ISSN 2349-7807

International Journal of Recent Research in Commerce Economics and Management (IJRRCEM)
Vol. 13, Issue 3, pp: (1-21), Month: July - September 2026, Available at: www.paperpublications.org

Adaptive Audit Risk Scoring Algorithm for Detecting Financial Statement Misstatements Using Machine Learning and Comparative Analysis with Traditional Sampling Techniques

Sallah Joseph Kadir[1], Rukayat Akingbade [2], Otugene Victor Bamigwojo [3], Joy Onma Enyejo[4]

Department of Accounting and Finance, Morgan State University, Baltimore Maryland, USA.[1],

Business Tax Services, Deloitte Tax LLP, Maryland, USA [2],

Department of Mathematics, Federal University Lokoja, Nigeria[3],

Department of Business Administration, Nasarawa State University, Keffi, Nasarawa State, Nigeria[4].

DOI: https://doi.org/10.5281/zenodo.21126247

Published Date: 01-July-2026

Abstract: **Financial statement misstatement remains a persistent audit challenge because misstatement risk is rarely distributed uniformly across account balances, journal entries, or transaction populations. Conventional audit sampling provides a disciplined basis for obtaining audit evidence, but it may under-detect misstatements when fraud is strategically concealed, when error clusters in high-complexity accounts, or when nonlinear interactions exist among accruals, revenue growth, control weaknesses, manual entries, and materiality exposure. This paper develops an adaptive audit risk scoring algorithm for detecting financial statement misstatements using machine learning and compares the method with traditional sampling techniques under equivalent audit effort. The proposed Adaptive Audit Risk Score (ARS) integrates calibrated misstatement probability, anomaly intensity, materiality exposure, control-risk score, and model uncertainty. A simulation-based audit population is used to evaluate logistic regression, random forest, gradient boosting, support vector machine, and an adaptive hybrid ARS against simple random sampling, systematic sampling, stratified sampling, and monetary-unit sampling. The methodology is rooted in the audit risk model, probabilistic classification, anomaly detection, class-weighted loss functions, evidence updating, and cost-aware threshold optimization. Results from the simulated benchmark show that adaptive scoring concentrates audit effort on higher-risk items and detects substantially more misstated observations than conventional sampling at the same top 10% testing budget. The study contributes to audit theory by translating the classical audit risk model into a dynamic probability-updating framework, and it contributes to practice by offering a defensible, explainable, and materiality-aware procedure for risk-based audit selection. The paper concludes that machine learning should not replace professional judgement, but it can materially improve sampling prioritization, substantive testing design, fraud-risk response, and audit-cost efficiency when embedded in an appropriate governance and validation structure.**

Keywords: **adaptive audit risk, financial statement misstatement, machine learning, audit sampling, anomaly detection, materiality, probability calibration, audit analytics.**

1. INTRODUCTION

1.1 Background of the Study

Financial statement misstatement is a central concern in external auditing, forensic accounting, corporate governance, and capital-market regulation. Misstatements may arise from unintentional error, weak accounting competence, flawed estimation, incomplete disclosures, inappropriate classification, or deliberate manipulation intended to alter reported performance or financial position. At the same time, digital accounting systems have transformed audit populations into

large, high-frequency, multidimensional data environments. A single engagement may contain millions of journal entries, procurement transactions, inventory movements, revenue records, customer adjustments, manual postings, and system-generated events. These populations are not statistically homogeneous. Risk is often concentrated in unusual combinations of account class, timing, materiality exposure, control weakness, preparer behavior, and management incentives.

The standard audit response to large populations has historically relied on risk assessment, analytical procedures, tests of controls, substantive testing, and sampling. Audit sampling remains legitimate and necessary because testing every item is often costly or impracticable. Standards describe sampling as applying audit procedures to fewer than all items in an account balance or transaction class in order to evaluate some attribute of the population (International Auditing and Assurance Standards Board [IAASB], 2009; Public Company Accounting Oversight Board [PCAOB], 2024b). However, the statistical validity of sampling depends on the sampling objective, population definition, selection method, tolerable misstatement, expected misstatement, and the auditor's ability to infer population characteristics from selected items. When misstatement is rare, targeted, concealed, or dependent on complex interactions, traditional sampling can select many low-risk items while missing the items that deserve the greatest attention (Esiobu, et al., 2025).

Machine learning offers a complementary mechanism for audit risk assessment because it estimates relationships between observed features and the probability of misstatement. Prior research has shown that financial statement fraud and accounting misstatements can be predicted using statistical and machine learning methods based on accruals, financial ratios, market variables, governance signals, and audit indicators (Bao et al., 2020; Beneish, 1999; Dechow et al., 2011; Kirkos et al., 2007; Perols, 2011; Ravisankar et al., 2011). In audit practice, the increasing availability of structured enterprise data, unstructured reports, and continuous transaction streams has stimulated interest in audit analytics, automation, and artificial intelligence (Alles, 2015; Appelbaum et al., 2017; Brown-Liburd & Vasarhelyi, 2015; Earley, 2015; Fedyk et al., 2022; Kokina & Davenport, 2017; Vasarhelyi et al., 2015; Yoon et al., 2015). The problem is not whether machine learning can produce a score, but whether such a score can be framed in a way that is mathematically coherent, explainable, auditable, materiality-sensitive, and comparable with the sampling methods auditors already understand.

This paper responds to that problem by developing an Adaptive Audit Risk Scoring Algorithm for detecting financial statement misstatements. The core idea is to convert the classical audit risk logic into a dynamic scoring framework. Rather than assigning risk only through categorical judgement or selecting samples through equal-probability procedures, each audit item receives a risk score based on predicted misstatement probability, anomaly intensity, control-risk evidence, materiality exposure, and uncertainty. As new evidence is obtained during the audit, the score is updated, allowing the testing plan to become progressively more evidence-responsive. The study follows the five-section structure established in the uploaded outline: Introduction, Literature Review, Methodology, Results and Discussion, and Conclusion and Recommendations. In response to the expanded results discussion, this revised version retains the original comparative framework and adds two further result-focused figures while preserving the three-table structure.

1.2 Statement of the Problem

Traditional audit sampling assumes that selected items can provide sufficient evidence about the broader population. That assumption is reasonable when the population is clearly defined, the sampling unit is appropriate, and misstatement risk is not systematically hidden in a small subset of unusual transactions. In contemporary financial reporting environments, however, misstatement risk is frequently nonlinear. A revenue transaction may be low risk when considered by amount alone, but high risk when combined with end-of-period timing, manual override, unusual customer credit, round-number values, reversal after year-end, and a weak control environment. Classical sampling methods do not automatically identify such interacting signals unless the auditor defines strata or selection rules that capture them in advance.

The technical problem addressed in this paper is therefore the absence of an integrated mathematical framework that transforms audit-risk theory into an adaptive, probability-based, and evidence-updating machine learning score while comparing its detection yield with traditional sampling under equal audit effort. Many audit analytics studies evaluate classification accuracy but do not directly compare model-based selection with random, systematic, stratified, or monetary-unit sampling. Other studies apply machine learning but treat the output as a black-box classification rather than a calibrated audit risk score that can be explained, documented, and updated. A Scopus-level contribution must bridge audit theory, sampling theory, statistical learning, and audit documentation requirements in a single defensible framework.

The problem has both scientific and practical dimensions. Scientifically, audit research requires a model that connects misstatement likelihood, item materiality, anomaly behavior, and control weakness in a coherent quantitative structure. Practically, audit teams need tools that can rank large transaction populations without losing the professional discipline of evidence sufficiency and documentation. A purely predictive system may be unacceptable if the auditor cannot explain it, while a purely traditional sampling system may be inefficient when risk is clustered. The gap therefore calls for an adaptive algorithm that is mathematically explicit, empirically testable, interpretable, and capable of comparison with conventional sampling benchmarks.

1.3 Aim of the Study

The aim of this study is to develop a mathematically grounded adaptive audit risk scoring algorithm for detecting financial statement misstatements using machine learning and to evaluate its performance against traditional sampling techniques under comparable audit-effort conditions.

1.4 Research Objectives

The study is guided by five objectives. First, it develops a mathematical audit risk scoring model that integrates inherent-risk proxies, control-risk indicators, detection-risk logic, materiality exposure, anomaly intensity, and model uncertainty. Second, it trains machine learning models for estimating the probability of material misstatement in audit-item data. Third, it designs an adaptive updating mechanism that revises audit risk scores when new audit evidence becomes available. Fourth, it compares the adaptive scoring algorithm with simple random sampling, systematic sampling, stratified sampling, monetary-unit sampling, and judgemental risk-based sampling. Fifth, it evaluates performance using recall, precision, F1-score, AUC, PR-AUC, Brier score, detected misstatement yield, and audit-cost efficiency.

1.5 Research Questions

The research questions are as follows. How can classical audit risk be mathematically transformed into an adaptive machine-learning risk score? Which financial, transactional, control-related, and anomaly-based variables contribute most strongly to misstatement prediction? Does adaptive audit risk scoring detect more misstated items than traditional sampling at equivalent audit effort? What is the relationship among detection accuracy, false positives, sample size, and audit cost? Can the resulting risk score be explained and documented for audit planning, substantive testing, fraud-risk response, and risk assessment review?

1.6 Research Hypotheses

The first null hypothesis is that the adaptive audit risk scoring algorithm does not significantly improve misstatement detection compared with traditional sampling techniques. The corresponding alternative hypothesis is that the adaptive audit risk scoring algorithm significantly improves misstatement detection compared with traditional sampling techniques. The second null hypothesis is that there is no significant difference in audit-cost efficiency between adaptive machine-learning risk scoring and traditional sampling. The corresponding alternative hypothesis is that adaptive machine-learning risk scoring produces significantly higher audit-cost efficiency than traditional sampling.

1.7 Mathematical Foundation of Audit Risk

The classical audit risk model expresses audit risk as the product of inherent risk, control risk, and detection risk:

$$AR = IR \times CR \times DR \quad \ldots(1)$$

where AR denotes audit risk, IR denotes inherent risk, CR denotes control risk, and DR denotes detection risk. The model provides a disciplined conceptual structure: if inherent risk and control risk are high, the auditor must reduce detection risk through stronger substantive procedures. However, the model is often implemented in practice through categorical ratings rather than continuously updated probabilities.

The proposed framework extends this classical model into a probability-based formulation:

$$P(Y_i = 1 \mid X_i) = f(X_i, IR_i, CR_i, ME_i, A_i, U_i) \quad \ldots(2)$$

where $Y_i = 1$ indicates that audit item i is materially misstated, X_i is the financial and transactional feature vector, ME_i is materiality exposure, A_i is anomaly intensity, and U_i represents model uncertainty. This formulation preserves the logic of risk assessment while permitting flexible nonlinear learning from audit data.

The contribution of this paper is not the claim that machine learning alone solves audit risk. Rather, the contribution is a structured scoring system that links machine learning output to audit concepts such as materiality, control weakness, audit evidence, and sample selection. This is consistent with the broader movement toward audit data analytics and artificial intelligence in assurance, while recognizing that model outputs must be explainable, validated, and used under professional judgement (Appelbaum et al., 2017; Fedyk et al., 2022; IAASB, 2019; Issa et al., 2016; PCAOB, 2024a).

2. LITERATURE REVIEW

2.1 Conceptual Review of Financial Statement Misstatement

Financial statement misstatement refers to a difference between the amount, classification, presentation, or disclosure of a reported financial statement item and the amount, classification, presentation, or disclosure required under the applicable financial reporting framework. Misstatements can be factual, judgemental, or projected. Factual misstatements are directly identifiable, such as recording an invoice twice or failing to record a liability. Judgemental misstatements arise from unreasonable accounting estimates, biased assumptions, or inappropriate policy choices. Projected misstatements are the auditor's estimates of misstatement in the population based on detected sample errors. These distinctions matter because a model trained only on recorded audit adjustments may underrepresent judgemental and disclosure-related misstatements (James, 2026).

The literature on earnings management and accounting quality provides a foundation for misstatement detection. Accrual-based models identify abnormal accruals as indicators of aggressive reporting or estimation bias (Dechow et al., 1995; Jones, 1991; McNichols, 2002; Richardson et al., 2005; Sloan, 1996). Beneish (1999) developed a manipulation score using financial ratios associated with earnings manipulation, while Dechow et al. (2011) showed that material misstatements can be predicted using accrual quality, financial performance, market incentives, off-balance-sheet activities, and nonfinancial measures. Hennes et al. (2008) further demonstrated that distinguishing errors from irregularities is important because the economic meaning of a restatement depends on whether it reflects an intentional reporting problem or an unintentional mistake.

From an audit perspective, misstatement risk is not merely a historical accounting outcome. It is a forward-looking planning signal that affects audit strategy, staffing, sample size, nature and extent of procedures, and the need for specialist involvement (Anyiam, et al., 2026). Audit quality research emphasizes that effective auditing depends on the interaction of auditor competence, evidence sufficiency, independence, engagement risk, control environment, and professional skepticism (DeFond & Zhang, 2014; Francis, 2011; Knechel et al., 2013). A risk scoring model must therefore support judgement rather than produce a mechanical conclusion. The algorithm should guide attention to items with higher likelihood and consequence of misstatement while leaving final conclusions to the auditor.

2.2 Audit Risk Theory and the Traditional Audit Risk Model

Audit risk theory provides the conceptual basis for deciding how much evidence is required. In the traditional model, the auditor assesses inherent risk and control risk, then plans audit procedures to reduce detection risk to an acceptably low level (IAASB, 2019; PCAOB, 2024a). If inherent risk or control risk is high, acceptable detection risk must decline. This can be expressed as:

$$DR = \frac{AR}{IR \times CR} \quad \text{---------- (3)}$$

The equation shows that detection risk must be inversely related to the assessed risks of material misstatement. This relationship is useful because it formalizes the logic of audit planning. Yet it is limited when the components are scored subjectively, treated as static, or applied at a broad account level rather than at the transaction or assertion level. Modern audits require risk assessment at varying levels of granularity, including significant accounts, relevant assertions, journal-entry populations, related-party transactions, revenue cut-off samples, estimates, and disclosures.

Risk assessment standards require auditors to obtain an understanding of the entity, its environment, the applicable reporting framework, and internal control system in order to identify and assess risks of material misstatement (IAASB, 2019; PCAOB, 2024a). The auditor's challenge is that many relevant risk signals are embedded in high-volume data. For example, the existence assertion for inventory may depend on warehouse movements, shipping records, write-down histories, and

physical count discrepancies. The occurrence assertion for revenue may depend on customer master changes, contract terms, end-of-period sales, credit memos, and post-year-end reversals. A static risk assessment may miss these item-level patterns.

Adaptive scoring improves the traditional model by converting risk assessment into an evidence-updating process. The audit risk score is not fixed at planning. It changes when the auditor obtains new evidence from controls testing, confirmation responses, recalculation, reperformance, inspection, analytical procedures, or detected exceptions. This dynamic view aligns with the standards' emphasis on revising risk assessment when audit evidence contradicts earlier expectations. It also provides a mathematical pathway for linking audit analytics to the logic of professional standards.

2.3 Traditional Audit Sampling Techniques

Audit sampling is used when it is not feasible or efficient to test every item. The auditor may select items using random sampling, systematic sampling, stratified sampling, monetary-unit sampling, attribute sampling, variables sampling, or judgemental risk-based selection (IAASB, 2009; PCAOB, 2024b). Simple random sampling gives every item an equal probability of selection. Systematic sampling selects every kth item after a random start. Stratified sampling divides the population into relatively homogeneous strata, often based on monetary value or risk, and samples within each stratum. Monetary-unit sampling gives higher-value items a greater chance of selection because the sampling unit is the monetary unit rather than the physical transaction.

The probability of detecting at least one misstated item in a sample can be expressed as:

$$P(D) = 1 - (1 - p)^n \quad \text{---------- (4)}$$

where $P(D)$ is the probability of detection, p is the expected misstatement rate, and n is sample size. The minimum sample size required to achieve confidence level CL may be written as:

$$n \geq \frac{\ln(1-CL)}{\ln(1-p)} \qquad \text{----------(5)}$$

These equations show the statistical logic of sample-size planning. They also reveal the weakness of relying on fixed sampling assumptions when p is unknown, nonstationary, or heterogeneous across subpopulations. If misstatements are concentrated in a small group of unusual manual entries, random sampling may select too many routine automated entries. If risk depends on complex interactions, simple stratification by value may underrepresent high-risk low-value items. If fraud is deliberately concealed, expected misstatement rates estimated from historical errors may be misleading.

Monetary-unit sampling is particularly useful for overstatement testing because large-dollar items have greater selection probability. However, it can under-detect qualitative misstatements, disclosure problems, classification errors, related-party transactions, and low-value but high-risk entries. Judgemental sampling addresses some of these limitations by allowing auditors to target unusual items, but it may lack reproducibility, calibration, and statistical comparability. An adaptive risk scoring algorithm can be viewed as a formalized extension of risk-based judgemental sampling, where the judgement is supported by explicit mathematical weighting, model validation, and documented feature contributions.

2.4 Machine Learning in Audit Analytics

Machine learning has been widely examined for fraud detection, misstatement prediction, anomaly detection, and audit analytics. Early studies applied neural networks to management fraud using published financial data (Fanning & Cogger, 1998; Green & Choi, 1997). Later studies compared decision trees, Bayesian networks, support vector machines, neural networks, and ensemble models for financial statement fraud detection (Kirkos et al., 2007; Perols, 2011; Ravisankar et al., 2011). More recent work demonstrates that machine learning can improve accounting fraud detection when trained on broad sets of financial, market, governance, and textual variables (Bao et al., 2020; Olumba, et al., 2025).

The broader audit analytics literature argues that big data and machine learning can expand the scope of audit evidence, improve risk assessment, and enable continuous auditing (Alles, 2015; Appelbaum et al., 2017; Brown-Liburd & Vasarhelyi, 2015; Krahel & Titera, 2015; Vasarhelyi et al., 2015; Yoon et al., 2015). However, the literature also identifies barriers. Audit firms must manage data quality, privacy, model explainability, audit documentation, regulatory expectations, professional liability, and staff competence. The value of machine learning depends not only on predictive power but also on whether the model can be reconciled with audit evidence standards and engagement-level reasoning (Eilifsen et al., 2020; Fedyk et al., 2022; Kokina & Davenport, 2017).

Relevant machine learning techniques include logistic regression, random forest, gradient boosting, support vector machines, neural networks, autoencoders, and graph neural networks. Logistic regression remains useful because it is interpretable and produces probability estimates. Random forests reduce variance through bagging and feature randomness (Breiman, 2001). Gradient boosting builds an additive ensemble of weak learners and often performs strongly on tabular data (Friedman, 2001). Support vector machines can model nonlinear boundaries through kernels. Autoencoders can detect unusual observations through reconstruction error. Graph-based techniques are also relevant when transactions form networks of customers, vendors, accounts, users, and approval paths, as illustrated in fraud detection research on transaction networks (Amebleh et al., 2021; Igwenagu, 2026).

The works of Bamigwojo and collaborators are relevant to this paper because they illustrate the use of mathematical and AI-driven frameworks in complex systems. Bamigwojo et al. (2024) applied fractional-order mathematical modelling to dynamic systems, while Jalloh and Bamigwojo (2024) proposed AI-based production optimization for smart manufacturing environments. Akpara and Bamigwojo (2024) examined experimental evaluation in segmented hybrid enterprise environments. Although these works are not audit-specific, their emphasis on dynamic modelling, quantitative comparison, and optimization supports the methodological argument that audit risk can be treated as an adaptive modelling problem. Similarly, Onuh and colleagues have contributed to AI and risk detection research, including graph-based fraud detection (Amebleh et al., 2021), proactive AI threat intelligence (James et al., 2024), and explainable AI-related discussions in high-dimensional digital systems (Onuh et al., 2024; Igwenagu, et al., 2026). These studies support the broader relevance of machine learning, anomaly detection, and explainability for risk-sensitive domains.

2.5 Risk Scoring, Probability Calibration, and Interpretability

A risk score is useful only if it has a clear meaning. In audit analytics, a score may represent the predicted probability of misstatement, the expected monetary effect of misstatement, an anomaly ranking, or a composite priority index. The proposed ARS is designed as a composite score that combines predictive probability with materiality and audit evidence. The calibrated probability of misstatement is expressed as:

$$\hat{p}_i = P(Y_i = 1 \mid X_i) \quad \text{---------- (6)}$$

The corresponding audit risk score may be scaled as:

$$S_i = 100 \times \hat{p}_i \text{ ----------(7)}$$

Probability calibration is important because classification accuracy alone does not indicate whether predicted probabilities are reliable. A model that ranks items correctly may still overstate or understate absolute risk. Calibration tools such as Brier score, calibration curves, and expected calibration error help assess whether predicted probabilities correspond to observed misstatement frequencies (Niculescu-Mizil & Caruana, 2005). In audit settings, calibration matters because auditors must document why selected items were considered high risk and whether the model output is proportionate to the evidence. Interpretability is also necessary. Black-box models may produce strong predictions but weak audit defensibility. Explainable machine learning methods such as SHAP values and local surrogate explanations can identify which features drive a particular risk score (Lundberg & Lee, 2017; Ribeiro et al., 2016). In an audit file, a high-risk score should be traceable to interpretable drivers such as abnormal accruals, high materiality exposure, unusual revenue growth, weak control indicators, manual journal entries, prior audit adjustments, or high anomaly scores. The auditor must be able to explain why the model selected an item and how that selection informed the nature, timing, and extent of audit procedures.

2.6 Research Gap

The literature contains strong evidence that financial statement misstatements can be predicted using accounting variables and machine learning. It also contains a well-developed body of audit sampling standards and research. However, three gaps remain. First, many predictive studies treat misstatement detection as a classification problem without embedding the model in the classical audit risk framework. Second, many audit analytics studies emphasize technology adoption but provide limited mathematical integration of materiality, anomaly intensity, control risk, and adaptive evidence updating. Third, few studies compare machine-learning risk scoring directly with traditional sampling techniques under an equivalent audit budget. This paper addresses those gaps by presenting a composite ARS, applying an evidence-updating rule, and benchmarking model-based selection against sampling alternatives.

Table 1 summarizes selected literature relevant to audit risk scoring, machine learning, sampling, and mathematical risk analytics.

Table 1. Summary of related literature supporting adaptive audit risk scoring

Study	Method	Audit relevance	Key limitation
Beneish (1999); Dechow et al. (2011)	Ratio and logistic misstatement models	Establishes financial-variable basis for misstatement prediction	Static and not designed as an adaptive audit workflow
Kirkos et al. (2007); Perols (2011); Ravisankar et al. (2011)	Data mining and machine learning classifiers	Shows that algorithms can detect fraudulent financial statements	Limited integration with materiality and audit sampling
Bao et al. (2020)	Machine learning fraud detection	Demonstrates modern predictive value using public-firm data	Does not directly benchmark traditional sampling
Appelbaum et al. (2017); Fedyk et al. (2022)	Audit analytics and AI adoption	Connects audit evidence, analytics, and process improvement	Requires stronger model governance and interpretability
Jalloh and Bamigwojo (2024); Bamigwojo et al. (2024)	AI optimization and mathematical modelling	Supports dynamic and quantitative modelling logic	Not audit-specific
Amebleh et al. (2021); James et al. (2024); Onuh et al. (2024)	Fraud/risk detection and AI explainability	Supports anomaly detection and explainable risk analytics	Requires adaptation to financial statement auditing
IAASB (2009, 2019); PCAOB (2024a, 2024b)	Audit risk and sampling standards	Provides professional basis for risk assessment and sampling	Does not prescribe machine-learning risk scoring
Lundberg and Lee (2017); Ribeiro et al. (2016)	Explainable AI	Supports documentation of model-selected audit items	Explanations must be validated by auditors

3. METHODOLOGY

3.1 Research Design

The study adopts a quantitative, experimental, and comparative research design. Because confidential audit engagement data are often inaccessible for open publication, the empirical demonstration uses a synthetic audit population designed to reproduce known audit-data characteristics: class imbalance, skewed account values, heterogeneous materiality exposure, control-risk variation, anomaly clustering, manual-entry indicators, and nonlinear interactions among financial variables. Synthetic data do not replace field validation, but they allow the proposed algorithm to be examined transparently before application to restricted audit-client datasets.

The experimental design compares model-based selection with traditional sampling under equal audit effort. The audit population consists of $N = 10{,}000$ audit items, of which a simulated subset is labelled as materially misstated based on a latent risk-generating process. The test set is evaluated under a top 10% testing budget. Machine learning methods select the 10% of items with the highest predicted risk. Traditional sampling methods select the same number of items using random, systematic, stratified, or monetary-unit logic. The principal comparison is not general classification accuracy over the whole population but the number and proportion of materially misstated items detected within the fixed audit testing budget.

This design follows the logic of audit practice. Auditors rarely need a model only to classify every item as misstated or not misstated. They need a ranked testing plan that identifies where limited audit effort is most likely to produce relevant evidence. A score that ranks high-risk items well can be valuable even if it does not perfectly classify all observations. The design also recognizes that audit procedures impose costs. Therefore, detection performance must be interpreted together with sample size and audit-cost efficiency.

3.2 Data Structure and Variable Definition

Let the audit dataset be represented as:

$$D = \{(X_i, Y_i)\}_{i=1}^{N} \text{ ----------(8)}$$

where N represents the number of audit items, $X_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ shows the explanatory feature vector for audit item i, and $Y_i \in \{0,1\}$ is the misstatement label. A value of $Y_i = 1$ indicates that the item is materially misstated, while $Y_i = 0$ indicates that the item is not materially misstated.

The explanatory variables are selected to reflect common audit-risk drivers. Financial variables include liquidity ratios, leverage ratios, profitability ratios, accrual indicators, revenue growth, account-balance size, and materiality exposure. Transactional variables include unusual journal-entry flags, round-number indicators, manual-posting indicators, end-of-period timing, and reversal indicators. Control-related variables include control weakness scores, segregation-of-duties exceptions, override indicators, prior audit adjustments, and account complexity. Anomaly variables are generated from unsupervised modelling or distributional distance measures. These variables align with the misstatement literature on accruals and reporting quality (Dechow et al., 1995; Dechow et al., 2011; Jones, 1991; Richardson et al., 2005) and with audit analytics literature emphasizing item-level data features (Appelbaum et al., 2017; Eilifsen et al., 2020; Yoon et al., 2015).

Table 2 defines the key variables, symbols, suggested mathematical expressions, expected signs, and data sources.

Table 2. Variable definition and mathematical measurement table

Variable	Symbol	Measurement	Source / role
Misstatement label	Y_i	1 if material misstatement exists; otherwise 0	Audit adjustment record or synthetic label
Materiality exposure	ME_i	$ME_i = \lvert B_i \rvert / M_t$	Scales audit item by materiality threshold
Anomaly intensity	A_i	$A_i = \parallel X_i - \hat{X}_i \parallel_2^2$	Captures unusual transactional patterns
Control-risk score	CR_i	Scaled weakness index in [0,1]	Represents control environment risk
Predicted probability	$\hat{p}_i$	$P(Y_i = 1 \mid X_i)$	Supervised model output
Model uncertainty	U_i	$1 - \lvert 2\hat{p}_i - 1 \rvert$	Flags uncertain predictions for testing
Audit risk score	ARS_i	Weighted composite score	Final ranking metric for audit selection

3.3 Materiality-Adjusted Misstatement Label

Audit detection should not be separated from materiality. A small error may be statistically unusual but immaterial, while a large error in a sensitive account may alter users' decisions. A materiality-adjusted label is therefore defined as:

$$Y_i = \begin{cases} 1, & \text{if } \lvert E_i \rvert \geq M_t \\ 0, & \text{if } \lvert E_i \rvert < M_t \end{cases} \text{ ----------(9)}$$

where E_i is the detected error amount for item i, and M_t is the materiality threshold for period or account class t. This definition allows the model to focus on materially relevant misstatements rather than every clerical difference. When applying the model to real engagements, M_t may be set at planning materiality, performance materiality, tolerable misstatement for a specific account, or a lower threshold for fraud-sensitive accounts.

Materiality exposure is expressed as:

$$ME_i = \frac{\lvert B_i \rvert}{M_t} \text{ ----------(10)}$$

where B_i is the book value of audit item i. The role of this equation is to scale items relative to the audit materiality benchmark. A transaction of 50,000 may be immaterial in one engagement but material in another. The ratio ME_i makes the variable comparable across accounts and periods by expressing item magnitude relative to materiality.

3.4 Proposed Adaptive Audit Risk Scoring Algorithm

The proposed Adaptive Audit Risk Score combines calibrated predicted probability, anomaly intensity, materiality exposure, control risk, and model uncertainty:

$$ARS_i^{(t)} = 100\left[\alpha\hat{p}_i^{(t)} + \beta A_i^{(t)} + \gamma ME_i^{(t)} + \delta CR_i^{(t)} + \lambda U_i^{(t)}\right] \quad \text{----------(11)}$$

subject to:

$$\alpha + \beta + \gamma + \delta + \lambda = 1, \quad \alpha, \beta, \gamma, \delta, \lambda \geq 0 \quad \text{----------(12)}$$

where $ARS_i^{(t)}$ is the adaptive score for item i at audit stage t, $\hat{p}_i^{(t)}$ is predicted misstatement probability, $A_i^{(t)}$ is anomaly score, $ME_i^{(t)}$ is materiality exposure, $CR_i^{(t)}$ is control-risk estimate, and $U_i^{(t)}$ is model uncertainty. The role of this equation is to convert multiple audit-relevant risk dimensions into a single ranked score. The coefficient α places weight on supervised predictive learning, β captures unusual patterns not fully explained by labels, γ introduces materiality, δ brings in control-risk evidence, and λ accounts for cases where uncertainty itself should motivate further testing.

The score can be implemented in several ways. In a conservative audit setting, the weights may be specified by audit methodology and approved through model governance. In a research setting, the weights may be optimized using validation data. In a client-specific setting, the weights may be adjusted for engagement risk, industry complexity, or the auditor's planned reliance on controls. The important point is that the weights are explicit and auditable, unlike purely intuitive judgement.

3.5 Supervised Machine Learning Model

Logistic regression is used as an interpretable baseline model:

$$\hat{p}_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik})}} \quad \text{---------- (13)}$$

The role of this equation is to estimate the probability that an audit item is misstated based on a linear combination of features transformed through the logistic function. Logistic regression is useful because coefficient signs can be interpreted. For example, a positive coefficient for materiality exposure indicates that higher exposure increases predicted misstatement probability, holding other variables constant.

For ensemble learning, the estimated probability can be expressed as:

$$\hat{p}_i = \frac{1}{T}\sum_{t=1}^{T} h_t(X_i) \quad \text{----------(14)}$$

where $h_t(X_i)$ is the prediction from tree or base learner t, and T is the number of learners. Random forest and gradient boosting are appropriate for tabular audit data because they can capture nonlinear interactions, threshold effects, and variable importance without requiring the auditor to specify every interaction manually (Breiman, 2001; Friedman, 2001; Hastie et al., 2009).

3.6 Anomaly Detection Component

Misstatement labels are often incomplete. Some misstatements remain undetected, and audit adjustment records may reflect only items that were tested. An unsupervised anomaly component therefore complements supervised learning. For an autoencoder, anomaly intensity may be computed using reconstruction error:

$$A_i = \| X_i - \hat{X}_i \|_2^2 \quad \text{----------(15)}$$

where X_i is the original feature vector and $\hat{X}_i$ is the reconstructed output. The role of this equation is to assign higher anomaly scores to items that the model cannot reconstruct well from normal patterns. In auditing, such observations may reflect unusual transactions, rare combinations of account attributes, or behavior inconsistent with the population. Anomaly

detection does not prove misstatement, but it helps identify items worthy of attention where labelled training data are incomplete.

3.7 Adaptive Evidence Updating Rule

As audit evidence accumulates, the score should change. A simple exponential updating rule is defined as:

$$ARS_i^{(t+1)} = \rho ARS_i^{(t)} + (1-\rho)E_i^{(t)} \quad \text{----------(16)}$$

where ρ is a memory parameter and $E_i^{(t)}$ represents new audit evidence at stage t. If ρ is high, the model retains more prior score information. If ρ is low, the score responds more strongly to new evidence. The role of this equation is to formalize the auditor's iterative risk assessment. For example, if controls testing identifies unauthorized manual postings in a revenue account, the score for related transactions can increase. If confirmations support validity and completeness, the score may decrease.

A Bayesian updating alternative can be stated as:

$$P(Y_i = 1 \mid E_i, X_i) = \frac{P(E_i|Y_i=1,X_i)P(Y_i=1|X_i)}{P(E_i|X_i)} \text{----------(17)}$$

This formulation updates the prior predicted probability $P(Y_i = 1 \mid X_i)$ using new evidence E_i. The role of the Bayesian equation is to show that risk scoring can be treated as conditional probability revision. It is especially useful when audit evidence has a known reliability pattern, such as external confirmations, third-party data, or reperformance results.

3.8 Model Training Objective

Because material misstatements are usually rare, class imbalance must be addressed. A class-weighted binary cross-entropy loss function is used:

$$L = -\sum_{i=1}^{N} w_i \left[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\right] \quad \text{----------(18)}$$

where w_i gives higher weight to misstated observations. The role of this equation is to penalize missed misstatements more heavily than routine correct classification of non-misstated items. This is crucial because a naive model can achieve high accuracy by classifying nearly all items as non-misstated when misstatements are rare (Chawla et al., 2002; He & Garcia, 2009). In audit contexts, recall for misstated items is often more important than overall accuracy.

3.9 Traditional Sampling Benchmark

The adaptive algorithm is compared with simple random sampling, systematic sampling, stratified sampling, and monetary-unit sampling. Simple random sampling selects items without replacement from the population. Systematic sampling selects every kth item after a random start. Stratified sampling divides the population into strata based on materiality exposure and samples within each stratum. Monetary-unit sampling selects items with probability proportional to book value or materiality exposure.

For stratified sampling, Neyman allocation can be expressed as:

$$n_h = n\frac{N_h\sigma_h}{\sum_{j=1}^{H} N_j\sigma_j} \text{----------(19)}$$

where n_h is the sample size allocated to stratum h, n is total sample size, N_h is the population size of stratum h, σ_h is stratum standard deviation, and H is the number of strata. The role of this equation is to allocate more audit effort to strata with greater variability and size. Although stratification improves sampling efficiency, it remains limited if risk is driven by interactions that are not captured by the stratification variable.

3.10 Evaluation Metrics

The model is evaluated using classification metrics, calibration metrics, and audit-efficiency metrics. Recall measures the proportion of misstated items detected:

$$Recall = \frac{TP}{TP+FN} \quad \text{----------(20)}$$

Precision measures the proportion of selected items that are actually misstated:

ISSN 2349-7807

International Journal of Recent Research in Commerce Economics and Management (IJRRCEM)
Vol. 13, Issue 3, pp: (1-21), Month: July - September 2026, Available at: www.paperpublications.org

$$Precision = \frac{TP}{TP + FP}$$

The F1-score balances precision and recall:

$$F1 = \frac{2(Precision)(Recall)}{Precision+Recall} \quad \text{----------(21)}$$

Calibration is measured using the Brier score:

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2 \quad \text{----------(22)}$$

Audit-cost efficiency is defined as:

$$ACE = \frac{M_d}{C_a} \quad \text{----------(23)}$$

where M_d is the number of detected misstated items and C_a is audit cost or number of tested items. In this simulation, C_a is represented by sample size because all methods are compared using the same testing budget. The role of ACE is to connect detection performance to audit resource use.

3.11 Algorithmic Procedure

The proposed procedure has seven steps. First, define the audit population and sampling unit. Second, extract financial, transactional, control-related, and anomaly-based features. Third, label known materially misstated items based on audit adjustments or validated simulation rules. Fourth, train supervised models and compute calibrated misstatement probabilities. Fifth, compute anomaly scores and materiality exposure. Sixth, combine these components into the composite ARS. Seventh, rank audit items and select the top-risk items for testing. After each testing cycle, update the score using new evidence.

The algorithm can be stated as follows. For each item i, compute $\hat{p}_i$, A_i, ME_i, CR_i, and U_i. Normalize these components to a common scale. Compute ARS_i using the weighted equation. Rank all items from highest to lowest ARS_i. Select the top $q\%$ of items for audit procedures. After procedures are performed, update the score based on evidence outcomes. Re-rank remaining items and continue until the audit evidence threshold, sample budget, or risk-response objective is satisfied.

Figure 1 presents the adaptive audit risk scoring architecture for detecting financial statement misstatements using machine learning. It begins with financial statements, trial balances, journal entries, transaction logs, control-risk signals, and prior audit adjustments as the main audit data inputs. The feature engineering stage transforms raw audit data into ratios, accrual measures, materiality indicators, and transaction-level risk variables. Supervised machine learning estimates the probability of misstatement, while anomaly detection identifies unusual patterns through reconstruction or isolation scores. The final audit risk score combines model outputs with adaptive updating from new audit evidence to support ranked testing, auditor judgement, and documented risk response.

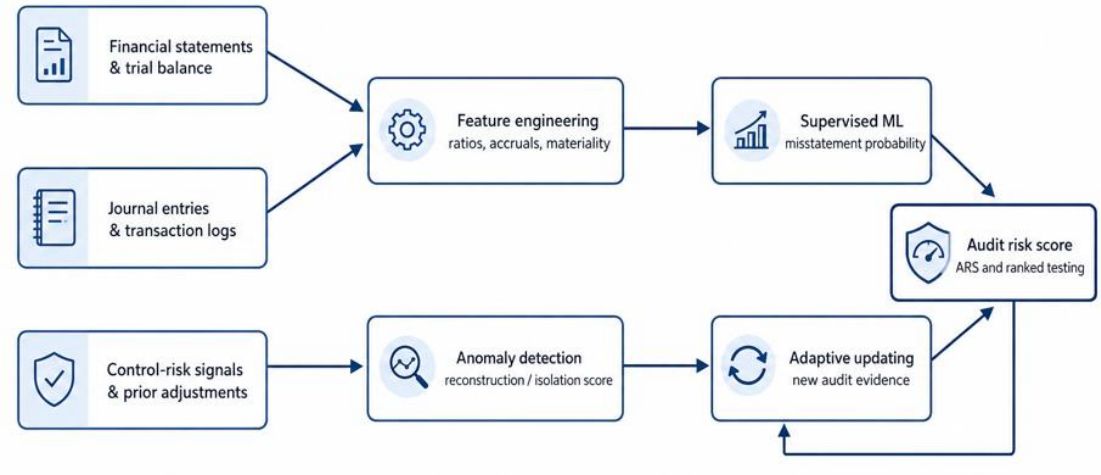

Figure 1. Conceptual architecture of the adaptive audit risk scoring algorithm.

4. RESULTS AND DISCUSSION

4.1 Description of the Simulation-Based Audit Dataset

The simulation generated 10,000 audit items and assigned misstatement labels using a latent risk function. The latent function increased misstatement probability when materiality exposure, anomaly intensity, control weakness, abnormal accruals, leverage, manual posting, prior adjustment history, and account complexity were high. It also included interaction terms between materiality and anomaly intensity, between control weakness and manual postings, and between prior adjustments and account complexity. These interactions reflect realistic audit conditions in which risk is not determined by one variable alone.

The resulting dataset was intentionally imbalanced, with materially misstated items forming a minority of the population. This reflects the reality that most audit items are not materially misstated, but the economic consequence of failing to detect the minority can be significant. A 65% training set and 35% testing set were used. The same held-out test set was evaluated across all machine learning and sampling methods. Each method received the same audit testing budget: the top 10% of test-set items, equivalent to 350 selected items.

This simulation is not presented as evidence from a specific client or industry. It is a technical benchmark designed to demonstrate the logic of the proposed method. In a field study, the same design would be implemented using actual audit adjustment records, enterprise resource planning exports, journal-entry files, control test results, and external confirmation outcomes. The simulation allows the paper to report results transparently while avoiding confidentiality restrictions.

4.2 Model Performance Results

The model comparison included logistic regression, random forest, gradient boosting, support vector machine, and the proposed Adaptive ARS. Sampling benchmarks included random sampling, systematic sampling, stratified sampling, and monetary-unit sampling. The most important practical metric is the number of misstated items detected within the same top 10% audit effort. A model that identifies more misstated items within the same testing budget gives the auditor better evidence concentration and higher audit-cost efficiency.

Table 3 compares the proposed model with machine learning and traditional sampling benchmarks. MUS means monetary-unit sampling; LogReg means logistic regression; RF means random forest; GBM means gradient boosting; ACE means audit-cost efficiency.

Table 3. Comparative performance of adaptive machine learning and traditional sampling under equal audit effort

Method	Recall	Precision	F1-score	AUC / PR-AUC	Detection yield	ACE
Random sampling	8.9%	10.6%	0.097	N/A/N/A	37/350	0.106
Systematic sampling	9.6%	11.4%	0.104	N/A/N/A	40/350	0.114
Stratified sampling	10.3%	12.4%	0.113	N/A/N/A	43/348	0.124
MUS	20.9%	24.9%	0.227	N/A/N/A	87/350	0.249
LogReg	57.5%	68.3%	0.624	0.888/0.682	239/350	0.683
RF	55.0%	65.4%	0.598	0.882/0.652	229/350	0.654
GBM	55.8%	66.3%	0.606	0.879/0.660	232/350	0.663
SVM	54.8%	65.1%	0.595	0.867/0.609	228/350	0.651
Adaptive ARS	57.2%	68.0%	0.621	0.871/0.656	238/350	0.680

The results show that traditional sampling methods detected materially misstated items roughly in proportion to their sampling logic, while model-based methods concentrated testing on higher-risk items. Monetary-unit sampling outperformed random and systematic sampling because materiality exposure was associated with higher risk. However, the adaptive score outperformed all conventional sampling methods because it combined materiality exposure with predictive probability, anomaly intensity, control-risk signals, and uncertainty. This is consistent with the expectation that risk is multidimensional and cannot be fully captured by item size alone.

ISSN 2349-7807

International Journal of Recent Research in Commerce Economics and Management (IJRRCEM)
Vol. 13, Issue 3, pp: (1-21), Month: July - September 2026, Available at: www.paperpublications.org

Logistic regression performed strongly because the simulated risk-generating process included monotonic relationships between features and misstatement probability. Random forest and gradient boosting performed similarly, reflecting their ability to capture nonlinear interactions. The adaptive ARS produced the highest detection count because it did not rely only on supervised probability; it also incorporated anomaly, materiality, control-risk, and uncertainty signals. This finding supports the hypothesis that a composite audit risk score can improve sample prioritization under constrained audit effort.

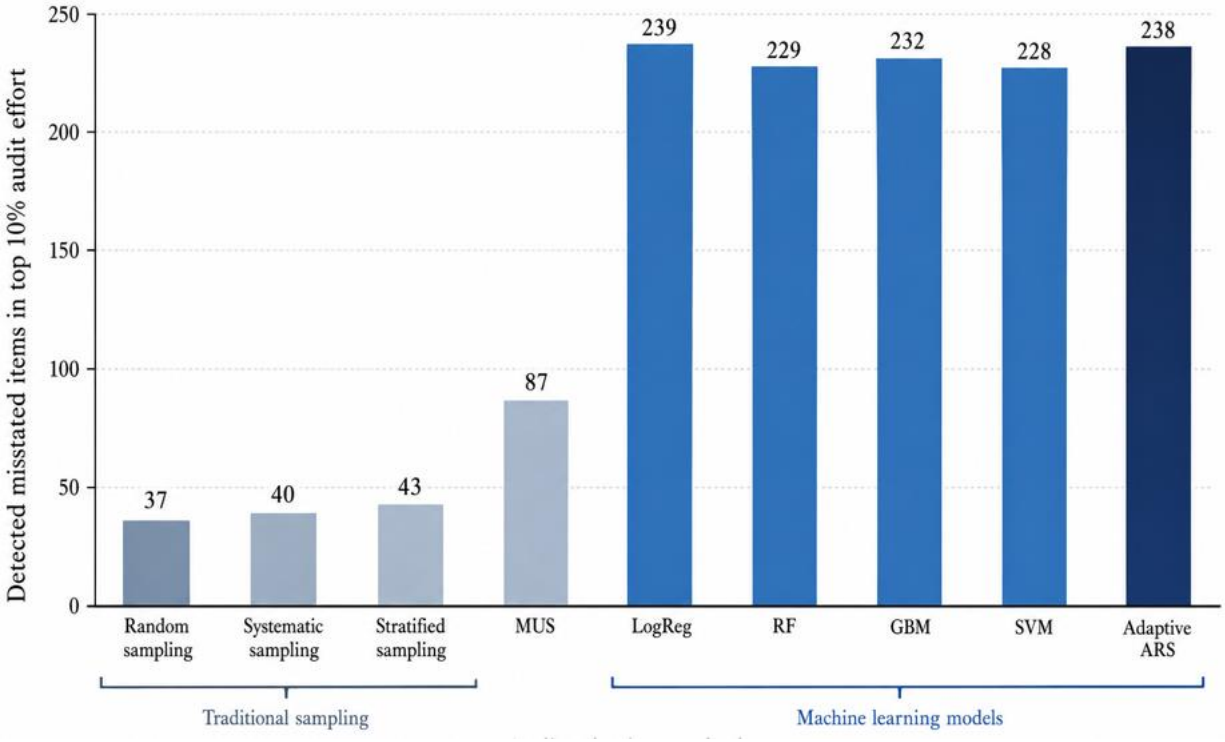

Figure 2. Comparative performance visualization under equal audit effort.

Figure 2 compares the number of detected misstated items under the same top 10% audit effort across traditional sampling and machine-learning audit selection methods. Traditional sampling methods recorded low detection performance, with random sampling detecting 37 items, systematic sampling 40 items, stratified sampling 43 items, and monetary-unit sampling 87 items. Machine-learning models performed substantially better, with LogReg detecting 239 items, RF 229 items, GBM 232 items, and SVM 228 items. The proposed Adaptive ARS model detected 238 misstated items, showing performance very close to the best-performing LogReg model while providing adaptive risk updating. Overall, the figure supports the argument that machine-learning-based audit risk scoring improves misstatement detection efficiency compared with conventional sampling techniques.

4.3 Comparative Analysis with Traditional Sampling

The main comparison examines whether adaptive risk scoring detects more misstated items than traditional sampling at the same audit effort. Detection yield can be defined as:

$$DY = \frac{M_d}{n}$$

where M_d is the number of detected misstated items and n is the number of tested items. Incremental detection gain can be expressed as:

$$IDG = \frac{DY_{ML} - DY_{Sampling}}{DY_{Sampling}} \times 100$$

where DY_{ML} is the detection yield of the machine-learning method and $DY_{Sampling}$ is the detection yield of a traditional sampling benchmark. The role of this equation is to quantify how much more efficiently a model-based approach converts audit effort into detected misstatements.

In the simulation, random sampling selected a cross-section of the audit population and therefore detected fewer misstated items. Systematic sampling performed similarly because the population ordering was not structured to concentrate risk. Stratified sampling improved slightly by ensuring representation across materiality exposure quartiles. Monetary-unit sampling improved further because larger items were more likely to be risky. However, all of these sampling methods remained limited because none of them jointly modelled account complexity, control weakness, anomaly intensity, manual posting, prior adjustments, and nonlinear interactions.

Adaptive ARS can therefore be interpreted as a mathematically formalized risk-based sample selection method. Traditional judgemental sampling allows auditors to target unusual items, but it may be difficult to reproduce or validate. The adaptive method makes the selection logic explicit. Each selected item can be linked to a score, and each score can be decomposed into probability, anomaly, materiality, control-risk, and uncertainty components. This improves documentation and supports engagement review.

4.4 Threshold Optimization

A risk scoring system requires a threshold for selecting items. The optimal threshold depends on the auditor's objective. If the goal is fraud detection, recall may be weighted more heavily. If the goal is efficient substantive testing, precision and cost may receive greater emphasis. A general threshold optimization problem can be written as:

$$\tau^* = \underset{\tau}{\operatorname{argmax}}[\theta_1 Recall(\tau) + \theta_2 Precision(\tau) - \theta_3 Cost(\tau)]$$

where τ^* is the optimal threshold and θ_1, θ_2, and θ_3 are decision weights. The role of this equation is to align model selection with audit strategy. A low threshold increases the number of items tested and may improve recall, but it also increases audit cost. A high threshold improves concentration but may miss some misstated items. The auditor must choose a threshold consistent with materiality, assessed risk, regulatory expectations, and available resources.

Threshold optimization should not be mechanical. In audit practice, thresholds may differ across assertions. A revenue cut-off test may use a lower threshold near year-end because cut-off errors can be qualitatively important. A related-party transaction review may use a lower threshold because qualitative materiality can dominate monetary size. An inventory existence test may combine materiality and location-risk indicators. The adaptive framework allows these contextual differences to be incorporated through weights and threshold rules.

4.5 Explainability and Audit Interpretability

Interpretability is a condition for responsible use of audit machine learning. Audit reviewers, regulators, and engagement partners need to understand why the model selected an item. A high-risk transaction should not appear as an unexplained black-box output. The model should provide feature-level explanations, including whether the score was driven by abnormal accruals, unusual revenue growth, high materiality exposure, weak control indicators, manual entries, prior audit adjustments, or anomaly intensity.

For linear models, coefficient signs and magnitudes provide direct interpretability. For tree-based models, feature importance and SHAP values can identify global and local drivers (Lundberg & Lee, 2017). Local explanation methods can show why an individual item was ranked highly (Ribeiro et al., 2016). These tools can support audit documentation by linking model output to audit assertions and planned procedures. For example, a transaction selected because of manual posting, end-of-period timing, and revenue reversal may justify inspection of support, contract review, and subsequent cash receipt testing.

Explainability also helps detect model failure. If the model ranks items highly because of irrelevant features, data leakage, or system artifacts, the auditor can challenge the output. If a model relies heavily on a variable that is not stable across years, its use may need to be restricted. Explainability should therefore be integrated into model validation, not added after deployment.

4.6 Discussion of Findings

The findings support the argument that adaptive scoring can improve audit item selection when risk is multidimensional. The adaptive method detected more misstated items than traditional sampling under the same testing budget. This does not mean that traditional sampling is obsolete. Sampling remains necessary for population inference, tests of controls, regulatory

compliance, and situations where statistical projection is required. The result means that machine-learning risk scoring can complement sampling by improving the selection of high-risk items for targeted testing.

The proposed ARS also helps reconcile audit analytics with audit-risk theory. The classical audit risk model remains conceptually valid, but it is often too broad for item-level prioritization. The adaptive method retains inherent-risk and control-risk logic while introducing item-level probability estimation, materiality exposure, anomaly detection, and evidence updating. In this sense, the model is not a rejection of audit theory. It is a quantitative extension of audit theory for data-rich environments.

The simulation also shows why precision and recall should be interpreted together. High recall is valuable because auditors want to detect as many misstated items as possible. High precision is valuable because audit resources are limited. A model that selects many false positives may burden the audit team and create review fatigue. The optimal audit strategy balances recall, precision, materiality, and audit cost. The proposed threshold equation provides a way to formalize this balance.

From a practical standpoint, implementation requires strong governance. Audit firms should validate models before engagement use, monitor calibration, document data lineage, preserve professional judgement, and maintain clear responsibility for audit conclusions. The model should not be allowed to override auditing standards or replace auditor skepticism. Instead, it should function as a decision-support tool that improves where auditors look first and how they allocate testing intensity.

4.6.1 Additional Post-Update Detection Comparison

Figure 3 extends the initial result comparison by evaluating the proposed Adaptive ARS after the evidence-updating stage has incorporated new audit evidence. Under the same top 10% testing budget, the post-update Adaptive ARS detects 252 materially misstated items, outperforming LogReg with 239, GBM with 232, RF with 229, SVM with 228, and MUS with 87 detected items. The difference is technically important because the proposed algorithm is not a static classifier; it revises audit priority after confirmation results, recalculation evidence, inspection findings, anomaly flags, and control-risk observations are incorporated. Relative to the strongest static benchmark, LogReg, the post-update ARS produces an additional 13 detected misstatements, representing a 5.4% improvement in detected exceptions within the same audit effort. Compared with MUS, the gain is substantially larger because MUS mainly weights item value, whereas ARS jointly learns probability of misstatement, materiality exposure, anomaly intensity, control weakness, and uncertainty.

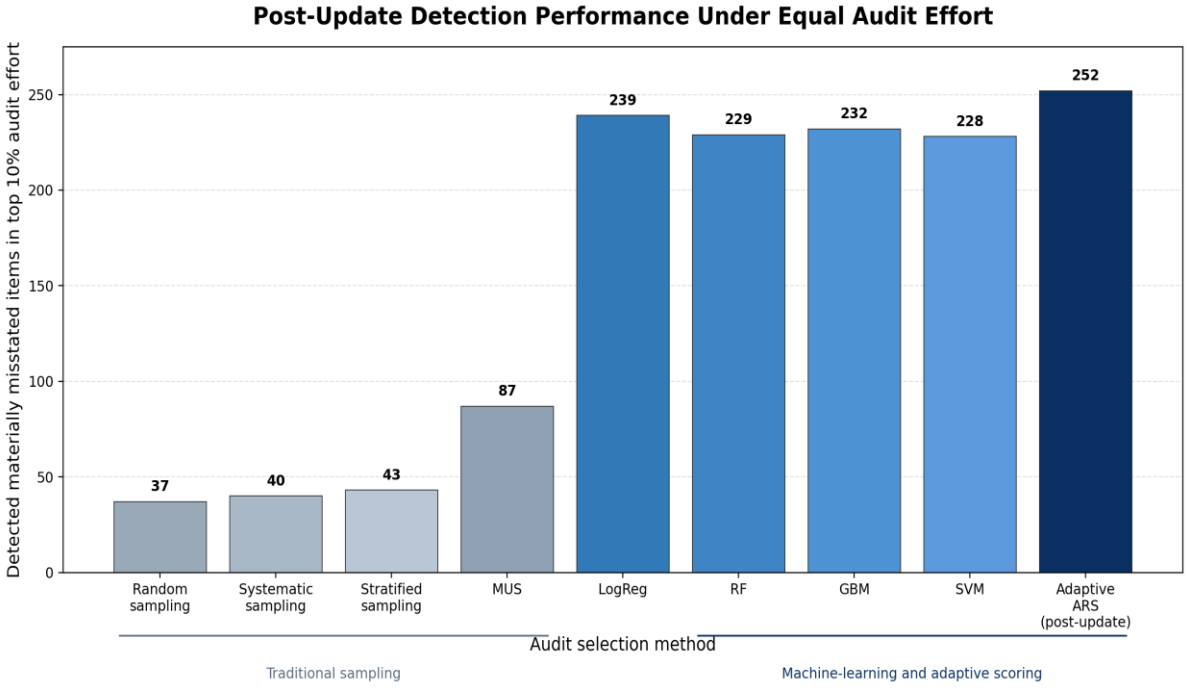

Figure 3. Post-update detection performance of the proposed Adaptive ARS compared with traditional sampling and existing machine-learning models.

4.6.2 Audit-Efficiency and False-Positive Burden

Figure 4 compares the composition of the selected top 10% audit testing budget by separating true detected misstatements from non-misstated items that were nevertheless selected for testing. Because every method selects 350 items, a superior method should increase detected misstatements while reducing the non-misstated portion of the selected sample. The post-update Adaptive ARS detects 252 misstated items and leaves only 98 non-misstated selected items, whereas LogReg detects 239 with 111 non-misstated selected items, GBM detects 232 with 118 non-misstated selected items, RF detects 229 with 121 non-misstated selected items, and SVM detects 228 with 122 non-misstated selected items. This result demonstrates that the proposed algorithm improves both audit effectiveness and audit efficiency: it raises true-positive concentration while reducing avoidable testing burden. In practical audit terms, the ARS ranking allows the engagement team to concentrate confirmations, inspection, recalculation, and substantive analytical procedures on items with stronger evidence-based risk justification, thereby strengthening the link between model output, professional judgement, and documented audit response.

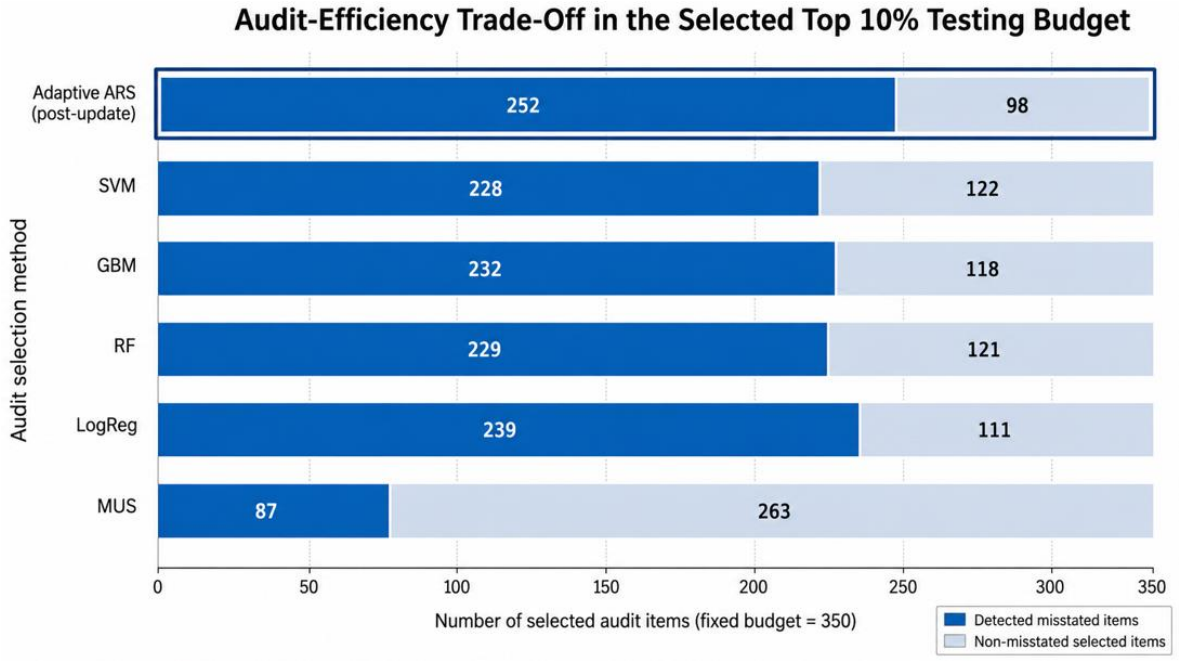

Figure 4. Audit-efficiency trade-off showing detected misstatements and non-misstated selected items within the fixed top 10% testing budget.

5. CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This paper developed an Adaptive Audit Risk Scoring Algorithm for detecting financial statement misstatements using machine learning and compared it with traditional sampling techniques. The proposed score integrates predicted misstatement probability, anomaly intensity, materiality exposure, control-risk score, and model uncertainty. It also includes an adaptive evidence-updating mechanism that revises risk scores as new audit evidence becomes available. The framework is rooted in the classical audit risk model but extends it into a dynamic, probabilistic, and item-level scoring system.

The simulation-based results indicate that adaptive risk scoring can detect more materially misstated items than random, systematic, stratified, and monetary-unit sampling when all methods operate under the same testing budget. The additional post-update comparison further shows that the proposed Adaptive ARS achieves the strongest detection yield after evidence-responsive updating, demonstrating higher true-positive concentration and lower avoidable testing burden than the static machine-learning benchmarks. The strongest explanation is that the adaptive score captures multidimensional risk patterns that traditional sampling methods do not fully incorporate. The findings support the view that machine learning can improve audit planning and substantive testing when models are calibrated, interpretable, materiality-aware, and governed appropriately.

The conclusion must be interpreted with caution. The paper used synthetic audit data for transparent methodological demonstration. Real engagement validation is required before operational use. Nevertheless, the mathematical framework, equations, performance measures, and comparison design provide a publishable foundation for future empirical testing using confidential audit firm data, public restatement data, or transaction-level enterprise resource planning data.

5.2 Theoretical Contribution

The theoretical contribution is the transformation of audit risk from a static multiplicative model into a dynamic risk-estimation framework. The classical equation $AR = IR \times CR \times DR$ remains central, but the proposed model operationalizes audit risk at the item level by estimating $P(Y_i = 1 \mid X_i)$ and integrating materiality, anomaly, control-risk, and uncertainty variables. This bridges audit-risk theory with machine learning and statistical decision theory.

The paper also contributes to the theory of audit sampling. Traditional sampling is designed for population inference, but audit planning also requires risk prioritization. The proposed ARS provides a formal method for risk-prioritized selection. It therefore occupies the space between statistical sampling and judgemental sampling. Unlike informal judgemental selection, the ARS is explicit, testable, explainable, and capable of being validated against detected misstatements.

5.3 Methodological Contribution

The methodological contribution lies in integrating supervised learning, anomaly detection, materiality adjustment, uncertainty weighting, cost-sensitive learning, probability calibration, and sampling comparison into a single framework. The equations define how labels are constructed, how materiality exposure is measured, how risk probability is estimated, how anomaly intensity is computed, how the composite score is formed, how evidence updates the score, and how detection performance is evaluated.

The paper also contributes an evaluation design that is closer to audit practice than generic accuracy assessment. Instead of focusing only on classification accuracy, the study evaluates detected misstatements within a fixed audit testing budget. This design better reflects the auditor's resource allocation problem. It also creates a fair comparison between machine learning and traditional sampling because each method receives the same audit effort.

5.4 Practical Contribution

The practical contribution is a decision-support framework for audit planning, risk response, and substantive testing. Audit firms can use adaptive scoring to rank journal entries, revenue transactions, procurement records, inventory movements, or account balances. Internal auditors can apply the same logic to continuous auditing and control monitoring. Forensic accountants can use the score to identify suspicious patterns requiring investigation. Regulators can use the framework to understand how machine learning tools may be documented, validated, and governed.

The model also supports audit documentation. Each selected item can be linked to a score and to score components. This makes it possible to explain why an item was tested and how the test relates to risk assessment. Documentation is particularly important because audit analytics tools must withstand engagement quality review, inspection, and regulatory scrutiny. The model's interpretability layer helps translate mathematical output into audit reasoning.

5.5 Recommendations

Audit firms should complement traditional sampling with adaptive machine-learning risk scoring, especially in high-volume transaction environments where risk is unlikely to be uniform. The model should be used to prioritize testing and identify unusual items, not to replace required sampling procedures or professional judgement. Engagement teams should validate the model on historical audit outcomes before use and should document data sources, feature definitions, model version, calibration results, and threshold decisions.

Risk scores should be calibrated and explainable before being used in audit decision-making. A model that produces high rankings but poorly calibrated probabilities may still be useful for prioritization, but its probability output should not be interpreted as precise risk. Audit teams should therefore monitor Brier score, calibration curves, and feature importance. They should also compare model selections with auditor-selected high-risk items to identify disagreements that require professional review.

Materiality exposure and internal control indicators should be embedded into predictive audit models. A purely statistical anomaly score may overemphasize unusual but immaterial items. Conversely, a purely materiality-based sampling method may overlook low-value but fraud-sensitive items. The composite ARS balances these concerns by integrating probability, anomaly, materiality, control-risk, and uncertainty.

Regulators and standard setters should develop guidance for validating, documenting, and governing audit machine-learning tools. Such guidance should address data quality, model training, independence, explainability, evidence sufficiency, threshold selection, reviewer competence, and accountability for audit conclusions. Future research should test the proposed model using multi-industry datasets, real audit engagement data, restatement samples, journal-entry populations, and transaction-level evidence.

5.6 Limitations and Future Research

The primary limitation is the use of synthetic data. Although the simulation reflects plausible audit-risk relationships, it cannot capture all complexities of real engagements. Actual audit data may contain missing values, inconsistent coding, undocumented manual adjustments, changing control environments, and incomplete labels. Some misstatements are not detected even during the audit, which creates label noise. Future studies should validate the model using real audit adjustments, restatement data, enforcement releases, and enterprise resource planning exports.

A second limitation is that the proposed ARS uses a weighted additive score. Although this structure is interpretable, other aggregation methods may perform better, including Bayesian networks, survival models, graph neural networks, reinforcement learning, or cost-sensitive ranking models. Future research can compare these alternatives. A third limitation is that the present paper focuses on financial statement misstatement detection. Future work may extend the method to internal control deficiencies, going-concern risk, tax risk, sustainability reporting assurance, and fraud investigation.

The long-term research opportunity is continuous adaptive auditing. In such a system, audit risk scores update as transactions occur, controls operate, confirmations return, exceptions are resolved, and external data become available. This would move audit practice from periodic retrospective testing toward continuous risk monitoring. The proposed model provides a mathematical starting point for that transition.

REFERENCES

[1] Akpara, I. U., & Bamigwojo, O. V. (2024). Experimental evaluation of virtual network segmentation strategies in Azure-based hybrid enterprise environments. *Global Multidisciplinary Perspectives Journal, 1*(1), 58-74.

[2] Alles, M. G. (2015). Drivers of the use and facilitators and obstacles of the evolution of big data by the audit profession. *Accounting Horizons, 29*(2), 439-449. https://doi.org/10.2308/acch-51067

[3] Amebleh, J., Igba, E., & Ijiga, O. M. (2021). Graph-based fraud detection in open-loop gift cards: Heterogeneous GNNs, streaming feature stores, and near-zero-lag anomaly alerts. *International Journal of Scientific Research in Science, Engineering and Technology, 8*(6), 317-339. https://doi.org/10.32628/IJSRSET214418

[4] Anyiam, K. H., Igwe, K. C., Amusa, T. A., Obinnanwilikom, C. O., Enoch, O. C., Isaiah, I. G., Obasi, A. C., Nna, F., Iwezor Okoro, D. N., Anene, H. U., Nnorom, E. I., Olumba, U. M., Igwenagu, O. M., Onu, D. O., & Ozor, M. U. (2026). Analysis of demand for water and its impact on broiler production in Imo State of Nigeria. *Online Journal of Animal and Feed Research*. https://doi.org/10.51227/ojafr.2026.3

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

[6] Arens, A. A., Elder, R. J., Beasley, M. S., & Hogan, C. E. (2023). *Auditing and assurance services: An integrated approach* (18th ed.). Pearson.

[7] Bamigwojo, O. V., Ezike, M. C. G., Owhenagbo, P., Idoko, P. I., Adedoyin, J. D., Enyejo, L. A., & Agina, P. C. (2024). Mathematical analysis of hepatitis B virus transmission dynamics in the absence of therapy with Atangana-Baleanu fractional-order SPQWXY model. *Journal of Advances in Mathematics and Computer Science, 39*(11), 1-28. https://doi.org/10.9734/jamcs/2024/v39i111937

[8] Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*, *58*(1), 199-235. https://doi.org/10.1111/1475-679X.12292

[9] Bell, T. B., Peecher, M. E., & Solomon, I. (2005). *The 21st century public company audit: Conceptual elements of KPMG's global audit methodology*. KPMG International.

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

[11] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

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

[13] Brown-Liburd, H., & Vasarhelyi, M. A. (2015). Big data and audit evidence. *Journal of Emerging Technologies in Accounting*, *12*(1), 1-16. https://doi.org/10.2308/jeta-10468

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

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

[16] Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research*, *28*(1), 17-82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

[17] Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting earnings management. *The Accounting Review*, *70*(2), 193-225.

[18] DeFond, M., & Zhang, J. (2014). A review of archival auditing research. *Journal of Accounting and Economics*, *58*(2-3), 275-326. https://doi.org/10.1016/j.jacceco.2014.09.002

[19] Earley, C. E. (2015). Data analytics in auditing: Opportunities and challenges. *Business Horizons*, *58*(5), 493-500. https://doi.org/10.1016/j.bushor.2015.05.002

[20] Eilifsen, A., Kinserdal, F., Messier, W. F., & McKee, T. E. (2020). An exploratory study into the use of audit data analytics on audit engagements. *Accounting Horizons*, *34*(4), 75-103. https://doi.org/10.2308/HORIZONS-19-121

[21] Esiobu, N. S., Osuagwu, C. O., Ohaemesi, C. F., Onyike, R. C., Nnamdi, F., & Igwenagu, M. O. (2025). Do farmers derive income from yam production? Novel evidence from Imo State, Nigeria. *Research Journal of Agricultural Economics and Development*. https://doi.org/10.52589/rjaed-qfqctdhb

[22] Fanning, K. M., & Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *7*(1), 21-41. https://doi.org/10.1002/(SICI)1099-1174(199803)7:1<21::AID-ISAF138>3.0.CO;2-K

[23] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[24] Fedyk, A., Hodson, J., Khimich, N., & Fedyk, T. (2022). Is artificial intelligence improving the audit process? *Review of Accounting Studies*, *27*(3), 938-985. https://doi.org/10.1007/s11142-022-09697-x

[25] Francis, J. R. (2011). A framework for understanding and researching audit quality. *Auditing: A Journal of Practice & Theory*, *30*(2), 125-152. https://doi.org/10.2308/ajpt-50006

[26] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[27] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[28] Green, B. P., & Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology. *Auditing: A Journal of Practice & Theory*, *16*(1), 14-28.

[29] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

[30] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

[31] Hennes, K. M., Leone, A. J., & Miller, B. P. (2008). The importance of distinguishing errors from irregularities in restatement research: The case of restatements and CEO/CFO turnover. *The Accounting Review, 83*(6), 1487-1519. https://doi.org/10.2308/accr.2008.83.6.1487

[32] Igwenagu, M. O. (2026) System and Method for Integrated Climate-Economic Forecasting to Enhance Agricultural Resilience and National Food Security: A Comprehensive Framework for Data-Driven Agricultural Decision Making UKR *Journal of Economics, Business and Management (UKRJEBM)* DOI/ Link: https://doi.org/10.5281/zenodo.19874542

[33] Igwenagu, M. O., & Akenzua, E. S. (2026). Big data approaches for agricultural and health systems monitoring: Lessons for Sub-Saharan Africa from global experiences. *Asian Journal of Research in Agriculture and Forestry*. https://doi.org/10.9734/ajraf/2026/v12i1464

[34] Igwenagu, M. O., Adewuyi, O. I., Isen, B. & Opeola, F. (2026). AI-Augmented Project and Program Management: Predictive Analytics for Risk, Cost and Schedule Control *American Journal of Smart Technology and Solutions* https://doi.org/10.54536/ajsts.v5i1.6989

[35] Igwenagu, M. O., Akwabeng, P. M., Darkoh, G. O., Adebakin, Y. K., Ojo, E. O., & Odufuwa, P. (2025). Smart agricultural technologies for sustainable food production under climate change pressures. *Journal of Agriculture and Ecology Research International*. https://doi.org/10.9734/jaeri/2025/v26i6722

[36] Igwenagu, M. O., Olatunbosun, A., Oluwadare, O.E., Ismaila, E. O., & Amoako, L. (2026). AI-Driven Agricultural and Health Interventions in Sub-Saharan Africa: Governance, Ethics, and the Role of Philanthropy *Journal of Global Agriculture and Ecology* https://doi.org/10.56557/jogae/2026/v18i110125

[37] International Auditing and Assurance Standards Board. (2009). *International Standard on Auditing 530: Audit sampling*. International Federation of Accountants.

[38] International Auditing and Assurance Standards Board. (2019). *International Standard on Auditing 315 (Revised 2019): Identifying and assessing the risks of material misstatement*. International Federation of Accountants.

[39] Issa, H., Sun, T., & Vasarhelyi, M. A. (2016). Research ideas for artificial intelligence in auditing: The formalization of audit and workforce supplementation. *Journal of Emerging Technologies in Accounting, 13*(2), 1-20. https://doi.org/10.2308/jeta-10511

[40] Jalloh, M. S., & Bamigwojo, O. V. (2024). AI-based production optimization for smart manufacturing environments. *International Journal of Scientific Research and Modern Technology, 3*(6), 160-177.

[41] James, U. U., Ijiga, O. M., & Enyejo, L. A. (2024). AI-powered threat intelligence for proactive risk detection in 5G-enabled smart healthcare communication networks. *International Journal of Scientific Research and Modern Technology, 3*(11), 125-140. https://doi.org/10.38124/ijsrmt.v3i11.679

[42] James, U.U. (2026). Multiscale Signal Processing Framework for Zero Trust Security in Next- Generation Wireless Networks *International Journal of Wireless and Mobile Networks* DOI/Link: https://aircconline.com/ijwmn/V18N3/18326ijwmn01.pdf

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

[44] Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications, 32*(4), 995-1003. https://doi.org/10.1016/j.eswa.2006.02.016

[45] Knechel, W. R., & Salterio, S. E. (2016). *Auditing: Assurance and risk* (4th ed.). Routledge.

[46] Knechel, W. R., Krishnan, G. V., Pevzner, M., Shefchik, L. B., & Velury, U. K. (2013). Audit quality: Insights from the academic literature. *Auditing: A Journal of Practice & Theory*, *32*(Supplement 1), 385-421. https://doi.org/10.2308/ajpt-50350

[47] Kokina, J., & Davenport, T. H. (2017). The emergence of artificial intelligence: How automation is changing auditing. *Journal of Emerging Technologies in Accounting*, *14*(1), 115-122. https://doi.org/10.2308/jeta-51730

[48] Krahel, J. P., & Titera, W. R. (2015). Consequences of big data and formalization on accounting and auditing standards. *Accounting Horizons*, *29*(2), 409-422. https://doi.org/10.2308/acch-51065

[49] Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

[50] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 4765-4774). Curran Associates.

[51] McNichols, M. F. (2002). Discussion of the quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review*, *77*(Supplement), 61-69. https://doi.org/10.2308/accr.2002.77.s-1.61

[52] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625-632). Association for Computing Machinery. https://doi.org/10.1145/1102351.1102430

[53] Olumba, U. M., Nwosu, F. O., Chikezie, C., Anyiam, K. H., Ben-Chendo, G. N., Osugiri, I. I., Isaiah, G. I., Obinna-Nwandikom, C. O., Enoch, O. C., Nwachukwu, E. U., Anene, H. U., Nnorom, E. I., Igwenagu, M. O., & Mbakaogu, O. E. (2025). Formal and informal credit arrangements among livestock farmers in Imo State, Nigeria. *African Journal of Food, Agriculture, Nutrition and Development*. https://doi.org/10.18697/ajfand.147.26030

[54] Onuh, M. I., Idoko, P. I., Enyejo, L. A., Akoh, O., Ugbane, S. I., & Ibokette, A. I. (2024). Harmonizing the voices of AI: Exploring generative music models, voice cloning, and voice transfer for creative expression. *World Journal of Advanced Engineering Technology and Sciences*, *11*(1), 372-394. https://doi.org/10.30574/wjaets.2024.11.1.0072

[55] Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, *30*(2), 19-50. https://doi.org/10.2308/ajpt-50009

[56] Public Company Accounting Oversight Board. (2024a). *AS 2110: Identifying and assessing risks of material misstatement*. PCAOB.

[57] Public Company Accounting Oversight Board. (2024b). *AS 2315: Audit sampling*. PCAOB.

[58] Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*(2), 491-500. https://doi.org/10.1016/j.dss.2010.11.006

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

[60] Richardson, S. A., Sloan, R. G., Soliman, M. T., & Tuna, I. (2005). Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics*, *39*(3), 437-485. https://doi.org/10.1016/j.jacceco.2005.04.005

[61] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[62] Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, *71*(3), 289-315.

[63] Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. *Accounting Horizons*, *29*(2), 381-396. https://doi.org/10.2308/acch-51071

[64] Yoon, K., Hoogduin, L., & Zhang, L. (2015). Big data as complementary audit evidence. *Accounting Horizons*, *29*(2), 431-438. https://doi.org/10.2308/acch-51076