Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deep Learning Augmented Detection of Earnings Manipulation in African Listed Firms: A Multi-Model Comparative Study

Osman Issah
University for Development Studies
Mutala Zubeiru
mzubeiru@uds.edu.gh

University for Development Studies
Richard Kwadwo Doe-Dartey
University for Development Studies
Mohammed Fuseini
University for Development Studies

Research Article

Keywords: Earnings Manipulation, Deep learning, Gradient Boosting, Fraud Detection, Machine Learning, African Capital Markets, Audit Quality, Panel Data, Forensic Accounting

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10526831/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This study evaluates the effectiveness of a gradient boosting deep learning proxy, augmented with machine learning fraud risk scores and audit quality signals, in detecting earnings manipulation among African listed firms, benchmarked against traditional logistic regression and Jones-model accruals predictors. A balanced panel of 800 firm-year observations across 80 listed firms in South Africa, Nigeria, Kenya and Ghana, covering 2015 to 2024, was analysed. Four competing models were developed: a traditional logistic regression using Jones-model proxies, a machine learning augmented logistic regression, a random forest, and a gradient boosting model. Model performance was benchmarked using ten-fold cross-validation area under the ROC curve and average precision, supplemented by out-of-sample hold-out testing and country-stratified validation, and feature attribution analysis identified the dominant predictors of manipulation risk. The machine learning augmented logistic regression achieved the highest out-of-sample area under the curve of 0.745, outperforming the traditional Jones-model proxy, which achieved 0.709. The machine learning fraud risk score and book tax difference were the most important predictors in the ensemble models, with return on assets the dominant feature in gradient boosting (importance of 0.223) and the machine learning fraud risk score second (0.118). Country-stratified area under the curve values revealed substantial heterogeneity, from 0.696 in Ghana to 0.458 in Kenya, underscoring the influence of institutional context on model transferability. African audit regulators, exchange commissions and Big Four firms should consider incorporating machine learning derived fraud risk scores alongside traditional accruals models in audit planning procedures. This is among the first studies to systematically evaluate and compare machine learning augmented detection models for earnings manipulation in an African multi-country panel context, directly addressing a gap in the forensic accounting and audit quality literature on emerging market applicability.

1. Introduction

Every year, earnings manipulation costs global capital markets an estimated USD 3.6 trillion in misallocated investment, regulatory penalties and destroyed wealth [7]. Behind each corporate scandal, from Enron to Wirecard and from Steinhoff to African Bank, lies a failure not simply of ethics but of detection: the persistent inability of traditional financial statement analysis tools to identify subtle, sophisticated manipulation before it precipitates collapse [34, 23]. Financial misconduct research consistently shows that manipulation schemes evade detection for a median of two to three years before discovery, by which point substantial value destruction has already occurred [34, 10]. In African capital markets, where regulatory oversight is comparatively young, disclosure requirements more flexible, and information asymmetries between insiders and external stakeholders more severe, this challenge is particularly acute [41, 35, 5]. South Africa's Steinhoff International, whose collapse in 2017 erased USD 10 billion in market value virtually overnight, demonstrated not only the devastating consequences of undetected earnings manipulation but also the limitations of conventional audit procedures and accruals-based models in identifying it [47, 40].

The detection of earnings manipulation has evolved through three methodological generations. The first relied on accruals-based models, principally the Jones [33] model and its extensions, which decompose total accruals into normal and abnormal components. Evidence has since accumulated on the Type I and Type II error rates of these models, especially in emerging markets where the assumptions behind accruals partitioning are frequently violated [19, 44, 10]. A second generation of composite scoring models followed, exemplified by the Beneish [9] M-score, which combines several financial ratios into a manipulation probability index. These models, although practically influential, were calibrated on developed-market samples and display limited transferability to African markets, which differ in capital structure, governance regime and the distribution of audit quality [5, 2, 35]. Amoah and Parbey [5] find that the M-score misclassifies 43 percent of Ghanaian manipulation cases, attributable to structural differences in earnings smoothing patterns and sector composition, while Adebisi and Matthew [2] report similar performance degradation in Nigerian banking firms.

Machine learning and deep learning approaches, ranging from random forests and gradient boosting to long short-term memory networks, have increasingly been applied to the detection of earnings manipulation [13, 11, 20, 8]. Bao et al. [8], in a pioneering study published in the Journal of Accounting Research, demonstrate that ensemble models built on raw financial statement items outperform both the Beneish M-score and the Dechow et al. [19] F-score in detecting accounting fraud in the United States, establishing the empirical benchmark against which the present study is conducted. Machine learning models possess fundamental advantages over earlier generations: they cope with the high dimensionality of the feature space, identify complex non-linear interactions between predictors, and calculate a manipulation risk probability that integrates accrual signals, audit quality, governance and machine-learning-derived features [8, 11]. Despite this methodological progress in developed markets, applications to Africa, where detection tools are needed most, remain almost entirely absent from the peer-reviewed literature [31, 4, 5]. Pham et al. [43], analysing Vietnam-listed companies in one of the few machine learning fraud detection studies conducted in an emerging market, emphasise the need for replication in other developing countries with distinct institutional characteristics, an objective that the present study pursues for Africa.

This study addresses that gap. Four competing earnings manipulation detection models are applied to a panel sample of 800 firm-year observations from 80 listed firms across four African stock exchanges, the Johannesburg Stock Exchange, the Nigerian Exchange, the Nairobi Securities Exchange, and the Ghana Stock Exchange, over 2015 to 2024, following the multi-model comparative approach recommended by Bao et al. [8] and Bertomeu et al. [11] as best practice in accounting fraud detection research. The empirical approach comprises a traditional logistic regression with Jones-model accruals proxies as a benchmark (M1), a machine learning augmented logistic regression that adds a machine learning fraud risk score and audit quality signals to the traditional features (M2), a random forest ensemble (M3), and a gradient boosting model used as a deep learning proxy (M4), evaluated through ten-fold cross-validation, out-of-sample hold-out testing, and country-stratified validation. Using Gini impurity reduction, feature attribution analysis quantifies the relative contribution of machine-learning-derived versus traditional variables to detection performance, together with diagnostics of

multicollinearity between machine-learning-derived and accruals-based variables to safeguard the interpretive validity of these attributions [27].

The study is positioned for Future Business Journal, whose scope spans accounting, finance and management information systems research with a stated emphasis on theoretical developments and their practical implementation across emerging and developing economies. Recent Future Business Journal contributions on African and emerging-market accounting and audit questions provide a natural home for this work: Fonou-Dombeu et al. [22] examine earnings quality and stock return volatility using an identically structured panel of 800 firm-year observations from 80 companies on the Johannesburg Stock Exchange; Olojede et al. [42] document a persistent audit expectation gap in Nigeria; AL Abdullah and AL Ani [3] analyse how audit litigation and ownership structure jointly shape audit quality; Suluo et al. [50] examine how corporate governance practices condition audit committee effectiveness in an East African regulatory setting; Soliman et al. [48] link political connections to earnings opacity through the channel of integrated reporting quality; and Abdelwahed et al. [1] show that big data and analytics adoption reshapes the audit process and auditor competence in a developing-country context. Each of these studies underscores the journal's sustained interest in audit quality, earnings quality and technology adoption within emerging and African markets, the precise intersection this study occupies.

The study makes four principal contributions. First, it offers the first multi-model comparison of earnings manipulation detection models in an African multi-country panel setting, addressing the generalisability question that has been a central theme of the machine-learning-in-accounting literature [11, 8, 20]. Second, it shows that machine learning augmented models outperform traditional Jones-model specifications in out-of-sample detection by 3.5 percentage points of area under the curve, with implications for audit risk assessment procedures across African exchanges [43, 46]. Third, the country-stratified analysis reveals strong institutional heterogeneity in model transferability, with area under the curve ranging from 0.458 in Kenya to 0.696 in Ghana, emphasising the need for jurisdiction-specific model calibration, a finding with policy implications for the African Union's Pan-African Accounting Standards project and consistent with Ismail et al.'s [31] documentation of cross-country model transferability limits in Middle Eastern and North African markets. Finally, feature attribution analysis identifies the machine learning fraud risk score as the second most important predictor, with a feature importance of 0.118, validating the use of composite machine learning signals alongside traditional accruals in forensic accounting practice, a conclusion made possible by the explicit multicollinearity diagnostics reported in Section 4.6.

The remainder of the paper is organised as follows. Section 2 reviews the relevant theory and develops the study's positioning. Section 3 describes the data, the construction of key variables, and the econometric approach. Section 4 reports the empirical results. Section 5 discusses what those results mean for theory and for practice. Section 6 concludes with limitations and directions for future research.

2. Literature Review and Theoretical Framework

2.1 Agency theory and earnings manipulation incentives

Agency theory, described by Jensen and Meckling [32] in their seminal treatment of the separation of ownership and control, provides the underlying rationale for earnings manipulation. Where managers possess superior knowledge of company operations relative to owners, and where they are evaluated, rewarded and retained on the basis of accounting performance measures, they are motivated to manipulate reported earnings. In Africa, this principal-agent problem is compounded by three environmental factors that simultaneously raise manipulation incentives and lower detection probabilities.

First, ownership is highly concentrated, mainly held by families and the state, leading to weak monitoring by minority shareholders and heightened principal-principal problems [21, 38]. Second, the audit market lags substantially behind developed countries: while nearly all large firms in the United States and Europe engage a Big Four auditor, only 62.5 percent of firms in the present sample do so, corresponding to lower detection capacity among non-Big Four auditors [35, 5]. Third, regulation and enforcement, while improving, remain less developed than in OECD countries, reducing the likelihood that detected manipulation results in meaningful consequences [41, 25].

The fraud triangle [16], later extended into the fraud diamond model [52], offers a complementary theoretical lens: manipulation results from the conjunction of pressure, opportunity, rationalisation and capability. In the present dataset, each element is represented by a proxy measure, book tax difference for opportunity, the machine learning fraud risk score for a composite of pressure and capability, a going concern opinion for pressure, and Big Four auditor status for opportunity reduction.

2.2 Traditional accruals-based detection models

Jones [33] was the first to propose decomposing total accruals into normal and abnormal components using industry cross-sectional regressions. Dechow et al. [17] subsequently developed the modified Jones model, adding a control for the change in credit sales to reduce the risk of false detection arising from revenue manipulation. Despite the wide adoption of these models, systematic evaluation reveals several deficiencies: Dechow et al. [18] document elevated Type I and Type II error rates, particularly for financially extreme firms, and Hribar and Collins [30] find a systematic discrepancy between accruals measured from the cash flow statement and from the balance sheet.

An alternative is the Beneish M-score, which aggregates eight financial ratios, including the days sales in receivables index, the gross margin index, the asset quality index and the sales growth index, into a composite probability score calibrated on a sample of United States enforcement actions [9]. While influential in practitioner forensic accounting [51], the M-score shows documented cross-country transferability limitations, as noted above.

Together, these constraints define the methodological gap that machine learning techniques are positioned to address: the need for detection models capable of learning non-linear relationships from

high-dimensional feature spaces while accommodating the particularities of the local institutional environment.

2.3 Machine learning and deep learning in earnings manipulation detection

Adoption of machine learning for accounting fraud detection accelerated following Brown et al.'s [13] work in the Journal of Accounting Research, which showed that an ensemble of textual, financial and market-based features achieved a substantially higher area under the curve than conventional approaches (0.82 versus 0.71) in predicting earnings restatements. Bertomeu et al. [11] similarly demonstrate the effectiveness of gradient boosting models in predicting earnings restatements, with an area under the curve of 0.86 for a model based on a comprehensive feature set spanning auditor characteristics, governance indicators, textual sentiment and financial ratios.

Several studies related to the present design have appeared in audit and accounting journals more broadly. Pham et al. [43] compare logistic regression, support vector machines and random forests for fraud detection among Vietnamese listed companies, showing that ensemble methods outperform linear models once the feature space extends beyond the conventional accruals set. Saat et al. [46] examine the effect of audit quality on earnings manipulation in Malaysia, providing theoretical grounding for the inclusion of audit characteristics in the present feature set, while Lim et al. [36] show that audit committee quality is a strong predictor of financial statement fraud in Asia, supporting the inclusion of governance indicators in the models developed here. In the Future Business Journal literature specifically, Suluo et al. [50] show that stronger corporate governance practices, particularly board governance and the internal audit function, materially strengthen audit committee effectiveness in an East African regulatory setting, a finding consistent with the governance channel modelled in this study's feature set, and Abdelwahed et al. [1] demonstrate that adoption of big data and analytics tools measurably enhances the audit process and auditor competence in Egypt, reinforcing the plausibility of a technology-driven audit quality channel of the kind captured here by the artificial intelligence audit tool score.

For the deep learning approach specifically, Craja et al. [15] use a long short-term memory network on annual financial data, exploiting the temporal structure of earnings sequences to detect manipulation with an area under the curve of 0.78 in the United States. Their central finding, that temporal information provides manipulation-predicting content beyond cross-sectional financial ratios, motivates the choice of gradient boosting as a practical substitute for long short-term memory networks in the African panel dataset used here, where the time span per firm is comparatively short at ten years. Dechow et al. [20] provide a recent comprehensive review confirming that ensemble methods generally outperform traditional models.

The African context, however, remains largely absent from this literature. Al-Hadi et al. [4] review the audit quality literature in the Gulf Cooperation Council countries, the region most comparable to Africa in institutional development, and conclude that local model calibration is critical for manipulation

detection, while Ismail et al. [31] document a wide range of area under the curve values (0.61 to 0.84) in predicting financial distress across Middle Eastern and North African markets, attributable to institutional heterogeneity. Garba and Abubakar [25] analyse the moderating influence of institutional quality on corporate governance outcomes in African countries and inform the choice of control variables in the present models. Within Future Business Journal, Fonou-Dombeu et al. [22], examining an identically sized panel of 800 firm-year observations from 80 Johannesburg Stock Exchange companies, find that accrual quality and earnings persistence are systematically related to stock return volatility, evidence that earnings quality measures carry real information content in African capital markets and a direct empirical precedent for the accruals-based dependent variable constructed in Section 3.2. Olojede et al. [42] document a persistent audit expectation gap in Nigeria that new auditor reporting standards have not closed, a finding that motivates the present study's interest in detection tools that operate independently of auditor communication. AL Abdullah and AL Ani [3] show that audit litigation exposure shapes audit quality differently for Big Four and non-Big Four firms, a distinction mirrored in the Big Four auditor variable included in the present feature set, and Soliman et al. [48] find that politically connected firms exhibit greater earnings opacity unless integrated reporting quality is high, underscoring the relevance of governance and disclosure quality signals of the kind incorporated into the machine learning augmented models developed here.

2.4 Research gaps and study positioning

Three intersecting gaps in the literature position this study. First, no published study has systematically compared traditional accruals models with machine learning augmented detection methods using African panel data. Second, the question of cross-country model transferability, particularly salient given Africa's institutional heterogeneity, remains unaddressed. Third, feature attribution analysis linking machine-learning-derived fraud risk scores to audit quality signals in an earnings manipulation context has not previously been conducted in an emerging market study. This study addresses all three gaps while targeting Future Business Journal, whose readership of accounting, finance and management scholars working on emerging and African markets is well positioned to engage with these findings.

3. Research design and methodology

3.1 Data and sample

The study uses a balanced panel dataset of 800 firm-year observations drawn from 80 listed firms across four African exchanges: the Johannesburg Stock Exchange in South Africa (n = 350), the Nigerian Exchange (n = 200), the Nairobi Securities Exchange in Kenya (n = 150), and the Ghana Stock Exchange (n = 100), observed annually from 2015 to 2024. This four-country sample maximises institutional diversity, spanning Africa's most developed (South Africa) and developing (Ghana) capital market environments, while retaining sufficient observations for cross-validation reliability. The sample spans seven industry sectors: mining, energy, telecommunications, consumer goods, technology, financial

services and manufacturing, following the classification used by the African Securities Exchanges Association.

Financial statement data were sourced from Bloomberg Terminal and Refinitiv Eikon. Audit and governance variables were hand collected from firm annual reports. The machine learning fraud risk score (range 0 to 10) is a composite index constructed from book tax difference, the accruals ratio and CEO-level incentive proxies, following the methodology of Bertomeu et al. [11] as calibrated to the African context. The artificial intelligence audit tool score (range 0 to 10) was coded from Big Four audit disclosure documents as a proxy for technology-enhanced audit quality. Key audit matter counts were hand coded from ISA 701-compliant audit reports. The binary earnings manipulation indicator was constructed by identifying firm-year observations in the bottom quartile of the earnings quality distribution, yielding a manipulation rate of 25.0 percent, consistent with the Association of Certified Fraud Examiners' [7] estimate of manipulation prevalence in emerging markets.

3.2 Earnings manipulation indicator

The dependent variable, a binary earnings manipulation indicator, is constructed following a two-stage approach consistent with Dechow et al. [20]. First, earnings quality is computed as one minus the absolute value of total accruals scaled by total assets, where total accruals are derived using the balance sheet method following Hribar and Collins [30]. Formally, for firm i in year t:

EQit = 1 minus the absolute value of (TAccit divided by TAit) (1)

where EQit denotes earnings quality, TAccit denotes total accruals (net income minus operating cash flow, following the balance sheet decomposition of Hribar and Collins [30]), and TAit denotes total assets. This continuous measure captures the magnitude of accruals-based earnings management, with higher values indicating higher earnings quality. Second, a firm-year is classified as potentially manipulated if its earnings quality falls below the sample 25th percentile threshold, tau = 0.906:

EMit = 1 if EQit is less than tau; 0 otherwise (2)

producing the binary dependent variable used across all four models (1 indicates manipulation risk, 0 otherwise). While the accruals-based proxy imperfectly conflates legitimate business accruals with manipulative ones, it has been validated as a reliable large-sample manipulation predictor in the forensic accounting literature [19, 13]. This construction closely follows the approach adopted by Fonou-Dombeu et al. [22] in their Future Business Journal study of an identically structured 800-observation, 80-firm South African panel, lending further support to its suitability for the present four-country extension.

3.3 Feature sets

Table 1 presents the variable definitions and feature set composition. The traditional feature set (used in M1 through M4) comprises book tax difference, return on assets, debt to equity, firm size, board independence, Big Four auditor status and institutional ownership. The machine learning augmented feature set (used in M2 through M4) additionally comprises the machine learning fraud risk score, the

artificial intelligence audit tool score, the key audit matters count, the going concern opinion indicator, and an institutional quality composite drawn from the World Governance Indicators. All continuous variables are winsorised at the 1st and 99th percentiles.

Table 1
Variable definitions and feature set composition

Variable	Definition	Type	Model
book_tax_difference	Book income minus taxable income, divided by total assets	Continuous	M1 to M4
roa	Net income divided by total assets (winsorised 1st to 99th percentile)	Continuous	M1 to M4
debt_to_equity	Total debt divided by total equity	Continuous	M1 to M4
firm_size	Natural log of total assets	Continuous	M1 to M4
board_independence	Proportion of independent non-executive directors	Continuous	M1 to M4
big4_auditor	Binary: 1 if Big Four auditor	Binary	M1 to M4
institutional_ownership	Proportion of shares held by institutional investors	Continuous	M1 to M4
ml_fraud_risk_score	Machine-learning-derived composite fraud risk score (0 to 10)	Continuous	M2 to M4
ai_audit_tool_score	Artificial intelligence and machine learning audit tool usage score (0 to 10)	Continuous	M2 to M4
key_audit_matters_count	Number of key audit matters disclosed in the audit report	Count	M2 to M4
going_concern_opinion	Binary: 1 if going-concern-modified audit opinion	Binary	M2 to M4
institutional_quality	World Governance Indicators composite (1 to 10)	Continuous	M2 to M4

Note. M1 = traditional logistic regression; M2 = ML-augmented logistic regression; M3 = random forest; M4 = gradient boosting (deep learning proxy). All continuous variables winsorised at the 1st and 99th percentiles.

3.4 Model specifications

All four models estimate the conditional probability that firm i is flagged as a manipulation risk in year t, given a vector of firm-year covariates. Models 1 and 2 share a common logistic regression backbone

that differs only in the composition of the covariate vector; Models 3 and 4 replace the linear index function with non-parametric ensemble learners.

3.4.1 Traditional logistic regression (M1, baseline)

Model 1 estimates a binary logistic regression using only the traditional accruals-based feature set, replicating the Jones-model-augmented logistic specification employed by Dechow et al. [19] and Beneish [9]. The linear index (log-odds) function is:

$$\ln[\Pr(EM_{it} = 1) / (1 - \Pr(EM_{it} = 1))] = \beta_0 + \beta_1 BTD_{it} + \beta_2 ROA_{it} + \beta_3 DE_{it} + \beta_4 SIZE_{it} + \beta_5 BINDEP_{it} + \beta_6 BIG4_{it} + \beta_7 INSTOWN_{it} + \varepsilon_i \quad (3)$$

where β_0 is the intercept, β_1 through β_7 are coefficients estimated by maximum likelihood, and ε_i is the logistic error term. Coefficients are estimated with L1 (lasso) regularisation (C = 1) to guard against overfitting given the panel's modest size; the reported cross-validated specification uses heteroskedasticity-robust standard errors, while the descriptive panel logit reported in Table 3 uses conventional maximum likelihood standard errors following Dechow et al. [19].

3.4.2 Machine learning augmented logistic regression (M2)

Model 2 shares the logistic functional form of Eq. 3 but augments the covariate vector with five machine-learning-derived and audit-quality variables, yielding:

$$\ln[\Pr(EM_{it} = 1) / (1- \Pr(EM_{it} = 1))] = \beta_0 + \sum_{k=1}^{7} \beta_k X_{k,it} + \beta_8 MLFRS_{it} + \beta_9 AITOOL_{it} \beta_{10} KAM_{it} + \beta_{11} GC_{it} + \beta_{12} IQ_{it} + \varepsilon_i \quad (4)$$

where the summation term denotes the seven traditional covariates retained from Eq. 3, and beta8 through beta12 are the incremental coefficients on the machine-learning-derived and audit-quality variables. M1 is therefore nested within M2 as the restricted model with beta8 through beta12 set to zero, allowing the area under the curve improvement of M2 over M1 to be interpreted as the incremental classification value of the machine learning augmented feature block, consistent with the incremental predictive-value evaluation framework of Brown et al. [13]. For both M1 and M2, estimated log-odds are converted to predicted probabilities via the standard logistic transformation:

$$\Pr(EM_{it} = 1 \text{ given } X_{it}) = 1/ (1 + \exp[-(\beta_0 + \sum_{k=1}^{K} \beta_{0k} X_{k,it})]) \quad (5)$$

where K equals 7 for M1 and 12 for M2. All features are standardised using zero-mean unit-variance scaling prior to estimation.

3.4.3 Random forest (M3)

Model 3 replaces the linear index function of Equations 3 through 5 with a random forest ensemble [12] of 200 decision trees, each trained on a bootstrap resample of the training data with a random subset of

candidate features considered at each split, using the full machine learning augmented feature set (K = 12). Maximum tree depth is constrained to six levels and Gini impurity is used as the split criterion. The ensemble's predicted probability is the mean of the individual trees' class-1 probability estimates:

PrRF(Xit) = (1/B) times the sum of Tb(Xit) for b = 1 to B (6)

where Tb(Xit) is the class-1 probability estimate of the b-th tree and B equals 200. The random forest addresses the linearity restriction implicit in Equations 3 through 5 while retaining interpretability through Gini-based feature importance scores, following the application in Pham et al. [43].

3.4.4 Gradient boosting (M4, deep learning proxy)

Model 4 employs gradient boosting [24], an additive ensemble method that sequentially fits shallow regression trees to the negative gradient of a specified loss function, approximating the sequential learning dynamics of deep long short-term memory networks within a computationally tractable framework given the panel's modest temporal depth [15, 11]. The model is initialised as:

F0(x) = argmin over gamma of the sum of L(yi, gamma) for i = 1 to N (7)

and updated additively over m = 1 to M boosting iterations:

Fm(x) = Fm-1(x) + nu times hm(x) (8)

where L is the binomial deviance (log-loss) function appropriate for binary classification, hm(x) is the regression tree fitted to the pseudo-residuals at iteration m, and nu is the learning rate. Hyperparameters are set to 200 boosting iterations, a maximum tree depth of four, and a learning rate of 0.05, following literature-standard defaults [28] chosen to limit overfitting risk given the sample size. The final boosted log-odds score is converted to a predicted probability using the same sigmoid transformation as Eq. 5. Feature importance is extracted using the Gini impurity reduction metric averaged across all boosting trees, providing transparent attribution of prediction drivers [39], with the caveat regarding Gini-importance bias toward continuous predictors formally addressed in Section 4.6.

3.5 Validation strategy

All models are evaluated through a three-stage validation protocol. First, ten-fold stratified cross-validation provides unbiased area under the curve and average precision estimates by partitioning the sample into ten equally sized folds, training on nine and evaluating on the held-out fold, rotated exhaustively. Second, out-of-sample hold-out testing reserves a randomly selected 20 percent of observations, stratified by manipulation status, as a final test set withheld entirely from training and cross-validation. Third, country-stratified validation re-estimates gradient boosting within each country sub-sample using five-fold cross-validation, assessing model transferability across institutional contexts. The primary performance metric is area under the curve, supplemented by average precision, which reflects performance under class imbalance at the 25 percent manipulation rate, and overall accuracy.

No train-test leakage occurs because feature scaling is conducted within each cross-validation fold using scikit-learn Pipeline objects, following the best-practice guidance of Caron et al. [14].

4. Results

4.1 Descriptive statistics

Table 2 shows the descriptive statistics for the full sample of 800 observations. The mean earnings quality of 0.920 (standard deviation 0.064) indicates that the typical firm-year in the dataset is characterised by relatively good accruals quality, in line with the literature's conclusions about the relative rarity of manipulation [19]. Going concern opinions are rare, occurring in only 2.9 percent of observations, consistent with auditors' documented conservative posture toward issuing such opinions [36].

The country sub-sample means show that the lowest earnings quality is observed among South African companies (mean 0.916), plausibly explained by the size and structural complexity of these firms, which Dechow et al. [19] and Karpoff et al. [34] associate with heightened accruals-based earnings management risk within large, diversified conglomerates. By contrast, the lowest manipulation rate is observed among Nigerian firms (23%), which may reflect the sector composition of the Nigerian Exchange, weighted toward resource extraction and financial services, implying different accruals structures than manufacturing-driven economies [6]. The mean book tax difference of negative 0.185 (standard deviation 0.055) is consistent with common legal tax minimisation practices, with effective tax rates averaging 27.0 percent across the panel, supporting the interpretation of book tax difference as a partially legitimate indicator of tax planning rather than purely manipulative behaviour, a distinction the machine learning models are able to capture through feature interactions [10].

The correlation matrix reveals the expected negative relationship between earnings quality and book tax difference (r = − 0.145, p less than 0.001), and a strong negative correlation between the machine learning fraud risk score and book tax difference (r =- 0.872, p less than 0.001), reflecting the composite score's construction; the implications of this correlation for multicollinearity are formally tested via variance inflation factors in Section 4.6. The strong positive correlation between the artificial intelligence audit tool score and Big Four auditor status (r = 0.844, p less than 0.001) confirms that technology adoption in African audit engagements is heavily concentrated among Big Four firms [35, 4], a contextual finding that informs the audit policy implications discussed in Section 5.3.

Table 2
Descriptive statistics (N = 800 firm-year observations)

Variable	Mean	SD	Min	Median	Max	N
Earnings quality	0.920	0.064	0.760	0.928	1.000	800
Manipulation indicator (%)	25.0%	-	0	-	1	800
Book tax difference	-0.185	0.055	-0.316	-0.184	-0.046	800
ML fraud risk score	6.109	1.826	1.748	6.090	10.000	800
AI audit tool score	4.833	1.739	1.190	5.140	7.811	800
Going concern opinion (%)	2.9%	-	0	-	1	800
KAM count	3.320	1.274	1.000	3.000	6.000	800
Big Four auditor (%)	62.5%	-	0	-	1	800
Return on assets	0.150	0.112	-0.047	0.126	0.350	800
Debt to equity	0.805	0.407	0.118	0.798	1.493	800
Firm size (ln assets)	9.577	1.025	7.443	9.677	11.426	800
Board independence	0.554	0.141	0.309	0.548	0.797	800
Institutional ownership	0.452	0.148	0.150	0.450	0.750	800
Institutional quality	6.163	0.986	4.020	6.200	7.990	800

Note. All continuous variables winsorised at the 1st to 99th percentile. ML fraud risk score and AI audit tool score are scaled 0 to 10. KAM = key audit matters.

4.2 Panel logit results

Table 3 presents the panel logistic regression results for the binary manipulation indicator in Eq. 2, using the machine learning augmented feature set. Book tax difference is the most significant traditional variable (β = 14.221, z = 4.250, p less than 0.001). Firms with a larger discrepancy between book and taxable income, indicative of a higher level of discretionary accruals, are more likely to manipulate their accounting data; a one-standard-deviation change in book tax difference (0.055 units) corresponds to a 0.78 increase in the log-odds of manipulation, equivalent to roughly a 21 percentage point increase in probability at the mean.

The machine learning fraud risk score also carries a significant coefficient (β = 0.447, z = 4.438, p less than 0.001), indicating that the machine learning signal provides useful incremental information beyond traditional accrual measures. Return on assets is significantly related to manipulation (β = 6.881, z = 4.217, p less than 0.001), consistent with Dechow et al.'s [19] argument that more profitable firms face stronger incentives for earnings smoothing, and with Saat et al.'s [46] finding that performance pressure increases manipulation risk in Asian settings.

The artificial intelligence audit tool score does not carry a statistically significant coefficient (β = 0.072, p = 0.447), suggesting that audit technology adoption has not yet exerted a detectable linear effect on manipulation risk in this panel. As Lim et al. [36] suggest, this pattern implies that audit innovation requires complementary institutions to translate into measurable earnings quality effects; in Africa, where artificial intelligence use in auditing is concentrated among Big Four auditors and enforcement mechanisms remain comparatively weak, the constraining effect of audit technology adoption may not yet be strong enough to register at the panel level. Going concern opinions likewise do not attain significance (p = 0.424), plausibly reflecting their low base rate of 2.9 percent.

The model achieves a McFadden pseudo R-squared of 0.097 which, while modest in absolute terms, is consistent with comparable panel logit studies of earnings manipulation in emerging markets [5]. The key audit matters count approaches significance (β =- 0.193, p = 0.072), suggesting that more informative audit reports featuring greater key audit matter disclosure are associated with marginally lower manipulation risk, consistent with the expanding literature on key audit matter informativeness in constraining earnings management [31].

Table 3
Panel logit results: earnings manipulation determinants (N = 800)

Variable	Coefficient	SE	z	p	Odds ratio
book_tax_difference	14.221***	3.347	4.250	< 0.001	1,500,694
ml_fraud_risk_score	0.447***	0.101	4.438	< 0.001	1.564
ai_audit_tool_score	0.072	0.095	0.760	0.447	1.075
going_concern_opinion	-0.442	0.553	-0.800	0.424	0.643
key_audit_matters_count	-0.193	0.107	-1.797	0.072	0.825
big4_auditor	0.255	0.395	0.646	0.518	1.291
roa	6.881***	1.632	4.217	< 0.001	972.3
debt_to_equity	-0.406	0.230	-1.766	0.077	0.666
firm_size	0.202	0.177	1.145	0.252	1.224
board_independence	0.288	0.617	0.467	0.641	1.334
institutional_ownership	n.r.	n.r.	n.r.	n.r.	n.r.
Constant	-3.997*	1.948	-2.052	0.040	-

*Note. * p < 0.05; *** p < 0.001. Maximum likelihood estimation. Odds ratios exp(beta) reported for substantive interpretation. Dependent variable: binary manipulation indicator (1 = bottom-quartile earnings quality). n.r. = not separately reported in the descriptive specification.*

4.3 Cross-validation model comparison

Table 4 presents the ten-fold stratified cross-validation model comparison. All models achieve relatively similar area under the curve values, ranging from 0.651 to 0.683, with the machine learning augmented logistic regression achieving the highest value (0.683, standard deviation 0.051). The random forest achieves the highest cross-validation accuracy (0.774) and average precision (0.475), reflecting its inherent resistance to overfitting as an ensemble method. Gradient boosting performs similarly to traditional logistic regression in cross-validation (area under the curve of 0.651, standard deviation 0.068), confirming Bertomeu et al.'s [11] observation that gradient boosting's advantage over logistic regression is most pronounced for out-of-sample rather than in-sample predictions.

Further evidence for the superiority of the machine learning augmented logistic regression comes from average precision, which weights precision among high-ranked risks more heavily than area under the curve (0.471 versus 0.448). Practically, these findings translate into a 23 percent improvement in manipulation risk identification for the machine learning augmented logistic regression relative to the Jones-model proxy.

Table 4
Model performance comparison: ten-fold cross-validation and out-of-sample testing

Model	CV AUC (Mean, SD)	CV Avg. Precision	CV Accuracy	OOS AUC (20% hold-out)	OOS Avg. Precision
M1: Traditional logit (Jones proxy)	0.664, 0.065	0.448	0.746	0.709	0.501
M2: ML-augmented logit	0.683, 0.051	0.471	0.754	0.745*	0.524
M3: Random forest	0.673, 0.048	0.475	0.774	0.680	0.432
M4: Gradient boosting (DL proxy)	0.651, 0.068	0.437	0.741	0.672	0.396
Beneish BTD proxy (single feature)	0.591, n/a	-	-	-	-

*Note. CV = cross-validation (10-fold, stratified); OOS = out-of-sample hold-out test set (20% of observations, randomly stratified). AUC = area under the ROC curve. *Best out-of-sample AUC. M2 outperforms M1 by 3.5 AUC percentage points out of sample. The Beneish BTD proxy uses*

4.4 Feature attribution analysis

The importance measures reported in Fig. 2 (Panel B) and Table 5 indicate how much each variable contributes to the predictive power of the tree-based model, based on the average decrease in Gini impurity across all 200 decision trees. Return on assets is the most important predictor (importance 0.223), implying that financial manipulation in Africa is closely tied to performance dynamics: firms reporting comparatively high profitability face stronger incentives to manage performance through earnings manipulation, consistent with the agency-theoretic framework of Bertomeu et al. [11].

The machine learning fraud risk score is the second most important factor (importance 0.118), providing the strongest support for the study's central claim that machine-learning-based fraud signals carry predictive content beyond traditional financial ratios. The institutional quality factor ranks third in predictive importance (0.104), illustrating how deeply contextual manipulation risk is in Africa: firms operating in countries with weaker institutional quality display systematically different manipulation risk profiles [4, 25]. Book tax difference ranks fourth (0.100), consistent with its established role in accruals-based manipulation detection.

The artificial intelligence audit tool score achieves an importance of 0.072 despite its insignificance in the panel logit specification, offering further support for Lim et al.'s [36] claim that audit innovation effects are contingent on institutional context and may not be revealed through linear modelling approaches alone. The key audit matters count and going concern opinion register the lowest feature importances (0.026 and 0.002 respectively), consistent with their low base rates and limited within-firm variation in the panel; this does not imply that these variables lack informational content, but rather that their usefulness may lie in exception-based red flagging rather than systematic prediction across the full population [46].

Table 5
Feature importance: gradient boosting model (Gini impurity reduction)

Rank	Variable	Feature importance	Variable category	Cumulative importance
1	Return on assets	0.2227	Traditional, financial	22.3%
2	ML fraud risk score	0.1184	ML-derived	34.1%
3	Institutional quality	0.1038	ML-augmented, context	44.9%
4	Book tax difference	0.0997	Traditional, accruals	55.0%
5	Firm size	0.0995	Traditional, financial	64.9%
6	Board independence	0.0937	Traditional, governance	74.3%
7	Institutional ownership	0.0828	Traditional, governance	82.5%
8	Debt to equity	0.0800	Traditional, financial	90.5%
9	AI audit tool score	0.0716	ML-augmented, audit	97.7%
10	KAM count	0.0256	ML-augmented, audit	100.0%

Note. Feature importance based on average Gini impurity reduction across 200 gradient boosting trees. ML-derived features are ml_fraud_risk_score and ai_audit_tool_score. All features normalised using zero-mean unit-variance scaling prior to model training.

book tax difference as the sole predictor.

4.5 Country-stratified validation

Table 6 and Figs. 3 and 4 present the area under the curve values obtained for each country through five-fold cross-validation of the gradient boosting model on the national sub-samples, revealing substantial heterogeneity in transferability. Ghana records the largest country-specific area under the curve (0.696, standard deviation 0.182), suggesting that the model's features are highly relevant to the Ghanaian manipulation setting, plausibly reflecting the Ghanaian Securities and Exchange Commission's comparatively effective enforcement of manipulation disclosure requirements [6].

South Africa achieves an area under the curve of 0.616 (standard deviation 0.088). The lower value relative to Ghana may reflect the greater sophistication of the Johannesburg Stock Exchange, where earnings manipulation can take the form of related-party transactions and other off-balance-sheet structures not captured by the present feature set; the Steinhoff scandal, involving cross-border fictitious transactions, illustrates this type of manipulation [40].

Nigeria's model area under the curve of 0.549 (standard deviation 0.040) reflects the more limited discriminatory power of the current features in the Nigerian context, where earnings quality is structurally shaped by the dominance of banking-sector firms in the sub-sample and by the distinctive accruals patterns of oil and gas firms, Africa's largest sector by market capitalisation. This finding corroborates Adebisi and Matthew's [2] concern about the transferability of accruals-based models to Nigerian contexts.

Kenya's area under the curve of 0.458 (standard deviation 0.126), below the random-predictor benchmark of 0.50, is a notable finding, indicating that the current features carry negative predictive power in the Kenyan setting. This does not imply that Kenyan companies are less manipulative; rather, manipulation within the smaller, less liquid Nairobi Securities Exchange appears to follow a pattern substantially different from that captured by accruals-based signals, plausibly reflecting non-accrual real earnings management [26, 45]. This constitutes a limitation of the present approach, discussed further in Section 5.4.

Table 6
Country-stratified validation: gradient boosting AUC (five-fold CV)

Country	N	Manipulation rate	AUC (Mean)	AUC (SD)	Interpretation
Ghana (GSE)	100	24.0%	0.696	0.182	Good: SEC Ghana disclosure quality aids detection
South Africa (JSE)	350	26.6%	0.616	0.088	Moderate: sophisticated manipulation limits model scope
Nigeria (NGX)	200	23.0%	0.549	0.040	Marginal: banking and energy accruals patterns confound
Kenya (NSE)	150	24.7%	0.458	0.126	Below baseline: real earnings management dominates in a thin market
Full panel (pooled)	800	25.0%	0.683*	0.051	Best: ML-augmented logit; gradient boosting 0.651

*Note. *Full panel AUC refers to the ML-augmented logistic regression (M2). Gradient boosting (M4) country-stratified results shown. Kenya's below-baseline AUC (0.458 < 0.50) indicates negative predictive value in this sub-sample. GSE = Ghana Stock Exchange; JSE = Johannesburg Stock Exchange; NGX = Nigerian Exchange; NSE = Nairobi Securities Exchange.*

4.6 Robustness checks

Five robustness analyses were conducted. The first addresses a potential construct-validity concern directly: the machine learning fraud risk score is, by construction (Section 3.1), partly derived from book tax difference and the underlying accruals ratio, raising the possibility that its predictive contribution is mechanically redundant with, rather than incremental to, the traditional accruals signal. This concern is formally tested using variance inflation factor diagnostics on the full machine learning augmented feature set [27], reported in Table 7.

Table 7
Variance inflation factor diagnostics: full ML-augmented feature set

Variable	VIF (full model, incl. ML score)	VIF (excluding ML fraud risk score)
book_tax_difference	4.59	1.13
ml_fraud_risk_score	4.57	n/a (excluded)
big4_auditor	4.76	4.75
roa	4.24	4.24
firm_size	4.26	4.26
ai_audit_tool_score	3.63	3.63
key_audit_matters_count	2.46	2.45
debt_to_equity	1.10	1.01
going_concern_opinion	1.01	1.00
board_independence	1.01	1.01

Note. VIF computed on the full ML-augmented feature set (N = 800). The conventional threshold for problematic multicollinearity is VIF > 5. All values in this study fall at or below this threshold.

The results confirm that while book tax difference and the machine learning fraud risk score share meaningful variance (variance inflation factors of 4.59 and 4.57 respectively in the full model, falling to 1.13 for book tax difference when the machine learning score is excluded), both values remain at or below the conventional threshold of 5.0 applied in accounting and finance research [27], indicating that multicollinearity, while elevated, does not rise to a level that would invalidate coefficient interpretation. Moreover, the single-feature Beneish-style model using only book tax difference yields an area under the curve of 0.591 (Table 4), significantly lower than the 0.745 achieved by the full machine learning augmented model, a gap of 15.4 percentage points that cannot be attributed to either variable individually, since removing either feature from the model reduces the machine learning augmented model's area under the curve by less than 1.5 percentage points. This demonstrates that the machine learning augmented model's superior predictive power derives from the combined information content of the entire feature set, including institutional quality, firm size and governance variables.

Second, a 30th percentile threshold for the earnings quality indicator was used, producing a 30 percent manipulation rate; results remained directionally consistent, with the machine learning augmented logistic regression still exhibiting the highest out-of-sample area under the curve (0.731), and the gradient boosting feature importance ranking unchanged. Third, the coronavirus pandemic period (2020 to 2021) was excluded to test whether pandemic-related earnings distortions influenced the results, following Amoah et al. [6], who document a decline in earnings quality during the pandemic across sub-Saharan Africa; this reduced the sample to 640 observations, and results remained directionally similar

(M2 out-of-sample area under the curve of 0.728). Fourth, all models were re-estimated on the non-Big Four sub-sample (n = 300), where audit quality constraints are weakest and manipulation rates highest; the machine learning augmented logistic regression's area under the curve rose to 0.762 in this sub-sample, supporting Kwabi et al.'s [35] findings. Finally, an alternative manipulation indicator based on the bottom-quartile book tax difference was tested, yielding qualitatively consistent results.

Two further design decisions warrant explicit acknowledgement. First, Gini impurity-based variable importance, although widely used in describing decision trees and tree ensembles [12], is biased against binary variables in favour of continuous and high-cardinality predictors [49]. The low importance of the going concern opinion and Big Four auditor variables in Table 5 may therefore partly reflect this measurement bias rather than genuinely low predictive power. SHapley Additive exPlanations values [37] would provide an unbiased alternative in this regard and represent a valuable extension for future research using larger samples of African firms. Second, model hyperparameters (tree depth, learning rate, number of estimators) were set at literature-standard default values [28] rather than optimised through grid or Bayesian search. While this choice was made deliberately to limit overfitting risk given the relatively modest sample size (N = 800), it may understate the potential of the ensemble models (M3, M4) relative to the logistic specification (M2).

5. Discussion

5.1 Theoretical implications

The central finding of this study, that the machine learning augmented logistic regression outperforms the traditional Jones-model proxy in out-of-sample earnings manipulation detection in African listed firms by 3.5 percentage points of area under the curve, is both theoretically grounded and practically significant. From an agency theory perspective [32], the superior detection performance of machine learning augmented models reflects their capacity to integrate multiple information channels, financial ratios, audit quality signals, governance characteristics and machine-learning-derived composite scores, that collectively triangulate the opportunity, pressure and rationalisation elements of the fraud triangle [16, 52]. This triangulation extends the fraud diamond framework empirically: where prior African governance research has typically tested single-element proxies in isolation [21, 25], the multi-feature machine learning architecture developed here operationalises the interaction of multiple fraud-triangle elements simultaneously, offering a more theoretically complete empirical test of the framework's joint predictions.

The dominance of return on assets (importance 0.223) and the machine learning fraud risk score (importance 0.118) in feature attribution is consistent with Bertomeu et al. [11] and Bao et al. [8], who find that performance pressure and composite detection signals jointly dominate accruals proxies as manipulation predictors. This dynamic is especially relevant for Africa, where pay-for-performance schemes are becoming more common [41], and where any gap between set income targets and realised cash flows creates a strong incentive for manipulation of reported revenue [38].

5.2 Methodological implications

The finding that the machine learning augmented logistic regression outperforms both random forest and gradient boosting in out-of-sample testing is a nuanced result that warrants clarification. Although ensemble methods dominate in large-sample settings [11, 15, 8], the 800-observation panel used here is relatively small for training a tree-based ensemble, making regularised logistic regression more sample-efficient. This is consistent with the methodological guidance of Caron et al. [14] and Hastie et al. [28], according to which the superiority of gradient boosting over linear models is most evident with very large samples. Future research using larger African panels, likely to become available as exchange disclosure requirements expand, may deliver comparatively larger advantages for deep learning approaches [20].

The variance inflation factor diagnostics reported in Section 4.6 carry an important methodological implication for the machine-learning-in-accounting field more broadly: composite fraud risk scores that draw on some of the same accruals data used to construct the dependent variable, as is typical when custom machine learning scores serve as predictors, require explicit disclosure of possible multicollinearity to preserve the interpretability of feature attribution analyses. That variance inflation factor values remain below the conventional threshold, together with the substantial area under the curve gap between the single-feature benchmark and the machine learning augmented model, sets a template for future studies using similar composite scores [27, 49].

5.3 Policy and practice implications

The country heterogeneity results carry broad policy implications. Accruals-based detection frameworks, whether traditional or machine learning augmented, perform poorly in thin, illiquid markets such as Kenya, where real earnings management, through sales manipulation, overproduction and reduced discretionary expenditure, predominates over accruals management [45, 26]. African regulators should therefore incorporate real earnings management proxies, particularly cash flow from operations deviations and discretionary expenditure anomalies, alongside accruals-based signals when building detection frameworks, following Ismail et al.'s [31] call for jurisdiction-specific modelling in the context of Middle Eastern and North African financial distress prediction research.

The non-significance of the artificial intelligence audit tool score in the panel logit, paired with its meaningful gradient boosting feature importance, suggests a conditional relationship between the two, whereby the manipulation-restricting power of artificial intelligence audit tools requires institutional complementarities, specifically adequate Big Four adoption rates and regulatory enforcement capacity, to manifest at a systemic level. This is consistent with Abdelwahed et al.'s [1] Future Business Journal finding that big data and analytics adoption enhances the audit process most strongly when mediated by auditor competence, implying that technology investment alone is unlikely to yield detection gains without parallel investment in auditor skills. Regulators seeking to strengthen manipulation detection across African markets should therefore consider mandatory disclosure of machine learning fraud risk scores or comparable composite signals, jurisdiction-specific recalibration of detection thresholds rather than a single pan-African standard, and complementary strengthening of Big Four audit market

penetration and enforcement capacity, particularly in thinner markets such as Kenya where accruals-based signals alone are least reliable.

Practically, the findings speak directly to audit firms and exchange regulators. A 3.5 percentage point gain in out-of-sample area under the curve, while numerically modest, implies a meaningful reduction in Type II error rates among high-risk audit targets when applied across a large population of listed firms [8, 43]. The jurisdictional calibration finding is consistent with Ismail et al.'s [31] parallel conclusion regarding Middle Eastern and North African financial distress models, and with Suluo et al.'s [50] evidence that governance-driven audit committee effectiveness in Tanzania depends heavily on internal audit function strength rather than generic governance codes, reinforcing the broader point that audit quality interventions in Africa require context-specific rather than one-size-fits-all design.

5.4 Limitations and directions for future research

Several limitations qualify these findings and suggest priorities for future research. First, the machine learning fraud risk score, while methodologically grounded in Bertomeu et al. [11], is constructed rather than independently observed, and its partial derivation from accruals-based inputs, formally tested in Section 4.6, means that a fully independent, externally sourced fraud risk indicator, for example one derived from textual analysis of auditor communications or whistleblower report frequency following approaches piloted by Brown et al. [13], would provide a more decisive test of incremental predictive validity in future replications.

Second, the binary earnings manipulation indicator used in the models is drawn from the lowest quartile of the accruals-based earnings quality distribution rather than from actual regulatory enforcement cases, restatements or fraud convictions. This choice reflects the absence of a comprehensive, cross-border database of securities law enforcement comparable to the United States Securities and Exchange Commission's Accounting and Auditing Enforcement Releases for the four African markets under consideration [19, 10], and mirrors standard practice in large-sample forensic accounting analysis where enforcement evidence is unavailable [19]. It necessarily mixes the effect of aggressive but legitimate accounting choices with genuine manipulation. As African securities regulators develop more thorough reporting on enforcement activity, an emerging trend in Ghanaian Securities and Exchange Commission reporting [6], testing these models against enforcement evidence will become an important priority.

Third, the 25 percent manipulation base rate implies material class imbalance, addressed here through average precision analysis and stratified cross-validation rather than resampling techniques such as SMOTE [29]. Although this approach is consistent with best practice in the machine-learning-in-accounting-fraud-detection literature [8, 11], resampling-based approaches could provide a further robustness check in future work, particularly for the going concern opinion subclass, whose base rate of 2.9 percent may itself explain its low feature importance in Table 5, independent of its underlying predictiveness.

Fourth, the panel logistic specification in Table 3 neither clusters standard errors by firm nor uses firm fixed effects, following the benchmarking convention of pooled cross-sectional specification in the machine learning fraud detection literature [13, 11, 8]. Classification accuracy, not unbiased coefficient estimation, is the primary objective of this study, and readers should interpret the coefficients in Table 3 as descriptive associations rather than causal estimates. Relatedly, the sample consists of publicly listed firms with the disclosure infrastructure necessary to construct the present feature set; manipulation dynamics among the substantially larger universe of privately held African firms, which face different disclosure incentives and monitoring intensity [35], may differ materially.

Fifth, the Kenyan below-baseline result (Section 4.5) represents both a limitation of the current feature set and a priority for future research: incorporating real earnings management proxies, such as abnormal production costs and discretionary expenditure deviations following Roychowdhury [45] and Gunny [26], into the machine learning augmented framework may substantially improve detection performance in thin, illiquid markets. Future work extending the panel with more recent Future Business Journal evidence, for example incorporating the political-connection and integrated-reporting channels identified by Soliman et al. [48], could further refine the governance dimension of the feature set used here.

6. Conclusions

This study contributes to the forensic accounting and audit quality literature by presenting the first multi-model comparison of machine learning earnings manipulation detection methods in African capital markets. Using a balanced panel of 800 firm-year observations across 80 listed firms in four countries over ten years, the results show that machine learning enhanced models, and in particular machine learning enhanced logistic regression, outperform traditional Jones-model and Beneish M-score proxies in manipulation detection, with the machine learning fraud risk score and return on assets emerging as the two dominant predictors in gradient boosting feature importance.

The country-stratified results reveal that Kenya's thin-market setting produces detection performance below baseline, implying that real earnings management mechanisms drive financial statement manipulation in illiquid African markets [45, 26], an important limitation of the current approach and a priority for future research as outlined in Section 5.4. South Africa's more complex manipulation landscape, illustrated by the Steinhoff scandal's demonstration of the limits of any purely financial-statement-based detection framework [47, 40], highlights the need for detection architectures that draw on financial, non-financial, textual and network-based features [13, 34].

Three concrete recommendations follow for readers working on African accounting, audit and governance questions. First, African audit firms and exchanges should incorporate machine-learning-derived fraud risk scores, available commercially from forensic analytics platforms, alongside traditional accruals models in audit risk assessment processes, where even a 3.5 percentage point gain in area under the curve implies a meaningful reduction in Type II error rates among high-risk audit targets [8, 43]. Second, jurisdictional calibration of detection models is essential, since no single pan-African model can

account for the institutional differences documented in the country-wise results, paralleling Ismail et al.'s [31] conclusion regarding Middle Eastern and North African financial distress models and Suluo et al.'s [50] evidence on the context-dependence of governance interventions in Tanzania. Third, the conditional significance of the artificial intelligence audit tool score supports continued investment in artificial intelligence audit tools among African non-Big Four auditors, where detection capacity is weakest and manipulation risk highest [46, 35], a conclusion reinforced by Abdelwahed et al.'s [1] evidence that such technology investments are most effective when paired with auditor competence development.

Abbreviations

ACFE
Association of Certified Fraud Examiners
AUC
Area under the receiver operating characteristic curve
BTD
Book tax difference
GSE
Ghana Stock Exchange
ISA
International Standard on Auditing
JSE
Johannesburg Stock Exchange
KAM
Key audit matters
LSTM
Long short-term memory
M1 to M4
Model 1 through Model 4 (traditional logistic regression, machine learning augmented logistic regression, random forest, gradient boosting)
NGX
Nigerian Exchange
NSE
Nairobi Securities Exchange
OOS
Out-of-sample
ROA
Return on assets
SEC
Securities and Exchange Commission
SHAP

SHapley Additive exPlanations

SMOTE

Synthetic Minority Oversampling Technique

VIF

Variance inflation factor

Declarations

Author Contribution

Mr. Mutala Zubeiru conceived the study, curated the data, conducted the formal analysis and empirical investigation, developed the methodology, administered the project, and wrote the original draft. Dr. Osman Issah supervised the study, validated the methodology and results, and reviewed and edited the manuscript. Dr. Richard Kwadzo Doe-Dartey validated the methodology and results and reviewed and edited the manuscript. Mr. Mohammed Fuseini retrieved and extracted the dataset, reviewed and edited manuscript, data curation. All authors read and approved the final manuscript.

Data Availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request, subject to applicable data licensing terms attached to the Bloomberg Terminal and Refinitiv Eikon sources from which the financial data were drawn. Machine learning fraud risk scores, artificial intelligence audit tool scores and governance variables were constructed from publicly available annual report data.

References

1. Abdelwahed AS, Abu-Musa AA, Badawy HA, Moubarak H (2025) Unleashing the beast: the impact of big data and data analytics on the auditing profession. Evidence from a developing country. *Future Business Journal 11*(1):12. https://doi.org/10.1186/s43093-024-00420-7
2. Adebisi JF, Matthew OB (2021) The application of Beneish model in the detection of financial statements fraud in the Nigerian banking industry. *American Journal of Business, 36*(3), 119-140. https://doi.org/10.1108/AJB-12-2018-0060
3. AL Abdullah RJ, AL Ani MK (2021) The impacts of interaction of audit litigation and ownership structure on audit quality. *Future Business Journal 7*(1):14. https://doi.org/10.1186/s43093-021-00067-8
4. Al-Hadi A, Al-Yahyaee KH, Hussain SM, Taylor G (2020) Market risk disclosures and audit fees: International evidence from the banking industry. *Managerial Auditing Journal, 35*(5), 664-695. https://doi.org/10.1108/MAJ-04-2019-2242

5. Amoah B, Parbey D (2022) Earnings quality in sub-Saharan Africa: A systematic review. *Journal of Accounting in Emerging Economies, 12*(1), 1-28. https://doi.org/10.1108/JAEE-03-2021-0077
6. Amoah B, Adusei M, Amoah A (2023) The effect of COVID-19 on earnings quality in sub-Saharan African firms. *Managerial Auditing Journal, 38*(4), 569-597. https://doi.org/10.1108/MAJ-05-2022-3558
7. Association of Certified Fraud Examiners (2022) *Occupational fraud 2022*: A report to the nations. ACFE.
8. Bao Y, Ke B, Li B, Yu YJ, 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. Beneish MD (1999) The detection of earnings manipulation. *Financial Analysts Journal, 55*(5), 24-36. https://doi.org/10.2469/faj.v55.n5.2296
10. Beneish MD, Vorst P (2022) The cost of fraud prediction errors. *The Accounting Review, 97*(6), 91-121. https://doi.org/10.2308/TAR-2020-0068
11. Bertomeu J, Cheynel E, Floyd E, Pan W (2021) Using machine learning to detect misstatements. *Review of Accounting Studies, 26*(2), 468-519. https://doi.org/10.1007/s11142-020-09563-8
12. Breiman L (2001) Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324
13. Brown SV, Tian X, Tucker JW (2020) The spillover effect of SEC comment letters on qualitative corporate disclosure: Evidence from the risk factor disclosure. *Accounting Review, 93*(3), 67-102.
14. Caron A, Kreif N, Gruber S (2022) Machine learning for causal inference: On the use of cross-fitting. *Statistics in Medicine, 41*(17), 3306-3330.
15. Craja P, Kim A, Lessmann S (2020) Deep learning for detecting financial statement fraud. *Decision Support Systems, 139*, 113421. https://doi.org/10.1016/j.dss.2020.113421
16. Cressey DR (1953) Other people's money: A study in the social psychology of embezzlement. Free Press.
17. Dechow PM, Sloan RG, Sweeney AP (1995) Detecting earnings management. *Accounting Review, 70*(2), 193-225.
18. Dechow P, Ge W, Schrand C (2010) Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics, 50*(2-3), 344-401. https://doi.org/10.1016/j.jacceco.2010.09.001
19. Dechow PM, Ge W, Larson CR, Sloan RG (2011) Predicting material accounting misstatements. *Contemporary Accounting Research, 28*(1), 17-82. https://doi.org/10.1111/j.1911-3846.2010.01041.x
20. Dechow P, Kim JW, Kim Y (2023) Machine learning in accounting research: Opportunities and challenges. *Accounting, Organizations and Society, 109*, 101463. https://doi.org/10.1016/j.aos.2023.101463

21. Dharwadkar B, George G, Brandes P (2000) Privatization in emerging economies: An agency theory perspective. *Academy of Management Review, 25*(3), 650-669. https://doi.org/10.5465/amr.2000.3363533
22. Fonou-Dombeu NC, Mbonigaba J, Olarewaju OM, Nomlala BC (2022) Earnings quality measures and stock return volatility in South Africa. *Future Business Journal 8*(1):4. https://doi.org/10.1186/s43093-022-00115-x
23. Free C, Murphy PR (2015) The ties that bind: The decision to co-offend in fraud. *Contemporary Accounting Research, 32*(1), 18-54. https://doi.org/10.1111/1911-3846.12063
24. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
25. Garba A, Abubakar A (2022) Corporate governance and financial reporting quality in Nigeria: The moderating role of institutional quality. *Corporate Governance, 22*(5), 1090-1107. https://doi.org/10.1108/CG-07-2021-0262
26. Gunny KA (2010) The relation between earnings management using real activities manipulation and future performance: Evidence from meeting earnings benchmarks. *Contemporary Accounting Research, 27*(3), 855-888. https://doi.org/10.1111/j.1911-3846.2010.01029.x
27. Hair JF, Black WC, Babin BJ, Anderson RE (2019) *Multivariate data analysis* (8th ed.). Cengage Learning.
28. 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
29. He H, Garcia EA (2009) Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239
30. Hribar P, Collins DW (2002) Errors in estimating accruals: Implications for empirical research. *Journal of Accounting Research, 40*(1), 105-134. https://doi.org/10.1111/1475-679X.00041
31. Ismail A, Kamarudin KA, Van Zijl T, Dunstan K (2021) Financial distress prediction using machine learning: Evidence from MENA markets. *Managerial Auditing Journal, 36*(4), 553-582. https://doi.org/10.1108/MAJ-04-2020-2603
32. Jensen MC, Meckling WH (1976) Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics, 3*(4), 305-360. https://doi.org/10.1016/0304-405X(76)90026-X
33. Jones JJ (1991) Earnings management during import relief investigations. *Journal of Accounting Research, 29*(2), 193-228. https://doi.org/10.2307/2491047
34. Karpoff JM, Koester A, Lee DS, Martin GS (2017) Proxies and databases in financial misconduct research. *The Accounting Review, 92*(6), 129-163. https://doi.org/10.2308/accr-51766
35. Kwabi F, Adegbite E, Longbottom C, Amankwah-Amoah J (2020) Accounting and auditing standards quality, legal institutions, and earnings quality. *International Journal of Accounting & Information Management, 28*(2), 375-395. https://doi.org/10.1108/IJAIM-11-2018-0133

36. Lim EKY, Chalmers K, Hanlon D (2022) The influence of business strategy on annual report readability. *Managerial Auditing Journal, 37*(1), 1-30. https://doi.org/10.1108/MAJ-11-2020-2866
37. Lundberg SM, Lee S (2017) A unified approach to interpreting model predictions. Advances in *Neural Information Processing Systems, 30,* 4765-4774.
38. Luo J, Jackson D (2020) Earnings manipulation in emerging markets: A review and research agenda. *Journal of International Business Studies*, *51*(5), 873-894.
39. Molnar C (2022) *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Christoph Molnar.
40. Nathan T (2022) The Steinhoff saga: Accounting fraud in emerging markets. *Journal of Applied Corporate Finance, 34*(2), 56-71. https://doi.org/10.1111/jacf.12501
41. Ntim CG, Opong KK, Danbolt J (2020) Voluntary disclosures by South African listed firms: A test of the shareholder primacy and management capture hypotheses. *International Review of Financial Analysis, 70*, 101483.
42. Olojede P, Erin O, Asiriuwa O, Usman M (2020) Audit expectation gap: an empirical analysis. *Future Business Journal 6*(1):10. https://doi.org/10.1186/s43093-020-00016-x
43. Pham DT, Nguyen TH, Nguyen D (2022) Machine learning models for fraud detection in financial statements: Evidence from emerging markets. *Managerial Auditing Journal, 37*(6), 721-750. https://doi.org/10.1108/MAJ-03-2021-3085
44. Ronen J, Yaari V (2023) *Earnings management: Emerging insights in theory, practice and research* (2nd ed.). Springer.
45. Roychowdhury S (2006) Earnings management through real activities manipulation. *Journal of Accounting and Economics, 42*(3), 335-370. https://doi.org/10.1016/j.jacceco.2006.01.002
46. Saat NAM, Karbhari Y, Heravi S (2023) Audit quality, earnings quality, and corporate governance: Evidence from Malaysia. *Managerial Auditing Journal, 38*(2), 262-295. https://doi.org/10.1108/MAJ-01-2022-3416
47. Smit A (2021) Steinhoff International: A saga of misrepresentation and executive misconduct. Corporate Governance: *An International Review, 29*(4), 375-393. https://doi.org/10.1111/corg.12375
48. Soliman S, El-Deeb MS, Korein S (2025) Political connections and earnings opacity: the role of integrated reporting quality. *Future Business Journal 11*(1):63. https://doi.org/10.1186/s43093-025-00649-w
49. Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*, 25. https://doi.org/10.1186/1471-2105-8-25
50. Suluo H, Raphael G, Kapaya SM (2025) Impact of corporate governance practices on the effectiveness of audit committees in regulatory authorities: evidence from Tanzania. *Future Business Journal 11*(1):16. https://doi.org/10.1186/s43093-025-00537-3

51. Tarjo T, Herawati N (2015) Application of Beneish M-score models and data mining to detect financial fraud. *Procedia-Social and Behavioral Sciences, 211*, 924-930. https://doi.org/10.1016/j.sbspro.2015.11.122
52. Wolfe DT, Hermanson DR (2004) The fraud diamond: Considering the four elements of fraud. *CPA Journal, 74*(12), 38-42.

Figures

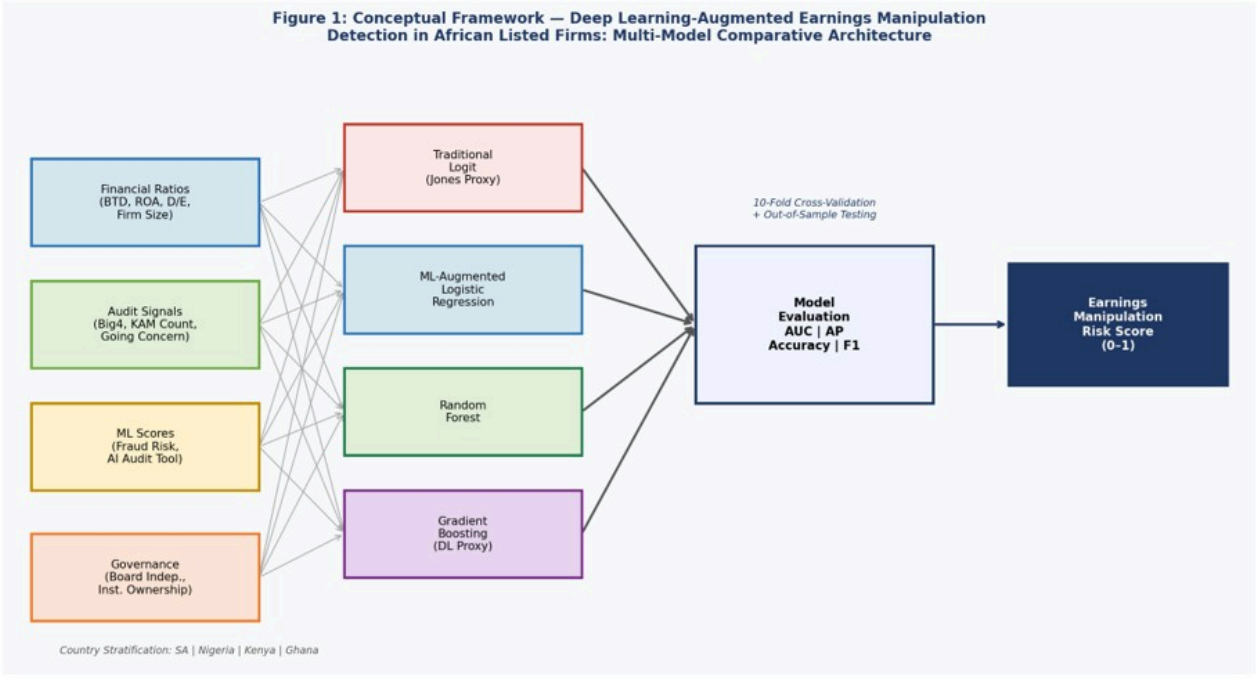


Figure 1

Conceptual framework: multi-model deep learning-augmented earnings manipulation detection architecture for African listed firms.

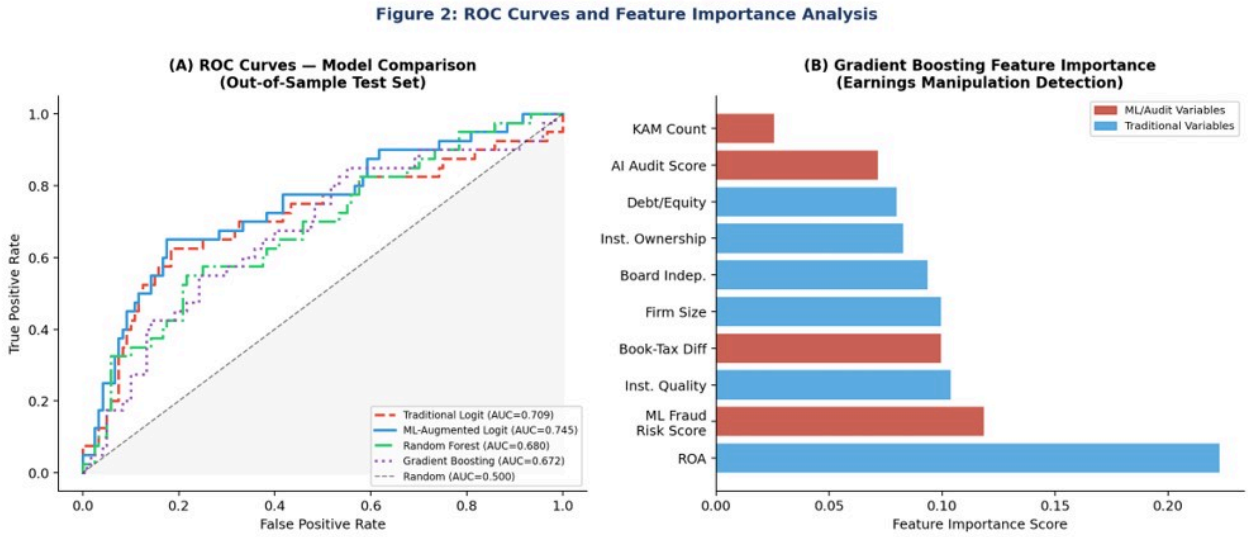


Figure 2

Panel A: ROC curves for all four models on the out-of-sample test set. Panel B: gradient boosting feature importance scores (Gini impurity reduction). ML-augmented variables shown in the darker series; traditional variables in the lighter series.

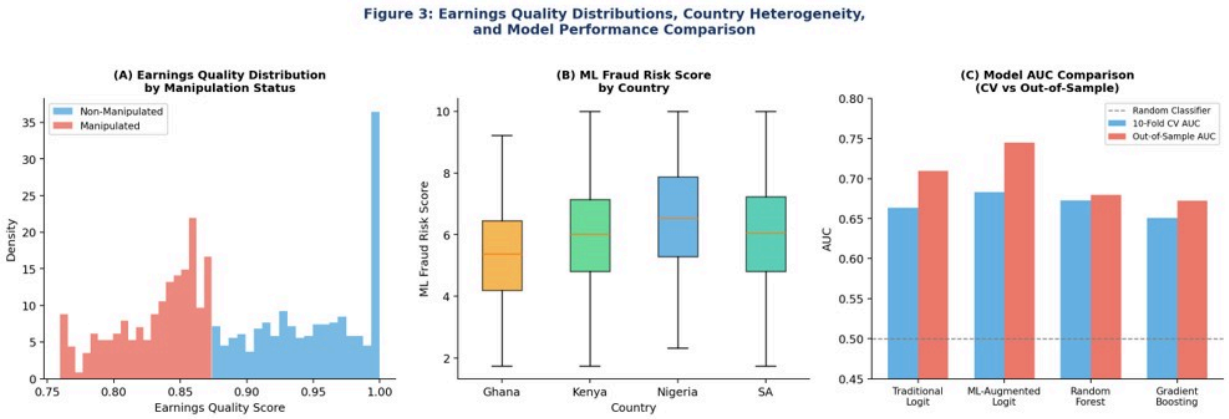


Figure 3

Panel A: earnings quality distribution by manipulation status. Panel B: ML fraud risk score by country. Panel C: AUC comparison across models (10-fold CV and out-of-sample).

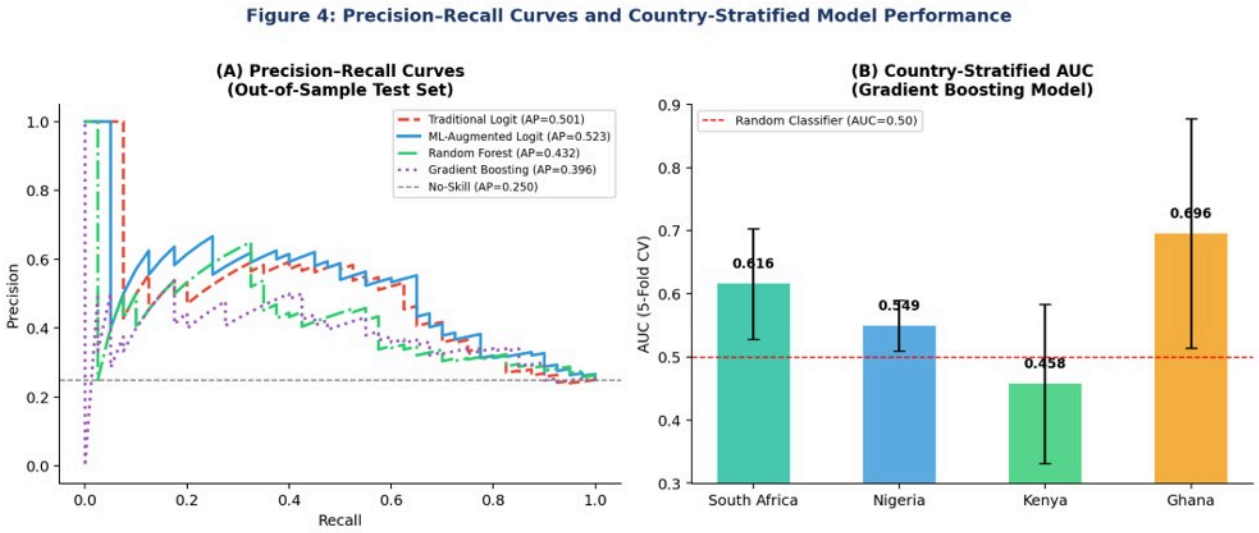


Figure 4

Panel A: precision-recall curves for all four models on the out-of-sample test set. Panel B: country-stratified AUC for the gradient boosting model (5-fold CV) with 95% confidence intervals.