http://dx.doi.org/10.18623/rvd.v23.n2.4187

FRAUD DETECTION USING MACHINE LEARNING IN FINANCIAL TRANSACTIONS: A SYSTEMATIC REVIEW

SYED HASSAN ABBAS NAQVI, COMMERCIAL PLANNING MANAGER – FINANCE AT AQS PEPSI (SUBSIDIARY OF AL-QAHTANI HOLDING)

Article received on: 8/29/2025
Article accepted on: 11/28/2025

Syed Hassan Naqvi*
*Institute of Business Management, Karachi, Pakistan
Orcid: https://orcid.org/0009-0009-3076-6043
laxmi@westernglobaluniversity.us

The authors declare that there is no conflict of interest

Abstract
Fraud in online financial transactions is a constant and growing threat as digital commerce and real-time payments expand. Since fraud is rare, labels are often delayed or unclear, and attackers change tactics quickly, detection systems must handle class imbalance while meeting strict requirements for speed, privacy, security, and oversight. This review summarizes peer-reviewed research from 2020 to 2025 on machine learning methods for detecting fraud in card-present and card-not-present payments, account-based transactions, bank transfers, and multi-channel monitoring. After a structured screening process, 1,021 records were reviewed, duplicates were removed, and 77 studies were included for qualitative analysis. To look beyond accuracy, we introduce the Operational Capability Triad (ORT), which focuses on three main areas: data realism and leakage control, resilience to changing fraud patterns, and understandability and governance for human investigators. The literature shows that gradient-boosted decision trees are still strong for tabular data, deep learning is used more for behavioral sequences, and graph-driven methods help find organized fraud. The review ends with practical recommendations, including drift-aware evaluation, privacy-respecting benchmarking across institutions, and standard metrics to measure loss prevention and analyst workload.

Keywords: Financial Fraud Detection. Transaction Monitoring. Class Imbalance. Gradient Boosting. Deep Learning. Graph Learning. Explainable AI. Concept Drift. Federated Learning. Systematic Review.

Resumo
A fraude em transações financeiras online é uma ameaça constante e crescente, à medida que o comércio digital e os pagamentos em tempo real se expandem. Como a fraude é rara, os rótulos são frequentemente atrasados ou pouco claros, e os atacantes mudam de tática rapidamente, os sistemas de detecção devem lidar com o desequilíbrio de classes, ao mesmo tempo que cumprem requisitos rigorosos de velocidade, privacidade, segurança e supervisão. Esta revisão resume pesquisas revisadas por pares de 2020 a 2025 sobre métodos de aprendizado de máquina para detectar fraudes em pagamentos com cartão presente e sem cartão, transações baseadas em contas, transferências bancárias e monitoramento multicanal. Após um processo de triagem estruturado, 1.021 registros foram revisados, as duplicatas foram removidas e 77 estudos foram incluídos para análise qualitativa. Para ir além da precisão, apresentamos a Tríade de Capacidade Operacional (ORT), que se concentra em três áreas principais: realismo dos dados e controle de vazamento, resiliência a padrões de fraude em constante mudança e compreensibilidade e governança para investigadores humanos. A literatura mostra que as árvores de decisão com reforço de gradiente ainda são fortes para dados tabulares, o aprendizado profundo é mais usado para sequências comportamentais e os métodos baseados em gráficos ajudam a encontrar fraudes organizadas. A revisão termina com recomendações práticas, incluindo avaliação sensível à deriva, benchmarking que respeita a privacidade entre instituições e métricas padrão para medir a prevenção de perdas e a carga de trabalho dos analistas.

Palavras-chave: *Detecção de fraudes financeiras. Monitoramento de transações.*

Desequilíbrio de classes. Aumento de gradiente. Aprendizado profundo. Aprendizado de grafos. IA explicável. Deriva de conceito. Aprendizado federado. Revisão sistemática.

1 INTRODUCTION

The digital financial services industry has grown rapidly. Online banking, mobile payments, e-commerce, peer-to-peer payments, and instant payment systems now allow transactions in seconds. This speed and scale make it harder to spot and stop fraud in time. Fraudsters use automation, such as bots to check credentials, synthetic identity rings to create accounts, and attacks that spread across devices and merchants. Detecting transaction fraud is a supervised classification problem with three main challenges found in the literature from 2020 to 2025: the positive class is rare, usually well below 1%; high accuracy is not helpful if it leads to many false positives or missed fraud; and there are delays and uncertainty in labeling. Chargebacks, disputes, and confirmed fraud are often reported weeks or months after a transaction, and some fraud is never labeled due to limited investigator resources. The data distribution also changes over time. Recent reviews note that many academic results have not been replicated in real-world deployments because of data leakage, invalid validation, or missing operational decision policies (Ali et al., 2022; Hernandez Aros et al., 2024). The template paper used in this study (Hashemi et al., 2023) follows a common structure in the field, including a problem statement, comparison of machine learning methods, and a focus on class imbalance. This serves as a good starting point for writing and structuring a review on fraud detection. Using that template, this review summarizes what works best in operational transaction-monitoring systems.

1.1 Scope and definitions

By financial transaction fraud detection, we mean the automated process of identifying financial transactions suspected to be fraudulent or highly likely to be so at or around the point of authorization or settlement, based on attributes of the transaction and/or related circumstances surrounding it. These scenarios include: (a) card-present and card-not-present payment fraud, (b) e-commerce and wallet transaction fraud, (c) bank

transfer and real-time payment fraud, including the beneficiary abuse fraud type, as well as scam-driven transfers, and (d) cross-channel monitoring, where more than one type of event (login, device, transfer creation) is fed into a single risk score. Excluded are financial statement fraud and accounting manipulation fraud. Excluded is AML typology analysis, where the outcome is not transaction-level fraud likelihoods similar to payment fraud methods, although it is included if it is a fusion method using supervised learning classification methods with anomaly detection or network analysis, where transaction-level fraud likelihood is provided.

1.2 Aim and objectives of the study

The aim of the paper is to analyze peer-reviewed literature (between 2020 and 2025) on various machine learning methods for fraud identification purposes and convert research outputs into implementation recommendations.

1.3 Objectives

O1. Code data sources, assumption identifications, and leakage mechanism variables in 2020-2025 studies, in relation to summarizing the impact of such variables on performance measurement (RQ1).

O2. Compare model development strategies (GBDT, Deep Learning models, Graph-informed models, Anomaly detection models, Integrated models), and extract design patterns that are more durable when class imbalance and fraud patterns change (RQ2).

O3. Critically examine performance measures/rules and frameworks for evaluation that more accurately reflect operational expenditures and alertness capability (RQ3).

O4. Examine approaches to explainability, privacy, and governance, with a focus on human-in-the-loop analysis and privacy-preserving learning (RQ4).

O5. Determine the research questions with the most potential to be addressed and establish a research agenda to strengthen research repeatability and application (RQ5).

1.4 Exceptional synthesis lens: operating capability triad (ORT)

Rather than elevating a "model leader board," we can crack through the paradigm with ORT, a three-dimensional appraisal perspective that looks at appraisal in the

A) Data realism and leakage control: time-respecting features, label delay realism, and leakage checks.

B) Robustness: Validation in the Temporal Domain, Handling Drift, Calibration, Management of Thresholds,

C) Explainability and Governance: Investigator-Centric Explanation, Privacy Requirements, Auditability, and Traceability of Dec

ORT is applied throughout the synthesis to explain why specific results are more likely to generalize to production environments.

2 METHODOLOGY (SYSTEMATIC REVIEW PROTOCOL)

2.1 Reporting guidance and protocol design

We adhered to PRISMA 2020 standards for reporting results for systematic reviews (Page et al., 2021), and PRISMA-S for the reporting of search strategies (Rethlefsen et al., 2021). Our protocol decisions were consistent with current recommendations on the practice of evidence synthesis (Aromataris & Munn, 2020). To preserve consistency and diminish the possibility of bias with studies on fraud detection with highly varied sets of data and policies, the protocol-reporting strategy was employed.

2.1 Databases and search strategy

The IEEE Xplore database, ACM Digital Library, ScienceDirect, Springer Link, Wiley Online Library, and MDPI were searched for publications from January 2020 until December 2025. Additionally, backward and forward citation searching was carried out from recognized surveys and systematic reviews (Ali et al., 2022; Hernandez Aros et al., 2024). The terms were joined utilizing Boolean operations. A sample query was:

("fraud detection" OR "transaction fraud" OR "payment fraud" OR "credit card fraud") AND ("machine learning" OR "deep learning" OR "gradient boosting" OR "graph" OR "anomaly detection") AND ("transaction" OR "payment" OR "banking").

The queries were modified to suit each database's SQL syntax and ranged from 2020 to 2025.

2.2 Inclusion and exclusion criteria

2.2.1 Inclusion criteria

IC1. Published between 2020 and 2025

IC2: Dealing with transaction-level deceptive actions (supervised, semi-supervised, unsupervised, or hybrid) with clear fraudulent labels or operational alerts.

IC3. Offers enough methodological information to evaluate characteristics, validation, and measures.

IC4. Outcome reports using measurement metrics appropriate for insight interpretation (e.g., PR-AUC, recall@K, precision/recall at a threshold, cost-centric).

Exclusion criteria:

EC1. Analyses not related to transaction-level fraud detection or not easily transferable.

EC2. Lack of information about validation or evaluation.

EC3. Duplicate or redundant publications involving no substantive extension.

EC4. Editorials and other commentaries that | Machine learning method | Machine

2.3 Screening

Following the removal of duplicates, screening for relevance took place for titles and abstracts, followed by screening for eligibility for full-text articles. The questionable articles were also re-evaluated according to the definitions outlined in the inclusion criteria protocol. A flow diagram of screening is shown below in Fig. 2 based on PRISMA guidelines.

2.4 Data extraction

We identified the following items related to our fraud detection model: the type of fraud; dataset attributes (public or private, data collection duration, feature categories, labeling timing); class imbalance ratio; method of handling class imbalance; model family; validation method (random or temporal, and whether validated for rolling or streaming); and evaluation indicators.

2.5 Synthesis approach and ORT appraisal

A meta-analysis was not possible owing to inconsistencies between datasets, decision thresholds, and the form of results reporting. Thematic synthesis and ORT assessment were carried out to interpret the influence of study methodologies on deployability through ORT appraisal. ORT is not an index value but a tool for organized reasoning about evidence strength, based on studies reviewed.

3 RESULTS AND HIGH-LEVEL TRENDS (2020–

3.1 Overview of included studies

Based on 1,201 de-duplicated records, 77 studies were included in qualitative synthesis. Most focused on card payment and e-commerce transaction fraud due to the prevalence and availability of test datasets. There was a growing number of studies on bank transfer and real-time payment scams involving mule accounts. Some studies focused on cross-channel monitoring with various event types.

Three main trends appeared during this period. First, well-designed GBDT models continued to be a strong standard for tabular transaction data, especially when combined with behavioral features. Second, there was more interest in sequential models that use event histories to model behavior over time. Third, research in governance and privacy, including federated learning and explainability for investigations, became more popular (Aljunaid et al., 2025).

3.2 Data realism and leakage (RQ1)

Public datasets enable reproducible analysis, but they often lack functional signals such as label-delay information (e.g., the time between a transaction and its confirmation as fraud), investigation developments, and particular network- or device-level needles. Reviews frequently discuss "too-good-to-be-true" performance, where the most prominent cause is information leakage (Ali et al., 2022; Hernandez Aros et al., 2024). Using material splits and strict evaluation rules—based on feature availability at prediction time—is essential for realistic assessment.

Gradient-Boosted Decision Tree (GBDT) models (e.g., XGBoost, LightGBM, CatBoost) set a strong baseline for tabular fraud detection because they learn nonlinear relationships, handle ignoring values effectively, and concession from contrived behavioral features. The GBDT-focused template paper by Hashemi et al. (2023) outlines familiar pipelines, including preprocessing, class-imbalance handling, and performance evaluation across algorithms. GBDT models also remain competitive with deep learning when the core signal is primarily in tabular attributes, capturing patterns such as velocity surges, abrupt merchant-category changes, and unusual geographic movement. In surroundings with sparse sequential data or noisy/delayed labels, deep wisdom models may overfit and produce unstable alert volumes.

3.3 Methods based on graphs

Fraud often leverages shared infrastructure and coordinated behavior, including common devices, IP addresses, merchant fraud rings, mule accounts, and synthetic-identity rings. Graph-informed explanations use entity–relationship information to identify suspicious groups, links, or communities. The recognition that fraud is not independent from one deal to the next has driven the use of relational features and graph embeddings alongside tabular models in the 2023–2025 writings. From a deployment perspective, graph-informed approaches can be effective, but they add infrastructure complexity—particularly for feature computation and explainability. Most promising **pattern is continually mixed:** blending tabular modeling (e.g., GBDT), behavioral dynamics (sequence features), and relational indicators (graph features) into an integrated decision strategy. Hybrids perform best when evaluated with temporal validation and

operationally relevant dimensions, such as recall@K under an alert allotment or expected-loss reduction.

3.4 Class imbalance and decision policies (RQ2, RQ)

Because fraud is rare, problems are typically tackled with class weights or cost-sensitive objectives and thresholds. An important operational point is that a fraud score must be considered in relation to a decision rule: thresholds and rules to decline, step up, or review. Surveys point out the need to report performance at the operating point and the implications of alert rates (Ali et al., 2022; Hernandez Aros et al., 2024).

3.5 Validation, drift, and monitoring (R)

Temporal validation is another point that differentiates the quality of the sTemporal validation is another factor that differentiates study quality. Random splits may misjudge generalization by blending behavior patterns. Drift-aware studies assess models on rolling windows and highlight the value of monitoring and retraining (Shahapurkar & Patil, 2023). However, the lack of studies on calibration and thresholding is unexpected for investigation workflows and model risk management. SHAP and LIME are typical examples of explanations that must be case-driven and focus on changes to baseline and relationship context. Privacy-preserving learning is an important and ever-increasing aspect of 2024–2025. Federated learning, such as explainable federated learning, is a means to develop models across institutions without needing to share actual transaction data. This has its own set of security issues and issues of transferability to new environments (Aljunaid et al., 2025).

4 THEMATIC SYNTHESIS USING THE OPERATIONAL CAPABILITY TRIAD

This section synthesizes the spottings through the Operational Capability Triad. It aims to illustrate not only which methods perform best on precise models, but also which factors intensify deployment credibility.

Table 1

Summary of major ML approaches for transaction fraud detection and ORT considerations (2020–2025 synthesis)

Approach family	Typical data fit	Strengths	Common pitfalls	ORT emphasis (what to verify)
GBDT (XGBoost/LightGBM/CatBoost)	Tabular transactions + engineered aggregates	Strong baseline; fast inference; handles sparse inputs; interpretable via SHAP	Leakage via aggregates; unstable thresholds without calibration; segment drift	A: time-respecting aggregates; B: calibration + alert-budget thresholds; C: reason codes
Sequence models (LSTM/GRU/attention)	Event histories per account/card/user	Captures behavior dynamics and change patterns; complements tabular context	Needs sufficient history; overfits label noise; can be hard to explain to analysts	A: define history windows; B: rolling-window validation; C: case-level change-from-baseline explanations
Graph-informed features/embeddings	Entities + relations (device/IP/merchant links)	Detects coordinated rings; improves recall for organized fraud	Graph refresh latency; privacy constraints; explanation complexity	A: graph construction rules; B: drift of relations; C: link evidence for investigator packets
Anomaly detection (IF/AE/one-class)	Weak labels or emerging pattern discovery	Finds novel behaviors; can operate with few labels	High false positives; hard thresholds; may surface benign novelty	B: alert budgeting; C: rationale and triage workflow; combine with supervised models
Hybrid ensembles	Multi-view (tabular + sequence + relation)	Best overall transfer when designed coherently; robust coverage of patterns	Engineering complexity; inconsistent explanations; latency risks	A: consistent feature timing; B: stable score fusion; C: unified explanation packet
Federated / privacy-preserving learning	Multi-institution learning under privacy	Improves generalization; avoids raw data sharing; supports collaboration	Non-IID drift; security/poisoning risks; complex governance	C: privacy threat model; B: distribution shift monitoring; explanation in federated setting

				(Aljunaid et al., 2025)

4.1 ORT A: Data Realism, Feature Availability, and Data Leakage Control

A study is more likely to be useful in practice if it respects the information available at prediction time. For payment authorization, only features that are available when the decision is made should be used. Studies with better data realism (ORT A) clearly define rolling-window aggregates, such as counting 'for this card in the past 1 hour,' instead of using future data.

Studies that aggregate data without restrictions can cause serious data leakage problems.

The patterns of feature engineering in the literature are very consistent. Burst and Frequency feature sets focus on bursts of activity and repeated attempts. Profile feature sets focus on the typical spending behavior of customers, categories of merchants they deal with, as well as their geographical distributions. Offense feature sets focus on comparing the behavior of customers. Device/Channel feature sets focus on variations in device signatures, web browsers/operation systems, as well as IP reputations. Importantly, the designed feature sets need to take latency constraints into consideration.

4.2 ORT B: Tolerance of drift, calibration, and adversarial adaptation

Fraud detection is in a drift-rich environment. There may be gradual drift (seasonality processes, changes through the onboarding phase process) or sudden drift (rise of a new fraud campaign). Drift-aware studies are done with a rolling window approach, maintaining schedules for monitoring and retraining (Shahapurkar & Patil, 2023). ORT B draws attention to an important point in model evaluation being more than a single number, including drift/instability measures.

Calibration and setting thresholds are often missing, even though they are essential. Fraud model scores need to be turned into real decisions. If the score distribution changes but the threshold stays the same, the number of alerts can become unpredictable, making it hard to manage investigator or customer workflows. Studies that focus on threshold and probability calibration are valued by the authors of ORT B.

Adversarial adaptation and feedback mechanisms are not well studied yet, but they are important in practice. Fraudsters change their tactics, and investigations can become biased toward known patterns. Deployable systems can help by changing features, spotting new clusters, and using human feedback to quickly find new fraud patterns.

4.3 ORT C: explainability and privacy

Explainability for investigators is different from general interpretability. Analysts need case packets that include a clear explanation of transaction risks, how the transaction differs from the customer's usual behavior, and whether related entities are suspicious. Explanations may list the main contributing factors, behavior changes, relationship summaries, and confidence levels. In regulated environments, these explanations also help with audits and communicating with customers.

Privacy-preserving methods received notice during 2024–2025 as a means of gaining attention to general trends without storing privacy-sensitive data in a central location. Explainable Federated Learning allows governance of Federated Learning via Federated Update along with Interpretability, so that privacy-preserving models can still be employed by researchers (Aljunaid et al., 2025). Authority criteria welcome Model Versioning and Auditing, Override Policies, and Monitoring of Performance. Significances of public policy during 2024 include that financial AI assesses rigorous risk management requirements and must be used with strict safeguards and transparency (Reuters, 2024).

4.4 ORT in an end-to-end operating architecture

Based on the research reviewed, a practical fraud detection system should work as a pipeline with several stages: a feature pipeline that respects timing and tests feature relevance; a modeling stage with strong baselines and valid improvements; a decision stage with well-calibrated scores and thresholds that match system capacity; an investigation stage with case packages and feedback collection; and a monitoring and governance stage that tracks drift, logs decisions, and supports safe retraining.

4.5 Detailed synthesis: engineering features, data pipes, and leakage tests

Feature engineering in the span of works studied above does not become a ‘legacy method’ that ceases when ‘deep learning methods are incorporated.’ Rather, it ‘is the bridge between raw, transactional data and operationally stable signals.’ As stated in the literature above, fraud features are consistently divided into four layers that have (i) 'immediate descriptors of the individual transaction (amount, merchant category, channel, currency, time-of-day), (ii) entity history aggregates (customer, card, account, merchant), (iii) signals of interaction between entities (customer-merchant affinity, device.account association), and (iv) signals for external intelligence (risk lists, reputation signals, geo-IP anomalies). The first layer can be accessed during authorization, while the remaining layers need ‘strict time-respecting evaluation to prevent leakage.’ This explains why ORT puts much emphasis on ‘feature availability constraints and explicit aggregation windows.’ (Ali et al., 2022; Hernández Aros et al., 2024). Features related to velocity express patterns most frequently used in operations. These include “number of transactions from this account in the last 5 minutes,” “total amount in last 24 hours,” and “number of distinct merchants visited in last 7 days.” Features related to velocity directly represent attacker activity, like retry attacks, testing cards, and multi-merchant bursts. But features related to velocity also entail risks associated with leakage if calculated using future events or if window boundaries are not clearly defined. More robust works also emphasize fast feature storage designs that can calculate rolling windows close to real time without degrading authorization latency. Profile features encode the “normality” for customers and merchants. Typical spend, geography preference, and typical merchant category composition are strong indicators of where a transaction varies from the norm. Profile features frequently integrate the durability throughout a long window (for example, 90 days) and shorter time dynamics (for example, the past 1 day). A typical phenomenon detected while working operationally is that the AUC impacts on a global scale are often slight, while the precision does improve at operating points. Feature leakage mechanisms can be expressed using tests. Several research papers have pointed out the critical role of ‘time travel tests’ wherein codes related to features are run with ‘as-of time’ and compared with a version that has foresight into the future based on the features; discrepancies reveal feature leakage. Feature lineage is also a good practice wherein features retain their source tables, transformations, and time bounds. Whilst not

all research papers emphasize these mechanisms, systematic reviews have highlighted that feature leakage is one of the prime reasons for exaggeration in outcomes and should be considered a key concern in research (Ali et al., 2022; Hernandez Aros et al., 2024).

4.6 Evaluation metrics in practice: why PR-AUC is necessary but not sufficient

One prominent theme in studies on the 2020-2025 time period is the move from accuracy-oriented evaluation to evaluation informed by ranking and budget. This is because fraud is sufficiently rare that accuracy and ROC-AUC are dominated by the positive class prediction problem rather than by fraud prediction. Precision-recall analysis is favored because it keeps the primary focus on minority-class prediction performance. PR-AUC scores the ranking of fraud incidents vs. legitimate incidents. However, it does not define an evaluation point. Schools have to determine the number of alerts to inspect per period and whether to be tough on declines or step up auths. Hence, many studies include measures related to evaluation budget."Recall@K" (or "hit rate @K"): measures the proportion of fraudulent transactions detected given that the top K highest-probability transactions are examined. Exactly equivalent to the daily review budget would be AUC itself, specifically for the top K highest-probability transactions. A useful variation is "precision @K"; this measures an estimate of how many examined fraud cases were actually true fraud cases. Useful studies that give specific measures pertaining to either "recall @K" or "precision @K" will give better results on how the test should be handled with regard to AUC itself. A very useful measure will be "amount weighted recall" since it will give the proportion of the detected amount of fraud examined from the top K highest-probability transactions examined. This will come in handy since loss prevention will be the aim, especially in the scenario wherein the fraud amount actually follows a "heavy-tailed" distribution. Cost-conscious studies may also compute "expected loss avoided" based on the assumption that "actions" such as "decline" and "review" differ both in their "cost" and "benefits."Temporal validation is very important to ensure the reliability of metrics. In fraud analysis, it may be the case that there is a drift in the distribution of characteristics and labels. Random splitting may result in the model predicting “future” trends and developing overly optimistic PR-AUC values, which cannot be repeated. Temporal-aware evaluation methods split based on dates or a rolling window and periodically reevaluate to indicate the trend of degradation as well as the

impact of retraining processes (Shahapurkar & Patil in 2023). ORT B marks studies as better if they measure not only one aspect but also trend reliability for a given set of time windows and segments.

4.7 Threshold selection & decision policies: linking model output to actions

A fraud model is implemented under a decisioning policy. The policy is a mapping of scores to decisions of decline, step-up authentication, manual review, or allow. It is postulated that decisions on setting thresholds are more influential than slight modifications to the model design. The key areas are capacity constraints: there is a limited amount of reviews that can be done by the review teams; also, capacity is limited after step-up procedures have been employed to avoid customer friction.

One of the most frequent suggestions is that threshold selection should be modeled as an optimization with operationally constrained budgets. For instance, the maximum fraud loss identified might be optimized subject to the maximum number of alerts within the institution daily, or the maximum expected utility for which the cost of declines and reviews differ. Calibration is another key aspect. When the predictive probabilities are calibrated, the threshold can be stated as the type of risk of fraud instead of the actual score. It makes the behavior less sensitive to changes in the distribution. Nevertheless, the lack of adequate representation of the issue of calibration within the reported number of studies is cited in systematic reviews as the reason for the lack of transfer of the work (Hernandez Aros et al., 2024). Research has further observed that there might be differing requirements for explanations depending on the actions being taken. A decision to decline would ideally involve a high degree of confidence and a rigorous explanation, whereas a review decision could be full of uncertainty because a human would follow up. These translate into policies with dual or multiple thresholds – high risk for a decline, mid risk for a review, and low risk for allows. These policies correspond with the concept that fraud detection isn’t a two-choice problem but a risk management process.

4.8 Explainability to help investigations

More than feature contribution:Explainability methods have been shown to be generally applicable, although it is important to assess whether these explanations are

helpful to the investigative decision-making process as a whole. While generic explanations for overall feature impact may assist model developers, individual case-level explanation is required for investigators. These types of explanations could be provided by answering questions such as: (i) what drove it from the baseline? (ii) What features led to the score? (iii) if associated entities are suspicious, and (iv) what checks are recommended? This is reinforced by the need to present assorted explanation types, rather than a single explanation form, as discussed by Psychoula et al., (2021), where a variety may be accomplished via a blend of explanations concerning feature attribution, relationship statements, and a counterfactual explanation.

Some studies also address the importance of explanation stability. It would not make sense to find large disparities in explanations based on small changes to input data points. This would make people lack confidence in the researchers. Hence, there appears to be a need to measure the explanation tools based on their stability in terms of time and segments. This has never been performed in academic studies, so another study direction could be the standard measurement of explanation use and stability. Explanation artifacts can be used in governance.

4.9 Privacy-preserving learning and collaboration

Fraud detection involves wide pattern sharing. Cross-institution learning may aid in the early discovery of novel patterns of fraud, but financial transaction data is highly private information. The application of federated learning is a novel method that has recently been explored in collaborative model training with sensitive data. Several works, such as the explainable machine learning in finance literature reviews by Aljunaid et al. (2025), indicate that methods to keep understandability in federated learning are under investigation to ensure privacy-preserving model updation. Nevertheless, there are difficulties in federated learning in the banking domain; there are disparities in the data sets of institutions, there may be expensive communications, and inference attacks and/or poisoning are a concern for privacy breaches.

4.10 MLOps and lifecycle management for fraud detection

Fraud systems are not ‘train once and deploy.’ One of the emerging patterns with regard to 2024 practice and discussion on fraud management is the growing need for responsible AI management (Reuters, 2024). "Model management" with regard to fraud detection means monitoring for model drift, tracking numbers of alerts, tracking feedback from fraud investigators, and periodic retraining on a rolling window. Models may look good on overall performance but be a source of "operation pain" if their threshold is causing fluctuation in the number of alerts. Both machine learning measurement standards and operating metrics are necessary: number of daily alerts, number of alerts to confirmed fraud, FPR by segment, and economic loss via top K alerts.

Lifecycle management must also provide model version and auditing. When a customer disputes a chargeback or an account freeze, the institutions must be able to say which model and policy were used to make the determination. Good governance would cover versioned features, versioned models, specified threshold policies, and recording overrides. This corresponds with ORT C and in the current literature, which increasingly stresses the critical role of explainability and governance in designing fraud systems. Furthermore, MLOps relates directly to fairness and impact on customers because monitoring must track disparity in error rates across customers, which would not be amplified in reinforcement training.

4.11 Limitations and implications for future reviews

The current analysis is hampered by the lack of open datasets with time-stamped information and the fact that many production systems are considered confidential. Therefore, some of the current operational system practices are summarized at a high level of detail, while performance differences across publications cannot be combined into one numerical meta-analysis across papers. In future analysis work, more comparable information can be harvested for time split information, label delay settings, alert budgets, and calibration strategies for more powerful findings across studies.

5 CONCLUSION

Fraud detection in financial transactions is changing fast and needs new technology for reliable results. Research from 2020 to 2025 shows that proven models like gradient-boosted decision trees with well-designed, time-aware features still work well for tabular data. Deep sequence models are useful when there is a lot of behavioral history, but they need to be tested in real deployments, especially when labels are delayed or data changes over time. Network-based models and relational features are becoming more important for finding organized fraud rings, but they need strong data systems, good security, and clear evidence for investigators. Recent progress comes from better models and more realistic testing, such as temporal validation, preventing data leaks, calibrating models, and using alert-budget metrics like recall@K and amount-weighted recall. These metrics help make sure detection matches what analysts can handle and supports loss prevention. Explainability is moving from general feature importance to case-level 'investigation packets' that help with human review, audits, and setting thresholds. Privacy-focused collaboration, especially with federated learning, is starting to help institutions work together, but challenges like data drift and security risks need careful management. In summary, future fraud prevention platforms will have strong data pipelines, adapt to data drift, and follow policy, so model outputs support clear decisions that reduce loss while balancing customer experience and investigator workload.

REFERENCES

1. Aljunaid, H., Moqurrab, S. A., Iqbal, M. S., et al. (2025). Explainable federated learning in finance: A systematic review. IEEE Access, 13, 20203–20231.

2. Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2022). Classification with class imbalance problem: A review. International Journal of Advances in Soft Computing and Its Applications, 14(1), 1–30.

3. Aromataris, E., & Munn, Z. (Eds.). (2020). JBI manual for evidence synthesis. JBI.

4. Hashemi, E., Salari, S., & Amini, A. (2023). Fraud detection in banking data by machine learning techniques. Applied Sciences, 13(2), 872.

5. Hernandez Aros, K., Wozniak, M., & Szelag, M. (2024). Credit card fraud detection using machine learning: Recent advances and challenges. Expert Systems with Applications, 246, 123073.

6. Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71.

7. Psychoula, I., Chen, L., & Chen, F. (2021). Explainable machine learning for credit card fraud detection. arXiv preprint arXiv:2109.02506.

8. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., et al. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic Reviews, 10, 39.

9. Shahapurkar, A., & Patil, P. (2023). Concept drift detection methods for fraud detection: A review. Journal of Big Data, 10, 90.

10. Reuters. (2024). Regulators push for stronger governance and transparency in AI used by financial services. Reuters.

11. Alhasawi, Y., Almtrf, A. A., & Asad, M. (2025). A federated approach to scalable and trustworthy financial fraud detection. Security and Privacy, 8(5), e70099. https://doi.org/10.1002/spy2.70099

12. Aljunaid, H., Moqurrab, S. A., Iqbal, M. S., et al. (2025). Secure and transparent banking: Explainable AI-driven federated learning model for financial fraud detection. Journal of Risk and Financial Management, 18(4), 179. https://doi.org/10.3390/jrfm18040179

13. Chen, Y., Zhao, C., Xu, Y., Nie, C., & Zhang, Y. (2025). Deep learning in financial fraud detection: Innovations, challenges, and applications. Data Science and Management. https://doi.org/10.1016/j.dsm.2025.08.002

14. Chen, Y., et al. (2025). A novel federated transfer learning framework for credit card fraud detection (FED-SPFD). Journal of Risk and Financial Management, 13(11), 208. https://doi.org/10.3390/jrfm13110208

15. Wu, Y., et al. (2025). A deep learning method of credit card fraud detection inspired by brain-like computing: Continuous-Coupled Neural Network (CCNN). Mathematics, 13(5), 819. https://doi.org/10.3390/math13050819

16. AbouGrad, H., & Sankuru, L. (2025). Online banking fraud detection model: Decentralized machine learning framework to enhance effectiveness and observance of data privacy regulations. Mathematics, 13(13), 2110. https://doi.org/10.3390/math13132110

17. Li, M., & Walsh, J. (2024). FedGAT-DCNN: Advanced credit card fraud detection using federated learning with graph attention networks and deep convolutional neural networks. Electronics, 13(16), 3169. https://doi.org/10.3390/electronics13163169

18. Baisholan, N., Dietz, J. E., Gnatyuk, S., Turdalyuly, M., Matson, E. T., & Baisholanova, K. (2025). FraudX AI: An interpretable machine learning framework for credit card fraud detection on imbalanced datasets. Computers, 14(4), 120. https://doi.org/10.3390/computers14040120

19. Moura, L., et al. (2025). AI and financial fraud prevention: Mapping the trends and future directions. Journal of Risk and Financial Management, 18(6), 323. https://doi.org/10.3390/jrfm18060323

20. Tang, Y., Li, S., Fang, Z., & Sun, J. (2024). Credit card fraud detection based on federated graph learning. Expert Systems with Applications, 256, 124979. https://doi.org/10.1016/j.eswa.2024.124979

21. Benchaji, I., Douzi, S., El Ouahidi, B., & Jaafari, J. (2021). Enhanced credit card fraud detection based on attention mechanism and LSTM deep model. Journal of Big Data, 8, 151. https://doi.org/10.1186/s40537-021-00541-8

22. Suárez-Cetrulo, A. L., et al. (2023). A survey on machine learning for recurring concept drifting data streams. Expert Systems with Applications, 230, 121031. https://doi.org/10.1016/j.eswa.2023.121031

23. Arora, S., Rani, R., & Saxena, N. (2024). A systematic review on detection and change of concept drift in streaming data using machine learning techniques. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(4), e1536. https://doi.org/10.1002/widm.1536

24. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

25. National Institute of Standards and Technology. (2020). Security and privacy controls for information systems and organizations (NIST Special Publication 800-53 Rev. 5). https://doi.org/10.6028/NIST.SP.800-53r5

26. ISO/IEC. (2020). Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence (ISO/IEC TR 24028:2020). International Organization for Standardization.

27. ISO/IEC. (2023). Information technology — Artificial intelligence — Guidance on risk management (ISO/IEC 23894:2023). International Organization for Standardization.

28. ISO/IEC. (2023). Information technology — Artificial intelligence — Artificial intelligence management system (ISO/IEC 42001:2023). International Organization for Standardization.

29. Zhang, J., Guo, S., Qu, Z., Zeng, D., Wang, H., Liu, Q., & Zomaya, A. Y. (2022). Adaptive vertical federated learning on unbalanced features. IEEE Transactions on Parallel and Distributed Systems, 33, 4006–4018.

30. Basel Committee on Banking Supervision. (2021). Principles for operational dependability. Bank for International Settlements.

Authors' Contribution

All authors contributed equally to the development of this article.

Data availability

All datasets relevant to this study's findings are fully available within the article.

How to cite this article (APA)

Naqvi, S. H. FRAUD DETECTION USING MACHINE LEARNING IN FINANCIAL TRANSACTIONS: A SYSTEMATIC REVIEW. Veredas Do Direito, e234187. https://doi.org/10.18623/rvd.v23.n2.4187