International Journal of Innovative Inventions in Social Science and Humanities

ISSN(print): 3050-6921, ISSN(online): 3050-693X

Volume 03 Issue 08 August 2026

DOI: 10.58806/ijiissh.2026.v3i8n07

Page No: 995-1002

From Sampling to Surveillance: Evaluating the Effectiveness of Artificial Intelligence in Detecting Complex Financial Fraud

Dr. Gaduga Godwin, Esq

Lecturer at Accra Institute of Technology Adjunct Lecturer at Blue Crest University College

ABSTRACT

For most of the twentieth century, the detection of financial fraud rested on an uncomfortable compromise: because no auditor or investigator could examine every transaction, assurance was built on samples, and fraud that fell outside the sample escaped notice. Artificial intelligence promises to dissolve that compromise by subjecting entire populations of transactions, disclosures, and communications to continuous algorithmic scrutiny. This article evaluates how far that promise has been kept. Drawing on three decades of empirical research in accounting, information systems, and computer science, it examines the performance of supervised classifiers, anomaly detection, natural language processing, and network analytics against complex schemes such as financial statement manipulation, collusive procurement fraud, and layered transaction fraud. The evidence supports a qualified conclusion. Machine learning models now outperform traditional ratio-based screens by meaningful margins, yet their effectiveness is constrained by severe class imbalance, biased training labels drawn only from detected fraud, adversarial adaptation by offenders, and opacity that sits awkwardly with evidentiary standards in criminal and regulatory proceedings. The article argues that artificial intelligence is best understood as an instrument of triage rather than adjudication, and it draws out the governance, forensic, and pedagogical consequences of that position for both mature and emerging markets, including African jurisdictions such as Ghana.

ARTICLE DETAILS

Published On:
15 August 2026

KEYWORDS: Financial Fraud, Artificial Intelligence, Machine Learning, Forensic Accounting, Audit Sampling, Continuous Monitoring, Anomaly Detection, Explainability.

Available on:
https://ijiissh.com/

1. INTRODUCTION

Fraud has always exploited the gap between what organizations do and what their overseers can feasibly inspect. The Association of Certified Fraud Examiners (ACFE, 2024), in its most recent global study of 1,921 occupational fraud cases, estimated that the typical organization loses about five per cent of its annual revenue to fraud, and reported a median loss of US$145,000 per case, with complex schemes involving senior management costing far more and lasting far longer before discovery. Those figures are estimates built from cases that were eventually uncovered, which means the true cost of fraud is by definition unknowable, and probably larger. What the figures do establish beyond serious dispute is that fraud is not a marginal irritant to commercial life but a persistent tax on it, and that the mechanisms societies have relied upon to detect it leave a great deal undetected for a very long time.

The dominant detection mechanism of the twentieth century was the sample. Auditors, examiners, and investigators drew statistically or judgmentally selected subsets of transactions, tested them against supporting evidence, and extrapolated their findings to the population as a whole. Sampling was never designed to catch fraud; it was designed to support an opinion on whether financial statements were materially misstated, on the charitable assumption that errors are distributed more or less randomly. Fraud is not distributed randomly. It is concealed deliberately, placed where scrutiny is weakest, and defended by the very managers whose representations the auditor must rely upon (International Auditing and Assurance Standards Board [IAASB], 2009). The result, documented consistently across two decades of ACFE surveys, is that formal audit procedures detect only a modest fraction of occupational fraud, while tips from employees and outsiders remain the leading detection channel, accounting for roughly forty-three per cent of cases in the most recent study (ACFE, 2024).

Over the past fifteen years, a different paradigm has taken shape. Falling data storage costs, the digitization of accounting records, and advances in machine learning have made it technically possible to move from inspecting samples to monitoring populations: every journal entry, every vendor payment, every insurance claim, every narrative disclosure, screened continuously by algorithms

trained to recognise the statistical shadows that fraud casts. I describe this movement as a shift from sampling to surveillance, and the word surveillance is chosen deliberately, because it captures both the analytical ambition of the new methods and the civil liberties anxieties they provoke. A system that watches everything to find the fraudulent few necessarily watches the honest many.

This article asks a question that is easy to pose and surprisingly hard to answer: how effective is artificial intelligence, in fact, at detecting complex financial fraud? The question matters for auditors deciding how much weight to place on algorithmic screens, for regulators and prosecutors deciding whether model outputs can ground enforcement action, for boards deciding what to buy, and for universities deciding what to teach. It matters with particular force in emerging markets, where supervisory resources are thin, and the appeal of automated oversight is correspondingly strong. Yet the vendor literature answers it with enthusiasm rather than evidence, and the academic literature, though rich, is scattered across accounting, information systems, statistics, criminology, and law, with each discipline holding a piece of the picture.

I aim to assemble those pieces into a single evaluative account. The article proceeds as follows. Section 2 reconstructs the sampling paradigm and explains, by reference to auditing standards and fraud theory, why it systematically underperforms against concealed wrongdoing. Section 3 traces the conceptual shift toward full-population, continuous monitoring. Section 4 surveys the principal families of artificial intelligence techniques applied to fraud detection and the empirical record of each. Section 5, the analytical core of the article, evaluates that record critically, attending to class imbalance, label bias, adversarial adaptation, cost asymmetries, and the explainability problem as it arises in forensic and evidentiary settings. Section 6 examines the governance and regulatory environment, including data protection and the risk-based approach of the European Union's Artificial Intelligence Act, with attention to African jurisdictions. Section 7 draws implications for forensic practice and professional education, Section 8 acknowledges limitations and charts a research agenda, and Section 9 concludes.

2. THE SAMPLING PARADIGM AND ITS LIMITS

Audit sampling emerged as a pragmatic response to scale. Once enterprises grew beyond what a single examiner could inspect item by item, the profession accepted that an opinion on financial statements would have to rest on partial evidence, and it developed statistical and judgmental sampling techniques to make that partial evidence defensible. The logic is sound for its intended purpose. If misstatements arise from error, and errors occur with some roughly stable frequency across a population, then a properly drawn sample supports an inference about the population at a known confidence level. Generations of auditors were trained on that logic, and the architecture of the modern audit, from materiality thresholds to tolerable misstatement, is built upon it.

Fraud defeats the logic at its foundation, because fraud violates the assumption of random distribution. Cressey's (1953) classic study of convicted embezzlers established that fraud arises from the conjunction of perceived pressure, perceived opportunity, and rationalization, and the opportunity element is decisive here: offenders locate their schemes precisely where controls are weakest, and inspection is least likely. Later theoretical work extended the triangle into richer models incorporating capability and arrogance, and re-centred attention on the predator who defrauds deliberately and repeatedly rather than the trusted employee who succumbs to pressure (Dorminey et al., 2012). Whichever variant one prefers, the practical consequence is the same. A fraudulent journal entry is not a random deposit of dirt awaiting a sweep of the floor; it is a needle hidden by someone who knows where the magnet will pass.

Auditing standards have long acknowledged the problem without being able to solve it within the sampling frame. International Standard on Auditing 240 requires the auditor to maintain professional skepticism, to treat management override of controls as a significant risk in every engagement, and to incorporate an element of unpredictability into the selection of procedures (IAASB, 2009), and the equivalent American standard imposes materially similar obligations (Public Company Accounting Oversight Board [PCAOB], 2002). These requirements are candid admissions that a predictable sample is an evadable sample. Yet unpredictability within a sampling regime only relocates the needle; it does not examine the whole haystack. The empirical record reflects this. External audit has never been the leading channel through which occupational fraud comes to light, trailing well behind tips and management review in every ACFE study of the past two decades (ACFE, 2024).

The profession's first responses to this weakness were analytical rather than technological. Beneish (1999) constructed a probit model, the M-score, from eight indices computed on published financial statements, capturing patterns such as deteriorating receivables quality and unusual sales growth, and showed that it identified a substantial majority of known earnings manipulators in holdout testing, at the cost of misclassifying a share of compliant firms. Dechow et al. (2011) built the F-score along similar lines from a large sample of enforcement actions by the United States Securities and Exchange Commission, producing a scaled measure of misstatement risk that remains a standard benchmark in the literature. Nigrini (2012) systematized the forensic use of Benford's law, which describes the expected distribution of leading digits in naturally occurring numerical data and flags populations whose digit frequencies depart from it. These tools mattered, and they still matter, but they are screens computed on aggregates, blunt by design, and every one of their input ratios is public knowledge available to the sophisticated manipulator.

3. FROM SAMPLING TO SURVEILLANCE

Three developments converted a methodological aspiration into an operational possibility. The first was the wholesale digitization of accounting, banking, and procurement records, which turned transactions into machine-readable data as a by-product of ordinary business. The second was the collapse in the cost of storing and processing that data, which removed the economic rationale for sampling: when the marginal cost of testing one more transaction approaches zero, the case for testing only four hundred of them weakens considerably. The third was the maturation of machine learning methods capable of finding structure in high-dimensional data without requiring the analyst to specify in advance what fraud looks like. Together these developments support what the literature calls full-population testing and continuous auditing, in which automated routines examine every record as it is created and raise exceptions in something close to real time (Gepp et al., 2018).

The shift is conceptual as much as technical. Under sampling, detection is episodic, retrospective, and inferential: the examiner visits the data periodically, looks backward, and generalizes from the part to the whole. Under surveillance, detection is continuous, contemporaneous, and exhaustive: the system never leaves, evaluates conduct as it occurs, and generalizes from the whole to the part, asking of each transaction how it compares with the population it belongs to. That inversion changes the economics of concealment. A fraudster no longer needs to be unlucky to be caught in a sample; the fraudster needs to produce records whose statistical fingerprint is indistinguishable from legitimate business across every dimension the model measures, which is a far more demanding task, though, as Section 5 will show, not an impossible one.

The word surveillance also signals a cost that the fraud detection literature has been slow to price. Continuous monitoring of all transactions is continuous monitoring of all transactors, most of whom are honest employees, customers, and counterparties whose conduct is being scored by systems they cannot see. Bolton and Hand (2002) observed at the very beginning of this literature that fraud detection is unusual among classification problems in that the objects being classified are people with legal interests, and the intervening two decades of data protection law have sharpened that observation into a binding obligation. Any evaluation of effectiveness that ignores this dimension is incomplete, and I return to it in Section 6.

4. ARTIFICIAL INTELLIGENCE TECHNIQUES AND THE EMPIRICAL RECORD

Artificial intelligence is an umbrella term, and the umbrella covers methods with very different strengths against very different schemes. It is useful to separate four families: supervised classification on structured financial data, unsupervised anomaly detection, natural language processing applied to narrative and communication data, and network analytics applied to relationships among entities. Complex frauds, which almost always combine falsified numbers, misleading narrative, and collusive relationships, are attacked most effectively by combinations of these families rather than by any one of them.

4.1 Supervised classification on structured data

Supervised methods learn a mapping from labelled examples: the researcher assembles a sample of firms or transactions known to be fraudulent, pairs it with a control sample presumed legitimate, and trains an algorithm to separate the two. Early studies applied decision trees, neural networks, and Bayesian classifiers to published financial statement data and reported encouraging discrimination between fraudulent and non-fraudulent firms (Kirkos et al., 2007; Ravisankar et al., 2011). Cecchini et al. (2010) advanced the field methodologically by designing a support vector machine kernel that encoded domain knowledge about financial ratios directly into the learning algorithm, and detected a large majority of fraud firms out of sample. Perols (2011) imposed a discipline the earlier literature had lacked, comparing six classifiers under realistic assumptions about the relative frequency of fraud and the asymmetric costs of misclassification, and found that logistic regression and support vector machines performed best under those conditions, a sobering result for anyone assuming that algorithmic complexity translates automatically into detective power.

Two later studies define the current frontier. Perols et al. (2017) showed that the severe scarcity of fraud observations, long treated as a nuisance, could be addressed directly by partitioning the data and the learning problem, materially improving prediction. Bao et al. (2020) then demonstrated, on a comprehensive sample of United States enforcement actions, that an ensemble method trained on raw accounting numbers outperformed both the traditional logistic regression approach and models built on theory-driven ratios, out of sample and by an economically meaningful margin. Their design choice is instructive: by feeding the algorithm raw data rather than researcher-constructed ratios, they allowed it to discover predictive structure that three decades of human theorizing had not specified. Abbasi et al. (2012) reached a complementary conclusion from the design science tradition, showing that a meta-learning framework combining multiple classifiers, feature sets, and time horizons outperformed any single constituent model.

The supervised record therefore supports a genuine but bounded claim. Against financial statement fraud of the kind that generates enforcement actions, machine learning classifiers now beat the classical screens by margins that matter, and ensembles beat single models. What the record does not show is anything approaching reliable detection, for reasons rooted in the structure of the problem rather than the sophistication of the algorithms, and those reasons occupy Section 5.

4.2 Unsupervised anomaly detection

Supervised learning presupposes labels, and labels presuppose that the frauds of the future will resemble the detected frauds of the past. Unsupervised methods drop that presupposition. Instead of learning what fraud looks like, they learn what normal looks like, and flag departures from it: transactions of unusual size or timing, accounts whose behaviour shifts abruptly, digit distributions that violate Benford's law, clusters of activity with no legitimate analogue (Bolton & Hand, 2002; Nigrini, 2012). The approach has two virtues that matter greatly in forensic work. It can, in principle, detect novel schemes for which no training examples exist, and it does not inherit the biases baked into historical enforcement data. Its corresponding vice is a high false positive rate, because the world of legitimate business is full of anomalies: quarter-end surges, one-off restructurings, seasonal effects, and simple eccentricity all depart from normal without being fraudulent. In practice, unsupervised outputs function as leads for investigators rather than accusations, which is, as I argue below, the correct epistemic status for algorithmic fraud detection generally.

4.3 Natural language processing

Complex fraud is committed with words as well as numbers, and a distinct research stream asks whether deception leaves recoverable traces in corporate language. Humpherys et al. (2011) found that the narrative sections of fraudulent annual filings differ measurably from honest ones in markers derived from deception theory. Larcker and Zakolyukina (2012) analyzed the unscripted answers of chief executives and chief financial officers on earnings conference calls and found that linguistic models identified deceptive discussions at rates modestly but reliably better than chance, with deceptive executives making greater use of references to general knowledge and less use of specific, verifiable detail. Purda and Skillicorn (2015) let the data select predictive vocabulary rather than imposing word lists drawn from psychology, and found that language-based models matched or exceeded the performance of models built on financial variables. Brown et al. (2020) applied topic modelling to the full text of annual reports and showed that what firms choose to discuss, and how much, adds predictive power over both financial ratios and prior linguistic measures. Hajek and Henriques (2017) confirmed, in a comparative study spanning many algorithms, that combining linguistic and financial indicators outperforms either alone.

The arrival of large language models has widened the aperture of this stream considerably, since modern systems can read filings, contracts, emails, and invoices at a scale no human review team can approach, and can encode subtler semantic regularities than word counts capture. The peer-reviewed evaluation of these newer systems against fraud detection tasks is still accumulating, and the prudent position is that the older literature establishes the principle, namely that deceptive financial language is statistically distinguishable from honest language at rates better than chance, while the effect sizes remain modest and the risk of models learning spurious stylistic correlates rather than deception itself remains real.

4.4 Network analytics

The schemes that most trouble forensic investigators, collusive procurement rings, related-party carousels, layered laundering structures, are relational by nature, and their individual transactions may look entirely innocent when inspected one at a time. Network methods represent entities as nodes and dealings as edges, and ask structural questions: which vendors share directors, bank accounts, or addresses with employees; which clusters of accounts move value in circles; which new counterparties sit suspiciously close, in network terms, to previously confirmed fraudsters. Van Vlasselaer et al. (2015) demonstrated the value of this representation in payment fraud, showing that features propagated through the network of cardholders and merchants improved detection significantly over models built on transaction attributes alone. For collusion and money laundering, network features are not merely helpful but close to indispensable, because collusion is invisible at the level of the individual record and visible only in the pattern of relationships, which is precisely what sampling could never observe and population-level analysis can.

5. EVALUATING EFFECTIVENESS: WHAT THE EVIDENCE PERMITS US TO CONCLUDE

It would be convenient if effectiveness could be read off a single number, and the literature does report such numbers: classification accuracies in the eighties and nineties of per cent, areas under the receiver operating characteristic curve well above chance, detection rates that improve benchmark by benchmark. Taken at face value, they suggest a solved problem. They should not be taken at face value. Five structural features of the fraud detection problem intervene between laboratory performance and field effectiveness, and an honest evaluation must work through each.

5.1 The base-rate problem

Fraud is rare relative to legitimate activity. In financial statement samples, confirmed fraud firms constitute a small fraction of one per cent of firm-years; in transaction streams, fraudulent items are rarer still. Rarity has two consequences that no algorithmic ingenuity fully escapes. The first is statistical: classifiers trained on wildly imbalanced data gravitate toward the majority class, and the standard remedies, such as the synthetic oversampling technique of Chawla et al. (2002) or the under-sampling embedded in the ensemble used by Bao et al. (2020), mitigate the problem at the price of distorting the training distribution. The second consequence is arithmetical and more stubborn. When the base rate of fraud is one in a thousand, even a screen that is ninety-nine per cent

accurate in both directions will generate roughly ten false alarms for every true detection, because the honest population to which the false positive rate applies is a thousand times larger than the fraudulent one. High accuracy on rare events is compatible with a low probability that any given alert is correct, and practitioners who do not internalize this arithmetic will either drown in alerts or, worse, treat alerts as findings.

5.2 Label bias and the shadow population

Supervised models learn from labelled fraud, and the labels come from enforcement: securities commission actions, restatements, prosecutions, confirmed chargebacks. Enforcement data are not a random sample of fraud; they are a sample of fraud that was caught, litigated, and documented, filtered at every stage by the priorities, resources, and blind spots of the detecting institutions (Amiram et al., 2018). A model trained on such labels learns to find the kinds of fraud that existing mechanisms already find, and is structurally blind to the shadow population of schemes that were never detected, which may differ systematically from the detected ones in exactly the respects that made them undetectable. The point is not that supervised learning is worthless, since the detected population is large and instructive, but that reported performance metrics measure agreement with past enforcement rather than coverage of actual fraud, and the two are not the same thing. Unsupervised and network methods partially escape this bias, which is a principled reason for combining families of techniques rather than crowning a single champion.

5.3 Adversarial adaptation and concept drift

Fraud detection is not pattern recognition against a passive background; it is a contest against intelligent opponents who observe the defences and adapt. This was documented at the very origin of the field, when Fawcett and Provost (1997) built adaptive systems for telecommunications fraud precisely because static rules decayed as offenders learned them, and every operational deployment since has confirmed the dynamic. The statistical name for the phenomenon is concept drift: the joint distribution of features and labels shifts over time, so that a model trained on last year's schemes degrades against this year's. In fraud, the drift is not merely environmental but strategic, since the more effective a detection feature becomes, the stronger the incentive to behave in ways that neutralize it, and features derived from public models, such as the Beneish indices, are available to manipulators as a checklist of what to avoid. Effectiveness is therefore a decaying asset that must be continuously renewed through retraining, red-teaming, and the deliberate retention of unpredictable, human-directed procedures of the kind auditing standards have always required (IAASB, 2009). Claims about model performance that lack a time dimension, which is to say most vendor claims, are claims about a snapshot of a moving target.

5.4 Cost asymmetry and the economics of alerts

The costs of the two possible errors are radically unequal and fall on different parties. A false negative means an undetected scheme that continues to compound, with median durations around a year and losses that scale with the seniority of the perpetrator (ACFE, 2024). A false positive means an investigation that consumes skilled hours, and, when the flagged party is a customer, an employee, or a counterparty, a blocked payment, a frozen account, a suspicion communicated or inferred, and a relationship damaged, costs borne substantially by innocent people. Perols (2011) showed that the ranking of algorithms reverses when realistic error costs replace symmetric ones, which means that an institution cannot choose a fraud detection system rationally without first deciding, explicitly, what it believes those relative costs to be. That decision is managerial and ethical rather than technical, and delegating it silently to a vendor's default threshold is an abdication dressed as procurement. Alert volumes must also respect the capacity of the humans downstream, because an investigation function receiving more alerts than it can competently examine does not become more effective; it becomes a lottery with paperwork.

5.5 Explainability, evidence, and the forensic threshold

The final constraint is the one most salient to forensic practice and least discussed in the computational literature. Detection is not the end of the process; it is the beginning of one that may terminate in disciplinary action, civil claims, regulatory sanction, or criminal prosecution, and each of those destinations imposes standards of proof and rights of challenge that an algorithmic score, standing alone, cannot satisfy. A gradient-boosted ensemble that assigns a firm a high fraud probability has produced a reason to investigate, not evidence of anything. The investigating accountant must still trace the flagged pattern to documents, witnesses, and admissions capable of surviving cross-examination, and the model's internal reasoning, distributed across thousands of parameters, is not itself examinable in any forensically meaningful sense. Rudin (2019) has argued forcefully that high-stakes decisions should rest on inherently interpretable models rather than on post hoc explanations of opaque ones, and the forensic context strengthens her argument, since post hoc explanation techniques produce plausible rationalizations whose fidelity to the model's actual computation is itself contested. Where an accused person's liberty or livelihood is engaged, due process requires that the basis of suspicion be articulable, disclosable, and testable, and detection architectures should be designed backwards from that requirement rather than retrofitted to it after deployment.

5.6 A synthesis

Assembling these strands yields a defensible overall judgment. Artificial intelligence has made population-wide screening feasible and has raised the quality of that screening measurably above the ratio-based tools it inherited; the gains reported by Bao et al. (2020), Perols et al. (2017), Brown et al. (2020), and Van Vlasselaer et al. (2015) are real, replicated in kind if not always in magnitude, and largest precisely where sampling was weakest, namely collusive and relational schemes visible only at population scale. At the same time, base rates guarantee that most alerts will be wrong, label bias guarantees that models see the past more clearly than the present, adaptation guarantees that performance decays, and evidentiary standards guarantee that no model output is a finding. The accurate description of current capability is therefore triage: artificial intelligence reorders the investigator's queue so that scarce human attention lands where suspicion is statistically densest, and it does this very well. It does not, and on present evidence cannot, determine that fraud has occurred. Institutions that deploy it as triage obtain a genuine and compounding advantage; institutions that deploy it as an oracle purchase false confidence at scale.

6. GOVERNANCE, REGULATION, AND THE ETHICS OF WATCHING EVERYONE

A detection architecture that monitors entire populations is, in legal terms, a large-scale processing operation over personal data, and the law of several jurisdictions now speaks to it directly. Within the European Union, the General Data Protection Regulation conditions such processing on identified lawful bases, proportionality, and safeguards around automated decision-making, and the Artificial Intelligence Act adds a risk-tiered regime under which systems used in contexts such as creditworthiness assessment attract mandatory requirements of data governance, transparency, human oversight, and accuracy (European Parliament & Council of the European Union, 2016, 2024). The direction of regulatory travel is unmistakable: the burden is shifting onto deploying institutions to demonstrate, in advance and on the record, that their surveillance is lawful, proportionate, documented, and contestable.

African jurisdictions are participants in this development, not bystanders to it. Ghana legislated comprehensively on data protection more than a decade ago, establishing a Data Protection Commission and binding processing to principles of lawfulness, purpose limitation, and data minimization (Data Protection Act, 2012), and has since erected a cybersecurity regime with its own regulator and licensing structure (Cybersecurity Act, 2020), alongside anti-money laundering legislation that obliges accountable institutions to monitor and report suspicious transactions (Anti-Money Laundering Act, 2020). The interaction of these statutes frames the Ghanaian deployment question precisely: transaction monitoring is not merely permitted but, in some sectors, mandated, while the same law that mandates it constrains how much may be collected, for what purposes, and with what rights of access and correction for the persons watched. For emerging markets generally, the practical constraint is less the statute book than the enforcement and audit capacity behind it, and there is a genuine risk of a two-tier world in which sophisticated surveillance is imported faster than the institutional muscle needed to supervise it. Regional coordination, investment in regulatory data science capability, and insistence on locally validated models, rather than models trained exclusively on foreign fraud patterns and foreign populations, are the sensible responses.

Two further governance obligations deserve emphasis because they follow directly from the analysis in Section 5. First, since error costs fall asymmetrically on innocent parties, institutions should measure and publish, at least internally, the false positive burden their systems impose, disaggregated by customer and employee groups, because a model whose alerts concentrate on particular communities is discriminating in effect whatever its designers intended. Second, since model outputs are triage rather than findings, governance frameworks should prohibit adverse action taken on an unreviewed score, preserving a human decision, made on articulable reasons, between the algorithm and the consequence. Neither obligation is technically demanding. Both are organizationally demanding, which is where governance usually fails.

7. IMPLICATIONS FOR FORENSIC PRACTICE AND PROFESSIONAL EDUCATION

For the practicing forensic accountant, the shift from sampling to surveillance changes the shape of the work without changing its object. The scarce skill is no longer the drawing of samples but the interrogation of alerts: understanding what a model can and cannot see, recognizing the signature of label bias and drift, converting a statistical anomaly into a documented evidential trail, and explaining to a court or tribunal, in ordinary language, why a pattern is suspicious and what was done to confirm or exclude innocent explanations. Investigators who treat the model as a colleague whose judgment must be checked, rather than as either an infallible oracle or an inscrutable rival, extract the most value from it. The unpredictability requirement of the auditing standards also acquires new content in this environment, since a monitoring system whose logic becomes known is a monitoring system that can be gamed, and the deliberate, human-directed departure from algorithmic routine remains one of the few defences against an adversary who has learned the routine (IAASB, 2009).

For educators, the implications are curricular and urgent. A forensic accounting or auditing graduate who cannot read a confusion matrix, reason about base rates, or ask pointed questions about training data is unequipped for the profession as it now exists, yet quantitative model literacy remains marginal in many accounting and law curricula. The remedy is not to turn accountants into

machine learning engineers but to teach the evaluative core: where the data came from, what the labels actually measure, how performance was estimated, what the error costs are, and who bears them. Equally, data scientists building these systems need exposure to evidence law and professional standards, because architectures designed without regard to the forensic threshold produce alerts that investigations cannot use. The most defensible systems will come from teams and classrooms, in which both literacies are present.

8. LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

This evaluation inherits the limitations of the literature it synthesizes. The strongest empirical studies concentrate on United States public company data and enforcement actions, and the transferability of their findings to private firms, to public sector procurement, and to emerging market institutional environments is asserted more often than it is demonstrated. Comparative work that trains and validates models on African, Asian, and Latin American enforcement and transaction data would test that transferability directly, and Ghana, with its combination of a maturing data protection regime, an active banking supervisor, and a documented fraud reporting practice, is a natural site for such work. Second, the field measures detection far more carefully than it measures deterrence, although the strongest argument for surveillance may be that credible, visible monitoring prevents fraud that never has to be detected; research designs capable of estimating that preventive effect is scarce and needed. Third, the evidentiary career of algorithmic alerts, meaning how courts and tribunals in different jurisdictions actually receive, weigh, and challenge model-derived material, is almost entirely unmapped, and doctrinal and empirical legal scholarship has a clear contribution to make. Finally, the evaluation of large language models against realistic, adversarial fraud tasks, under peer review rather than in vendor material, has barely begun.

9. CONCLUSION

The movement from sampling to surveillance is the most consequential change in fraud detection since the professionalization of auditing, and its central promise, that concealment should no longer be able to hide behind the arithmetic of partial inspection, has been substantially delivered. The evidence assembled here shows machine learning outperforming the classical screens on structured data, deceptive language proving statistically distinguishable from honest language, and relational schemes that sampling could never see becoming visible in network structure. The same evidence shows every one of these gains bounded by base rates, biased labels, adaptive adversaries, asymmetric error costs, and the unyielding requirements of due process and proof. The mature conclusion is neither enthusiasm nor dismissal but role clarity. Artificial intelligence has become an excellent answer to the question of where investigators should look, and it remains no answer at all to the question of what happened. Institutions, regulators, and educators who hold that line will find these systems powerful allies; those who blur it will discover, at the expense of the innocent and to the benefit of the guilty, why the line was there.

REFERENCES

1) Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). MetaFraud: A meta-learning framework for detecting financial fraud. MIS Quarterly, 36(4), 1293–1327. https://doi.org/10.2307/41703508
2) Amiram, D., Bozanic, Z., Cox, J. D., Dupont, Q., Karpoff, J. M., & Sloan, R. (2018). Financial reporting fraud and other forms of misconduct: A multidisciplinary review of the literature. Review of Accounting Studies, 23(2), 732–783. https://doi.org/10.1007/s11142-017-9435-x
3) Anti-Money Laundering Act, 2020 (Act 1044) (Ghana).
4) Association of Certified Fraud Examiners. (2024). Occupational fraud 2024: A report to the nations. https://www.acfe.com/report-to-the-nations
5) Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. Journal of Accounting Research, 58(1), 199–235. https://doi.org/10.1111/1475-679X.12292
6) 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
7) Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235–255. https://doi.org/10.1214/ss/1042727940
8) Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. Journal of Accounting Research, 58(1), 237–291. https://doi.org/10.1111/1475-679X.12294
9) 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
10) Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
11) Cressey, D. R. (1953). Other people's money: A study in the social psychology of embezzlement. Free Press.
12) Cybersecurity Act, 2020 (Act 1038) (Ghana).

13) Data Protection Act, 2012 (Act 843) (Ghana).
14) 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
15) Dorminey, J., Fleming, A. S., Kranacher, M.-J., & Riley, R. A., Jr. (2012). The evolution of fraud theory. Issues in Accounting Education, 27(2), 555–579. https://doi.org/10.2308/iace-50131
16) European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). Official Journal of the European Union, L 119, 1–88.
17) European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.
18) Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. Data Mining and Knowledge Discovery, 1(3), 291–316. https://doi.org/10.1023/A:1009700419189
19) Gepp, A., Linnenluecke, M. K., O'Neill, T. J., & Smith, T. (2018). Big data techniques in auditing research and practice: Current trends and future opportunities. Journal of Accounting Literature, 40(1), 102–115. https://doi.org/10.1016/j.acclit.2017.05.003
20) Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud: A comparative study of machine learning methods. Knowledge-Based Systems, 128, 139–152. https://doi.org/10.1016/j.knosys.2017.05.001
21) Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems, 50(3), 585–594. https://doi.org/10.1016/j.dss.2010.08.009
22) International Auditing and Assurance Standards Board. (2009). International Standard on Auditing 240: The auditor's responsibilities relating to fraud in an audit of financial statements. International Federation of Accountants.
23) 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
24) Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50(2), 495–540. https://doi.org/10.1111/j.1475-679X.2012.00450.x
25) Nigrini, M. J. (2012). Benford's law: Applications for forensic accounting, auditing, and fraud detection. John Wiley & Sons.
26) 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
27) Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. (2017). Finding needles in a haystack: Using data analytics to improve fraud prediction. The Accounting Review, 92(2), 221–245. https://doi.org/10.2308/accr-51562
28) Public Company Accounting Oversight Board. (2002). AS 2401: Consideration of fraud in a financial statement audit.
29) Purda, L., & Skillicorn, D. (2015). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. Contemporary Accounting Research, 32(3), 1193–1223. https://doi.org/10.1111/1911-3846.12089
30) Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems, 50(2), 491–500. https://doi.org/10.1016/j.dss.2010.11.006
31) Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
32) Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems, 75, 38–48. https://doi.org/10.1016/j.dss.2015.04.013