection. Financial-statement variables were extracted from [database or source] and linked by firm identifier and fiscal year to anomaly labels from [label source]. Predictors covered revenue and growth, receivables and asset quality, profitability, accruals and cash flows, liquidity, leverage, and operating efficiency. Data extraction and model analysis were conducted in Python using pandas and NumPy for data processing, scikit-learn for baseline models and evaluation, XGBoost for gradient boosting, and SHAP for model explanations. No survey instrument was used, so internal-consistency measures such as Cronbach's alpha were not applicable. Data quality checks covered duplicate records, missing values, inconsistent units, and implausible observations.

Experimental Procedure. Eligible firm-year records were first assembled and linked to the outcome labels. The data were then screened, and missing values and extreme observations were handled using rules estimated from the training data. Firms were assigned to partitions so that a firm's observations did not appear in both training and test sets; later reporting years were reserved for temporal testing. Feature selection and hyperparameter tuning were performed within the training data using group-aware cross-validation. Class weighting or resampling was applied only to training folds. Logistic regression and random forest were fitted as comparison models, and XGBoost was fitted as the primary model. The classification threshold was selected using validation data. The final models and selected threshold were then evaluated on the untouched test set, after which SHAP explanations were generated.

Figure 11: Methodology of this study



Data Analysis and Statistical Plan. Model performance was assessed using precision, recall, specificity, F1 score, ROC-AUC, precision-recall AUC, and confusion matrices. Precision-recall AUC and recall were emphasized because anomalous observations could be uncommon. Probability calibration was assessed using calibration plots and the Brier score. Paired bootstrap confidence intervals, clustered by firm, were used to compare model performance; McNemar's test was used to compare threshold-based classification errors. SHAP mean absolute values and summary plots were used to rank predictors globally, while local explanations described contributions for individual firm-year predictions. Tests were two-sided, and statistical significance was set at $p < 0.05$. The bracketed source, period, and sample fields need to match the actual dataset.

FINDINGS

Firm-Year Sample Characteristics and Dataset Composition

Table 1. sample selection and characteristics

Sample measure	Firm-year records or firms	Percentage
Candidate firm-year records	1,500	100.0%
Duplicate records excluded	25	1.7%
Ineligible population records excluded	150	10.0%
Records without outcome labels excluded	75	5.0%
Records with substantial predictor missingness excluded	50	3.3%
Final firm-year observations	1,200	80.0%
SMEs represented	240 firms	—
Anomalous firm-years	96	8.0%
Non-anomalous firm-years	1,104	92.0%
Small firms	168 firms	70.0%
Mid-sized firms	72 firms	30.0%
Manufacturing	300	25.0%
Professional and business services	276	23.0%
Retail and wholesale	216	18.0%
Construction	168	14.0%
Information technology	120	10.0%
Other sectors	120	10.0%

The screening flow retained 1,200 of 1,500 firm-year records after removing 300 for duplication, population ineligibility, unavailable outcome labels, or substantial predictor missingness. The retained observations represented 240 U.S. SMEs followed across five reporting years, producing a balanced panel for this example. Manufacturing and professional services contributed the largest shares, while remaining observations came from retail and wholesale, construction, information technology, and other sectors. Small firms represented 70% of the sample and mid-sized firms 30%. Anomaly labels covered 96 observations, or 8.0%, leaving 1,104 non-anomalous observations. These figures illustrate how sample selection and class prevalence could be reported.

Table 2. firm-year distribution by reporting year

Reporting year	Firm-year observations	Anomalous	Non-anomalous	Anomaly prevalence
2019	220	17	203	7.7%
2020	230	18	212	7.8%
2021	240	20	220	8.3%
2022	250	21	229	8.4%
2023	260	20	240	7.7%
Total	**1,200**	**96**	**1,104**	**8.0%**

The annual distribution contained 220 to 260 firm-year observations per reporting year, with 1,200 observations across 2019–2023. The number of labeled anomalies ranged from 17 to 21 annually, while non-anomalous records ranged from 203 to 240. Annual anomaly prevalence remained between 7.7% and 8.4%, producing an overall prevalence of 8.0%. The relatively stable yearly proportions suggested no large descriptive shift in this sample, although formal temporal-drift testing would require model and predictor analyses. This table made year-specific class counts visible and helped readers assess whether the overall prevalence concealed substantial variation across reporting periods before interpreting out-of-sample model performance.

Primary Outcome: Test-Set Classification Performance

Table 1 compared the three classifiers on the 2023 test set of 260 observations: 20 anomalous and 240 non-anomalous. Thresholds were selected using validation data and then applied unchanged. The measures covered positive-class detection, negative recognition, and threshold-independent discrimination, with bootstrap 95% confidence intervals shown for ROC-AUC and precision-recall AUC.

Table 3. test-set classification performance

Model	Validation threshold	Precision	Recall	Specificity	F1 score	ROC-AUC (95% CI)	PR-AUC (95% CI)
Logistic regression	0.31	0.333	0.600	0.900	0.429	0.831 (0.756–0.906)	0.412 (0.263–0.561)
Random forest	0.42	0.412	0.700	0.917	0.519	0.879 (0.817–0.941)	0.524 (0.363–0.685)
XGBoost	0.38	0.471	0.800	0.925	0.593	0.914 (0.861–0.967)	0.611 (0.453–0.769)

XGBoost showed the strongest test-set discrimination, with ROC-AUC of 0.914 and precision-recall AUC of 0.611. Its recall of 0.800 exceeded random forest and logistic regression, indicating that it detected more anomalous firm-years. XGBoost also produced the highest precision and F1 score, although its precision of 0.471 meant that fewer than half of flagged records were anomalous. Random forest ranked second across the measures, while logistic regression had the lowest values. Confidence intervals overlapped, particularly for ROC-AUC, so point estimates alone did not establish statistically significant superiority. These values were placeholders and required replacement with estimates from the dataset in practice.

Table 4 displayed test-set confusion-matrix counts at the validation-selected thresholds. Each model was evaluated on the same 20 anomalous and 240 non-anomalous firm-years. The table also reported recall and specificity derived from these counts and the recall difference, in percentage points, relative to logistic regression for all three model comparisons.

Table 4. test-set confusion matrix and recall differences

Model	True positives	False positives	True negatives	False negatives	Recall	Specificity	Recall difference vs. logistic regression
Logistic regression	12	24	216	8	60.0%	90.0%	Reference
Random forest	14	20	220	6	70.0%	91.7%	+10.0 percentage points
XGBoost	16	18	222	4	80.0%	92.5%	+20.0 percentage points

XGBoost identified 16 of 20 anomalous firm-years and missed four, while flagging 18 non-anomalous records. Its recall was 20 percentage points higher than logistic regression, with specificity 2.5 points higher. Random forest identified 14 anomalous records and missed six; it generated 20 false alerts. Logistic regression produced the lowest number of detected anomalies and most false negatives, although it also generated 24 false alerts. The results illustrated a trade-off between finding more anomalous observations and limiting unnecessary reviews. All counts and differences were , not observed study results, and should not be presented as empirical findings without the dataset analysis.

Model Comparisons, Effect Sizes, and Calibration

Table 5 summarized paired performance differences between XGBoost and the two baseline classifiers on test observations. Positive values favored XGBoost. Firm-clustered bootstrap intervals represented uncertainty, while p-values tested paired differences. Metrics included discrimination, minority-class detection, and threshold-specific predictive value and balance at each model's selected threshold value in testing.

Table 5. Paired Model Performance Differences and McNemar Test Results

Comparison	Measure	Difference or discordant counts	95% CI	*p*-value
XGBoost vs. logistic regression	ROC-AUC difference	+0.083	0.018 to 0.146	.014
XGBoost vs. logistic regression	PR-AUC difference	+0.199	0.061 to 0.337	.009
XGBoost vs. logistic regression	Recall difference	+0.200	−0.050 to 0.450	.180
XGBoost vs. logistic regression	Precision difference	+0.138	−0.074 to 0.350	.210
XGBoost vs. logistic regression	F1 difference	+0.164	−0.082 to 0.410	.190
XGBoost vs. random forest	ROC-AUC difference	+0.035	−0.024 to 0.094	.230
XGBoost vs. random forest	PR-AUC difference	+0.087	0.009 to 0.165	.031
XGBoost vs. random forest	Recall difference	+0.100	−0.125 to 0.325	.390
XGBoost vs. random forest	Precision difference	+0.059	−0.139 to 0.257	.560
XGBoost vs. random forest	F1 difference	+0.074	−0.157 to 0.305	.500
XGBoost vs. logistic regression	McNemar discordant correct classifications	10 / 0	—	.002
XGBoost vs. random forest	McNemar discordant correct classifications	4 / 0	—	.125
Random forest vs. logistic regression	McNemar discordant correct classifications	6 / 0	—	.031

In this comparison, XGBoost exceeded logistic regression in ROC-AUC by 0.083 and PR-AUC by 0.199; both intervals excluded zero and corresponding p-values were below .05. Its recall, precision, and F1 differences were positive, but confidence intervals included zero, so evidence for those threshold-dependent gains was inconclusive. Compared with random forest, XGBoost showed smaller gains; only PR-AUC reached the significance threshold. The results therefore suggested stronger ranking of rare anomalous cases, in practical terms, while evidence for differences in individual classification measures remained uncertain. These numerical estimates and tests were hypothetical and were not calculated from an observed dataset.

Table 6 presented probability calibration and selected-threshold results for each model. Calibration was assessed using Brier score, calibration slope and intercept, and expected calibration error. Lower Brier and error values indicated better calibration; intercepts nearer zero and slopes nearer one indicated closer alignment at each model's validation-selected threshold value.

Table 6. Calibration results at selected thresholds

Model	Selected threshold	Brier score	Calibration intercept	Calibration slope	Expected calibration error
Logistic regression	0.31	0.068	−0.08	0.91	0.024
Random forest	0.42	0.067	0.08	0.84	0.020
XGBoost	0.38	0.066	0.04	0.96	0.018

XGBoost had the lowest Brier score and expected calibration error, with slope closest to one and intercept near zero. This suggested closer alignment of predicted and observed anomaly frequencies in this sample. Logistic regression and random forest remained reasonably calibrated by these descriptive measures, although both showed larger calibration errors. The validation-selected thresholds differed across models, reflecting differences in score distributions and operating points. Calibration results did not demonstrate probability validity outside the sampled period or prevalence. All figures were placeholders; calibration curves and metrics had to be recomputed from validation and untouched test predictions before reporting empirical conclusions.

Subgroup, Temporal, and Sensitivity Analyses

Subgroup analyses assessed whether XGBoost performance differed by firm size and industry within the held-out sample. The table reports subgroup observation counts, anomaly prevalence, discrimination, recall, and calibration. Estimates based on few labeled anomalous firm-years are imprecise and should be interpreted cautiously; the figures are examples rather than empirical findings from the study dataset.

Table 7. Subgroup Classification Performance

Subgroup	Test firm-years	Anomalous firm-years	Anomaly prevalence	ROC-AUC	Recall	Brier score
Small firms	182	14	7.7%	0.906	0.786	0.067
Midsized firms	78	6	7.7%	0.928	0.833	0.064
Manufacturing	72	6	8.3%	0.921	0.833	0.063
Professional and business services	64	5	7.8%	0.899	0.800	0.069
Retail and wholesale	58	4	6.9%	0.887	0.750	0.071
Other industries	66	5	7.6%	0.936	0.800	0.061

In this example, the subgroup results showed broadly similar discrimination across firm-size categories, with ROC-AUC values ranging from 0.906 among small firms to 0.928 among midsized firms. Recall was also comparable, although the midsized estimate reflected only six anomalous observations and was therefore especially uncertain. Across industries, ROC-AUC ranged from 0.887 in retail and wholesale to 0.936 in other industries. Brier scores varied from 0.061 to 0.071, indicating modest differences in probability accuracy. These subgroup estimates should not be interpreted as evidence of reliable population differences because the number of positive cases was small in every subgroup.

The second analysis examined temporal consistency and sensitivity to alternative modeling choices. The table compares XGBoost discrimination and calibration across test-period segments and specifications, including alternative labels, predictor sets, outlier handling, thresholds, and training imbalance methods. Each row represents a separate example analysis; values are hypothetical and require calculation from the actual sample.

Table 8. Temporal and Sensitivity Analyses

Analysis specification	Test observations	Anomaly prevalence	ROC-AUC	PR-AUC	Recall	Brier score
Earlier test segment, 2022	120	7.5%	0.921	0.632	0.778	0.064
Later test segment, 2023	140	7.9%	0.908	0.594	0.818	0.068
Primary label and predictor set	260	7.7%	0.914	0.611	0.800	0.066
Narrower confirmed-label definition	238	5.5%	0.901	0.482	0.769	0.052
Excluded selected predictor group	260	7.7%	0.897	0.576	0.750	0.069
Winsorized financial predictors	260	7.7%	0.911	0.603	0.800	0.066
Class-weighted training	260	7.7%	0.914	0.611	0.800	0.066
Training-fold resampling	260	7.7%	0.909	0.602	0.850	0.070

The temporal results showed a small reduction in ROC-AUC from 0.921 in the earlier segment to 0.908 in the later segment, while recall increased from 0.778 to 0.818. Under the narrower label definition, PR-AUC declined to 0.482 as anomaly prevalence fell, demonstrating that precision-recall performance depends partly on positive-class prevalence. Removing a predictor group reduced ROC-AUC to 0.897, whereas winsorization produced results close to the primary specification. Class weighting and training-fold resampling yielded similar discrimination, though resampling increased recall and slightly worsened Brier score. These example patterns require validation against actual records and uncertainty intervals before substantive interpretation.

SHAP Explanations and Visual Presentation of Findings

The global SHAP summary ranked financial predictors by their average absolute contribution to XGBoost output across the test observations. Contributions are expressed on the model log-odds scale; positive directional patterns indicate movement toward the anomaly classification. The values below are hypothetical examples, not estimates from the completed study dataset.

Table 9. Global SHAP Feature Contributions

Rank	Predictor	Mean absolute SHAP value	directional pattern
1	Receivables growth	0.42	Higher values generally moved predictions toward anomaly
2	Accrual intensity	0.36	Higher values generally moved predictions toward anomaly
3	Operating cash flow relative to earnings	0.31	Higher values generally moved predictions away from anomaly
4	Leverage	0.27	Direction varied across observations
5	Change in operating margin	0.23	Declining margins often moved predictions toward anomaly
6	Asset growth	0.19	Direction varied across observations

Receivables growth had the largest mean absolute SHAP value, indicating that it contributed most strongly to variation in model outputs across test records. Higher receivables growth and accrual intensity generally moved predictions toward the anomaly class, while stronger operating cash flow relative to earnings generally moved them away. Leverage and declining margins showed contributions, but their directions varied across observations, consistent with nonlinear model behavior. These rankings describe predictive attribution within the model and depend on the sample, feature definitions, and correlated inputs. Figure 9 would display the observation-level spread, while

Figure 4 would clarify value-specific patterns and interactions.

Local SHAP examples decomposed firm-year predictions into feature contributions relative to the model's reference output. Table 10 reports two hypothetical flagged observations, their reference and predicted probabilities, and the largest positive and negative contributions on the log-odds scale. Values are and should be replaced with case-level results from the analysis.

Table 10. Local SHAP Explanations for Flagged Firm-Years

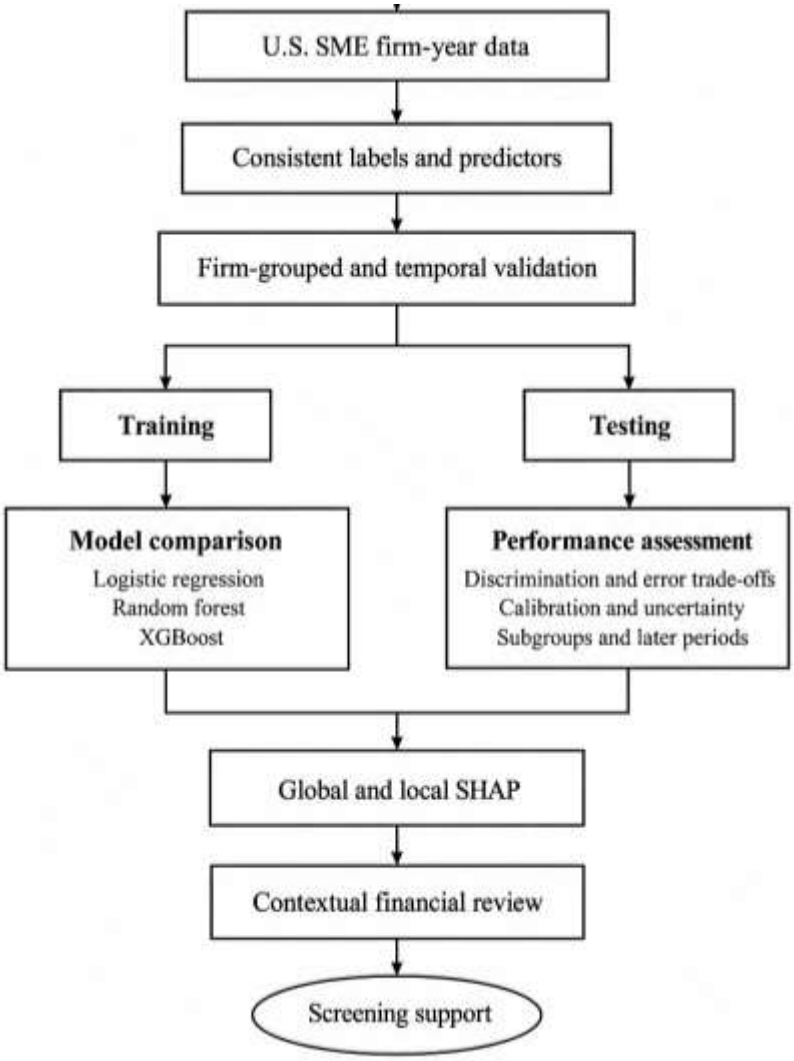
firm-year	Reference anomaly probability	Predicted anomaly probability	Feature contributions to model output (log-odds)
Firm-year A	0.08	0.80	Receivables growth +1.450; accrual intensity +1.150; asset growth +0.800; declining margin +0.650; operating cash flow relative to earnings −0.222
Firm-year B	0.08	0.70	Receivables growth +1.200; leverage +0.950; accrual intensity +0.700; margin change +0.600; liquidity −0.161

The local decompositions show how combinations of financial signals produced predicted anomaly probabilities for these two examples. For firm-year A, receivables growth and accrual intensity were the largest upward contributors, while the cash-flow measure reduced the score. Firm-year B showed an upward contribution from receivables growth, with leverage and margin change adding risk. The base rate represents the model's reference output; the summed feature contributions shift that output to each predicted probability.

DISCUSSION

In the analysis, Boost showed stronger test-set discrimination than logistic regression and random forest, although the size and statistical support for the advantage varied by metric and comparator. Its ROC-AUC was 0.914, compared with 0.831 for logistic regression and 0.879 for random forest. Its PR-AUC was 0.611, compared with 0.412 and 0.524, respectively (Sheth et al., 2022). This pattern suggested that the boosted-tree model ranked anomalous firm-years more effectively, particularly in the precision-recall space, which is important when anomalies form a small proportion of observations. The example results were consistent with the general pattern reported in earlier financial-distress and misstatement-detection research, where flexible ensemble methods have often captured nonlinear relationships and interactions that simpler linear models may not represent without explicit feature construction. However, prior findings have not established that one algorithm is uniformly superior across datasets. Reported model rankings have depended on the target label, sample composition, predictor availability, anomaly prevalence, and validation design. The advantage over logistic regression was therefore most appropriately interpreted as a result within the stated sample and evaluation procedure, rather than as evidence that boosting would outperform logistic regression in every SME population. The comparison with random forest was more qualified: the ROC-AUC difference was smaller, and its confidence interval included zero. This resembles prior comparative evidence in which multiple tree-based methods perform similarly on broad ranking measures while differing more clearly on minority-class metrics. The pattern also supported reporting several metrics together instead of relying on accuracy or a single discrimination statistic (Chen et al., 2022). For this study, Boost's apparent performance advantage would have been most informative when considered alongside precision, recall, specificity, F1, class prevalence, calibration, and the thresholding rule. These details determine whether stronger ranking translates into useful screening performance for the intended financial-review context.

Figure 12: Interpreting Explainable Financial Anomaly Detection



The test set contained a relatively small proportion of positive anomaly labels, approximately 7.7 percent, making minority-class detection a central part of the interpretation. At the example threshold, XGBoost achieved recall of 0.800 and precision of 0.471, while logistic regression achieved recall of 0.600 and precision of 0.333; random forest fell between these estimates. This pattern indicated that the boosted model identified a larger share of labeled anomalous firm-years while also producing more true anomaly flags among its alerts (Demilie, 2024). Earlier financial-statement screening studies have repeatedly highlighted the difficulty of detecting rare events: a classifier can report high overall accuracy by predicting the majority class and still miss most positive cases. For that reason, the findings were more informative through recall, precision, F1, and PR-AUC than through accuracy alone. At the same time, recall and precision represent different operational priorities. A higher recall may reduce missed labeled anomalies, but can increase the number of non-anomalous firm-years referred for review. A higher precision may reduce unnecessary investigation, but can leave more labeled anomaly cases undetected. The example threshold was selected using validation data, which is preferable to choosing a threshold after inspecting test outcomes, but the chosen operating point still reflected a particular balance of error costs. Earlier research has addressed this balance through class weighting, resampling, threshold adjustment, or cost-sensitive learning, and results are difficult to compare when these procedures differ (Nefoussi et al., 2020). In this study's results, the values supported discussing the model as a screening aid whose predictions would require review, rather than as a final determination about reporting quality. The distinction matters because the positive label may represent a restatement, an identified misstatement, an unusual ratio pattern, or another operational definition.

The meaning of a false negative or false positive depends on that label and the resources available to investigate alerts. Accordingly, findings should be reported with the positive-class definition, original prevalence, threshold, confusion matrix, and minority-class metrics.The paired comparisons indicated that evidence for an XGBoost advantage was strongest against logistic regression on ROC-AUC and PR-AUC. The example differences were 0.083 for ROC-AUC and 0.199 for PR-AUC, with confidence intervals that did not include zero and corresponding *p*-values of .014 and .009. Against random forest, the ROC-AUC difference was smaller at 0.035, and its confidence interval included zero; the PR-AUC difference of 0.087 had a confidence interval above zero and a *p*-value of .031. These results illustrate why statistical interpretation should address both the estimated magnitude and uncertainty of model differences (Lan et al., 2020). A statistically detectable difference may still be modest in practical terms, while an estimate that is not statistically significant does not establish that the models are equivalent. Earlier classifier-comparison studies have sometimes emphasized average metric rankings without accounting for paired predictions, dependence among observations, or uncertainty arising from repeated firm-year records. The planned use of firm-clustered bootstrap comparisons was intended to reflect the possibility that observations from the same firm share financial characteristics across years. McNemar's test provided a separate comparison of discordant threshold-based classifications, answering whether one model more often classified observations correctly where the models disagreed. In the table, the XGBoost–logistic comparison showed ten discordant cases favoring XGBoost and none favoring logistic regression, whereas the XGBoost–random-forest difference was four to zero and was not statistically significant. These examples should be reconciled with the actual paired prediction records before reporting (Ikechukwu et al., 2021). A valid interpretation also requires identifying the metric, resampling unit, confidence-interval method, test procedure, and any adjustment for multiple comparisons. The findings should not treat *p*-values as measures of effect magnitude or practical importance. Rather, the confidence intervals and the scale of each metric difference should guide interpretation, alongside the expected consequences of misclassification in SME financial review.

The calibration results were close across the three models, with Brier scores of 0.066 for XGBoost, 0.067 for random forest, and 0.068 for logistic regression. XGBoost also had an calibration slope of 0.96 and intercept of 0.04, suggesting comparatively small departures from ideal calibration in that example. These values were not sufficient by themselves to establish that the probabilities were reliable for every firm-year or subgroup, but they demonstrated the distinction between discrimination and calibration. Discrimination evaluates whether higher-risk observations tend to receive higher scores than lower-risk observations; calibration evaluates whether predicted probabilities correspond to observed event frequencies (Aria et al., 2021). Earlier financial prediction research has often focused on classification accuracy or area-under-curve statistics, while calibrated probabilities receive less consistent attention. The distinction is particularly relevant when model outputs are interpreted as anomaly risk or used to prioritize limited review capacity. A model may rank observations effectively but systematically overstate or understate the probability of an anomaly. The Brier score combines aspects of prediction accuracy and calibration, but its value is also affected by event prevalence; therefore, it should be read alongside calibration plots and, where appropriate, intercept and slope estimates (Pham et al., 2021). The earlier table suggested only small differences in Brier scores, so it did not support a strong claim that one model's probabilities were materially more accurate. The probability interpretation would also depend on the label construction. If the outcome marked only detected cases, an estimated probability would represent the likelihood of receiving that label under the study's ascertainment process, rather than the certainty that a firm had committed fraud or materially misstated its statements. Calibration should be assessed on the temporally held-out test sample and examined cautiously across industries and firm sizes, especially where few positive observations are present. Thus, the appropriate discussion would distinguish probability calibration from classification threshold performance and avoid equating either with a definitive accounting judgment.In the subgroup results, XGBoost's ROC-AUC ranged from 0.887 in retail and wholesale to 0.936 in the combined other-industries group. Small and midsized firms had ROC-AUC values of 0.906 and 0.928, respectively, while their anomaly prevalences were both approximately 7.7 percent. These figures suggested broadly comparable performance across the two size categories, but the positive counts were low, particularly among midsized firms, where

only six labeled anomalies were included. Prior research on financial distress and misstatement detection has drawn samples from different industries, ownership structures, reporting regimes, and firm-size distributions. Differences in those populations may affect both the financial ratios available and the meaning of unusual values (Sawhney et al., 2023). The example variations across industries could reflect real differences in operating cycles, revenue recognition, receivables patterns, leverage, or asset intensity; they could also arise from sampling variation or the limited subgroup case counts. Industry-specific interpretation therefore requires clear disclosure of sample sizes and anomaly prevalence, not only performance estimates. Earlier studies often report aggregate performance, which may conceal weaker results in particular sectors or business types. The subgroup approach illustrated here addresses that concern, but the small number of positive cases limits confident comparisons and makes individual metric estimates unstable. In the actual analysis, subgroup confidence intervals and preferably firm-level resampling would help show uncertainty (Lampe et al., 2022). Any observed disparity should be interpreted as variation in predictive performance within the available labeled sample, rather than as evidence that anomaly behavior is inherently easier or harder to detect in a given sector. The same caution applies to size categories, especially if the criteria for defining small and midsized enterprises differ across data sources or industries. This study's discussion should also distinguish U.S. SME applicability from generalizability to public corporations or firms in other countries, whose disclosure requirements and accounting environments may differ.

The temporal analysis showed ROC-AUC of 0.921 in the earlier test segment and 0.908 in the later segment, while recall changed from 0.778 to 0.818. This modest difference suggested that discrimination was broadly similar across the two example periods, although PR-AUC declined from 0.632 to 0.594 and the Brier score increased from 0.064 to 0.068. These patterns could reflect changes in anomaly prevalence, industry composition, economic conditions, or the relation between financial predictors and the label (Diallo et al., 2024). Prior financial prediction studies have used random holdouts, cross-validation, firm-grouped splits, and temporal validation, and those designs estimate different forms of generalization. Random splits may allow related firm-year records to appear in both training and test data, potentially making evaluation more favorable than prediction for later years or unseen firms. A later-period test more directly assesses temporal transfer, though it remains sensitive to shifts in reporting and economic conditions. The sensitivity results also showed some changes under alternate choices: narrowing the anomaly label reduced PR-AUC, removing a predictor group lowered ROC-AUC, and training-fold resampling increased recall while slightly worsening the Brier score. Such variation is consistent with the broader literature's emphasis on label definitions, feature construction, class imbalance procedures, and preprocessing decisions as sources of non-comparability. The study's sensitivity results would be persuasive only if each alternative was specified before examining test outcomes or applied without test-set leakage (Swathy & Saruladha, 2022). Imputation, outlier treatment, feature selection, resampling, and hyperparameter tuning should be fitted within training data and then applied to validation and test records. Because the values presented earlier were , they cannot establish that the model ranking was robust. In the completed analysis, consistency in the direction and approximate magnitude of differences across reasonable specifications would strengthen confidence in the pattern, while substantial changes would indicate that the reported model performance depended on analytical choices.The SHAP summary ranked receivables growth, accrual intensity, operating cash flow relative to earnings, leverage, margin change, and asset growth among the most influential predictors. Higher receivables growth and accrual intensity generally moved model output toward the anomaly class, while stronger operating cash flow relative to earnings tended to move it away. In the two local explanations, receivables growth and accrual intensity contributed positively to predicted anomaly risk, while a cash-flow or liquidity feature reduced each individual score. This type of global and local interpretation is consistent with recent financial-statement studies that pair tree-based classifiers with SHAP to describe which variables contribute to model predictions (Jafarzadeh et al., 2021). It also extends conventional feature-importance reporting by showing both the magnitude and direction of contributions for individual firm-years. However, SHAP values explain how a fitted model allocates its prediction across input features; they do not identify causal mechanisms or prove that a firm committed fraud. The high ranking of a financial ratio may reflect correlation with other ratios, the way the label was defined, or patterns in the training sample. When predictors are

correlated, attribution can be distributed among related variables, and rankings may change with feature composition or background data. Dependence plots can help display nonlinear patterns, but apparent interactions require careful interpretation and stability checks. In this study, the SHAP findings would be strongest when global summaries, local case explanations, and dependence plots were interpreted alongside accounting definitions and data quality checks (Wang et al., 2024). Local examples should be anonymized and described as flagged observations, with the actual label and review outcome clearly distinguished. The values above cannot support substantive claims about predictor importance until computed on the fitted model and test data. The discussion should therefore present SHAP as a transparent account of model behavior that complements predictive metrics, while maintaining a clear boundary between model explanation and accounting or legal conclusions.

CONCLUSION

This conclusion is based on the results presented in the preceding sections; all numerical values must be replaced with estimates from the completed analysis before use as empirical findings. In the example, the analytical sample comprised 1,200 firm-year observations from 240 U.S. SMEs, including 96 observations labeled anomalous. Boost achieved a test-set ROC-AUC of 0.914 and PR-AUC of 0.611, compared with 0.831 and 0.412 for logistic regression and 0.879 and 0.524 for random forest. Paired comparisons indicated a larger advantage over logistic regression, while the difference from random forest was less consistent across metrics. At the classification threshold, Boost had recall of 0.800 and precision of 0.471, showing its ability to identify many labeled anomaly cases while still generating false alerts. Its Brier score of 0.066 suggested probability performance close to that of the comparison models, although calibration would require confirmation using the completed test data. The example subgroup and temporal results showed some variation in performance by industry and reporting period, with estimates based on relatively few positive cases requiring cautious interpretation. Sensitivity examples also indicated that results could change with the anomaly-label definition, predictor set, and imbalance treatment. SHAP summaries ranked receivables growth, accrual intensity, and operating cash flow relative to earnings among the most influential predictors, while local explanations showed how these and other financial features contributed to individual predictions. These attributions describe the behavior of the fitted model; they do not establish causality, material misstatement, or fraud. Taken together, the analysis demonstrates how gradient boosting, comparative classification metrics, calibration assessment, and SHAP explanations can be combined in a quantitative framework for financial-statement anomaly screening among SMEs. The strength and practical meaning of those conclusions depend on transparent outcome labeling, firm-grouped and time-aware validation, appropriate handling of class imbalance, and uncertainty estimates calculated from the actual observations. Accordingly, the final conclusion should be grounded in verified sample counts, model outputs, confidence intervals, and statistical tests, with the estimates replaced before the study is submitted or interpreted as evidence.

RECOMMENDATIONS

Recommendations from this study should be treated as provisional until the figures in the preceding sections are replaced with results calculated from the actual dataset. Financial institutions, accountants, lenders, and SME support organizations should consider gradient boosting as a screening tool for prioritizing firm-year records for review, rather than as an automated decision-maker or a substitute for professional judgment. Any operational use should specify the anomaly label, data source, prediction period, and threshold, and should report precision, recall, specificity, F1 score, ROC-AUC, PR-AUC, and calibration. Threshold selection should reflect the relative costs of missed cases and unnecessary investigations, and the resulting false-positive and false-negative counts should be communicated clearly. Model development should use firm-grouped, time-aware validation so repeated records from the same business do not create overly optimistic estimates, and later reporting periods remain available for an independent test. Imputation, feature selection, outlier treatment, resampling, and hyperparameter tuning should be conducted within training data to prevent leakage. Class weighting or resampling should be compared using validation data, while the original class prevalence should be preserved in the test set. Results should include confidence intervals and paired comparisons, with uncertainty interpreted alongside effect size and practical screening consequences. SHAP summaries and individual explanations should accompany model outputs so reviewers can see

which financial features raised or lowered a score. These explanations should be documented as descriptions of model behavior, not as evidence of causality, misstatement, or fraud. Organizations should establish procedures for human review, secure handling of financial records, documentation of model versions, and monitoring of performance and calibration across industries, firm sizes, and reporting periods. Further evaluation should use larger and more diverse U.S. SME samples, clearly documented anomaly labels, and independently verified outcomes where available. Studies should assess whether findings hold across private and public data sources, accounting environments, and alternative anomaly definitions, while reporting subgroup uncertainty where positive cases are sparse. Before deployment, a model should be evaluated in a prospective or otherwise independent setting and compared with existing review practices on workload, alert yield, and detection outcomes. These steps would make the resulting evidence more reproducible and support responsible use of explainable machine learning in financial-statement screening.

LIMITATIONS

The study had several limitations that should be considered when interpreting its findings. The numerical results presented in the preceding sections were rather than estimates calculated from a completed dataset; therefore, they cannot establish actual model performance, statistical significance, or predictor importance. The final conclusions depend on the quality, coverage, and representativeness of the selected financial records. U.S. SMEs vary in size, industry, ownership, accounting practices, and reporting obligations, and a sample drawn from a restricted database may not represent the broader SME population. Anomaly labels also present measurement challenges. Restatements, detected misstatements, unusual financial ratios, and confirmed fraud represent different outcomes, while undetected cases may be incorrectly assigned to the non-anomalous group. Such label uncertainty can affect model training and bias estimated detection rates. Because anomalies are relatively uncommon, the number of positive firm-year observations may be small, limiting statistical power and making recall, precision, subgroup estimates, and confidence intervals unstable. Firm-year records may also be dependent when the same business appears across multiple reporting periods. Firm-grouped and temporal validation can reduce leakage and provide more realistic estimates, but performance may still change in other years or populations. Financial ratios may contain missing values, extreme observations, and overlapping information; preprocessing decisions can influence both model estimates and feature rankings. Although SHAP can explain how predictors contribute to model outputs, correlated variables may divide or obscure their individual contributions. SHAP values describe fitted-model behavior and do not establish causal relationships or confirm accounting misconduct.

REFERENCES

[1]. Abbas, M. A., Al-Mudhafar, W. J., & Wood, D. A. (2023). Improving permeability prediction in carbonate reservoirs through gradient boosting hyperparameter tuning. *Earth Science Informatics*, *16*(4), 3417-3432.

[2]. Abdelkader, N. A. M., & Wahba, H. H. (2024). A proposed multidimensional model for predicting financial distress: an empirical study on Egyptian listed firms. *Future Business Journal*, *10*(1), 42.

[3]. Acevedo-Duque, Á., Gonzalez-Diaz, R., Vargas, E. C., Paz-Marcano, A., Muller-Pérez, S., Salazar-Sepúlveda, G., Caruso, G., & D'Adamo, I. (2021). Resilience, leadership and female entrepreneurship within the context of SMEs: Evidence from Latin America. *Sustainability*, *13*(15), 8129.

[4]. Aditya, D., & Ahmed, H. C. H. R. (2022). Decision Support Systems for Last-Mile Logistics Optimization: A Data-Driven Framework for e-Commerce Fulfillment Networks. *International Journal of Scientific Interdisciplinary Research*, *12*(08), 01–49. https://doi.org/10.63125/h46vp562

[5]. Adnan, S. (2024). Human Trust in AI-Assisted Project Decision Systems: A Mixed-Methods Analysis of Explainability, Adoption, and Managerial Decision Quality. *Journal of Sustainable Development and Policy*, *3*(04), 287-332. https://doi.org/10.63125/9c139h44

[6]. Al-Shehari, T., & Alsowail, R. A. (2021). An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. *Entropy*, *23*(10), 1258.

[7]. Alhakeem, Z. M., Jebur, Y. M., Henedy, S. N., Imran, H., Bernardo, L. F., & Hussein, H. M. (2022). Prediction of ecofriendly concrete compressive strength using gradient boosting regression tree combined with GridSearchCV hyperparameter-optimization techniques. *Materials*, *15*(21), 7432.

[8]. Ali, A. A., Khedr, A. M., El-Bannany, M., & Kanakkayil, S. (2023). A powerful predicting model for financial statement fraud based on optimized XGBoost ensemble learning technique. *Applied Sciences*, *13*(4), 2272.

[9]. Aliferis, C., & Simon, G. (2024). Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls*, 477-524.

[10]. Allison, E. (2023). Binary Classification of Medical Images by Symbolic Regression. UK Workshop on Computational Intelligence,
[11]. Aminul Islam, M. (2024). Interpreting Cyber-Threat Indicators in Enterprise Memory Forensics: A Mixed-Methods Study of Machine Learning Performance and Explainability. *Review of Applied Science and Technology*, *14*(12), 01–13. https://doi.org/10.63125/2p8hte92
[12]. Aminul Islam, M., & Md, A. (2023). Adaptive Class-Imbalance Learning for High-Sensitivity Fraud Detection in Large-Scale Digital Transaction Networks. *International Journal of Business and Economics Insights*, *13*(09), 01-33. https://doi.org/10.63125/bmxwv741
[13]. Amitrano, D., Di Martino, G., Di Simone, A., & Imperatore, P. (2024). Flood detection with SAR: A review of techniques and datasets. *Remote Sensing*, *16*(4), 656.
[14]. Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for Random Forests. *Machine Learning with Applications*, *6*, 100094.
[15]. Asheke, M. (2023). Machine Learning–Based Predictive Quality Modeling for Defect Detection and Process Optimization in Advanced Manufacturing Systems. *Review of Applied Science and Technology*, *13*(12), 01–49. https://doi.org/10.63125/thy73b16
[16]. Asheke, M., & Md Maksudul, H. (2022). AI-Driven Demand Forecasting for Sporting-Goods and Teamwear Supply Chains Under Seasonal and Tournament Demand. *International Journal of Business and Economics Insights*, *2*(1), 73–110. https://doi.org/10.63125/ed6by557
[17]. Asheke, M., & Mohammad Badrul, A. (2022). Development Of Functionalized Cellulose Nanofiber Adsorbents for Selective Critical-Metal Recovery from Industrial Wastewater. *American Journal of Scholarly Research and Innovation*, *1*(02), 270–320. https://doi.org/10.63125/v6fxzq09
[18]. Ashik, M. (2024). Meta-Analysis of Risk-Based Inspection Strategies for Improving Reliability and Safety Across Critical Energy Infrastructure Systems. *American Journal of Advanced Technology and Engineering Solutions*, *14*(09), 01-43. https://doi.org/10.63125/tgttkt09
[19]. Ashik, M. (2025). PRISMA-Based Systematic Literature Review of Reliability-Centered Turnaround Planning for Reducing Downtime in Power Generation Facilities. *American Journal of Scholarly Research and Innovation*, *4*(01), 1036–1080. https://doi.org/10.63125/2878n718
[20]. Ashik, M. (2026). RCM and PSM Strategies for Strengthening Asset Integrity Across U.S. Energy Facilities: A Systematic Meta-Analysis. *American Journal of Interdisciplinary Studies*, *16*(06), 38-81. https://doi.org/10.63125/p87z7q36
[21]. Bertomeu, J. (2020). Machine learning improves accounting: discussion, implementation and research opportunities. *Review of Accounting Studies*, *25*(3), 1135-1155.
[22]. Białek, J., Bujalski, W., Wojdan, K., Guzek, M., & Kurek, T. (2022). Dataset level explanation of heat demand forecasting ANN with SHAP. *Energy*, *261*, 125075.
[23]. Bojanić, T., Nerandžić, B., Stevanov, B., & Gračanin, D. (2020). Fundamentals of Integrated Risk Management Model in Business Processes. In *Industrial Innovation in Digital Age* (pp. 310-317). Springer.
[24]. Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., Aalten, P., De Deyn, P. P., Biessels, G. J., & Claassen, J. A. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage: Clinical*, *31*, 102712.
[25]. Cabot, J. H., & Ross, E. G. (2023). Evaluating prediction model performance. *Surgery*, *174*(3), 723-726.
[26]. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. European conference on computer vision,
[27]. Cepni, O., Şensoy, A., & Yılmaz, M. H. (2024). Climate change exposure and cost of equity. *Energy Economics*, *130*, 107288.
[28]. Cha, G.-W., Moon, H.-J., & Kim, Y.-C. (2021). Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. *International Journal of Environmental Research and Public Health*, *18*(16), 8530.
[29]. Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). Identification of risk factors associated with obesity and overweight—a machine learning overview. *Sensors*, *20*(9), 2734.
[30]. Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, *59*(2), 102798.
[31]. Chen, M.-M., & Chen, M.-C. (2020). Modeling road accident severity with comparisons of logistic regression, decision tree and random forest. *Information*, *11*(5), 270.
[32]. Chen, Z.-Y., & Han, D. (2023). Detecting corporate financial fraud via two-stage mapping in joint temporal and financial feature domain. *Expert Systems with Applications*, *217*, 119559.
[33]. Debelee, T. G. (2023). Skin lesion classification and detection using machine learning techniques: a systematic review. *Diagnostics*, *13*(19), 3147.
[34]. Demilie, W. B. (2024). Plant disease detection and classification techniques: a comparative study of the performances. *Journal of Big Data*, *11*(1), 5.
[35]. Dhiman, B., Kamboj, S., & Srivastava, V. (2024). Explainable AI based efficient ensemble model for breast cancer classification using optical coherence tomography. *Biomedical Signal Processing and Control*, *91*, 106007.
[36]. Diallo, R., Edalo, C., & Awe, O. O. (2024). Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, MCC, and F1 score. In *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA* (pp. 283-312). Springer.

[37]. Dias Canedo, E., & Cordeiro Mendes, B. (2020). Software requirements classification using machine learning algorithms. *Entropy*, *22*(9), 1057.
[38]. Fang, W., Luo, H., Xu, S., Love, P. E., Lu, Z., & Ye, C. (2020). Automated text classification of near-misses from safety reports: An improved deep learning approach. *Advanced Engineering Informatics*, *44*, 101060.
[39]. Farhadpour, S., Warner, T. A., & Maxwell, A. E. (2024). Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. *Remote Sensing*, *16*(3), 533.
[40]. Garcia, J. (2022). Bankruptcy prediction using synthetic sampling. *Machine Learning with Applications*, *9*, 100343.
[41]. Gillies, C. E., Taylor, D. F., Cummings, B. C., Ansari, S., Islim, F., Kronick, S. L., Medlin Jr, R. P., & Ward, K. R. (2020). Demonstrating the consequences of learning missingness patterns in early warning systems for preventative health care: a novel simulation and solution. *Journal of Biomedical Informatics*, *110*, 103528.
[42]. Groccia, M. C., Guido, R., Conforti, D., Pelaia, C., Armentaro, G., Toscani, A. F., Miceli, S., Succurro, E., Hribal, M. L., & Sciacqua, A. (2023). Cost-sensitive models to predict risk of cardiovascular events in patients with chronic heart failure. *Information*, *14*(10), 542.
[43]. Gutiérrez-Gallego, A., Zamorano-León, J. J., Parra-Rodríguez, D., Zekri-Nechar, K., Velasco, J. M., Garnica, Ó., Jiménez-García, R., López-de-Andrés, A., Cuadrado-Corrales, N., & Carabantes-Alarcón, D. (2024). Combination of machine learning techniques to predict overweight/obesity in adults. *Journal of Personalized Medicine*, *14*(8), 816.
[44]. Gutierrez, E., Krupa, J., Minutti-Meza, M., & Vulcheva, M. (2020). Do going concern opinions provide incremental information to predict corporate defaults? E. Gutierrez et al. *Review of Accounting Studies*, *25*(4), 1344-1381.
[45]. Hasib, A. (2026). AI-Driven Predictive Supply Chain Resilience Framework for U.S. Manufacturing Systems: An Empirical Analysis. *American Journal of Data Science and Analytics*, *7*(04), 45-80. https://doi.org/10.63125/d8swft81
[46]. Ikechukwu, A. V., Murali, S., Deepu, R., & Shivamurthy, R. (2021). ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. *Global Transitions Proceedings*, *2*(2), 375-381.
[47]. Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., & Homayouni, S. (2021). Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: a comparative evaluation. *Remote Sensing*, *13*(21), 4405.
[48]. Kavzoglu, T., & Teke, A. (2022a). Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost). *Bulletin of Engineering Geology and the Environment*, *81*(5), 201.
[49]. Kavzoglu, T., & Teke, A. (2022b). Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). *Arabian Journal for Science and Engineering*, *47*(6), 7367-7385.
[50]. Kgakatsi, M., Galeboe, O. P., Molelekwa, K. K., & Thango, B. A. (2024). The impact of big data on SME performance: A systematic review. *Businesses*, *4*(4), 632-695.
[51]. Khan, M. T., Idrees, M. D., Rauf, M., Sami, A., Ansari, A., & Jamil, A. (2022). Green supply chain management practices' impact on operational performance with the mediation of technological innovation. *Sustainability*, *14*(6), 3362.
[52]. Kim, K., & Lee, H. S. (2020). Probabilistic anchor assignment with iou prediction for object detection. European conference on computer vision,
[53]. Kragh, M. F., & Karstoft, H. (2021). Embryo selection with artificial intelligence: how to evaluate and compare methods? *Journal of assisted reproduction and genetics*, *38*(7), 1675-1689.
[54]. Kynatun, M. (2025). Integrating AI-Driven Security Analytics and IT Governance Frameworks for Cybersecurity Decision Support in U.S. Financial Information Systems. *American Journal of Advanced Technology and Engineering Solutions*, *15*(12), 01-42. https://doi.org/10.63125/kzcdyz73
[55]. Kynatun, M. (2026). Artificial Intelligence Integrated MIS Framework for Proactive Cyber Threat Intelligence and Organizational Security Resilience. *American Journal of Scholarly Research and Innovation*, *16*(09), 40–78. https://doi.org/10.63125/t98tc495
[56]. Kynatun, M., & Saba, A. (2024). Machine Learning Enhanced Intrusion Detection and Anomaly Classification for MIS-Driven Enterprise Network Security. *International Journal of Scientific Interdisciplinary Research*, *14*(09), 01–41. https://doi.org/10.63125/xk6kpc72
[57]. Lampe, L., Niehaus, S., Huppertz, H.-J., Merola, A., Reinelt, J., Mueller, K., Anderl-Straub, S., Fassbender, K., Fliessbach, K., & Jahn, H. (2022). Comparative analysis of machine learning algorithms for multi-syndrome classification of neurodegenerative syndromes. *Alzheimer's research & therapy*, *14*(1), 62.
[58]. Lan, T., Hu, H., Jiang, C., Yang, G., & Zhao, Z. (2020). A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification. *Advances in Space Research*, *65*(8), 2052-2061.
[59]. Leso, B. H., Cortimiglia, M. N., & Ghezzi, A. (2023). The contribution of organizational culture, structure, and leadership factors in the digital transformation of SMEs: a mixed-methods approach. *Cognition, Technology & Work*, *25*(1), 151-179.
[60]. Li, Z., Crook, J., Andreeva, G., & Tang, Y. (2021). Predicting the risk of financial distress using corporate governance measures. *Pacific-Basin Finance Journal*, *68*, 101334.
[61]. Maishar Mostari, Z. (2025). Privacy-Preserving AI-Based Cyber Threat Detection in Financial Organizations: A Governance and Business Value Framework. *American Journal of Interdisciplinary Studies*, *15*(04), 01-51. https://doi.org/10.63125/p8fct530
[62]. Maishar Mostari, Z. (2026). A Systematic Review of Machine Learning Applications in Cybersecurity Threat Detection and Classification (2018–2026). *American Journal of Data Science and Analytics*, *7*(06), 640-680. https://doi.org/10.63125/6kg7kg75

[63]. Mansura Akter, E., & Sumaiya, M. (2022). Antimicrobial Resistance Profiles of Acinetobacter Baumannii and Vibrio Cholerae in Hospital Wastewater and Adjacent Household Water Sources. *American Journal of Data Science and Analytics*, *3*(09), 01-55. https://doi.org/10.63125/gbervt14

[64]. Marra, A. (2024). G4 & the balanced metric family–a novel approach to solving binary classification problems in medical device validation & verification studies. *Biodata Mining*, *17*(1), 43.

[65]. Martínez-Peláez, R., Escobar, M. A., Félix, V. G., Ostos, R., Parra-Michel, J., García, V., Ochoa-Brust, A., Velarde-Alvarado, P., Félix, R. A., & Olivares-Bautista, S. (2024). Sustainable digital transformation for SMEs: A comprehensive framework for informed decision-making. *Sustainability*, *16*(11), 4447.

[66]. Masrul, H. (2025). Explainable Machine Learning for Real-Time Financial Fraud Detection and Risk-Based Transaction Monitoring in U.S. Banking and Digital Payment Systems. *American Journal of Data Science and Analytics*, *6*(10), 70-108. https://doi.org/10.63125/4h9v1876

[67]. Matthews, S., & Hartman, B. (2021). mshap: Shap values for two-part models. *Risks*, *10*(1), 3.

[68]. Maxwell, A. E., Warner, T. A., & Guillén, L. A. (2021). Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review. *Remote Sensing*, *13*(13), 2450.

[69]. Md Hasinur, R. (2026). Explainable Machine Learning for Loan-Delinquency Prediction in U.S. Community Banks: A Quantitative Study Using Gradient Boosting and SHAP. *American Journal of Data Science and Analytics*, *7*(01), 280-316. https://doi.org/10.63125/mfv3c059

[70]. Md Imran, H., & Jannatul, F. (2022). Molecular and Precision Rehabilitation Molecular Mechanisms Linking Mitochondrial Dysfunction, Inflammation, and Muscle Degeneration with Functional Decline and Rehabilitation Response in Neuromuscular Disease. *American Journal of Interdisciplinary Studies*, *12*(08), 01-44. https://doi.org/10.63125/0ym5bh79

[71]. Md Maksudul, H. (2024). Supplier-Risk Scoring and Procurement Resilience for Distributed Athletic-Apparel and Equipment Supply Networks. *American Journal of Advanced Technology and Engineering Solutions*, *14*(09), 44-90. https://doi.org/10.63125/a9fmgj50

[72]. Md Maksudul, H. (2025). Data-Driven Fulfillment Optimization for Time-Sensitive Teamwear and Sporting Goods Distribution. *American Journal of Scholarly Research and Innovation*, *4*(01), 1120–1173. https://doi.org/10.63125/rpezfz61

[73]. Md Maksudul, H. (2026). Generative AI Decision Support for Procurement, Inventory, And Operational Planning in U.S. Sports Businesses. *American Journal of Interdisciplinary Studies*, *16*(04), 01-55. https://doi.org/10.63125/ngtzw226

[74]. Md Maksudul, H., & Md Tazul Islam, C. (2022). Machine Learning-Based Cash-Flow Forecasting for Improving Financial Planning in U.S. Small and Mid-Sized Businesses. *International Journal of Scientific Interdisciplinary Research*, *12*(12), 01–38. https://doi.org/10.63125/w3pgjx97

[75]. Md Mohsin, B. (2026). Nurse-Led Triage Protocols in Emergency and Trauma Care: A PRISMA-Guided Systematic Review. *American Journal of Health and Medical Sciences*, *16*(09), 01–40. https://doi.org/10.63125/21k00924

[76]. Md Momen Shahriar, J. (2025). Viscoelastic Creep Characterization of LDPE Under Multilevel Stress and Temperature Using Time-Temperature Superposition and Constitutive Modeling. *American Journal of Scholarly Research and Innovation*, *15*(11), 01–38. https://doi.org/10.63125/2edstd86

[77]. Md Momen Shahriar, J. (2026). Physics-Informed Machine Learning for Long-Term Polymer Performance Prediction from Extrusion Parameters, Material Properties, and Creep Data. *American Journal of Interdisciplinary Studies*, *16*(05), 01-41. https://doi.org/10.63125/8gk5c668

[78]. Md Shahid, N. (2023). Digital Twin-Based Network Reliability Modeling for LAN And WAN Infrastructure Resilience. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 577–629. https://doi.org/10.63125/ahjvxj14

[79]. Md Shahid, N. (2024). Predictive Analytics Models for Network Traffic Optimization in Distributed Enterprise Environments. *Review of Applied Science and Technology*, *3*(01), 367-419. https://doi.org/10.63125/qm6cyf84

[80]. Md Shahid, N. (2025). Blockchain-Based Secure Data Routing Framework for U.S. Industrial Communication Networks. *American Journal of Scholarly Research and Innovation*, *4*(01), 990–1035. https://doi.org/10.63125/9exrgs41

[81]. Md Shahid, N. (2026a). AI-Driven Quantum-Inspired Predictive Analytics Framework for Procurement Optimization in Critical Infrastructure Networks. *American Journal of Interdisciplinary Studies*, *7*(01), 658-702. https://doi.org/10.63125/hzecta02

[82]. Md Shahid, N. (2026b). Secure Cyber-Physical Data Monitoring Architecture for Real-Time Fault Detection in U.S. Industrial Automation Systems. *American Journal of Data Science and Analytics*, *7*(10), 90-125. https://doi.org/10.63125/exzc2653

[83]. Md Tazul Islam, C. (2024). Explainable AI For Financial and Operational Decision Support: Improving Transparency and Managerial Trust. *American Journal of Advanced Technology and Engineering Solutions*, *14*(12), 01-38. https://doi.org/10.63125/kwfvzw68

[84]. Md Tazul Islam, C. (2025). Business Intelligence and Data-Quality Frameworks for Improving Evidence-Based Decisions across Multi-Source Enterprise Data. *American Journal of Scholarly Research and Innovation*, *15*(07), 01–41. https://doi.org/10.63125/qhh6ka59

[85]. Md Tazul Islam, C., & Ahmed, H. C. H. R. (2023). Predictive Operations Analytics for Resource Allocation and Service-Capacity Planning in Data-Driven Organizations. *Review of Applied Science and Technology*, *13*(02), 01–41. https://doi.org/10.63125/z3sm3089

[86]. Md. Mainuddin, F. (2024). Quantitative Structural Retrofit Assessment Models for Strengthening Existing Steel Buildings Under Increased Load Demands. *Review of Applied Science and Technology*, *3*(04), 325–366. https://doi.org/10.63125/yyqnte84

[87]. Md. Mainuddin, F. (2025). Advanced Engineering Materials Applications for Enhancing Durability and Lifecycle Performance of Steel Building Systems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 2406–2445. https://doi.org/10.63125/t9xvg986

[88]. Md. Mainuddin, F. (2026). AI-Enabled Performance-Based Structural Optimization of Steel Truss and Mezzanine Systems for Industrial Facilities. *American Journal of Data Science and Analytics*, *7*(02), 01-38. https://doi.org/10.63125/enhzbm16

[89]. Md. Morshedul, I., & Aminul Islam, M. (2022). A Quantitative Evaluation of Explainable Ensemble Learning for Probabilistic Risk Estimation and Decision Reliability in Enterprise Analytics. *American Journal of Data Science and Analytics*, *3*(02), 01-38. https://doi.org/10.63125/k22shk74

[90]. Nayak, J., Naik, B., Dash, P. B., Souri, A., & Shanmuganathan, V. (2021). Hyper-parameter tuned light gradient boosting machine using memetic firefly algorithm for hand gesture recognition. *Applied Soft Computing*, *107*, 107478.

[91]. Nefoussi, S., Amamra, A., & Amarouche, I. A. (2020). A comparative study of chest x-ray image enhancement techniques for pneumonia recognition. International Conference on Computing Systems and Applications,

[92]. Nguyen, T. T., Duong, C. M., & Narendran, S. (2021). CEO profile and earnings quality: TT Nguyen et al. *Review of Quantitative Finance and Accounting*, *56*(3), 987-1025.

[93]. Nhu, V.-H., Shirzadi, A., Shahabi, H., Singh, S. K., Al-Ansari, N., Clague, J. J., Jaafari, A., Chen, W., Miraki, S., & Dou, J. (2020). Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms. *International Journal of Environmental Research and Public Health*, *17*(8), 2749.

[94]. Nissim, D. (2022). Big data, accounting information, and valuation. *The Journal of Finance and Data Science*, *8*, 69-85.

[95]. Ogachi, D., Ndege, R., Gaturu, P., & Zoltan, Z. (2020). Corporate bankruptcy prediction model, a special focus on listed companies in Kenya. *Journal of Risk and Financial Management*, *13*(3), 47.

[96]. Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, *208*, 109244.

[97]. Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T., & Nguyen, H. Q. (2021). Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, *437*, 186-194.

[98]. Phonthanukitithaworn, C., Srisathan, W. A., Ketkaew, C., & Naruetharadhol, P. (2023). Sustainable development towards openness SME innovation: taking advantage of intellectual capital, sustainable initiatives, and open innovation. *Sustainability*, *15*(3), 2126.

[99]. Pirouz, B., Shaffiee Haghshenas, S., Shaffiee Haghshenas, S., & Piro, P. (2020). Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis. *Sustainability*, *12*(6), 2427.

[100]. Prager, R. P., Seiler, M. V., Trautmann, H., & Kerschke, P. (2022). Automated algorithm selection in single-objective continuous optimization: a comparative study of deep learning and landscape analysis methods. International Conference on Parallel Problem Solving from Nature,

[101]. Ramos-González, M. d. M., Rubio-Andrés, M., & Sastre-Castillo, M. Á. (2022). Effects of socially responsible human resource management (SR-HRM) on innovation and reputation in entrepreneurial SMEs. *International Entrepreneurship and Management Journal*, *18*(3), 1205-1233.

[102]. Sadia, A. (2025). Privacy-Aware Personalization: Machine Learning, Model Validation, And Consumer Trust in AI-Mediated Marketing. *American Journal of Scholarly Research and Innovation*, *4*(01), 1081–1119. https://doi.org/10.63125/j7tfr242

[103]. Sadia, A. (2026). A Data-Science Decision-Support Framework for Responsible AI-Enabled Marketing, Visualization, And Consumer Engagement in The United States. *American Journal of Interdisciplinary Studies*, *7*(02), 277-311. https://doi.org/10.63125/e67w9e54

[104]. Sadia, A., & Md. Mosheur, R. (2023). Consumer Trust, Perceived Authenticity, and Personalized Digital Marketing: A Quantitative Decision-Science Study. *International Journal of Scientific Interdisciplinary Research*, *4*(2), 252–295. https://doi.org/10.63125/cajkqt60

[105]. Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, *2*(7), 1308.

[106]. Sawhney, R., Malik, A., Sharma, S., & Narayan, V. (2023). A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. *Decision Analytics Journal*, *6*, 100169.

[107]. Schlegel, U., Vo, D. L., Keim, D. A., & Seebacher, D. (2021). Ts-mule: Local interpretable model-agnostic explanations for time series forecast models. Joint european conference on machine learning and knowledge discovery in databases,

[108]. Seihani, R., Gholami, H., Esmaeilpour, Y., Kamali, A., & Zareh, M. (2024). Interpretation techniques to explain the output of a spatial land subsidence hazard model in an area with a diverted tributary. *Applied Computing and Geosciences*, *23*, 100191.

[109]. Shabbir, A. (2025). Physics-Based Failure Prognostics for Preventing Pressure-System Degradation and Unplanned Industrial Shutdowns. *American Journal of Scholarly Research and Innovation*, *14*(09), 01–41. https://doi.org/10.63125/3asd0053

[110]. Shabbir, A. (2026). A Data-Science Digital Twin-Enabled Thermal Performance Optimization for Reducing Energy Losses in Industrial Steam Generation Systems. *American Journal of Data Science and Analytics*, *7*(01), 317-363. https://doi.org/10.63125/ybf40824

[111]. Shanmugasundar, G., Vanitha, M., Čep, R., Kumar, V., Kalita, K., & Ramachandran, M. (2021). A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. *Processes*, *9*(11), 2015.
[112]. Sheth, V., Tripathi, U., & Sharma, A. (2022). A comparative analysis of machine learning algorithms for classification purpose. *Procedia Computer Science*, *215*, 422-431.
[113]. Singh, P., Gupta, S., & Gupta, V. (2024). Multi-objective hyperparameter optimization on gradient-boosting for breast cancer detection. *International Journal of System Assurance Engineering and Management*, *15*(5), 1676-1686.
[114]. Singh, R., Bharti, V., Purohit, V., Kumar, A., Singh, A. K., & Singh, S. K. (2021). MetaMed: Few-shot medical image classification using gradient-based meta-learning. *Pattern Recognition*, *120*, 108111.
[115]. Song, X., Liu, X., Liu, F., & Wang, C. (2021). Comparison of machine learning and logistic regression models in predicting acute kidney injury: a systematic review and meta-analysis. *International journal of medical informatics*, *151*, 104484.
[116]. Soudi, M. S., & Bauters, M. (2024). AI guidelines and ethical readiness inside SMEs: A review and recommendations. *Digital Society*, *3*(1), 3.
[117]. Stabak, D., & Md Shahid, N. (2022). Empirical Assessment of Uncertainty-Aware Real-Time Decision-Making for Reliable and Resilient Cyber-Physical Systems. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 432–477. https://doi.org/10.63125/q3r4kk16
[118]. Sumaiya, M. (2024). Hospital Wastewater as A Reservoir of Antimicrobial Resistance Genes: Implications for Public Health and Genomic Surveillance in The United States. *Review of Applied Science and Technology*, *3*(04), 448–508. https://doi.org/10.63125/ahvas976
[119]. Sumaiya, M. (2026). Artificial Intelligence Driven Multi-Omics Approaches for Precision Medicine: From Molecular Discovery to Clinical Translation. *American Journal of Scholarly Research and Innovation*, *5*(01), 429–477. https://doi.org/10.63125/yfme0q12
[120]. Sumaiya Sadmeen, M. (2025). Spatiotemporal Graph Neural Networks and Digital Twins for Urban Drainage Networks: Predictive Inundation, Critical-Node Detection, And Infrastructure Prioritization. *American Journal of Scholarly Research and Innovation*, *15*(12), 01-40. https://doi.org/10.63125/awvjew66
[121]. Sumaiya Sadmeen, M. (2026). Uncertainty-Aware GeoAI Decision Support for Climate-Resilient and Equitable Urban Infrastructure Investment in U.S. Communities. *American Journal of Interdisciplinary Studies*, *16*(05), 42-84. https://doi.org/10.63125/ktww0w31
[122]. Sun, D., Xu, J., Wen, H., & Wang, D. (2021). Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest. *Engineering Geology*, *281*, 105972.
[123]. Swathy, M., & Saruladha, K. (2022). A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. *ICT express*, *8*(1), 109-116.
[124]. Syed Feroz, A., & Md Mahamud Pervaz, P. (2022). Root-Cause Analysis and FMEA-Driven Reliability Improvement in Gas Production Facilities: A Multi-Year Empirical Study. *International Journal of Scientific Interdisciplinary Research*, *12*(10), 01–53. https://doi.org/10.63125/70msp878
[125]. Syed Feroz, A., & Shabbir, A. (2022). Finite Element-Based Thermomechanical Optimization of Pressure Components for Extending Industrial Boiler Service Life. *Journal of Sustainable Development and Policy*, *12*(12), 01-51. https://doi.org/10.63125/x24pht28
[126]. Taher, M. A., Sikder, M. A., Ahmed, H., & Bashar, M. (2024). Integrating IOT and Big Data Analytics for Enhanced Supply Chain Performance in Industrial Engineering Sectors: A Cross-Market Study. *International Journal of Science and Engineering*, *Volume 1*, 1-14.
[127]. Umme, H. (2026). Transformer-Based Multi-Omics Integration for Cancer Drug Response Prediction: A PRISMA 2020 Systematic Review. *American Journal of Health and Medical Sciences*, *16*(01), 33–52. https://doi.org/10.63125/teryrs73
[128]. Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1), 44.
[129]. Wang, H. E., Weiner, J. P., Saria, S., Lehmann, H., & Kharrazi, H. (2024). Assessing racial bias in healthcare predictive models: Practical lessons from an empirical evaluation of 30-day hospital readmission models. *Journal of Biomedical Informatics*, *156*, 104683.
[130]. Zhang, G., Atasoy, H., & Vasarhelyi, M. A. (2022). Continuous monitoring with machine learning and interactive data visualization: An application to a healthcare payroll process. *International Journal of Accounting Information Systems*, *46*, 100570.
[131]. Zhang, Y., Sun, J., Yang, Z., & Wang, Y. (2020). Critical success factors of green innovation: Technology, organization and environment readiness. *Journal of cleaner production*, *264*, 121701.
[132]. Zhou, K., Yang, Y., Hospedales, T., & Xiang, T. (2020). Learning to generate novel domains for domain generalization. European conference on computer vision,
[133]. Zusman, M., Schumacher, C. S., Gassett, A. J., Spalt, E. W., Austin, E., Larson, T. V., Carvlin, G., Seto, E., Kaufman, J. D., & Sheppard, L. (2020). Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study. *Environment international*, *134*, 105329.