Financial Technology and Innovation

♦ ISSN: 2836-5372 Vol: 05, No. 01, 2025 ♦ PP. 01–09

Article DOI: https://doi.org/10.54216/FinTech-I.050101

REVIEW ARTICLE

Explainable Artificial Intelligence for Real-Time Financial Fraud Detection: A Systematic Literature Review

Ulugbek Inoyatov[1,*] **Eugene Q. Castro**[1]

[1] Department of Computer Science, Central Asian University, Tashkent, Uzbekistan

Emails: 220407@centralasian.uz · e.castro@centralasian.uz

Received: February 02, 2025 Revised: April 04, 2025 Accepted: July 05, 2025 ⋆ Corresponding author

ABSTRACT

Financial fraud detection systems increasingly rely on machine learning to identify suspicious transactions at scale. However, the opacity of many high-performing models raises significant concerns regarding trust, regulatory compliance, and practical deployment in real-time financial environments. Explainable Artificial Intelligence (XAI) has emerged as a promising solution to enhance transparency and accountability, yet its feasibility under real-time constraints remains unclear. This systematic literature review examines empirical studies on explainable AI approaches for financial fraud detection, with explicit focus on real-time applicability. Following PRISMA guidelines, nineteen peer-reviewed empirical studies were selected and analyzed based on fraud domain, model type, explainability technique, evaluation metrics, and evidence of real-time performance. Results show that post-hoc explanation methods, particularly SHAP and LIME, dominate the literature, while intrinsic explainability and deployment-level latency reporting remain limited. Despite frequent claims of real-time applicability, only one study provides quantitative runtime evidence. The findings highlight critical gaps: absence of explanation latency evaluation, lack of deployment-oriented validation, and insufficient regulatory compliance integration. This review reveals a systematic disconnect between real-time claims and empirical evidence, establishing the need for standardized latency benchmarking in explainable fraud detection research.

Keywords: Explainable AI ▪ Financial Fraud Detection ▪ Systematic Literature Review ▪ PRISMA ▪ Real-Time Systems ▪ Model Interpretability

1. INTRODUCTION

Financial fraud poses a persistent and evolving threat to banking institutions, payment systems, insurance providers, and digital marketplaces. As transaction volumes grow and fraud strategies become increasingly sophisticated, machine learning (ML) techniques have become central to automated fraud detection systems. These models are capable of identifying complex patterns in large-scale transactional data and have demonstrated strong predictive performance across various fraud domains, including credit card fraud, insurance fraud, and online payment fraud.

Despite these advances, the deployment of machine learning models in financial decision-making contexts introduces critical challenges. Many high-performing models, particularly ensemble methods and deep learning architectures, operate as black boxes, offering limited transparency into how predictions are generated. In regulated financial environments, such opacity is problematic. Financial institutions are required to justify automated decisions to regulators, auditors, and customers, while fraud analysts must understand model outputs to validate alerts and minimize false positives. As a result, accuracy alone is insufficient; explainability has become a fundamental requirement for trustworthy fraud detection sys-

tems.

Explainable Artificial Intelligence (XAI) aims to address this challenge by providing human-interpretable explanations of model behavior and individual predictions. Techniques such as SHAP, LIME, feature importance analysis, and intrinsically interpretable models have been widely proposed to enhance transparency in fraud detection. Prior studies suggest that explainability can improve user trust, facilitate regulatory compliance, and support analyst decision-making. However, many explainability methods introduce additional computational overhead, raising concerns about their suitability for real-time fraud detection environments, where decisions must often be made within milliseconds.

Although the literature increasingly emphasizes the importance of "real-time" explainable fraud detection, it remains unclear to what extent existing empirical studies evaluate real-time feasibility in practice. Many works claim real-time applicability without reporting inference latency, explanation generation time, or deployment constraints. This gap complicates the assessment of whether current XAI techniques can realistically operate in production-level fraud detection systems.

Identified Research Gaps:

- **Gap 1:** Lack of quantitative explanation latency evaluation in claimed "real-time" systems
- **Gap 2:** Absence of deployment-ready validation under production constraints
- **Gap 3:** Limited human-centered evaluation with domain experts (fraud analysts, auditors)
- **Gap 4:** Insufficient integration of regulatory compliance requirements in XAI design

To address these gaps, this paper conducts a systematic literature review of empirical studies on explainable AI applied to financial fraud detection, with particular attention to real-time considerations.

Review Objectives:

- Identify dominant machine learning and explainability techniques in empirical fraud detection studies
- Examine how explainability methods are evaluated and integrated into fraud detection systems
- Assess the extent to which real-time feasibility and computational efficiency are empirically validated
- Synthesize limitations and research gaps to inform future work on trustworthy real-time explainable fraud detection

By synthesizing existing evidence and highlighting key limitations, this review aims to clarify the current state of research and inform future work on trustworthy and real-time explainable fraud detection systems.

2. BACKGROUND AND RELATED WORK

2.1 Financial Fraud Detection Using Machine Learning

Machine learning-based fraud detection systems typically frame fraud identification as a binary or probabilistic classification problem, where transactions are labeled as fraudulent or legitimate. Commonly used models include logistic regression, decision trees, random forests, gradient boosting machines, and, more recently, deep learning and graph neural networks. These models are often trained on highly imbalanced datasets, where fraudulent transactions represent a small minority of observations, necessitating specialized evaluation metrics such as precision-recall curves and cost-sensitive measures.

While advanced models can achieve high predictive performance, their complexity often limits interpretability. This trade-off between accuracy and transparency has become a central concern in financial applications, where automated decisions may have legal, financial, and ethical consequences.

2.2 Explainable Artificial Intelligence in Finance

Explainable AI encompasses a range of techniques designed to make machine learning models more transparent and understandable to humans. Broadly, XAI approaches can be categorized into post-hoc explanation methods and intrinsic interpretability methods. Post-hoc techniques, such as SHAP and LIME, generate explanations after a model has produced predictions, without altering the underlying model. Intrinsic methods, by contrast, rely on inherently interpretable models or architectures that embed explainability directly into the learning process.

In financial contexts, explainability is closely linked to regulatory compliance, auditability, and user trust. Regulators increasingly require institutions to justify automated decisions, while fraud analysts rely on explanations to prioritize alerts and reduce false positives. Consequently, XAI has gained significant attention in recent years as a mechanism to bridge the gap between model performance and accountability.

2.3 Limitations of Existing Reviews

Several prior reviews have examined the application of explainable AI in finance more broadly. However, many of these reviews focus on conceptual frameworks, surveys of techniques, or non-empirical analyses. Moreover, few reviews explicitly assess the real-time feasibility of explainability methods in fraud detection systems. In particular, the lack of standardized reporting on runtime, latency, and deployment constraints limits the practical applicability of existing findings.

This SLR differentiates itself by focusing exclusively on empirical studies of explainable AI for financial fraud detection and by explicitly analyzing the presence or absence of real-time evidence. By adopting a PRISMA-based methodology and systematically extracting data related to models, explainability techniques, evaluation metrics, and computational considerations, this review provides a structured and practice-oriented synthesis of the literature.

3. METHODOLOGY

This study follows a Systematic Literature Review (SLR) methodology in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The review process was designed to be transparent, reproducible, and verifiable, as required for academic research in scientific communication courses. The overall methodology consists of protocol definition, literature search, study selection, quality assessment, and data extraction.

3.1 Research Questions

The review is guided by the following research questions:

RQ1: What machine learning models and explainable AI techniques are empirically used for financial fraud detection?

RQ2: How are explainability methods evaluated and integrated into fraud detection systems?

RQ3: To what extent do empirical studies provide evidence of real-time applicability, including runtime or latency considerations?

RQ4: What limitations and research gaps are identified in existing explainable fraud detection studies?

3.2 Review Protocol

A review protocol was defined prior to conducting the literature search to minimize selection bias and ensure methodological rigor. The protocol specifies the databases to be searched, search strings, inclusion and exclusion criteria, screening stages, and data extraction strategy. All steps were executed sequentially and documented through search logs and a PRISMA flow diagram.

3.3 Literature Search Strategy

3.3.1 Databases

The literature search was conducted using three scholarly databases: IEEE Xplore, ACM Digital Library, and ScienceDirect (Elsevier). These databases were selected due to their comprehensive coverage of peer-reviewed research in machine learning, artificial intelligence, and financial technologies.

3.3.2 Search Strings

The search strings were constructed by combining keywords related to fraud detection, explainable AI, and financial systems. Boolean operators were used to ensure comprehensive coverage. The exact search strings used for each database are:

IEEE Xplore:

```
("explainable AI" OR XAI) AND ("fraud detection" OR
"financial fraud") AND (SHAP OR LIME OR GNN)
```

ACM Digital Library:

```
("explainable AI") AND ("fraud detection") AND
(SHAP OR LIME OR GNN)
```

ScienceDirect:

```
("explainable AI") AND ("fraud detection")
```

3.3.3 Search Dates and Filters

- Search period: January 2020 – March 2025
- Language: English
- Access type: Open-access articles (where applicable)
- Document type: Journal articles and conference proceedings

The decision to focus on open-access literature was made to ensure full-text availability and reproducibility. This constraint is acknowledged as a limitation in Section 6.

3.3.4 Search Log

Table 1 provides the exact search strings, filters, dates, and result counts from each database.

3.4 Inclusion and Exclusion Criteria

Clear inclusion and exclusion criteria were defined and applied consistently throughout the screening process.

Inclusion Criteria:

- Peer-reviewed journal article or conference paper
- Empirical study with experiments on real or synthetic datasets
- Focused on financial fraud detection
- Applied at least one explainable AI or interpretable modeling technique
- Reported quantitative evaluation metrics

Exclusion Criteria:

- Review papers, surveys, editorials, or conceptual frameworks
- Non-financial fraud domains
- No empirical evaluation
- No explainability or interpretability component
- Duplicate publications

Note: Runtime or latency reporting was not used as an exclusion criterion during screening, as the objective was to assess the extent to which empirical XAI fraud detection studies address real-time considerations. This methodological choice allows the review to identify and quantify the gap in runtime reporting across the included literature.

3.5 Study Selection Process

The study selection process was conducted in three stages:

1. Title Screening: Obvious irrelevance to fraud detection or explainable AI was removed.

2. Abstract Screening: Studies that did not meet inclusion criteria based on abstract content were excluded.

3. Full-Text Screening: Remaining articles were read in full to verify empirical contribution, explainability usage, and relevance.

All screening decisions were logged, and reasons for exclusion were documented.

Table 1. Search Log – Verifiable and Reproducible

Database	Exact Search String	Filters Applied	Search Date	Results
IEEE Xplore	("explainable AI" OR XAI) AND ("fraud detection" OR "financial fraud") AND (SHAP OR LIME OR GNN)	2020–2025, English, Open Access	11/05/2025	8
ACM Digital Library	("explainable AI") AND ("fraud detection") AND (SHAP OR LIME OR GNN)	2020–2025, Research Articles	11/05/2025	16
ScienceDirect	("explainable AI") AND ("fraud detection")	2020–2025, Research articles, Open Access	11/05/2025	131

3.6 PRISMA Flow Summary

The initial database search returned 155 records across all three databases. After deduplication, 54 duplicate records were removed. An additional 20 records were excluded for other predefined reasons (non-English language, inaccessible full text, or preliminary conference versions superseded by journal publications), leaving 81 unique records for title and abstract screening. Following title and abstract screening, 51 full-text articles were assessed for eligibility. Based on the inclusion and exclusion criteria, 19 empirical studies were selected for final analysis. The complete PRISMA flow is shown in Figure 1.

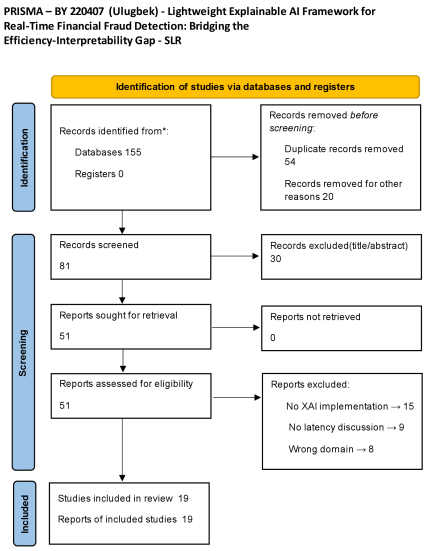

Figure 1. PRISMA flow diagram for study selection.

3.7 Quality Assessment

Each study was assessed using a quality checklist adapted from established SLR guidelines. The checklist evaluated:

- Peer-reviewed venue clearly identified
- Dataset described (name/source and type)
- Empirical evaluation conducted
- Model(s) explicitly specified
- XAI method(s) explicitly specified
- Evaluation metrics reported
- Limitations/future work stated
- Runtime/latency/compute evidence stated (if present)

Studies were scored on a 10-point scale. The quality scores for the 19 included studies ranged from 4 to 9, with a median score of 7, indicating generally high-quality empirical research.

3.8 Data Extraction Strategy

A structured data extraction form was designed to ensure consistency across studies. For each included paper, the following information was extracted:

- Publication year and venue
- Fraud domain (e.g., credit card, insurance, blockchain)
- Dataset type and size
- Machine learning model(s) used
- Explainable AI technique(s) applied
- Evaluation metrics reported
- Evidence of real-time applicability or runtime reporting
- Stated limitations or future research directions

Note that runtime or latency reporting was not used as an exclusion criterion during study selection; rather, it was treated as an analytical dimension for post-inclusion assessment. This approach allows the review to identify and quantify the gap in real-time evidence reporting across included empirical studies.

This extraction schema was applied uniformly to all included studies and recorded in a master extraction table.

3.9 Data Analysis

The extracted data were analyzed using thematic synthesis, identifying patterns across studies related to XAI techniques, model architectures, evaluation practices, and real-time considerations. Key themes include: (1) dominant XAI techniques in financial fraud detection, (2) real-time performance and latency optimization, (3) privacy-preserving explainable AI, and (4) research gaps and challenges.

4. RESULTS

This section presents the findings of the systematic literature review based on the 19 empirical studies selected through the PRISMA process. The results are organized to address the research questions by summarizing study characteristics, fraud domains, machine learning models, explainable AI techniques, evaluation metrics, and evidence of real-time applicability.

4.1 Overview of Included Studies

The final dataset consists of 19 peer-reviewed empirical studies published between 2020 and 2025. The studies include both journal articles and conference proceedings, reflecting a mix of applied and research-oriented contributions. The publication venues primarily include IEEE Access, Finance Research Letters, Journal of Risk and Financial Management, ACM conference proceedings (CIKM, HCII, CAiSE), and related peer-reviewed outlets. The selected studies cover a range of financial fraud domains and modeling approaches, with varying degrees of emphasis on explainability and deployment considerations.

4.2 Fraud Domains Addressed

The included studies focus on multiple fraud detection contexts. Credit card and payment fraud represent the most frequently studied domain with 3 dedicated studies, while banking fraud is also examined in 3 studies. Additional domains include automobile insurance fraud (1 study), e-commerce transaction fraud (1 study), blockchain/cryptocurrency fraud (1 study), financial statement fraud (2 studies), and regulatory compliance contexts (1 study). Several review papers address multiple fraud domains simultaneously. Overall, credit card and payment fraud dominate the empirical literature, while insurance, blockchain, and compliance-oriented fraud remain less frequently studied but are gaining attention. Table 2 presents the complete distribution of studies across fraud domains.

Table 2. Study Distribution by Fraud Domain (n=19)

Domain	Count
Credit card / payment fraud	3
Banking fraud	3
Insurance fraud	1
E-commerce / marketplace fraud	1
Blockchain / cryptocurrency fraud	1
Financial statement fraud	2
Compliance / audit-related	1
Multi-domain (review papers)	7

4.3 Machine Learning Models Used

A wide variety of machine learning models are employed across the reviewed studies. Traditional and linear models such as logistic regression and ridge regression are primarily valued for interpretability and baseline comparisons. Tree-based and ensemble models including random forests, gradient boosting machines, XGBoost, LightGBM, CatBoost, and stacking ensembles are the most commonly used due to their strong performance on tabular financial data. Deep learning models including deep neural networks and convolutional neural networks appear particularly in blockchain and federated learning settings. Graph-based models such as Graph Neural Networks (GNNs), including architectures like GCN, GAT, and specialized designs for fraud detection on transaction graphs, are used primarily in large-scale transaction networks and e-commerce platforms.

4.4 Explainable AI Techniques Applied

Explainability techniques used in the reviewed studies fall into two main categories: post-hoc explainability and intrinsic explainability.

Post-hoc Explainability: The majority of studies rely on post-hoc explanation methods. SHAP (Shapley Additive Explanations) is the most frequently used method, appearing in 9 studies and applied for both local and global explanations. LIME is used in 5 studies to provide instance-level explanations, often in combination with SHAP. Feature importance and permutation-based methods appear in 3 studies to rank feature contributions at a global level. Partial Dependence Plots (PDP) are used in 1 study to analyze the marginal effect of individual features. These techniques are commonly applied to ensemble and deep learning models to improve interpretability without altering the underlying model.

Intrinsic Explainability: A smaller subset of studies employs intrinsically interpretable approaches, including interpretable tree-based models where model structure itself provides transparency, architectural explainability in graph models such as two-stage or time-aware graph neural networks designed to separate batch and real-time inference, and self-explainable graph models where explanation emerges from learned graph structures or meta-graph mechanisms. Intrinsic explainability approaches are less common but are often motivated by efficiency and deployment considerations. Table 3 summarizes the frequency of model families and XAI methods across all reviewed studies.

Table 3. Most Common Model Families and XAI Methods

Model family (count)	XAI method (count)
Tree-based ensembles (12)	SHAP (9)
Linear models (5)	LIME (5)
Deep learning (4)	Feature importance (3)
Graph neural networks (5)	PDP (1)
	Intrinsic (1)

4.5 Evaluation Metrics

Across the included studies, evaluation focuses primarily on predictive performance. The most commonly reported metrics include Accuracy (8 studies), Precision (8 studies), Recall/Sensitivity (8 studies), F1-score (7 studies), Area Under the ROC Curve (AUC-ROC) (9 studies), and Precision-Recall AUC (AUPR) in imbalanced settings (1 study). While these metrics assess classification performance, quantitative evaluation of explanation quality is rare. Only a limited number of studies report metrics related to explanation fidelity, coverage, or human-centered evaluation. Table 4 summarizes the most common evaluation metrics used across the reviewed studies.

Table 4. Common Evaluation Metrics in Included Studies

Metric	Count
AUC-ROC	9
Precision	8
Recall / Sensitivity	8
Accuracy	8
F1-score	7
AUPR / PR-AUC	1

4.6 Evidence of Real-Time Applicability

Real-time applicability is frequently claimed across the reviewed studies; however, the nature of supporting evidence varies significantly. Only one study reports explicit quantitative inference latency measurements, specifically demonstrating end-to-end inference latency at the 99th percentile (P99) with sub-100ms constraints using a lightweight Graph Neural Network architecture (BRIGHT). Several studies propose architectural designs intended to support online screening, such as separating batch and real-time inference stages, but do not provide measurable runtime or deployment benchmarks. Many studies describe their systems as "real-time" or "online" without providing concrete runtime data. A substantial number of empirical studies (14 out of 19) do not address runtime, latency, or deployment constraints at all. For studies lacking runtime data, the absence of latency reporting was explicitly recorded as "not reported" during data extraction. Table 5 categorizes the type of runtime evidence provided across all reviewed studies.

Table 5. Runtime Evidence Reporting Across Studies

Runtime evidence category	Count
Explicit inference latency (ms reported)	1
Training time only	0
Qualitative/implicit ("real-time" claim only)	4
None reported	14

5. DISCUSSION

This section interprets the results presented in Section 4 in relation to the research questions and synthesizes findings across the included empirical studies.

5.1 Dominant Modeling and Explainability Practices (RQ1)

The reviewed literature demonstrates a strong preference for tree-based ensemble models, such as XGBoost, Random Forest, and gradient boosting variants, for financial fraud detection. These models are consistently selected due to their strong performance on tabular and imbalanced datasets, which are characteristic of financial transaction data. Deep learning and graph-based models appear less frequently but are increasingly adopted in large-scale transaction networks and e-commerce settings where relational information between entities (such as users, merchants, and accounts) provides valuable signal for fraud detection.

With respect to explainability, post-hoc explanation methods dominate empirical practice, particularly SHAP and LIME. SHAP appears in 9 of the 19 studies, making it the most prevalent XAI technique. LIME is used in 5 studies, often complementing SHAP to provide instance-level explanations. These techniques are favored for their model-agnostic nature and ease of integration with high-performing black-box models. However, reliance on post-hoc explanations reflects a broader trend in which explainability is treated as an auxiliary layer rather than a core system component. Intrinsically interpretable models and self-explainable architectures remain underrepresented, despite their potential advantages in efficiency and transparency. Only one study employs intrinsic explainability through meta-graph architectures in Graph Neural Networks, suggesting that intrinsic approaches remain an area for future exploration.

5.2 Evaluation Focus and Explainability Assessment (RQ2)

Across the reviewed studies, evaluation primarily emphasizes predictive performance, using metrics such as AUC-ROC (9 studies), Precision (8 studies), Recall (8 studies), Accuracy (8 studies), and F1-score (7 studies). While these metrics are appropriate for assessing fraud detection accuracy, they provide limited insight into the effectiveness or usefulness of explanations. Only a small subset of studies evaluates explanation fidelity, coverage, or alignment with human decision-making. One study explicitly examines the Average Prediction Switching Point (ASP) to measure explanation quality, but such quantitative explainability metrics remain rare.

Furthermore, human-centered evaluation—such as user studies involving fraud analysts, auditors, or compliance officers—is extremely limited. Only one study in the review conducted a prototype evaluation with fraud detection experts from a banking partner to assess the usability of local and global feature importance explanations. As a result, it remains unclear how explanations are interpreted or acted upon in real operational settings. This gap suggests that explainability is often validated theoretically or visually through plots and feature rankings, rather than empirically assessed in real decision workflows. Future research should prioritize user validation studies to determine whether XAI outputs genuinely support analyst decision-making, reduce alert fatigue, and improve trust in automated fraud detection systems.

5.3 Real-Time Feasibility and Deployment Readiness (RQ3)

A central finding of this review is the disconnect between real-time claims and empirical validation. Although many studies describe their systems as "real-time" or suitable for online fraud detection, only one study provides quantitative evidence of deployment-ready performance. Specifically, the BRIGHT study (Lu et al., 2022) reports end-to-end inference latency at the 99th percentile (P99), demonstrating feasibility under sub-100ms constraints through a lightweight Graph Neural Network architecture that separates batch feature extraction from real-time scoring. This study reduces P99 latency by over 75% compared to baseline GNN approaches while maintaining strong predictive performance.

Other works propose architectural designs intended to support online screening—such as federated learning frameworks that distribute computation, stacking ensembles optimized for efficiency, or two-stage GNN pipelines—but stop short of reporting concrete runtime metrics such as inference time per transaction, explanation generation overhead, or end-to-

end system latency. In many cases, real-time feasibility is implied through qualitative claims (e.g., "suitable for online environments") rather than demonstrated through quantitative benchmarks.

This lack of standardized reporting makes it difficult to assess whether current explainability techniques can realistically operate under strict time constraints. Post-hoc methods such as SHAP, while informative, are known to introduce significant computational overhead due to their reliance on repeated model evaluations or coalition sampling to compute Shapley values. LIME similarly requires multiple perturbations and local model fitting. Without explicit latency measurements, it is unclear whether these techniques can meet the millisecond-level response requirements of production fraud detection systems, particularly at high transaction volumes. Future work should adopt standardized latency benchmarking practices and report both model inference time and explanation generation time separately to clarify deployment feasibility.

5.4 Identified Research Gaps (RQ4)

Based on cross-study synthesis, several key research gaps emerge:

1. Lack of explanation latency evaluation: Most studies fail to report the computational cost of generating explanations, despite frequent real-time claims. This gap prevents accurate assessment of whether XAI techniques can operate at scale in production environments.

2. Limited deployment-oriented evaluation: Empirical validation often occurs in offline or simulated environments, with minimal consideration of production constraints such as distributed system architectures, concurrent user requests, or integration with existing fraud monitoring infrastructure.

3. Insufficient human-centered validation: Few studies empirically assess how explanations support analysts, auditors, or regulators in practice. User studies involving domain experts are needed to evaluate whether XAI outputs improve decision quality, reduce false positives, or enhance trust.

4. Overreliance on post-hoc methods: Intrinsic and architecturally explainable models remain underexplored relative to post-hoc techniques. Intrinsic approaches may offer computational advantages and more faithful explanations but are less commonly adopted.

5. Weak alignment with regulatory requirements: Although compliance and transparency are frequently cited motivations, concrete regulatory evaluation remains rare. Studies rarely discuss how explanations meet specific regulatory standards such as GDPR's "right to explanation" or financial sector guidelines on algorithmic accountability.

6. Class imbalance effects on explanation quality: While many fraud datasets exhibit extreme class imbalance (often <1% fraud rate), few studies investigate how class imbalance affects explanation stability, fidelity, or interpretability. It remains unclear whether explanations remain reliable across majority and minority class predictions.

Addressing these gaps is critical for advancing explainable fraud detection from experimental prototypes to trustworthy real-world systems that meet both operational and regulatory requirements.

6. LIMITATIONS

This review has several limitations that should be acknowledged. First, the search was restricted to open-access articles from three academic databases (IEEE Xplore, ACM Digital Library, and ScienceDirect), which may have excluded relevant subscription-based studies or grey literature such as technical reports and industry white papers. Second, the analysis relies on information reported by authors; absence of runtime or deployment metrics does not necessarily imply infeasibility—some studies may have conducted runtime evaluations that were not reported in the published manuscript. Third, the focus on empirical studies excludes conceptual frameworks and theoretical contributions that may offer valuable insights into XAI design principles. Fourth, dataset reuse across studies (e.g., the Kaggle credit card fraud dataset appearing in multiple studies) may limit generalizability of findings. Fifth, as a solo researcher project, the screening and data extraction process lacked independent dual review, which could introduce subjective bias despite systematic documentation. Finally, the review is limited to English-language publications from 2020-2025, potentially missing earlier foundational work or non-English contributions.

These limitations are acknowledged to maintain transparency and support reproducibility. Future reviews could expand the search scope, include grey literature, and incorporate multiple independent reviewers to enhance robustness.

7. CONCLUSION

This paper presented a systematic literature review of empirical studies on explainable artificial intelligence for financial fraud detection, with a particular focus on real-time applicability. Nineteen peer-reviewed studies were analyzed following PRISMA guidelines, examining fraud domains, machine learning models, explainability techniques, evaluation metrics, and deployment considerations.

The findings indicate that while explainable AI is widely adopted in fraud detection research, explainability is predominantly implemented through post-hoc methods—particularly SHAP and LIME—and rarely evaluated under real-time constraints. Tree-based ensemble models such as XGBoost and Random Forest dominate the modeling landscape, while Graph Neural Networks are increasingly applied in domains involving relational transaction data. Explicit reporting of inference or explanation latency remains uncommon, with only one study providing quantitative P99 latency measurements. Intrinsic explainability and user-centered validation are also underrepresented, limiting understanding of how explanations support operational decision-making in practice.

The review highlights critical gaps in the literature. Most notably, there is a disconnect between frequent claims of real-time suitability and the scarcity of empirical evidence demonstrating deployment-ready performance. Computational overhead of explanation generation is rarely quantified, and human-centered evaluations with fraud analysts or compliance officers are largely absent. Additionally, the impact of class imbalance on explanation quality, adversarial robustness of explanations, and alignment with regulatory frameworks remain underexplored.

To advance the field, future research should integrate ex-

plainability more deeply into model design through intrinsic approaches, systematically evaluate explanation efficiency through standardized latency benchmarks, conduct user studies with domain experts to validate operational utility, and explicitly align XAI techniques with regulatory and compliance requirements. By addressing these gaps, researchers can contribute to the development of trustworthy, transparent, and operationally viable explainable fraud detection systems capable of meeting the demands of real-time financial environments.

The reviewed studies include work on insurance fraud [1], privacy-preserving and federated learning approaches [2, 3], credit card fraud detection [4, 6, 7], real-time GNN frameworks [5], blockchain fraud [8, 9], financial statement fraud [10, 13], ensemble methods [11], meta-graph search [12], user-centric design [15], regulatory compliance [16, 17, 20], and systematic reviews [14, 18, 19].

ETHICS STATEMENT

This study adheres to ethical research and academic integrity standards. All reviewed studies were obtained from peer-reviewed scholarly sources. No data fabrication, plagiarism, or misrepresentation was conducted. Search logs, inclusion/exclusion criteria, and data extraction procedures are documented to support transparency and reproducibility. The review methodology follows PRISMA 2020 guidelines and established systematic review protocols to ensure rigorous and unbiased synthesis of existing evidence. Generative AI tools were used solely as assistive resources for formatting, organization, and language refinement, with all substantive content, analysis, and interpretation produced by the author under full editorial oversight.

REFERENCES

[1] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene, "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection," *Journal of Risk and Insurance*, vol. 69, no. 3, pp. 373–421, 2002.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.

[4] V. Arora, R. S. Leekha, K. Lee, and A. Kataria, "Facilitating User Authorization from Imbalanced Data Logs of Credit Cards Using Artificial Intelligence," *Mobile Information Systems*, 2020.

[5] M. Lu et al., "BRIGHT: Graph Neural Networks in Real-Time Fraud Detection," in *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22)*, 2022.

[6] J. Jurgovsky et al., "Sequence Classification for Credit-Card Fraud Detection," *Expert Systems with Applications*, vol. 100, pp. 234–245, 2018.

[7] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection," *Information Sciences*, vol. 557, pp. 317–331, 2021.

[8] S. Farrugia, J. Ellul, and G. Azzopardi, "Detection of Illicit Accounts over the Ethereum Blockchain," *Expert Systems with Applications*, vol. 150, 2020.

[9] Y. Kang, W. Kim, H. Kim, M. Lee, M. Song, and H. Seo, "Malicious Contract Detection for Blockchain Network Using Lightweight Deep Learning Implemented through Explainable AI," *Electronics*, vol. 12, no. 18, 2023.

[10] P. Fukas, J. Rebstadt, L. Menzel, and O. Thomas, "Towards Explainable Artificial Intelligence in Financial Fraud Detection: Using Shapley Additive Explanations to Explore Feature Importance," in *Advanced Information Systems Engineering (CAiSE 2022)*, Lecture Notes in Computer Science, 2022.

[11] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," *IEEE Access*, vol. 6, pp. 14277–14284, 2018.

[12] Z. Qin, Y. Liu, Q. He, and X. Ao, "Explainable Graph-based Fraud Detection via Neural Meta-graph Search," in *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22)*, 2022.

[13] Y. Zhou, H. Li, Z. Xiao, and J. Qiu, "A user-centered explainable artificial intelligence approach for financial fraud detection," *Finance Research Letters*, vol. 58, 2023.

[14] J. West and M. Bhattacharya, "Intelligent Financial Fraud Detection: A Comprehensive Review," *Computers & Security*, vol. 57, pp. 47–66, 2016.

[15] D. Cirqueira, M. Helfert, and M. Bezbradica, "Towards Design Principles for User-Centric Explainable AI in Fraud Detection," in *Human-Computer Interaction. Design and User Experience (HCII 2021)*, Lecture Notes in Computer Science, 2021.

[16] P. Bracke, A. Datta, C. Jung, and S. Sen, "Machine Learning Explainability in Finance: An Application to Default Risk Analysis," *Bank of England Staff Working Paper*, no. 816, 2019.

[17] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[18] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances," *Expert Systems with Applications*, vol. 193, 2022.

[19] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[20] S. Ahmadi, "Advancing Fraud Detection in Banking: Real-Time Applications of Explainable AI (XAI)," *Journal of Electrical Systems*, 2022.