Role of Artificial Intelligence in Financial Fraud Investigations: A Critical Review

Prof. (Dr.) Anand G. Jumle

Dr. Deepak Jain

Mansee Daga

Sheetal Deshmukh

Abstract: AI significantly influences the financial fraud investigations across different parts of the workflow related to detection, automation, transactions prioritization, anomalous discovery, and fraud investigation explanations. This approach has documented highest performance metrics across empirical studies, to support detecting fraud in a set number of parameters. Random assessment gave up to 96% Accuracy (with a 98.9% Area under the curve); fuzzy logistic regression provides an AUC up to >0.99, with up to 0.90+ Sensitivity and Specificity on credit-card. Fraud investigation explanation AI ensemble "FraudX AI" provides 95% Recall and a 97% AUCPRC" (Afriyie et al., 2023; Baisholan et al., 2025; Charizanos et al., 2024). Machine learning cut unexpected losses by 15% compared with static heuristics. Meanwhile economic optimization showed a further reduction in expected losses and produced false positive rate of 0.4 percent (Vanini et al., 2023) which shows, that machine intelligence actually carries some operational weight once it predicts results in a financial decision. Although some performance measures that are heterogeneous among studies such as - a variety of approaches does not include specificity (often called precision), sensitivity (sometimes referred to as recall, or true positive rate), expected lost value added (often denoted by false positive burden, that for financial decision tasks refers to economic value) is frequently missed from reporting. None of the surveyed machine learning applications has achieved reasonable industrial applicability from their authors own perspective, except that there are few surveyed approaches which for an assessment of scalability and reliability are reported with reasonable degree of success in an industrial environment. (Ryman-Tubb et al., 2018). Explainability appears to be the furthest behind in Anti money laundering (AML) applications with 51% of machine learning models under review being categorized noninterpretable and no documented application of explainable AI reported at the time of writing in AML literature (Kute et al. 2021). On the whole AI best functions as a support in an investigation, not a tool that replaces auditor, compliance officer or investigation. Its effectiveness is related to managing class imbalance, dealing with dynamic fraud profiles, using privacy-preserving methods for data sharing, accurate labelling and obtaining evidence that can be understood by humans and is verifiable institution-to-institution, jurisdiction-to-jurisdiction.

INTRODUCTION

Financial fraud includes any deceptive practice related to transactions, financial statements, insurance claims, banking, securities, cryptocurrency and money laundering. The growing size and variety of technology used in these transactions have consequently outpaced existing methods of prevention and investigation. Technology use has enabled fraudster transactions through electronic commerce, mobile payments, contactless systems, Internet banking, data leaks, multi-level transactions; institutions must be able to detect suspected fraud quickly while providing an actionable alert which is not overwhelming to the investigator (Abdallah et al., 2016; Ryman-Tubb et al., 2018).

Financial fraud detection systems are becoming an integral part of prevention systems with the goal of distinguishing abnormal behavior or actions occurring after the event, allowing investigation resources to focus on the largest threats (Abdallah et al., 2016; Hilal et al., 2022).

For our purpose of AI, we use a general definition: Computational techniques to discover patterns in financial or non-financial data, classify new observations, detect patterns, model relationship, create synthetic data or provide an explainable answer. Machine learning uses patterns in supervised classification and prediction, semi-supervised and unsupervised for detecting patterns when honest and cheat labels are limited. Deep learning methods apply these concepts to sequential text, graph-based transactions and high dimensional embeddings.

Explainable AI (XAI) is a class of models and their explanation to present a model output in terms of human understandable reasoning for its decision. In context of forensic of financial investigation activities, the use case for such models may lie in financial transaction monitoring, Alert prioritisation, forensics examination, examination of financial reports or statement screening as well as organised fraud identification & decision making by auditors.

The relevant research evolved in different fraud applications instead of a generic investigative approach. It has shown a growing interest in ensemble methods, deep learning, graph learning, federal learning, natural language processing (NLP), and XAI as well as continued interest in earlier developments of logistic, neural networks, Bayesian learning, support vector machine, anomaly detection and trees decision (Al-Hashedi & Magalingam, 2021; Hafez et al., 2025; Ngai et al., 2010). However, there are considerable difficulties in turning a model with predictive ability into an investigation model since several challenges such as severe class imbalance, dynamic behavior of fraud, small and rare labels, privacy protection constraints, opacity in decision making and lack of consistency between academic datasets and business scenarios still persist.

Consequently, we must move towards critical synthesis: not just whether AI can be fraud, but how AI informs investigation, when AI aids in decision, and the operational, evidentiary and governance gaps which may persist. In our review, we review AI's applications in fraud and detection and investigation in financial

transaction fraud, financial-statement fraud and anti-money-laundering settings, focusing on efficacy, methodological compromises, interpretability, practical constraints in its adoption and suitability in actual investigation operations.

METHODS

2.1 Search Strategy

A total search strategy across some academic papers. We used a hybrid semantic and keyword-based retrieval approach to achieve the optimum search space.

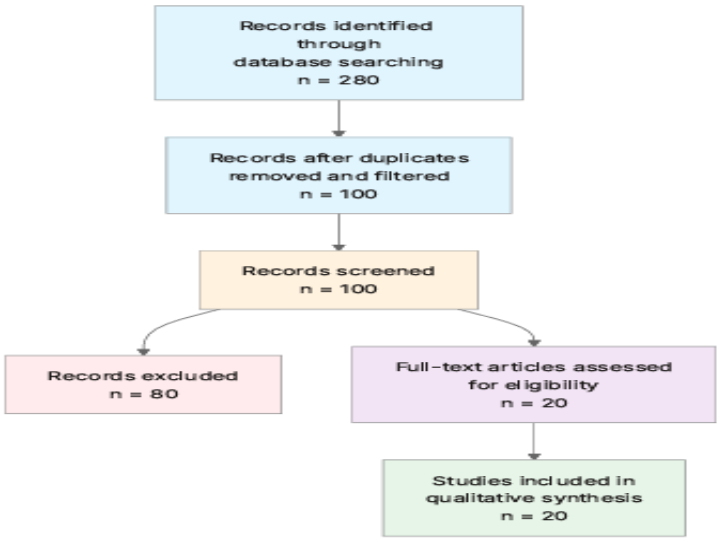
Search queries included:

- Artificial Intelligence for Financial Fraud Detection and Analysis.
- The Application of Machine Learning for Fraud Detection in Banking and Finance.
- AI-based Anomaly Detection in Credit Card and Transaction Fraud.
- Machine Learning, Deep Learning, and Fraud Analytics in Anti-Money Laundering.
- Explainable AI (XAI) for Financial Fraud Detection and Forensics.
- Artificial Intelligence in Financial Fraud: A Survey.
- The Data-mining and predictive techniques for detection and prevention of financial services.

2.2 Study Selection

Initially 280 records were identified via database search. 100 duplicate records were removed and screened. 20 records being retained when filtered for relevance and evaluated against selection criteria.

Figure-1: PRISMA Flow Diagram



Sources: Primary Data

Eligibility Criteria Included:

- **Human Study** - Does the study use real financial fraud cases, transactions, investigations, or institutions instead of synthetic examples or toy data?

- **AI Method** - Does the study evaluate an artificial intelligence, machine learning, deep learning, anomaly detection, NLP, graph, or explainable AI technique for fraud research?

- **Fraud Focus** - Does the study concentrate on fraud, money laundering, suspicious transactions detection, or forensic fraud analysis in finance services?

- **Investigation Relevance** - Is there a focus on the role of the investigation process? This refers to detection, triage, monitoring, alerting, priority selection, pattern extraction, or case analysis.

- **Quantitative Evaluation** - Does the study contain an empirical evaluation using standard performance metrics, case outcomes, or comparative benchmarks?

- **Explainability** - Is interpretability, transparency, explainability, auditability or regulatory compliance addressed for the method?

- **Operational Limits** - Does the study discuss issues like deployment concerns, bias, quality data, performance scalability issues or adaptation of opponents (adversarial)?

- **Review Type** - Does the work provide a survey, systematic review, framework or critical commentary beyond the scope of a specific model evaluation?

The remaining review and survey and also taxonomy and critical review papers don't on its own give the model level quantitative performance, explainability methodology or concrete operational problems, yet the papers have been selected for not presenting this to directly address the research questions research but to establish the evidence on model's role, investigation role, methodological shortcomings, conditional application, of different models used and on different data to study.

With regard to the financial statement GAN paper which is included even though it used data concerning ten financial statements of Iranian banks along with Synthetic Fraud Data which is not real-data the GAN model has been used in order to include both financial statements data and a comparison between

different models to evaluate fraud detection so this is disclosed for transparency of used data and comparison.

2.3 Data Extraction and Synthesis

Data extraction Focused on the Following Variables:

- **AI Approach**: Most investigated and adopted AI or machine learning approach for fraud investigation (including supervised learning, anomaly detection, deep learning, explainable AI, graph methods and natural language processing).

- **Fraud Context:** Fraud setting in investigation (including credit-card fraud, transaction fraud, AML, insurance fraud, accounting fraud, banking fraud and their relevant contexts).

- **Investigation Role**: What functions does AI perform in investigation (including detection, alert prioritization, case triage, pattern detection, forensic analysis, monitoring, explanation etc.).

- **Data Sources**: Which data sources or features do they use primarily (including transaction records, customer behaviors, network connections, finance reports, audit data, textual data, etc.).

- **Evaluation Metrics:** What performance measures or outcome they reported (including accuracy, precision, recall, F1-score, AUC, False-Positive Rate (FPR), loss, performance comparison).

- **Main Conclusions**: Advantages, findings, and limitations of each approach. Explainability: Methods to render the AI transparent for interpretability and investigative or compliance purposes.

- **Implementation Challenges**: Technical deployment challenges, data quality issues, ethical concerns on bias, regulatory compliance, scalability challenges, privacy requirements, and adversarial attacks.

Fraud evidence of different contextual areas and families of methodology was contrasted via thematic analysis. Evidentially weight of the studies analysed was made by relating their consistencies, study designs, comparability of findings to one another, and significance to fraud investigation contextually.

RESULTS

3.1 Characteristics of Included Studies

Table-1

Study and Year	Study Type	Fraud Context	Dataset or Data	Source Method	Investigative
Ngai et al. (2010)	Academic Review	Bank, Insurance, Securities, and Related Fraud	Empirical Studies Across Financial Fraud Settings	Classification, Regression, Clustering, Prediction, Outlier Detection, Visualization	Detection, Classification, and Pattern Discovery
Ravisankar et al., 2011	Comparative Empirical Study	Financial Statement Fraud	202 Chinese Companies	MLFF, SVM, GP, GMDH, LR, PNN	Company Level Fraud Classification
Gray and Debreceny, 2014	Taxonomy and Conceptual Review	Financial Statement Audit Fraud	Financial Statements, Journal Entries, Annual Reports, MD&A, Email	Quantitative, Text, and Email Mining	Audit-Oriented Pattern Discovery
Abdallah et al., 2016	Survey	Electronic Commerce Fraud	Credit Card, Telecommunications, Insurance, and Auction Systems	Fraud Detection Systems	Detection and Prevention Integration
Ryman-Tubb et. al., 2018	Survey And Industry Benchmark	Payment-Card Fraud	Payment-Card Transactions and 2017 Transaction Volumes	AI and ML Methods	Operational Deployment Assessment
Craja et al., 2020	Empirical Model Study	Financial Statement Fraud	Financial Ratios and MD&A Text	Hierarchical Attention Network	Classification and Red-Flag Identification
Jullum et al., 2020	Empirical Institutional Study	Money Laundering	DNB Transaction, Alert, Customer, & Report Data	Supervised ML	Manual Investigation Prioritization
Al-Hashedi and Magalingam, 2021	Comprehensive Review	Bank, Insurance, Statement, and Crypto Currency Fraud	Reviewed Empirical Studies	Thirty-Four Data-Mining Techniques	Method Selection and Detection
Ashtiani and Raahemi, 2022	Systematic Literature Review	Financial Statement Fraud	Ratios, Statements, MD&A, Social-Media, and Vocal Data	Supervised, Unsupervised, Deep, Hybrid, and Evolutionary Methods	Auditor Decision Support

Hilal et al., 2022	Critical Review	Broad Financial Fraud	Multiple Financial Fraud Domains	Statistical, Supervised, Semi Supervised, and Unsupervised Anomaly Detection	Post-Event Detection
Kute et al., 2021	Critical Review	AML	Banking Transactions, Graphs, Alerts, SMRs, and Synthetic Data	Deep Learning, Graph Learning, NLP, Anomaly Detection, XAI	Alert Review, Forensic Analysis, and Reporting
Afriyie et al., 2023	Comparative Empirical Study	Credit-Card Fraud	Credit-Card Transaction Records	Logistic Regression, Random Forest, Decision Trees	Automated Classification and Monitoring
Shahana et al., 2023	Systematic Review and Bibliometric Analysis	Financial Statement Fraud	Fraud-Firm Databases and Control Groups	Data Analytics, Dimension Reduction, Optimization	Auditor and Regulator Decision Support
Aftabi et al., 2023	Empirical Model Study	Banking Financial Statement Fraud	Statements From Ten Iranian Banks	GAN and Ensemble Models	Synthetic Data-Assisted Classification
Vanini et al., 2023	Empirical Risk Management Study	Online Payment Fraud	Three Aggregated Payment Channels	ML, Economic Optimization, Risk Modelling	Loss Reduction and False-Alarm Control
Cheng et al., 2023	Empirical Graph-Learning Study	Organized AML	Real-World Bank Card-Alliance Transaction Graphs	Group-Aware Deep Graph Learning	Gang-Level Detection, Offline and Online
Awosika et al., 2024	Empirical Framework Study	Financial Transaction Fraud	Realistic Distributed Transaction Datasets	Federated Learning And XAI	Privacy Preserving Detection and Review
Charizanos et al., 2024	Empirical Framework Study	Credit-Card Fraud	Credit-Card Transaction Data	Online Fuzzy Logistic Regression	Real-Time Monitoring and Classification
Hafez et al., 2025	Systematic Review	Credit-Card Fraud	Reviewed Recent Credit-Card Studies	AI, ML, DL, and Meta-Heuristic Optimization	Comparative Method Assessment
Baisholan et al., 2025	Empirical Ensemble Study	Credit-Card Fraud	Naturally Imbalanced European Credit-Card Dataset	Random Forest, XGBoost, Threshold Optimization, SHAP	Detection, False-Positive Reduction, and Explanation

Sources: Primary Data

It covers an evolution from simple data-mining and supervised classification approaches, to the use of multi-modality, on-line updates, graph approaches, collaborative private training, synthetic data creation, explainable ensembles

etc. Transactional and financial statement fraud seem to be the most mature and well-studied use-case scenarios from an empirical standpoint. Although AML is more rapidly advancing in using relationship and group effects, challenges remain regarding both availability of data and transparency. As a consequence, the reviewed evidence varies widely in methodologies employed: we evaluate based on metrics like accuracy, AUC, recall, sensitivity, specificity, loss savings, FPs rate, and pragmatic criteria such as descriptive evaluation of how such model might be deployed in practice, where the impact is measured less by metrics than a qualitative assessment of a successful introduction to real life operations, and where often we try, where available to complement measures with evidence from expert judgment and insights of current fraud analysts and investigators for each application domain, by which we could estimate our relative ranking more accurately than based on those numbers.

3.2 Thematic Findings

3.2.1 AI Improves Automated Detection, but Reported Effectiveness is Highly Context-Dependent

The strongest empirical support comes from fraud detection applied to both financial transactions and reports where reported performances are high. For credit card settings, a random forest provided 96% accuracy with 98.9% AUC while outperforming logistic regression and decision trees in the reported experiment (Afriyie et al., 2023). An online fuzzy logistic-regression framework showed sensitivity >0.90, specificity >0.90, MCC >0.80, and accuracy >0.99 with both small sample and extreme class imbalance (Charizanos et al., 2024). FraudX AI reported 95% recall and 97% AUC-PR retaining naturally unbalanced dataset from a European credit card dataset (Baisholan et al., 2025).

These results are directionally consistent; however, they are not directly comparable. It should be acknowledged that accuracy becomes problematic in rare-event environments, where specificity, false-positive rate, recall, and AUC-PR may serve as more indicative measures of investigative effort, and in the financial-statement domain high accuracy rates have previously been documented, with a high-performance deep dense multilayer perceptron achieving 93.7%, logistic regression 93.33%, boosting 91.8%, self-organizing-map plus k-means 91%, random forest 87.5%, bagging 87.9%, and stacking 83% (Ashtiani & Raahemi, 2022). Nevertheless, the datasets and validation regimes from which these findings were drawn are varied, and better performance was achieved in the context of balanced fraud/non-fraud datasets.

These papers used credit-card data, financial statements of firms or bank records, which only partially fit the question population of financial fraud investigations across financial services. These findings should be interpreted taking the difference into consideration. Confidence: Moderate, given the strong classification performance as observed in multiple empirical studies, however the

outcome definitions, data populations, classes, and validation methodologies are very diverse.

3.2.2 Investigative Value is Greatest When AI Optimizes Human Workload Rather Than Prediction Alone

AI is applicable not merely to binary classification. In the online banking domain, a combined approach lowered both expected and unexpected losses by 15% compared to static if-then rules; with a profit optimization of the model outputs, expected losses were decreased by a further 52% with a false positive rate kept at 0.4% (Vanini et al. 2023), demonstrating that investigative impact relies as much on translating alerts into economic decisions as on forecasting precision.

A similar focus on operations can be seen in AML where supervised learning methods are used to rank transaction risks of generating report-worthy activities in the supervision system and is demonstrated to perform better than the bank's current AML method by an AML-focused performance metric (Jullum et al., 2020). The training design indicates that investigated-but-not-reported alerts carry useful information, and that naive-negative examples (un-investigated normal transactions) produce inferior models. This redirect focusses from models to constructing labels and redesigning auditor processes.

Financial-statement models favor triage over fully automated decision making. Models based on hierarchical attention networks detect "red-flag" sentences in MD&A passages, allowing relevant parties to decide if more time is needed for an analysis (Craja et al., 2020). The literature regarding reviews uses AI as support for decision making by auditors rather than a substitute for auditor judgment (Ashtiani & Raahemi, 2022).

These studies surveyed the teams engaged in bank compliance investigations, auditors, and those investigating fraud within an institution; their relevance needs to be interpreted within the context of investigators being targeted across distinct groups, not an aggregated representative of all fraud investigators. Confidence level: High regarding claims of AI being useful for prioritization/resourcing decisions; Moderate regarding claims of impact on investigative real outcomes because deployment outcomes and investigator-level indicators are rarely disclosed.

3.2.3 Class Imbalance, Label Scarcity, and Changing Fraud Behavior Are Central Determinants of Performance

Extremely rare events pose a structural issue for supervised learning. An overwhelming consensus in literature highlight issues regarding unbalanced datasets, few labelled samples, and that the reported high accuracy is derived from the fact that most transactions are in fact real transactions (Abdallah et al., 2016; Hilal et al., 2022; Kute et al., 2021). More concretely, in the literature

concerning financial-statement analysis, it is noticed that model's performance is better on the balanced set of data whereas the literature on AML is mainly concerned by the poor number of labelled suspicious transactions and its use of outdated or synthesized data (Ashtiani & Raahemi, 2022; Kute et al., 2021).

These are some solutions to the issue. GAN's produce synthetic samples prone to fraud and are as good as any supervised model yet better than any unsupervised model (Aftabi et al., 2023), FraudX AI continues natural imbalance and is a good recall/false positive choice by implementing random forest, XGBoost, probability averaging and threshold optimization (Baisholan et al., 2025). Fuzzy logistic regression solves both class imbalance, complete separation and non-stationary fraud by using a continuously improving online model (Charizanos et al., 2024). The review guides to use more unsupervised, semi-supervised, reinforcement learning, anomaly-detection methods as label information are incomplete and fraud techniques are constantly changing (Hilal et al., 2022; Kute et al., 2021).

The findings suggest that there is a difference between statistical rarity and behavioural novelties. Synthetic over-sampling will presumably provide good classification of the recognized fraud signatures and a solution like unsupervised or graph-based methods has a better chance at discovery of previously unseen, unlabelled signatures. Yet, the data in the provided papers does not indicate that any of the solutions perform overall better than the others. Confidence: Strong for imbalance, lack of labels, and concept-drift problems; Limited for definitive superiority of any solution.

3.2.4 Multimodal and Relational AI Expands Investigative Visibility

AI also becomes more investigatory when it combines multiple forms of evidence. The HAN combines financial ratios with MD&A text and captures word and sentence level structure, where the textual features can largely increase the explanatory power of financial ratios (Craja et al., 2020). Financial statement review looks at structured ratios, accounting variables, MD&A text, analyst reports, social media chatter, management disclosures, and voice information as supplementing information, but structured information is still prevalent (Ashtiani & Raahemi, 2022). The taxonomic approach in audit research also broadens data mining from quantitative statements and journal entry data to annual reports, e-mail, and MD&A text (Gray & Debreceny, 2014).

An additional weakness of account-level models, as a group, is that they fail to capture cooperative behaviour. Thus, a technique referred to as group-aware deep graph learning that aggregates similar users to form gang like communities and models' adjacent behaviour achieves above-par performance on both offline and online detection of AML, relative to existing methods (Cheng et al., 2023). Other techniques deemed useful in AML reviews include graph mining, link analysis and graph convolutional network methods for examining high density transaction networks in visual ways to support decision making (Kute et al., 2021).

While these studies were of corporate filings, bank transaction data and institution-based AML data respectively, and not all financial fraud investigations – their results need to be tempered with this nuance. Confidence level Moderate, primarily because the line of evidence is consistent with an interpretation of this information source but there is limited numerical information comparing modalities. And, in each case, they use institution specific data.

3.2.5 Explainability Is an Uneven but Essential Condition for Investigative Adoption

Most reviewed methods focused on performance, without mentioning how a researcher could extract an interpretation for specific decisions of a learned model. Among these, those based on the following methods failed to address explainability in this context: random-forest, fuzzy-logistic, GAN-ensemble, graph-learning, and economic-optimization research (Afriyie et al., 2023; Aftabi et al., 2023; Charizanos et al., 2024; Cheng et al., 2023; Vanini et al., 2023). The AML-review reported that 51% of surveyed ML models were non-interpretable and no usage of XAI to AML was made as for the time the work had been compiled (Kute et al., 2021).

"Better explained" "Explanatory": The HAN model outputs red flag statements outlining "Which managerial statements need to be investigated" (Craja et al., 2020). The FraudX AI model uses SHAP to determine feature impact for both individual and aggregated predictions enabling them to be used to aid in evaluation of predictions by domain experts (Baisholan et al., 2025). Federated learning using XAI is utilized in a bid to protect customer data whilst still ensuring explanations can be interpreted by humans (Awosika et al., 2024).

Overall, there appears to be support that explanations go "way beyond an interface feature" In AML and financial-statement investigations "it can underpin evidence building, escalate a suspicious activity, report, auditability and defending against scrutiny of a regulatory". Nevertheless, the research reviewed do not clarify if the explanations offered are reliable, sensitive to shifts in the data, informative and actionably enough, or if the explanations provided are habitually employed by investigators, Confidence: Medium for the relevance and Limited for its documented practical effectiveness.

3.2.6 Privacy, Scalability, and Deployment Conditions Constrain the Transition from Research to Practice

The use of AI necessitates having a large volume of up-to-date labelled data, but financial data are sensitive information and institutions can be legally bound from sharing their customer data. Federated learning makes a compromise by enabling distributed training to be applied on customer data which remain separated on a institutional repository (Awosika et al., 2024). This leads to a significant governance advantage, but communication overheads, scalability, a

legal and governmental approval, as well as mechanisms dealing with bias and adversarial attacks were missing.

The bridge from "prediction" to "implementation" is evident in a payment card benchmark which concluded that only eight payment-card methods would realistically perform under the 2017 business volume transaction rates if deployed within industry Ryman Tubb et al., 2018). Studies highlight consistent implementation challenges such as real-time capabilities, huge volumes, data change/drift and integration with prevention systems (Abdallah et al., 2016). AML assessments add challenges integrating AI with existing rule-based systems that remain necessary for threshold and cross border transaction analysis, although can be prone to a 98% false-positive rate (Kute et al., 2021).

The assessment pertains to deployments that occur in an institutional setting such as within banks, on payment rails, and on top of AML infrastructure; these environments only partially intersect with the populations of financial services firms in Scope and the financial institutions/ jurisdictions considered in financial fraud; a more conservative confidence for each dimension may be warranted. Confidence: Strong for the existence of deployment constraints; Strong for the existence of research-to-practice gaps; Limited for the effectiveness of the considered governance and infrastructure solutions.

3.3 Summary of Evidence

Table-2

Theme	Key Finding	Population Applicability	Effect Direction	Confidence Level	Supporting Studies
Automated Detection Performance	Random forest achieved 96% accuracy and 98.9% AUC; fuzzy logistic regression achieved accuracy greater than 0.99; FraudX AI achieved 95% recall and 97% AUC-PR	Credit-card and financial statement datasets; partially applicable to broader financial investigations	Positive	Moderate	Afriyie et al. 2023, Charizanos et al. 2024, Baisholan et al. 2025
Investigative Prioritization and Loss Reduction	Losses declined by 15% against static rules, with an additional 52% expected-loss reduction after optimization and a 0.4% false-positive rate	Online banking and AML institutions; partially applicable to financial investigators generally	Positive	Strong	Vanini et al. 2023, Jullum et al., 2020

Class Imbalance and Label Scarcity	AML models face scarce labels; financial-statement models perform better on balanced datasets; GANs and threshold methods address rarity	Transaction, AML, & financial statement datasets; broadly relevant but context dependent	Mixed	Strong	Kute et al. 2021, Ashtiani & Raahemi, 2022, Aftabi et al., 2023
Adaptation to Evolving Fraud	Concept drift, non-stationary patterns, and changing fraud vectors undermine static models	Electronic commerce, payment cards, online credit-card systems; applicable to dynamic investigations	Negative for static models; positive for adaptive approach	Strong	Abdallah et al. 2016, Charizanos et al. 2024, Hafez et al. 2025
Multimodal and Graph Based Analysis	Text reinforces financial ratios; group-aware graph learning improves organized AML detection	Corporate reports & bank transaction networks; partially applicable to broader financial investigations	Positive	Moderate	Craja et al. 2020, Cheng et al., 2023
Explainability	SHAP, attention based red flags, And XAI improve interpretive potential, but 51% of AML Models were non-interpretable and AML-specific XAI applications were not identified in the reviewed literature	Auditors, AML analysts, and financial experts; directly relevant but incompletely validated	Mixed	Moderate	Craja et al., 2020, Kute et al., 2021, Baisholan et al., 2025
Privacy and Collaborative Learning	Federated learning supports joint training without direct customer-data sharing while preserving interpretability through XAI	Financial institutions using distributed data; applicability depends on infrastructure and regulation	Positive with implementation uncertainty	Limited	Awosika et al., 2024
Research-To-Practice Translation	Only eight surveyed methods were judged practically deployable at 2017 transaction volumes	Payment-card industry; narrower than the full financial investigation population	Negative or constrained	Moderate	Ryman-Tubb et al., 2018

Sources: Primary Data

DISCUSSION

4.1 Principal Findings and Their Interpretation

The key message is that AI is most useful in assisting financial fraud investigation when used within a human decision loop. Classification score alone is not enough to provide value in an investigation; of far greater consequence is work where it is used for transaction prioritization, prediction of expected loss, detection of red flag text, and formalization of network structure for analyst consumption. This explains why the loss-optimization framework is particularly enlightening; it incorporates false alarms and financial impact into the investigation decision rather than as a secondary technical concern (Vanini et al., 2023). By analogy AML prioritization shows that the quality of the training labels and the handling of investigated-yet-unreported cases may be more influential than the learner (Jullum et al., 2020).

The synthesis also suggests a methodological move from single-record analysis toward modelling contextual representation. The use of attention mechanisms offers doc- level context and graph learning models the ability to model the correlated action across accounts that an accounting model is ill-suited to do (Cheng et al., 2023; Craja et al., 2020). This all seems mechanically possible in the computational sense as fraudulent behavior seems plausible to involve not just one peculiar marker, but patterns across combinations of financial signals, language use, temporal patterns, and account interaction. But no papers offer biological or Physiological mechanism and no research is available as pertinent to this specific query. Thus, the primary missing mechanism must be an explanatory one; models could identify correlations without being in a position to say why that correlation means what fraudsters intends.

The highest confidence is attributed to the conclusion that AI increases detection and triage under a specific set of circumstances. Lower confidence is given to the broadest conclusions of superiority since the metrics studied are varied and many datasets tested are internal and synthetically balanced or lack the necessary metadata. Top numbers achieved-95% recall and 97% AUC-PR for FraudX AI, greater than 0.99 accuracy for fuzzy logistic regression, and 96% accuracy and 98.9% AUC with random forest-should be considered study specific successes instead of indicative performance of a system broadly (Afriyie et al., 2023; Baisholan et al., 2025; Charizanos et al., 2024). The synthesis points out that operational utility is only obtained when predictions, costs, explanations, privacy and the investigator workflow align.

4.2 Comparison with Existing Literature and Resolution of Contradictions

The reported results are consistent with previous reviews, the use of mining/ML in fraud classification is not new. However, they help illustrate why past occurrences cannot be presented as proof that they would work universally.

Previous surveys mention logistic, neural networks, Bayesian belief network and decision trees-more modern studies also reference ensembles, deep learning, graph-based classification, federated learning, XAI etc (Al-Hashedi & Magalingam, 2021; Ngai et al., 2010). The increasing range of methods is relevant as more modern techniques do not implicitly assume that issues like high imbalance, relational behavior, privacy constraints and human interpretable data may not be important-they specifically seek to overcome them.

The fact that a method obtained high predictive measures but failed to be deployable in practice is also not explainable by one group of studies being "wrong". The studies have simply measured different characteristics. Metrics such as accuracy and AUC are concerned with performance within the dataset in which they are trained and tested; a model's deployability additionally requires consideration of factors like transaction volume, latency, load placed on the system by the rate of false-positive signals, data available in operational settings, infrastructure for integration and the overhead incurred in model maintenance. That the payment-card benchmarks found eight models deployable, despite any of them could have achieved extremely good accuracy in study, is due to the fact that the requirements for suitable operation are not the same as those for classification over historical data (Ryman Tubb et al., 2018). There is no information in the submitted data to determine what criteria failed for each specific method.

Another dilemma has to do with supervision and whether models should be trained with labelled data (supervised) or not (unsupervised/semi-supervised learning). Most financial statement and anti-money laundering research that utilize data has applied supervised methodologies that have shown good classification performance. Nonetheless researchers typically call for either unsupervised or semi-supervised models to address the lack of fraud labelled data and increase detection of newly appearing behaviours (Ashtiani & Raahemi, 2022; Hilal et al., 2022). These could be empirically reconcilable outcomes - supervised methods might perform better to detect known labelled anomalies while unsupervised methods might achieve this on previously unheard-of anomalies. The problem is that the literature does not provide enough direct comparisons across different models, and different levels of supervision to confidently assess which approach adds more value to investigative work. Since publishing failed deployments or bad outcomes seems more difficult and would diminish a researcher’s credibility (there is a propensity to report success metrics), a substantial publication bias could be present.

4.3 Practical Implications

The initial application for banks and payment institutions of AI should be in a triage & monitoring function to prioritize the alerted list on risk of fraud, economic value and investigational benefit; a situation that exists when static rules generate too many false positives, such as is common for AML systems where

false positive rates can as high as 98% (Kute et al., 2021). Banks and payment institutions in rapidly changing credit card & online payment fraud environments, should prioritize dynamic, non-stationarity robust models; institutions should emphasize recall, precision, specificity, AUC-PR, false positive rates, monetary losses and analyst load as metrics over simple accuracy (Charizanos et al., 2024; Vanini et al., 2023).

There could be value for auditors and financial-reporting investigators to use a multimodal system with ratio variables plus text from MD&A statements, journal entries, disclosures, etc. The presence of attention-based red-flag sentences and of SHAP-based feature contributions provides a clearer explanation for model outputs, but should be framed in such a way to aid professional discretion not prove an investigator's claims about a fraud (Baisholan et al., 2025; Craja et al., 2020). More specifically, there are strong implications for institutional investigators and auditors as well as compliance specialists and experts familiar with the subject matter. There was not one, general class of investigators.

For the regulators, documentation of trained labels, class imbalance, false positive burden, model updating mechanisms, explanation methods, and human-review procedures. Federated learning is a method of joint detection in cases where laws prohibit central sharing of data; however, the literature fails to verify the communication cost, fairness, scalability, or regulatory approval of federated learning (Awosika et al., 2024). There is no dose response issue with this policy question; the most appropriate policy principle is that no performance threshold provides a safe system under evolving fraud trends, evolving labels, and changing contexts.

4.4 Strengths and Limitations

This review is of high quality because it includes a wider array of papers due to a wider initial search, incorporates empirical studies along with previously published review articles, synthesizes results across transaction fraud, financial-statement fraud and AML, not just detection accuracy, and not just predictive accuracy (e.g., also includes investigative usage, explainability, privacy concerns, class imbalance, concept drift and deployability). Limitations, such as those with heterogeneous, or incomplete provenance of datasets, incomplete reporting of measurement variables (e.g., lack of precision, recall, F1 score), reliance on firm or time labelled data, class imbalance, poor model-explainability to investigators, presentation of models at a high-level for some existing review articles, and the reliance of some existing empirical studies on small geographic samples, restricted populations or data synthesis limits may be applicable, but for many papers the reporting for these are sufficiently covered. The broadness of the application area, combined with the breadth of issues considered, makes it an excellent and comprehensive starting point for those embarking on or performing research on the topic of machine learning applications within forensic accounting.

The current review is intrinsically limited by a reliance on the downloaded description paper only, plus additionally extracted from the descriptions for some records, and only by available abstract data only for others. A formally performed risk-of-bias appraisal was performed, but a full pooled meta-analysis could not be executed. Thus, the estimated performance measure cannot be conceived as a an estimate of a shared effect nor be understood as system benchmarks to which those published elsewhere can be compared.

GAPS AND FUTURE DIRECTIONS

Future studies would need to test AI in prospective, institutionalized, prospective rather than the present retro classification. Workload for investigators, alert to case conversion, loss prevention in monetary terms, workload per false positive, time to detect and model calibration and post model update stability would be the key variables to assess and provide direct evidence for better investigation rather than labelling of records.

One specific aspect deserving consideration is the definition of labels. Existing research in AML demonstrates the misuse of non-reported and investigated alerts, and normal transactions, that have not been investigated, as straightforward training data labels, and accounting-research of fraud is challenged by infrequent fraud occurrences and control group formulation (Jullum et al., 2020; Shahana et al., 2023). The evaluations in future should ideally include a comparative analysis among supervised, semi-supervised, unsupervised, graph, and reinforcement learning approach over similar datasets and realistic class distributions.

The need for replication beyond institutions, jurisdictions, types of fraud and over time is also not satisfied by literature. Much of the existing evidence comes from credit card transactions, few banks, company reports and a handful of AML networks. We require the testing of hypotheses through changes in concepts (concept drift), the validation of the correlation between SHAP / attention measures and investigator thought-process, and the examination of performance in heterogeneous federated learning settings, across diverse institutions. There is also a particular need for the development of specific research focusing on organized fraud, money laundering, mortgage fraud and securities fraud, in underrepresented non-English or non-Western geographies / institutions, due to the poor category / geographical coverage revealed by existing review literature (Al-Hashedi & Magalingam, 2021; Ngai et al., 2010).

CONCLUSION

There is an important and evolving role for AI in financial fraud investigations, but the strongest and most justifiable application is augmenting, not

autonomously deciding, upon fraud itself. Across credit-card, payment, financial-statement and AML use-cases, AI improves detection, risk scoring, anomaly detection and profiling where the data is suitably informing and the model matches human review. The best results cited herein include 96% accuracy and 98.9% AUC when using random forest, accuracy >0.99 with >0.90 sensitivity and specificity using fuzzy logistic regression and 95% recall and 97% AUC-PR on an interpretable ensemble system (Afriyie et al., 2023; Baisholan et al., 2025; Charizanos et al., 2024). With respect to online payment fraud investigations ML improved both expected and unexpected loss by 15% over static rules whereas economic optimization would lower expected loss by an additional 52% and a false-positive rate of 0.4% (Vanini et al., 2023).

This analysis is based on partially overlapping datasets containing a set of banks, payment records, companies' information, networks, etc. The prediction results can't be generalized to every financial investigation setting yet. It remains open as to whether measured predictive gains carry over into operational future deployment across several institutions while controlling for levels of false positive rates and maintaining evidence's intuitiveness, privacy, equity and capability of adaptation to new fraud schemes. AI will most influence practitioners with a mixture of contextual and relationship info, evidence of explicability, accounting for costs, and the control of experts, thereby it should be used as a focal point. More investigation is thus needed on the results in reality and the governing procedures, not on the predictive capabilities.

REFERENCES

1. Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90–113. https://doi.org/10.1016/j.jnca.2016.04.007
2. Afriyie, J. K., Tawiah, K., Pels, W. A., Addai-Henne, S., Dwamena, H. A., Owiredu, E. O., Ayeh, S. A., & Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decision Analytics Journal, 6, 100163. https://doi.org/10.1016/j.dajour.2023.100163
3. Aftabi, S. Z., Ahmadi, A., & Farzi, S. (2023). Fraud detection in financial statements using data mining and GAN models. Expert Systems with Applications, 227, 120144. https://doi.org/10.1016/j.eswa.2023.120144
4. Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review, 40, 100402. https://doi.org/10.1016/j.cosrev.2021.100402
5. Ashtiani, M. N., & Raahemi, B. (2022). Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. IEEE Access, 10, 72504–72525. https://doi.org/10.1109/access.2021.3096799

6. Awosika, T., Shukla, R. M., & Pranggono, B. (2024). Transparency and privacy: The role of explainable AI and federated learning in financial fraud detection. IEEE Access, 12, 64551–64560. https://doi.org/10.1109/access.2024.3394528
7. Baisholan, N., Dietz, J. E., Gnatyuk, S., Turdalyuly, M., Matson, E. T., & Байшоланова, К. (2025). FraudX AI: An Interpretable Machine Learning Framework for Credit Card Fraud Detection on Imbalanced Datasets. Computers, 14, 120–120. https://doi.org/10.3390/computers14040120
8. Charizanos, G., Demirhan, H., & İçen, D. (2024). An online fuzzy fraud detection framework for credit card transactions. Expert Systems with Applications, 252, 124127. https://doi.org/10.1016/j.eswa.2024.124127
9. Cheng, D., Ye, Y., Xiang, S., Ma, Z., Zhang, Y., & Jiang, C. (2023). Anti-Money laundering by group-aware deep graph learning. IEEE Transactions on Knowledge and Data Engineering, 35(12), 12444–12457. https://doi.org/10.1109/tkde.2023.3272396
10. 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
11. Gray, G. L., & Debreceny, R. S. (2014). A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. International Journal of Accounting Information Systems, 15(4), 357–380. https://doi.org/10.1016/j.accinf.2014.05.006
12. Hafez, I. Y., Hafez, A. Y., Saleh, A., Abd El-Mageed, A. A., & Abohany, A. A. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. Journal of Big Data, 12(1). https://doi.org/10.1186/s40537-024-01048-8
13. Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. Expert Systems with Applications, 193, 116429. https://doi.org/10.1016/j.eswa.2021.116429
14. Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., & Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. Journal of Money Laundering Control, 23(1), 173–186. https://doi.org/10.1108/jmlc-07 2019-0055
15. Kute, D. V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. IEEE Access, 9, 82300–82317. https://doi.org/10.1109/access.2021.3086230
16. Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2010). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50, 559–569. https://doi.org/10.1016/j.dss.2010.08.006
17. Ravisankar, P., Ravi, V., Raghava Rao, G., & 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

18. Ryman-Tubb, N., Krause, P., & Garn, W. (2018). How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. Engineering Applications of Artificial Intelligence, 76, 130–157. https://doi.org/10.1016/j.engappai.2018.07.008
19. Shahana, T., Lavanya, V., & Bhat, A. R. (2023). State of the art in financial statement fraud detection: A systematic review. Technological Forecasting and Social Change, 192, 122527. https://doi.org/10.1016/j.techfore.2023.122527
20. Vanini, P., Rossi, S., Zvizdic, E., & Domenig, T. (2023). Online payment fraud: from anomaly detection to risk management. Financial Innovation, 9(1). https://doi.org/10.1186/s40854-023-00470-w