



(RESEARCH ARTICLE)



ARTIFICIAL INTELLIGENCE IN FINANCIAL STATEMENT FRAUD DETECTION: A CRITICAL SYNTHESIS OF EMERGING TRENDS, METHODOLOGIES, AND FUTURE DIRECTIONS FOR U.S. PUBLICLY TRADED COMPANIES

Daniel Adu Boateng ^{1,*}, Adwoa Agyeiwaa Ampomah-Britwum ², Valerie Colley ² and Mary Magdalene Yeboah ³

¹ Carolina University.

² Department of Statistics and Actuarial Science, Knust.

³ University of Ghana Business School.

* Corresponding Author

Magna Scientia Advanced Research and Review, 2026, 18(01), 136–144

Publication history: Received on 10 August 2026; revised on 20 September 2026; accepted on 22 September 2026

Article DOI: <https://doi.org/10.30574/msarr.2026.18.1.0191>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Financial statement fraud causes systemic damage to U.S. capital markets, investor confidence, and the credibility of public company disclosures. Artificial intelligence (AI) offers promising new detection capabilities, yet its evidence base and operational readiness remain unevenly developed across the literature. This structured narrative review critically synthesizes peer-reviewed articles published between 2020 and 2026 on AI-based fraud detection in SEC-reporting publicly traded U.S. companies, with focused attention on methodological rigor, strength of evidence, and deployment readiness. The thematic synthesis shows five emerging phenomena: (1) ensemble and deep learning models significantly outperform classical ratio screens in structured data; (2) NLP models based on transformer architectures identify fraud signals from narrative disclosures, which are independent of the structured data; (3) multimodal fusion architectures result in the highest detection performance (AUC 0.86–0.91); (4) AI is starting to transform audit risk assessment and regulatory surveillance; and (5) operational deployment is still hindered by challenges with explainability, temporal validation, label quality, and adversarial robustness. Persistent gaps remain in model explainability, temporal validation, adversarial robustness, and regulatory deployment protocols. The NIST AI Risk Management Framework is proposed as an organizing governance structure. Explainable AI, federated learning, and blockchain-based audit trails are identified as priority research investments, with targeted recommendations offered for regulators, auditors, and capital market participants.

Keywords: Financial Statement Fraud; Machine Learning; Large Language Models; SEC Filings; Anomaly Detection

1 INTRODUCTION

Financial statement fraud is among the most consequential forms of white-collar crime, imposing direct economic losses on investors, creditors, and the broader public. Publicly traded companies in the United States are subject to rigorous disclosure obligations under the Securities Exchange Act of 1934 and the Sarbanes-Oxley Act of 2002. The Association

of Certified Fraud Examiners (ACFE) estimates that organizations lose approximately five percent of annual revenue to fraud, with financial statement fraud generating the largest median losses per case; a pattern consistently confirmed by U.S. enforcement data (Association of Certified Fraud Examiners, 2024). High-profile failures such as Enron and WorldCom, alongside more recent accounting anomalies at major issuers, underscore the potentially catastrophic consequences of undetected misrepresentation. Classical rule-based approaches, including the Beneish M-Score and Dechow F-Score, are limited to structured financial data and generally achieve AUC values in the 0.68–0.73 range on experimental benchmarks (Li et al., 2024; Craja et al., 2020). Beneish & Vorst (2022) also showed that even the best classical models achieve false positive rates above 100:1, which makes them unsuitable for use. Bao et al. (2020) demonstrated that these classical baselines are significantly outperformed by machine learning ensemble methods, with the same accounting variables motivated by theories. The digitalization of financial reporting through eXtensible Business Reporting Language (XBRL) tagging and automated EDGAR submission has simultaneously expanded data availability and increased interpretive complexity, creating conditions in which AI has become a transformational detection tool (Wang et al., 2025a). Machine learning algorithms, deep neural networks, and large language models (LLMs) have demonstrated the capacity to identify anomalous patterns in structured financial data, unstructured narrative disclosures, and temporal reporting sequences, surfacing latent fraud indicators beyond the reach of human review at scale (Craja et al., 2020; Wang et al., 2025a). Significant gaps nevertheless persist: much of the existing literature focuses on non-U.S. settings or proprietary datasets; the translation of laboratory metrics into operational audit workflows remains poorly documented; and questions of algorithmic transparency, fairness, and regulatory liability have received insufficient attention (Bhagat, 2024; Qatawneh, 2025). This review synthesizes the 2020–2026 literature across five thematic axes: (1) ML and deep learning on structured data; (2) NLP and LLMs on qualitative disclosures; (3) multimodal and ensemble frameworks; (4) AI in audit and regulatory compliance; and (5) cross-cutting methodological challenges. Rather than cataloguing studies individually, the review evaluates methodological paths, assesses evidence quality, situates findings within the U.S. regulatory context, and advances a structured research and governance agenda grounded in a fraud-construct taxonomy.

2 REVIEW METHODOLOGY

This review follows a structured narrative approach to synthesize the 2020–2026 empirical and conceptual literature on AI-based financial statement fraud detection in U.S. publicly traded companies. The following subsections describe the search strategy, inclusion and exclusion criteria, thematic coding process, and the measures taken to address potential conflicts of interest arising from co-author self-citation.

2.1 Search Strategy

Literature was identified through systematic database searches of Web of Science, Scopus, and SSRN conducted in early 2026. Core search strings combined terms from three concept clusters: (i) fraud constructs: “financial statement fraud,” “earnings management,” “accounting fraud,” “SEC enforcement”; (ii) AI/ML methods: “machine learning,” “deep learning,” “large language model,” “natural language processing,” “transformer”; and (iii) reporting context: “SEC filings,” “EDGAR,” “10-K,” “XBRL”. Searches were restricted to publications from January 2020 to March 2026. Reference lists of retrieved articles were manually checked for additional eligible studies (citation chaining).

2.2 Inclusion and Exclusion Criteria

Studies were included if they: (a) applied at least one AI or ML method; (b) addressed financial statement fraud detection, earnings management detection, or anomaly detection in financial disclosures; and (c) were published as a peer-reviewed journal article, conference proceeding, or substantive working paper with sufficient methodological detail. Studies were excluded if they: (a) focused exclusively on non-financial fraud (e.g., insurance fraud, payment card fraud) without applicability to financial reporting; (b) were purely conceptual without empirical evaluation; or (c) were unavailable in English. No geographic restriction was imposed on inclusion, but the synthesis foregrounds U.S.-relevant findings and notes where results may not transfer to the U.S. regulatory context.

2.3 Study Selection and Thematic Coding

Two authors performed initial screening of titles and abstracts independently, with disagreements resolved by discussion. Thirty-six studies met all inclusion criteria and were retained for full-text synthesis. Each study was coded on the following dimensions: AI/ML modality (structured ML, deep learning, NLP/LLM, and multimodal/ensemble); dataset and label source (AAER, restatements, Audit Analytics, proprietary); evaluation metrics reported (AUC, F1, precision, recall, AUPRC); and degree of U.S. applicability. Thematic codes were developed inductively from the initial set of included studies and refined iteratively until saturation. The five thematic axes that structure the thematic-coding scheme below represent the final coding structure.

2.4 Self-Citation Disclosure

Six studies in the reference list share one or more authors with this review (Amoako et al., 2025a; Amoako et al., 2025b; Amoakoh et al., 2026; Boateng et al., 2025; Nartey et al., 2025; Yeboah et al., 2026). These works are cited where their empirical content is directly relevant to the thematic discussion and evaluated using the same evidentiary criteria applied to all other included studies. Readers should be aware of this potential conflict of interest when assessing the review’s conclusions.

3 TAXONOMY: DISTINGUISHING ANOMALIES, EARNINGS MANAGEMENT, AND FRAUD

One persistent source of confusion in the AI fraud detection literature is the conflation of anomaly detection, earnings management, and outright financial statement fraud. These constructs differ in their labeling regimes, base rates, legal consequences, and appropriate evaluation procedures. A governing taxonomy is presented in Table 1 and applied consistently throughout this review.

Table 1: Taxonomy of Financial Reporting Irregularities

Construct	Definition	Label Sources	AI Modalities
Anomaly	Statistical deviation from expected reporting patterns	Unsupervised; no ground-truth required	Autoencoders, isolation forests, LSTM-AE
Earnings Management	Discretionary accrual manipulation within GAAP	Restatements, Audit Analytics, accrual models	Supervised ML on financial ratios; NLP on MD&A
Financial Statement Fraud	Intentional material misrepresentation in filings	SEC AAERs, DOJ/SEC enforcement releases	Deep learning, LLMs, multimodal fusion

Anomaly detection is unsupervised with no currently labeled frauds and is based on statistical inconsistency between distributions. The earnings management is in the legally gray area, which is discretionary but falls under GAAP and is generally defined by the accrual-based proxies (Bakumenko & Elragal, 2022). Financial statement fraud involves deliberate misrepresentation that must be substantiated by regulating authorities; hence SEC AAERs are the most predominant source of labeling. Different evaluation procedures would be needed between the two tasks: in fraud detection, precision-recall tradeoffs are the most significant due to extreme class imbalance (average prevalence of typically below 2% of observations. This reflects the rarity of confirmed fraud cases), whereas in anomaly detection, reconstruction error thresholds and precision-at-k metrics indexed to the capability of the review committee would be more important (Lin, 2024; Craja et al., 2020).

4 EMERGING TRENDS AND THEMATIC ANALYSIS

4.1 Machine Learning and Deep Learning for Structured Financial Data

The foundational strand of AI-based fraud detection applies supervised ML classifiers to structured numeric financial data; balance sheet ratios, income statement measures, cash flow indicators, and derivatives of the Beneish M-Score elements. Wyrobek (2020) demonstrated that standard classifiers can surface fraud patterns not captured by ratio screens alone. Li et al. (2024) reported that XGBoost ensemble classifiers incorporating theorized financial indicators achieved precision of approximately 0.81 and recall of approximately 0.76 on time-split SEC filing samples, with outputs interpretable enough for forensic accountants. Deep learning has yielded further gains: Craja et al. (2020) found that deep feedforward networks and convolutional architectures surpassed traditional classifiers by 5–8 AUC percentage points on AAER-linked balanced datasets. A critical methodological issue that has emerged alongside these performance gains is class imbalance: confirmed fraud represents a small fraction of all filings, and naive training inflates apparent accuracy while suppressing sensitivity. Wijaya et al. (2026) and others have established SMOTE-based oversampling and cost-sensitive loss weighting as standard preprocessing steps. Li et al. (2025) proposed an attention-based LSTM capable of detecting slow multi-period manipulation that single-period models miss. Bakumenko & Elragal (2022) showed that unsupervised approaches improve detection by approximately 10–12 percentage points overrule-based screens, particularly when labeled fraud samples are scarce. Lin (2024) identified generalizability as the principal constraint on operational adoption and recommended domain-specific feature engineering combined with rolling-origin temporal cross-validation to avoid look-ahead bias, a leakage risk particularly acute in EDGAR/XBRL pipelines where restatement labels may be forward-filled into historical training windows.

4.2 Natural Language Processing and Large Language Models for Qualitative Disclosures

Most of the information in SEC filings, including MD&A disclosures, risk factors, and earnings call transcripts is unstructured text. NLP and LLMs have opened new avenues for mining these disclosures for fraud indicators. Linguistic features such as sentiment polarity, syntactic complexity, and hedging density have been shown to improve AUC by approximately 0.05–0.07 (Velloor Sivasubramanian & Skillicorn, 2024; Zhu et al., 2025) relative to ratio-only models on 10-K fraud prediction tasks, consistent with the hypothesis that fraudulent managers employ characteristic linguistic-evasion strategies (Ketelaar & Mickovic, 2025; Bhattacharya & Mickovic, 2024). BERT and domain-adapted variants have emerged as the strongest text-based performers. Bhattacharya & Mickovic (2024) reported an F1 of approximately 0.79 using contextual BERT versus approximately 0.61 for TF-IDF baselines on accounting fraud detection, confirming the importance of domain-specific pre-training corpora. Ketelaar & Mickovic (2025) applied similar methods to 10-K filings, detecting linguistic anomalies, unusual boilerplate specificity, and shifts in hedging patterns that were associated with subsequent SEC enforcement actions. Generative LLMs (GPT-4, LLaMA variants) have introduced both new opportunities and new risks. Ergun & Sefer (2025) found that fine-tuned LLMs achieved AUC of 0.86 on fraud benchmarks versus 0.71 for zero-shot prompting, confirming that domain adaptation is a critical performance driver. Wu & Liu (2025) introduced FDLLaMa, a finance-aware LLaMA variant, demonstrating further gains through task-specific fine-tuning. A material risk in compliance contexts, however, is LLM hallucination: models may generate plausible but factually incorrect assessments of filing content. Mitigations include retrieval-augmented generation (RAG) grounded in the literal filing text, constrained decoding with citation requirements, and human-in-the-loop verification of LLM-flagged anomalies before any regulatory action (Iqbal, 2025). Zhu et al. (2025) demonstrated that abrupt shifts in risk disclosure language precede adverse restatements; Hajek et al. (2026) showed that topic-driven sentiment analysis improves F1 by 0.09 over unidimensional sentiment measures. Yu et al. (2025) and Liu et al. (2023) independently developed representation learning approaches to track disclosure change trajectories, identifying fraud risk as anomalous deviation from baseline disclosure patterns over time. Liang (2026) extended this to deep learning-based detection of SEC filing discrepancies for regulatory compliance validation.

4.3 Multimodal and Ensemble Learning Frameworks

Converging evidence suggests that no single data modality is sufficient for reliable fraud detection. Wang et al. (2023) proposed a multimodal attention model combining structured financial data with unstructured text, achieving an AUC of approximately 0.89 compared with unimodal baselines near 0.82; a gain of roughly 7 percentage points attributable to cross-modal information fusion. This finding has a direct theoretical implication: structured and textual fraud signals are partially independent, which should motivate fusion architectures as a primary research direction. Rezaei et al. (2024) found that ensembles combining tree-based models with neural architectures consistently outperformed single-modality models on accuracy, recall, and AUPRC, with the largest gains on highly imbalanced datasets, precisely the realistic conditions of fraud detection. Lu et al. (2025) extended this approach by incorporating earnings conference call transcripts alongside 10-K filings in a contrastive multimodal dialogue network, achieving AUC of 0.91 by detecting inconsistencies between management oral commentary and reported financial data. Wang et al. (2025b) concluded that integrated multi-source architectures represent the next generation of operationally viable fraud detection, while calling for standardized evaluation protocols to enable meaningful cross-study comparison.

4.4 AI in Regulatory Compliance, Audit, and SEC Oversight

Nartey et al. (2025) documented AI benefits across audit risk assessment, analytical procedures, and sampling optimization. Amoakoh et al. (2026) found that intelligent systems are particularly effective at surfacing audit risks in large, diversified enterprises where human review capacity is the binding constraint. Qatawneh (2025) demonstrated that NLP substantially enhances fraud detection utility by enabling automated analysis of management commentary and contractual documentation at scale. The SEC's existing Accounting Quality Model (AQM, colloquially "RoboCop") represents an early generation of quantitative surveillance; the current literature indicates that far more capable systems are achievable. Md Shahadat Hossain et al. (2025) showed that integrating SEC filings with IRS tax records significantly improves detection rates, a finding with direct implications for future SEC-IRS cross-agency collaboration. Amoako et al. (2025a; 2025b) identified a significant gap between the technical capabilities now available and regulatory guidance on model risk management, explainability requirements, and liability assignment for false positives. Boateng et al. (2025) emphasized that cross-institutional data sharing agreements and regulatory sandboxes are key enablers of the scale of AI deployment needed to meaningfully protect the financial system.

4.5 Controversies and Unresolved Methodological Challenges

Four substantive controversies warrant candid acknowledgment. First, label reliability: SEC AAERs are subject to enforcement selection bias; the SEC pursues cases meeting specific evidentiary thresholds, producing a labeled population that is not representative of the broader fraud universe. Technical mitigations include positive-unlabeled (PU) learning, semi-supervised methods, and noise-robust loss functions such as symmetric cross-entropy (Lin, 2024). Propensity score matching offers a complementary approach by adjusting for observable firm characteristics that predict enforcement probability. Second, interpretability: the highest-performing models, transformer LLMs, deep

graph networks are architecturally opaque, and their outputs resist translation into accounting or legal reasoning. SHAP and LIME provide post-hoc approximations, but these may be unfaithful to the underlying model’s true decision logic. When AI outputs must support enforcement actions or audit qualifications, faithful interpretability is not optional. More promising directions include inherently interpretable architectures, sparse linear models with accounting-theory-constrained features, prototype-based explanations grounded in 10-K section structure, and attention mechanisms with section-level attribution (Bhagat, 2024; Qatawneh, 2025). Third, adversarial robustness: as AI-driven detection systems become more widely deployed, sophisticated actors may learn to modify disclosures to evade algorithmic flags, creating a detection of arms race. Dzuyo et al. (2025) proposed a framework for detecting concept drift and evasion strategies, but red teaming of narrative disclosures, simulation of strategic managerial responses, and adversarial training on both tabular and textual inputs remain underdeveloped and warrant organized research investment. Fourth, temporal validation rigor: look-ahead bias in EDGAR/XBRL pipelines, where restatement labels are forward-filled into historical training windows; artificially inflates reported performance metrics. Best practices include rolling-origin evaluation with expanding training windows, nested time-series cross-validation, and strict label assignment based on original filing dates rather than restatement announcement dates (Lin, 2024; Wijaya et al., 2026). Yeboah et al. (2026) further noted that fraud pattern concept drift requires periodic model retraining supported by formal drift detection procedures.

5 CONSOLIDATED EVIDENCE SUMMARY

Table 2 is a reference summary of representative studies organized by method and reported performance metrics. It is not meant to be a display of comparative results, but rather a navigational tool for the reader; performance numbers are only experimental conditions from each study cited and cannot be directly compared from row to row because the datasets vary in label source, class balance, time window and train-test construction. A common EDGAR benchmark that includes time-split MD&A text and financial information would be a great contribution from the community, which would allow for standardizing the comparison between studies.

Table 2: Consolidated Evidence Summary of Key AI Fraud Detection Studies (2020–2026)

Study	Method	Key Metric(s)	Reported Finding
Craja et al. (2020)	Deep FF/CNN	AUC ~0.74 vs. LR ~0.68	Deep models outperform LR/SVM on balanced samples
Bakumenko & Elragal (2022)	Isolation Forest, Autoencoder	F1 improvement >12% vs. rules	Unsupervised ML exceeds rule-based anomaly detection
Velloor S. & Skillicorn (2024)	LSTM/BERT	AUC +0.06 vs. ratio-only	Text adds independent signal beyond financial ratios
Li et al. (2024)	Ensemble (XGBoost+LR)	Precision ~0.81, Recall ~0.76	Theorized features improve interpretability & accuracy
Bhattacharya & Mickovic (2024)	Contextual BERT	F1 ~0.79 (BERT) vs. ~0.61 (TF-IDF)	Contextual LM >> bag-of-words for fraud language
Wang et al. (2023)	Attention fusion	AUC values around 0.88–0.90 (multimodal) vs. roughly 0.80–0.83 (unimodal)	Multimodal fusion outperforms any single modality
Ergun & Sefer (2025)	Fine-tuned LLM	AUC values around 0.85–0.87 fine-tuned substantially outperforming non-fine-tuned baselines	Domain fine-tuning critical for LLM fraud detection
Hajek et al. (2026)	LDA + Sentiment	F1 +0.09 vs. unidimensional sentiment	Topic-driven sentiment outperforms standard sentiment
Lu et al. (2025)	Contrastive dialogue net	AUC values approaching 0.90–0.91 in experimental evaluations	Conference call contradictions expose narrative fraud

Rezaei et al. (2024)	Ensemble (ML+DL+LLM)	AUPRC best among all ablations	Ensemble diversity reduces majority-class overfitting
----------------------	-------------------------	--------------------------------	--

Table 2 summarizes performance metrics as reported in the included studies. AUC: Area Under the ROC Curve; AUPRC: Area Under the Precision-Recall Curve; FF: Feedforward; LM: Language Model; LLM: Large Language Model. Multimodal outputs are indications of a combination of tabular and textual modalities. Benchmark heterogeneity (dataset, time window, class balance) prevents numerical cross-row comparisons.

Limitations and Deployment Barriers

The primary binding constraints are data scarcity and label quality. AAER-based fraud datasets are small; typically, fewer than 300 confirmed cases, temporally clustered, and subject to enforcement selection bias. Models trained on these datasets may fail to generalize real-time EDGAR data due to distributional shifts and concept drift. Addressing drift requires continuous retraining supported by documented drift monitoring (e.g., population stability indices), MLOps pipelines with data lineage tracking and feature store versioning, and post-deployment performance monitoring consistent with financial services Model Risk Management (MRM) practices under SR 11-7. The economic costs of false positives are also material: erroneous fraud flags can trigger reputational damage, higher audit fees, and adverse market reactions for companies that have not violated any rule. Translating raw accuracy metrics into operational value therefore requires decision-analytic frameworks with cost-sensitive metrics calibrated to enforcement capacity constraints, expected utility analysis under asymmetric misclassification loss, and precision-at-k calibration aligned to regulator review throughput.

Ethical Considerations and AI Governance

The NIST AI Risk Management Framework (AI RMF 1.0, 2023) provides a practical governance structure for AI fraud detection deployment. Its four core functions; Govern, Map, Measure, and Manage, map directly onto regulatory AI requirements: establishing accountability lines (Govern), defining risk contexts and stakeholder impacts (Map), assessing model performance and bias (Measure), and implementing risk controls (Manage). Adoption of AI RMF-conformant documentation standards by the SEC and PCAOB would create a systematic basis for vetting AI tools used in enforcement and audit contexts. Systemic bias risks require explicit attention: fraud detection models trained on historical AAER data may carry industry, firm-size, or accounting-complexity biases that produce elevated false-positive rates for firms in sectors with complex revenue recognition, such as technology or life sciences or complex corporate structures. Fairness audits stratified by industry and firm size should be conducted before any deployment. Privacy and data governance considerations also arise because effective AI systems may incorporate non-public regulatory databases, alternative data sources, and third-party aggregators; the legal permissibility of using such data in enforcement contexts must be carefully analyzed under applicable securities and privacy law. Legitimate deployment further requires transparency measures including disclosure of AI tool use in regulatory proceedings and the provision of meaningful appeal pathways for entities flagged by algorithmic outputs (Amoako et al., 2025a; Qatawneh, 2025; Bhagat, 2024).

6 FUTURE DIRECTIONS AND RESEARCH GAPS

The most pressing near-term priority is explainability and interpretable AI. Beyond post-hoc approximations such as SHAP and LIME, which may be unfaithful to the underlying model, the field needs architectures that are intrinsically interpretable: sparse logistic models with accounting-theory constrained feature sets, prototype-based explanation systems that ground fraud flags in specific accounting standard provisions, and attention mechanisms with section-level attribution aligned to 10-K/10-Q filing structure. Crucially, explanation quality should be evaluated empirically through auditor comprehension studies and enforcement case simulations, not only by technical approximation fidelity.

Federated learning offers a promising path through the data scarcity problem. Federated architectures allow banks, audit firms, and regulatory agencies to jointly train fraud detection models without exchanging sensitive records. Governance of federated consortia requires: (i) differential privacy guarantees with published privacy budgets; (ii) secure aggregation protocols resistant to gradient-based inference attacks; (iii) auditability of federated model updates; and (iv) interagency memoranda of understanding (MoUs) covering data minimization requirements, model change management, and exit provisions. A minimal viable pilot, a two-party federated arrangement between the SEC and a major audit network is likely feasible within existing legal frameworks and would provide invaluable real-world evidence on federated fraud detection performance.

Temporal and longitudinal modeling deserves a dedicated investment. Rolling-origin evaluation with expanding training windows, nested time-series cross-validation, and strict filing-date-based labeling should be established as non-negotiable field norms. More ambitiously, models that track dynamic disclosure trajectories across multiple

reporting periods, tracing the evolution of MD&A language, accrual patterns, and footnote specificity in the quarters leading up to enforcement action could substantially improve early-warning detection capability.

Blockchain-integrated audit trails represent an emerging frontier. Immutable, time-stamped blockchain records of financial transactions and reporting events would provide AI detection systems with a manipulation-resistant evidentiary foundation, resolving data integrity concerns that currently undermine trust in AI outputs. The intersection of blockchain data infrastructure with AI analytics engines in the SEC reporting context is a largely unexplored research niche with both technical and policy dimensions.

Research on generative AI dual-use risks is urgently needed. As LLMs become capable of producing highly plausible financial narrative text, the potential for AI-assisted preparation of fraudulent disclosures is a near-term threat. Adversarial training of detection models on generative adversarial examples, watermarking of AI-generated financial content, and automated identification of LLM-generated text in regulatory filings are research directions that are both technically tractable and directly policy-relevant. For LLM-based fraud screening specifically, retrieval-augmented generation (RAG) architectures grounded in the literal filing text, combined with constrained decoding to limit hallucination represent the most defensible deployment approach for compliance contexts (Iqbal, 2025; Ergun & Sefer, 2025).

7 CONCLUSION

The results are clear: AI techniques, especially those that combine supervisory ML, deep learning, NLP, LLM, and multimodal fusion, perform significantly better than classical approaches like the Beneish M score and the Dechow F score, especially in areas that require an analysis of unstructured disclosures and multi-period temporal patterns. The state-of-the-art model is currently transformer-based language models and multimodal ensemble architectures, achieving AUC of 0.86-0.91 with domain adaptation. This review has three structural features which make the evidence base transparent and replicable: a documented search and inclusion process, a fraud-construct taxonomy used consistently throughout the included studies, and a disclosure about the co-authorship overlap between the review team and some of the included studies in the form of a self-citation disclosure. The biggest challenge is to figure out how to translate performance in the lab to operational impact. The four priorities stand out. First, federated learning consortia of audit firms, financial institutions and regulators provide a viable vehicle for training bigger, more representative fraud populations without sacrificing data privacy. Second, interpretability goes beyond post-hoc approximations: Interpretable architectures, attention mechanisms grounded in the sections are needed if AI outputs are to underpin audit qualifications or enforcement proceedings. Thirdly, the use of a strict filing date label needs to be a non-negotiable field norm for rolling-origin evaluation. Fourth, formal connection to the NIST AI RMF, fully engaged with the SEC and PCAOB, is needed to build the accountability structures needed to scale the risks of false positives. The evidence now available offers a scientific basis for strategic investments in artificial intelligence and financial supervision, systems that need to be constructed and managed with maximum scientific rigor and transparency, and which should be accountable for any damage to the integrity of U.S. capital markets.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript.

REFERENCES

- [1] Amoako, D., Boboye, C. O., Boateng, V., & Nii-Laryea Laryea, J. E. (2025a). Artificial intelligence and machine learning for fraud detection in the U.S. banking industry: Regulatory frameworks, implementation, and challenges. *EPRA International Journal of Economics, Business and Management Studies*. <https://doi.org/10.36713/epra23647>
- [2] Amoako, E. K., Okeke, P. T., Boateng, V., Bonsu, M. A., & Oman-Amoako, M. (2025b). Enhancing risk management and fraud detection in the U.S. financial industry through machine learning algorithms: Applications, challenges, and future directions. *World Journal of Advanced Research and Reviews*, 28(3). <https://doi.org/10.30574/wjarr.2025.28.3.4031>
- [3] Amoakoh, C. K., Nwinyi, I. P., & Ampofo, I. (2026). AI-enabled assurance frameworks: A review of intelligent systems enhancing accuracy and independence in U.S. financial auditing. *EPRA International Journal of Economics, Business and Management Studies*. <https://doi.org/10.36713/epra26251>

- [4] Bakumenko, A., & Elragal, A. (2022). Detecting anomalies in financial data using machine learning algorithms. *Systems*, 10(5), 130. <https://doi.org/10.3390/systems10050130>
- [5] Bhagat, N. (2024). Artificial intelligence challenges and its impact on detection and prevention of financial statement fraud: A theoretical study. In *Demystifying the Dark Side of AI in Business* (pp. 60–80). IGI Global Scientific Publishing.
- [6] Bhattacharya, I., & Mickovic, A. (2024). Accounting fraud detection using contextual language learning. *International Journal of Accounting Information Systems*, 53, 100682. <https://doi.org/10.1016/j.accinf.2024.100682>
- [7] Boateng, V., Amoako, E. K., Ajay, O., & Adukpo, T. K. (2025). Harnessing artificial intelligence for combating money laundering and fraud in the U.S. financial industry: A comprehensive analysis. *Finance & Accounting Research Journal*.
- [8] 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>
- [9] 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>
- [10] Dzuyo, G. S. W., Guibon, G., Cerisara, C., & Belmar-Letelier, L. (2025). Read between the lines: A robust financial statement fraud detection framework. *ArXiv preprint*.
- [11] Ergun, Z., & Sefer, E. (2025). Financial statement fraud detection via large language models. *Intelligent Systems in Accounting, Finance and Management*, 32(4), e70021. <https://doi.org/10.1002/isaf.70021>
- [12] Hajek, P., Novotny, J., & Munk, M. (2026). Financial statement fraud detection using topic-driven financial sentiment analysis. *Decision Support Systems*, 114615. <https://doi.org/10.1016/j.dss.2026.114615>
- [13] Iqbal, Z. (2025). The impact of generative AI adoption on corporate financial reporting accuracy: A secondary data analysis of SEC filers. *SSRN Working Paper* 6011234. <https://ssrn.com/abstract=6011234>
- [14] Ketelaar, F., & Mickovic, A. (2025). Artificial intelligence in fraud detection: Textual analysis of 10-K filings. *Maandblad voor Accountancy en Bedrijfseconomie*, 99(2), 61–71.
- [15] Li, B., Yen, J., & Wang, S. (2024). Uncovering financial statement fraud: A machine learning approach with key financial indicators and real-world applications. *IEEE Access*, 12, 194859–194870. <https://doi.org/10.1109/ACCESS.2024.3522156>
- [16] Li, J., Gan, Q., Liu, Z., Chiang, C., Ying, R., & Chen, C. (2025). An improved attention-based LSTM neural network for intelligent anomaly detection in financial statements. *Working paper*.
- [17] Liang, D. (2026). Detecting disclosure discrepancies in SEC filings: A deep learning approach for regulatory compliance verification. *Journal of Sustainability, Policy, and Practice*, 2(1), 101–114.
- [18] Lin, D. (2024). Key considerations to be applied while leveraging machine learning for financial statement fraud detection: A review. *IEEE Access*, 12, 168213–168228. <https://doi.org/10.1109/ACCESS.2024.3498213>
- [19] Liu, R., Huang, J., & Zhang, Z. (2023). Tracking disclosure change trajectories for financial fraud detection. *Production and Operations Management*, 32(2), 584–602. <https://doi.org/10.1111/poms.13896>
- [20] Lu, Q., Du, W., Yang, S., Xu, W., & Zhao, J. L. (2025). Can earnings conference calls tell more lies? A contrastive multimodal dialogue network for advanced financial statement fraud detection. *Decision Support Systems*, 189, 114381. <https://doi.org/10.1016/j.dss.2025.114381>
- [21] Md Shahadat Hossain, M. S. H., Mohammad Ali, M. A., & Md Whahidur Rahman, M. W. R. (2025). Machine learning-based analytics framework for detecting tax evasion and financial misconduct in US enterprises. *International Journal of Emerging Technologies in Learning*, 2(12), 114–138.
- [22] Nartey, O. L., Nwinyi, P., Yeboah, M. M., & Anim, B. (2025). Artificial intelligence in financial auditing. *Zenodo*. <https://doi.org/10.5281/zenodo.18188728>
- [23] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- [24] Qatawneh, A. M. (2025). The role of artificial intelligence in auditing and fraud detection in accounting information systems: Moderating role of natural language processing. *International Journal of Organizational Analysis*, 33(6), 1391–1409. <https://doi.org/10.1108/IJOA-01-2024-4000>

- [25] Rezaei, Z., Samghabadi, S. S., Amini, M. A., & Banad, Y. M. (2024). The power of ensemble methods: A comparative study of machine learning, deep learning, and LLMs for financial fraud detection. In 2024 International Conference on AI x Data and Knowledge Engineering (AIXDKE) (pp. 125–126). IEEE.
- [26] Velloor Sivasubramanian, S., & Skillicorn, D. (2024). Predicting fraud in MD&A sections using deep learning. *Journal of Business Analytics*, 7(3), 197–206.
- [27] Wang, G., Ma, J., & Chen, G. (2023). Attentive statement fraud detection: Distinguishing multimodal financial data with fine-grained attention. *Decision Support Systems*, 167, 113913. <https://doi.org/10.1016/j.dss.2023.113913>
- [28] Wang, M., Zhang, X., & Han, X. (2025b). AI driven systems for improving accounting accuracy fraud detection and financial transparency. *Frontiers in Artificial Intelligence Research*, 2(3), 403–421.
- [29] Wang, Y., Chiu, T., & Vasarhelyi, M. A. (2025a). Financial statement fraud prediction system: A deep learning-based approach. *Journal of Forensic Accounting Research*, 10(1), 239–259. <https://doi.org/10.2308/JFAR-2023-0030>
- [30] Wijaya, J. R. T., Rahmatika, D. N., & Herwiyanti, E. (2026). Machine learning techniques in detecting fraud: A state of the art review in the Scopus database. *Jurnal Bina Akuntansi*, 13(1).
- [31] Wu, Q., & Liu, A. (2025). FDLLaMa: Unlocking the potential of finance-aware LLaMA in combating fraud. In 2025 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC) (pp. 1–6). IEEE.
- [32] Wyrobek, J. (2020). Application of machine learning models and artificial intelligence to analyze annual financial statements to identify companies with unfair corporate culture. *Procedia Computer Science*, 176, 3037–3046. <https://doi.org/10.1016/j.procs.2020.09.234>
- [33] Yeboah, M. M., Nartey, O. L., Agyei, E., & Anim, B. (2026). Artificial intelligence in fraud detection and financial risk mitigation. *International Journal of Finance, Management and Research*, 8(1). <https://doi.org/10.36948/ijfmr.2026.v08i01.63239>
- [34] Yu, Y., Wu, Z., Han, Y., Li, Z., & Wei, W. (2025). Unlocking financial statement fraud detection: Tracking disclosure changes via representation learning. In ICASSP 2025 (pp. 1–5). IEEE.
- [35] Zhu, X., Wu, H., Chang, Y., & Li, J. (2025). Accounting fraud detection through textual risk disclosures in annual reports: From the perspective of SEC guidelines. *Accounting & Finance*, 65(2), 1837–1862. <https://doi.org/10.1111/acfi.13293>