

Harnessing Machine Learning to Detect Financial Statement Fraud in Real Time

Christian Iyafekhe

Edo State College of Education, Igueben, Edo State, Nigeria

Corresponding author: Christian Iyafekhe

ABSTRACT

Financial statement fraud remains difficult to detect because manipulators exploit accrual discretion, related-party structures, and delayed audit cycles. Classical ratio models such as the Beneish M-score and the Dechow F-score remain useful screens, but they use fixed coefficients estimated on older samples. This article reviews machine-learning approaches to financial statement fraud detection and specifies a real-time screening design suitable for a Nigerian college of education research project. Evidence from cross-country and market-specific studies shows that ensembles combining gradient boosting with distance-based classifiers, and models that reuse the inputs of the Altman, Beneish, Montier, and Dechow scores, can exceed the roughly 78% average accuracy reported across earlier algorithm comparisons. A 2025 cross-country ensemble reached 85.69% overall accuracy on a held-out set and flagged part of the Wirecard record retrospectively. The article argues that “real time” should mean continuous ingestion of interim filings, board packs, and bank-reconciliation exceptions, not a claim of perfect contemporaneous detection. Implications for auditors, audit committees, and teacher-education programmes in accounting are discussed. This paper does not report primary data collected at the college.

Keywords: *financial statement fraud; machine learning; Beneish M-score; earnings manipulation; Nigeria; audit analytics*

1. Introduction

Financial statement fraud is the deliberate misrepresentation of an entity's financial position or performance. It is less common than asset misappropriation, but it is the most expensive category of occupational fraud in median loss (Association of Certified Fraud Examiners [ACFE], 2024). Detection is late because the same managers who prepare the statements control many of the documents auditors sample. The fraud triangle still organises motive, opportunity, and rationalisation (Cressey, 1953), but opportunity has widened with complex group structures, non-GAAP metrics, and digital close processes.

Nigerian preparers and auditors work under the Companies and Allied Matters Act 2020, Financial Reporting Council of Nigeria rules, and International Standards on Auditing, including ISA 240 on the auditor's responsibilities relating to fraud. Colleges of education that train business and accounting teachers need a research base that connects these standards to tools students will meet in practice. The approved project, “Harnessing Machine Learning to Detect Financial Statement Fraud in Real-Time,” is a five-month Institution-Based Research award to Edo State College of Education, Igueben. This article states the problem, synthesises verifiable evidence, and sets out a design that can be executed without fabricating local results.

The research questions are as follows. First, which published machine-learning designs improve on fixed-coefficient manipulation scores? Second, what data latency is realistic for a real-time screen in a jurisdiction where many entities still file annually? Third, how should a college of education study be specified so that a later empirical paper can meet a journal impact-factor threshold of at least 3, as required for release of the 40% second tranche?

2. Literature Review

2.1 Classical scores

Beneish (1999) estimated a probit model on eight ratios, days' sales in receivables, gross margin, asset quality, sales growth, depreciation, sales and general administration, accruals, and leverage, and reported that the model identified a large share of manipulators in a hold-out sample while misclassifying a smaller share of non-manipulators. The eight-variable M-score remains a common benchmark. Altman (1968) was built for bankruptcy, not fraud, but distress and manipulation co-occur often enough that the Z-score is reused as a feature. Dechow et al. (2011) developed an F-score from Accounting and Auditing Enforcement Releases, using accrual quality, performance, and off-balance-sheet proxies. These models are transparent. Their weakness is stationarity: coefficients estimated on U.S. manufacturing firms in the 1980s and 1990s need not rank Nigerian or 2020s digital firms correctly (Özari et al., 2025).

2.2 Machine learning

Kirkos et al. (2007) compared decision trees, neural networks, and Bayesian belief networks on a small Greek sample and showed that data-mining classifiers could separate fraudulent from non-fraudulent statements. Perols (2011) warned that fraud data are rare, so accuracy alone misleads. Support-vector machines and logistic regression behaved differently once class imbalance and the cost of false negatives were modelled. Cecchini et al. (2010) used a support-vector machine with a custom financial kernel. Bao et al. (2020) reported that an ensemble of raw accounting numbers outperformed a logistic model built on the Dechow ratio set in out-of-sample U.S. tests, which matters because ratio construction itself embeds researcher choices.

Reviews converge on three points. Data-mining and machine-learning methods dominate the fraud-detection literature after 2000, but many papers optimise on one market and one split (Ngai et al., 2011; West & Bhattacharya, 2016). Performance claims are not comparable when authors report accuracy on balanced test sets and omit precision at a realistic base rate. Feature sets that reuse established scores transfer more easily than opaque text embeddings when labelled fraud cases are scarce.

2.3 Recent evidence

Steingen and Löw (2025) assembled 2,014,827 firm-years from 1988 to 2019, of which 1,145 were labelled fraudulent, then trained on a balanced set of 1,046 fraudulent and 1,046 non-fraudulent years. Inputs were the Altman Z-score, Beneish M-score, Montier C-score, and Dechow F-score. An ensemble of gradient boosting and k-nearest neighbours classified 82.03% of manipulated and 89.88% of non-manipulated firm-years on the held-out set (overall accuracy 85.69%). Applied to Wirecard, the model flagged 7 of 17 firm-years. The authors noted that the average accuracy of 84 previously analysed algorithms was 78.37%. Özari et al. (2025) dropped the original coefficients of the Z-score and M-score and fed the underlying ratios to a random forest, checked against Borsa Istanbul fines from 2018 to 2022. Both papers support coefficient-free or re-weighted use of classical inputs rather than abandonment of accounting theory.

A smaller case application of the M-score to three U.S. Securities and Exchange Commission matters found that the score flagged Luckin Coffee in the peak fraud window and was less decisive for other filers (Journal of Lifestyle and SDGs Review, 2025). That pattern matches Beneish's own caution: the score is a screen, not a verdict.

2.4 Real time

No published model detects fraud at the instant a false journal entry is posted unless it sits inside the entity's general ledger. What the literature can support is near-real-time screening: quarterly or monthly recomputation of ratios, exception queues

from reconciliations, and narrative inconsistency checks on management discussion. ISA 240 already requires auditors to treat journal entries and significant estimates as fraud risks. A college project can operationalise that duty as a dashboard specification without claiming production deployment in five months.

3. Conceptual Framework

The framework has three layers. The incentive layer uses the fraud triangle and agency conflict between preparers and capital providers (Cressey, 1953; Jensen & Meckling, 1976). The signal layer uses accrual and margin ratios from Beneish (1999) and Dechow et al. (2011), plus distress from Altman (1968). The learning layer treats fraud as a rare binary label and prefers models that report recall on the fraud class, precision, and area under the precision-recall curve, not accuracy alone (Perols, 2011). Real time is defined as a refresh no slower than the entity's interim close.

4. Proposed Method

The design is a supervised classification study on publicly filed financial statements, with a Nigerian sub-sample if labels can be built from Financial Reporting Council sanctions, Securities and Exchange Commission Nigeria actions, and court-reported cases. Where local labels are fewer than 50 entity-years, the study will train on a published cross-country set and test transfer to Nigerian IFRS filers.

Inputs are the eight Beneish ratios, Altman inputs, Dechow F-score inputs, and a small set of governance flags (auditor change, qualified opinion, related-party note length). Raw totals will be retained so that a raw-number ensemble in the style of Bao et al. (2020) can be compared with a score-input ensemble in the style of Steingen and Löw (2025). Models are logistic regression as the baseline, random forest, gradient boosting, and a boosting plus k-nearest-neighbours ensemble. Class imbalance will be handled by class weights, not by reporting only a balanced test set.

Validation uses a temporal split: train on earlier years, test on later years. Metrics are fraud-class recall, precision, specificity, and precision-recall area. A confusion matrix at a threshold chosen for high recall will be reported because missing fraud is costlier than a false alarm that triggers a review (Beneish & Vorst, 2022). Only public filings and public enforcement labels will be used. No non-public client data from audit firms will be requested in the five-month window. This method section is a protocol. The results below are a synthesis of published numbers, not a pilot from Igueben.

5. Synthesis of Published Results

Fixed-coefficient screens remain informative. Beneish's original out-of-sample work identified about three quarters of manipulators with a false-positive rate near 17.5% among non-manipulators, a trade-off later papers still use as a reference point. Machine-learning studies improve ranking when they keep accounting features and respect time. Steingen and Löw (2025) exceed the 78.37% average they calculated from earlier algorithms, but their balanced evaluation set overstates precision at a real base rate of far below 1% fraudulent firm-years. Bao et al. (2020) show gains from ensembles on raw numbers in U.S. data; transfer to Nigerian GAAP-to-IFRS history is untested. Özari et al. (2025) show that random forests can use Z-score and M-score ratios without the original weights and still align with an enforcement label in one emerging market.

For a real-time claim, the binding constraint is filing lag, not algorithm speed. A model can score a ledger extract in seconds. Nigerian listed companies file interim results on an exchange calendar; most colleges, schools, and small entities do not. A college of education study should therefore split the product into an annual-filing screen for listed firms and a monthly reconciliation-exception screen for the college's own bursary,

framed as an internal-control teaching case rather than as a fraud accusation.

6. Discussion

The approved budget (₦3,700,000 total; ₦2,220,000 first tranche) can cover data licences or open-source stacks, a small labelling exercise, and article processing if a target journal charges fees. It cannot cover a production surveillance system. The scientific contribution available in five months is a transparent benchmark: classical scores versus coefficient-free machine learning on a documented label set, with an explicit base-rate adjustment.

Three limitations follow. First, enforcement labels miss undetected fraud, so recall is downward biased. Second, Nigerian cases are sparse, so a paper that claims a Nigeria-only accuracy will be under-powered. Third, impact factor is a funder condition, not a scientific one. Journals with an impact factor of 3 or higher in accounting analytics are few, and desk rejection is common when results are simulated. The second-tranche note in the 5 December 2025 TETFund letter requires at least one article from the project findings in a refereed journal with an impact factor of at least 3. That bar should shape journal choice early, but it should not be met by manufactured local findings.

7. Conclusion

Machine learning can raise the sensitivity of financial statement fraud screens when it is anchored to Beneish, Dechow, and Altman inputs, validated out of time, and judged on rare-event metrics. Real-time detection is feasible only to the extent source systems refresh. For Edo State College of Education, Igueben, the executable project is a reproducible benchmark and a bursary teaching case, written up with public data and with published performance figures attributed to their original authors.

8. Funding

This article was prepared for the 2024/2025 TETFund Institution-Based Research intervention at Edo State College of Education, Igueben. It does not report primary data collected at the college. Figures cited from other studies are those authors' published results.

References

- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- Association of Certified Fraud Examiners. (2024). *Occupational fraud 2024: A report to the nations*. ACFE.
- Bao, Y., Ke, B., Li, B., Yu, 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>
- Beneish, M. D. (1997). Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy*, 16(3), 271–309. [https://doi.org/10.1016/S0278-4254\(97\)00023-9](https://doi.org/10.1016/S0278-4254(97)00023-9)
- 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>
- Beneish, M. D., & Vorst, P. (2022). The cost of fraud prediction errors. *The Accounting Review*, 97(6), 91–121. <https://doi.org/10.2308/TAR-2020-0068>
- 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>
- Cressey, D. R. (1953). *Other people's money: A study in the social psychology of embezzlement*. Free Press.
- 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>
- International Auditing and Assurance Standards Board. (2021). *ISA 240: The auditor's responsibilities relating to fraud in an audit of financial statements*. IFAC.
- Jensen, M. C., & Meckling, W. H. (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](https://doi.org/10.1016/0304-405X(76)90026-X)
- 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>
- Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569.
<https://doi.org/10.1016/j.dss.2010.08.006>
- Özari, Ç., Can, E. N., & Demirkale, Ö. (2025). Financial fraud detection with Altman Z-score and Beneish M-score via random forest: Verified by Borsa İstanbul fines (2018–2022). *SAGE Open*, 15(4).
<https://doi.org/10.1177/21582440251386174>
- 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>
- Price, R. A., Sharp, N. Y., & Wood, D. A. (2011). Detecting and predicting accounting irregularities: A comparison of commercial and academic risk measures. *Accounting Horizons*, 25(4), 755–780.
<https://doi.org/10.2308/acch-50064>
- Steingen, L., & Löw, E. (2025). Using machine learning to detect financial statement fraud: A cross-country analysis applied to Wirecard AG. *Journal of Risk and Financial Management*, 18(11), 605.
<https://doi.org/10.3390/jrfm18110605>
- West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57, 47–66.
<https://doi.org/10.1016/j.cose.2015.09.005>