Explainable AI in Insider Financial Fraud Detection Models: A Review of Transparency and Trust

Hillary Kwame Ofori[1], William Leslie Brown-Acquaye[1], Forgor Lempogo[1], Kwame Bell-Dzide[1] and Israel Edem Agbehadji[1,2]

[1]Ghana Communication Technology University, Accra North, Ghana
[2]Durban University of Technology, Durban, South Africa

hofori@gctu.edu.gh
wbrown@gctu.edu.gh
lforgor@gctu.edu.gh
kbell-dzide@gctu.edu.gh
israeldel2006@gmail.com

Abstract: Financial and insider fraud increasingly intersect with broader cybercrime ecosystems, creating attack vectors that undermine national cyber resilience and the integrity of digital financial infrastructures. As organizations turn to machine learning (ML) and deep learning (DL) models for automated fraud and insider-threat detection, the opacity of these systems presents strategic risks for cyber defense: unexplainable alerts weaken analyst trust, complicate incident response, and challenge regulatory and forensic accountability. This study presents a systematic review of 107 empirically validated works (2015–2025) examining how Explainable Artificial Intelligence (XAI) techniques enhance transparency, trustworthiness, and operational readiness in AI-driven fraud detection systems. Using a mixed bibliometric–thematic methodology, the review maps the evolution of ML/DL architectures, XAI adoption patterns, evaluation practices, and dataset limitations within security-critical environments. The findings highlight a sector-wide dependence on post-hoc feature attribution and reveal emerging shifts toward intrinsic interpretability through attention mechanisms and hybrid temporal models. Despite progress, gaps persist: limited use of sequential behavioral models, narrow evaluation metrics, and overreliance on structured datasets weaken real-world resilience against adaptive adversaries. To address these challenges, the paper proposes a Three-Pillar Framework: Algorithmic Transparency, Evaluation Accountability, and Data Traceability that positions explainability as a foundational architectural property for cyber defense systems. By aligning model interpretability with security operations, regulatory requirements, and analyst cognition, the framework strengthens organizational readiness against insider threats, financial fraud, and AI-targeted adversarial manipulation, key considerations in modern cyber warfare and security operations.

Keywords: Explainable AI, Transparent model, Model interpretability, Financial fraud detection, Insider fraud detection

1. Introduction

Financial fraud, insider threats, and transactional manipulation have become key attack methods in modern cyber warfare and cybercrime networks. As national economies rely more on digital financial infrastructure, adversaries exploit large-scale transactional processes to disrupt operations, steal funds, and undermine public confidence (Tian, 2025; Zhu and Chen, 2024). These developments elevate fraud detection from a financial risk issue to a matter of cybersecurity and national security, as emphasized by regulatory agencies and global security groups (Adetumi et al., 2024; Zhu et al., 2025). In this environment, organizations are increasingly deploying machine learning (ML) and deep learning (DL) models to detect complex, adaptable threats that traditional rule-based systems often miss (Aljunaid et al., 2025; Zhang et al., 2025).

However, the widespread use of opaque ML/DL models introduces new vulnerabilities in cyber defense. Black-box fraud detection systems hide decision logic precisely when analysts and incident responders need actionable intelligence. This tension is well-understood in socio-technical theory, which stresses that technological outputs must remain interpretable to the human defenders responsible for operational responses and risk decisions (Al-hchaimi et al., 2024; Qin et al., 2022; Roshan and Zafar, 2021). In cyber defense environments, where forensic traceability, adversary attribution, and rapid triage are critical, unexplainable model outputs damage trust, slow response coordination, and threaten regulatory compliance.

Explainable Artificial Intelligence (XAI) has become a key tool in cybersecurity, enhancing both transparency and mission preparedness. Foundational work by Gunning and Aha (2019) and the development of Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) methods (Ribeiro, Singh, and Guestrin, 2016) highlight the importance of explainability for making AI-driven predictions auditable, verifiable, and meaningful in operations. Regulations like the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence (EUAI) Act (The European Parliament, 2016, 2024) reinforce these needs through legal requirements for transparency and accountability, especially crucial for

national financial systems targeted by cyber threats. Recent studies also show that transparent Artificial Intelligence (AI) improves security by increasing analyst confidence, reducing false positives, and strengthening human-AI teamwork in high-stakes decision-making (Abbas, Hilal, and Jabbar, 2025; Praveenraj et al., 2023; Vivek et al., 2025).

In parallel, emerging explainability-driven architectures such as eXplainable Security Policy-Induced Data Engineering (X-SPIDE) for smart contract fraud (Pennella, Pinelli, and Galletta, 2025) and hybrid Bidirectional Long Short-Term Memory – Variational Autoencoder (BiLSTM-VAE) models for behavioral anomaly detection (Ofori et al., 2025) reflect a shift toward security-oriented, intrinsically interpretable systems. From the perspective of cognitive fit theory, these approaches better support the mental models of cybersecurity analysts; enabling them to identify anomalous sequences, validate alerts, and anticipate adversarial behavior.

Despite these advancements, significant gaps still exist in cybersecurity research. Many studies on fraud and insider-threat detection continue to focus on predictive accuracy rather than interpretability, making analysts rely on opaque reasoning processes. There is limited evidence on which XAI techniques offer the most stable, reliable, and domain-appropriate explanations in high-stakes cyber defense settings (Vivek et al., 2025). Furthermore, the prevalence of post-hoc explainers like SHAP and LIME often raises concerns about explanation fidelity and their vulnerability to adversarial attacks (Agomuo et al., 2025). The lack of standardized methods and fragmented evaluation practices makes it harder for security operations centers (SOCs), auditors, and regulators to adopt transparent and trustworthy AI systems (Mozolewski, Robek, and Nalepa, 2024).

To address these challenges, this study conducts a systematic review of 107 empirically validated works published between 2015 and 2025 on explainable fraud detection (XFD) within cyber defense contexts. Using bibliometric mapping, thematic analysis, and technical synthesis, the review examines five key areas critical to cyber warfare resilience: (i) the evolution of XFD research, (ii) ML/DL algorithm foundations, (iii) integration of XAI techniques, (iv) evaluation practices related to explanation quality, and (v) the impact of datasets on transparency, generalizability, and operational trust. Based on these insights, the study proposes a Three-Pillar Framework that conceptualizes explainability not just as a technical feature but as a fundamental socio-technical capability for secure, auditable, and resilient AI systems in financial and insider-threat environments.

The following research questions guide the investigation:

- What patterns of scholarly output, co-authorship clustering, and thematic co-occurrence define the intellectual structure of XFD research between 2015 and 2025?
- Which ML/DL models dominate XFD, and how have their usage and hybridization evolved in response to adversarial and cyber warfare–driven requirements?
- What XAI techniques are most frequently employed, and how do they support transparency, trust, and regulatory or forensic accountability in cyber defense?
- Which evaluation metrics and datasets are prioritized, and how do they shape interpretability, fairness, and real-world operational effectiveness of XFD models?

Section 2 details the methodological approach; Section 3 shows the analytical results; Section 4 discusses these findings in the context of cybersecurity and cyber warfare; and Section 5 highlights research gaps and suggests future directions for creating transparent and resilient AI-powered fraud detection systems.

2. Methodology

This study uses a structured approach to explore how XAI techniques are integrated into the development of transparent financial fraud detection models. The methodology follows a clear protocol emphasizing transparency, reproducibility, and rigor. The process includes four main stages: database selection, search query formulation, screening and eligibility assessment, and bibliometric–thematic analysis.

2.1 Data Sources and Search Strategy

Three reputable databases, Scopus, Web of Science (WoS), and Google Scholar (Gs), were used to ensure comprehensive coverage of peer-reviewed research. The search was limited to the period 2015–2025, reflecting the emergence and rapid evolution of XAI research following the Defense Advanced Research Projects Agency's (DARPA's) formal introduction of the concept in 2016 (Gunning and Aha, 2019) and its subsequent application in financial analytics. The search query combined controlled terms and Boolean operators ("AND," "OR") to capture variations in explainability and fraud detection terminology.

("explainable AI" OR "XAI" OR "model interpretability" OR "model transparency" OR "explainability") AND ("fraud detection" OR "financial fraud" OR "transaction fraud" OR "insurance fraud" OR "banking fraud")

The query was used on article titles, abstracts, and keywords. The initial search returned 575 publications across all databases (Scopus: 374, WoS: 132, Gs: 69).

2.2 Inclusion, Exclusion, and Data Cleaning

Publications were included if they: (i) were peer-reviewed journal articles, conference papers, or book chapters; (ii) were written in English; (iii) examined XAI techniques applied to fraud detection; and (iv) provided empirical validation through experiments or performance evaluation. Studies that were non-peer-reviewed, conceptual-only, or unrelated to fraud detection were excluded. Fifty-three (53) duplicate records were removed using DOI matching and title normalization, followed by manual checks for accuracy. Figure 1 illustrates the screening process.

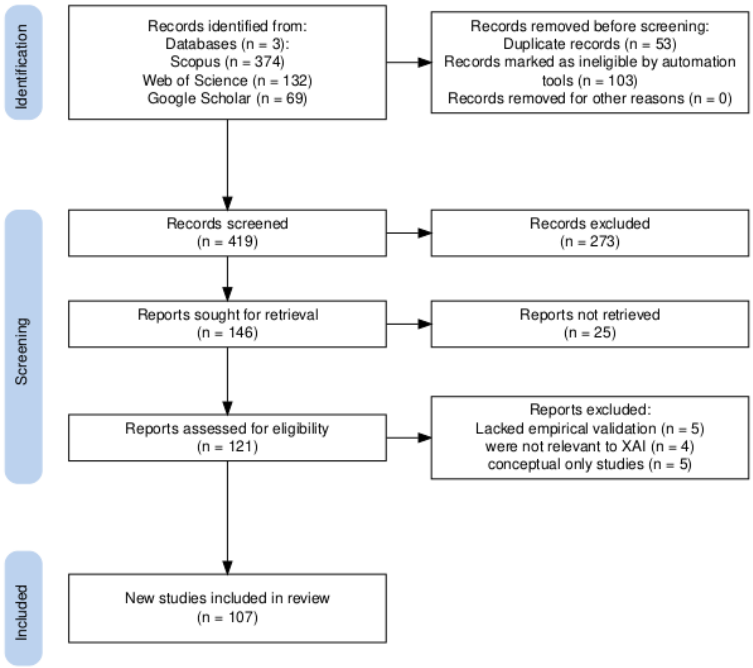

Figure 1: Flowchart for the identification and screening of the dataset

2.3 Screening and Dataset Formation

To ensure that only practical and reproducible studies were included, each publication was screened for empirical validation - that is, the use of real or simulated data rather than purely conceptual or theoretical analysis. Titles and abstracts of all 419 records were reviewed, leading to the exclusion of 273 studies without empirical evidence and 25 subscription-restricted papers. The remaining 121 publications were assessed in full. Fourteen were further excluded: five lacked empirical validation, four were unrelated to XAI, and five were conceptual-only. This process resulted in a final dataset of 107 empirically validated studies (Scopus: 90, WoS: 10, Gs: 7), representing the most relevant and robust contributions.

2.4 Study Quality Assessment

All 107 studies were assessed using the Mixed Methods Appraisal Tool (Nha HONG et al., 2018), which evaluates five criteria: clarity of research questions, suitability of data collection, validity of findings, analytical rigor, and relevance. Each study was rated as High (Yes), Moderate (Partial), or Low (No) quality. Low-quality studies were retained only when they offered unique methodological or technical value.

2.5 Data Extraction

A structured Excel-based form was used to capture key attributes from each study. These include publication details (year, venue, authors, country), fraud domain and dataset used, ML/DL model(s) applied, XAI technique(s) implemented, evaluation metrics used, key findings and contributions, and limitations and future recommendations

2.6 Data Synthesis and Analysis

A mixed bibliometric–thematic approach was employed.

- Bibliometric analysis (using VOS viewer v1.6.20) identified publication trends, collaboration networks, and keyword co-occurrences.
- Thematic analysis categorized studies by intellectual structure, model foundations, XAI adoption, evaluation metrics, and dataset influence on transparency.

A minimum keyword frequency of five was used, with clustering at the default resolution (1.0). Co-authorship networks were also generated to map collaboration patterns.

3. Analysis of the Results

3.1 Intellectual Structure and Collaboration Patterns in XFD Research

Figure 2 shows consistent growth in research on explainable fraud detection (XFD) from 2019 to 2025, with publication volumes accelerating sharply from 2021 onward. Early work focused on basic model comparisons, while studies from 2023–2025 increasingly explored attention mechanisms and intrinsic interpretability. The surge in 2024–2025 signals a shift toward integrating transparency into mainstream fraud analytics, reflecting the sector's heightened emphasis on algorithmic accountability.

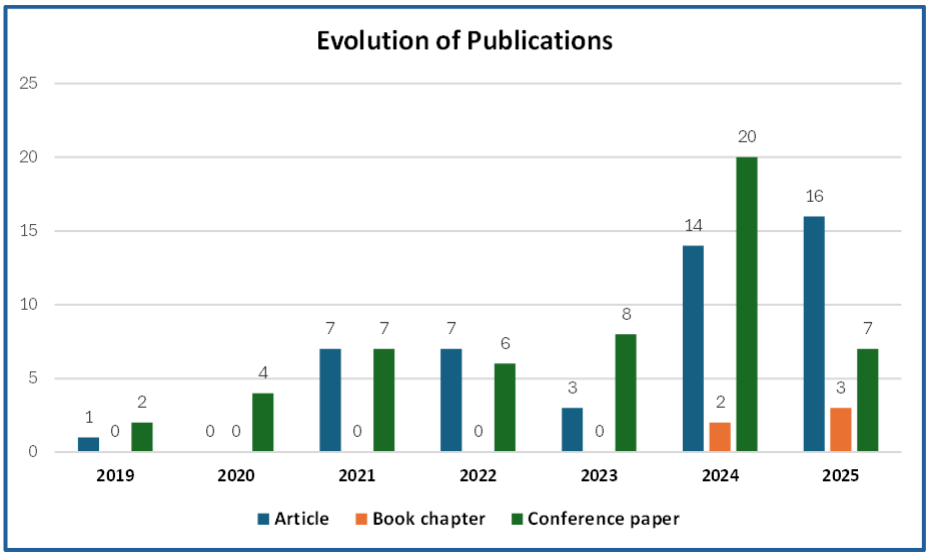

Figure 2: Evolution of Publications

Keyword co-occurrence patterns highlight four thematic clusters: (i) model-agnostic XAI (SHAP, LIME) applied to tabular fraud data; (ii) ML/DL methods for performance–transparency balancing; (iii) privacy-preserving approaches such as federated learning; and (iv) neural architectures leveraging attention and GNNs for relational reasoning. The convergence of "fraud detection," "anomaly detection," "deep learning," and "SHAP" indicates a maturing consensus that explainability must evolve alongside model complexity.

Intellectual activity in XFD reflects growing recognition that transparency, trust, and regulatory alignment must be embedded into fraud detection systems, not added post-hoc.

3.2 Algorithmic Foundations of ML and DL in Fraud Detection

Across the 107 studies, ML models remain dominant (76 occurrences) due to their compatibility with tabular data and interpretability tools. Random Forest (10 studies), XGBoost (8), Decision Trees (7), and Logistic Regression (6) continue to underpin many explainable pipelines. DL methods appear in 54 studies, led by Neural Networks (39), Autoencoders (7), CNNs (6), and LSTMs (4). Table 1 shows the progressive rise of DL from 2019 to 2025, with significant growth in CNNs and Transformers as computational capacity and sequence-modelling needs increased.

Table 1: Trend of ML and DL Techniques Adoption

TECHNIQUE	ML/DL	2019	2020	2021	2022	2023	2024	2025
Transformer	DL	0	0	0	0	0	1	0
Autoencoder	DL	1	0	2	1	0	0	1
Convolutional Neural Network	DL	0	0	0	1	0	2	3
Long Short-Term Memory (LSTMM)	DL	0	0	0	0	1	2	1
Neural Network	DL	2	0	6	5	2	9	9
Recurrent Neural Network (RNN)	DL	0	0	0	1	0	1	0
Transformer	DL	0	0	0	0	1	1	3
Decision Tree	ML	0	1	2	0	1	5	5
Gradient Boosting	ML	0	1	0	1	0	2	6
K-Nearest Neighbors (KNN)	ML	0	0	2	0	0	0	1
Light Gradient Boosting Machine (LightGBM)	ML	0	0	0	0	1	1	2
Logistic Regression	ML	0	1	1	1	2	3	4
Naive Bayes	ML	0	0	0	0	0	0	2
Random Forest	ML	0	1	2	0	3	7	10
Support Vector Machine SVM)	ML	0	0	2	1	1	3	7
Extreme Gradient Boosting XGBoost	ML	0	0	2	0	0	6	8

A dual-track methodological evolution is observed:

- ML models remain widely used due to their transparency and auditability, key requirements in regulated financial environments.
- DL architectures address the rising complexity of fraud patterns, particularly temporal and behavioral dependencies.

Hybrid ML–DL architectures exemplify this socio-technical balancing. Studies such as Roshan and Zafar (2021) and Sakil et al. (2025) combine deep feature extraction, such as Autoencoders and CNNs, with interpretable classifiers, including Random Forest and SHAP-enhanced ML models. This layered strategy mirrors enterprise information systems, where accuracy and interpretability must coexist to support human analysts, auditors, and compliance officers.

3.3 The Role of XAI in Enhancing Transparency and Trust

XAI functions as the bridge between high-performance analytics and the transparency expectations of financial institutions, regulators, and customers. Across the corpus, SHAP, LIME, and Attention Mechanisms dominate, reflecting both their accessibility and alignment with regulatory guidelines such as GDPR's "right to explanation" (The European Parliament, 2016) and Basel's model governance principles (Financial Stability Institute, 2011).

SHAP (34 studies) leads due to its model-agnostic nature and ability to provide both global and local explanations. LIME (15 studies) remains popular for lightweight interpretability, while attention-based models (16 studies) enable intrinsic transparency in sequential and graph-based fraud detection. Figure 3 shows sharp increases in SHAP and LIME after 2022 as financial institutions moved toward auditable AI.

XAI adoption patterns vary by data type:

- SHAP dominates credit card and banking fraud, where features are structured, and explanations must be audit-ready.
- LIME supports logistic regression and Support Vector Machine (SVM) architectures commonly used in insurance claim fraud.
- Attention mechanisms are used for blockchain, identity fraud, and sequential transaction data, where relational dependencies are crucial.

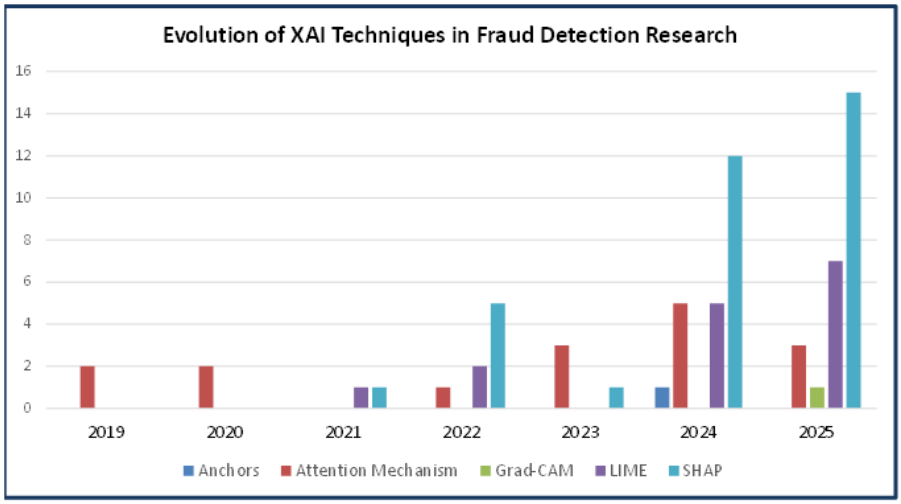

Figure 3: Evolution of XAI Techniques in Fraud Detection Research

However, methods such as counterfactuals, causal explanations, and interactive dashboards remain underused, despite their substantial value for organizational decision-making. Their limited adoption signals an opportunity for further research to enhance end-user comprehension, regulatory reporting, and operational trust.

3.4 Evaluation Metrics and Their Influence on XFD System Design

The literature indicates that Accuracy (48 studies) and Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) (45 studies) dominate XFD evaluations. Their continued use reflects historical habits but fails to capture operational realities in financial institutions, where false positives, false negatives, and the quality of explanations determine trust, compliance, and resource demands.

More operationally meaningful metrics; Recall, Precision, True Positive Rate (TPR), False Positive Rate (FPR), and Mean Squared Error (MSE); remain underused despite their relevance for customer impact, analyst workload, and human–AI decision alignment. This reveals a methodological gap: few studies adopt multi-metric frameworks that integrate predictive performance with explanation fidelity. Such approaches are essential for transparency, governance, auditability, and cross-institutional comparability.

3.5 Dataset Influence on Transparency and Interpretability

Dataset properties significantly affect model transparency. The literature reveals reliance on structured datasets, Financial Datasets (48 studies), Institute of Electrical and Electronics Engineers – Computational Intelligence Society (IEEE-CIS) (36 studies), University of California, Irvine (UCI) (21 studies), and Credit Card Fraud (20 studies), which support interpretable models and effective XAI methods valued by compliance and risk teams.

In contrast, blockchain, mobile money, smart contracts, and insurance datasets are rarely used due to access, standardization, and proprietary constraints, limiting methodological generalizability to emerging digital-finance ecosystems.

Hybrid approaches are gaining traction by integrating structured data with unstructured or relational sources, for example, by combining attention-based blockchain models with SHAP explanations. The findings underscore the need for broader dataset diversity, more precise documentation, and new benchmarks to enhance the transferability and transparency of XFD systems.

3.6 Methodological Limitations

This review focuses solely on peer-reviewed empirical studies published from 2015 to 2025, potentially excluding conceptual or emerging XAI approaches. The English-only criterion introduces linguistic bias, and keyword-based searches may miss studies using unconventional terminology. Although thematic synthesis maintained coherence, some interpretations remain subjective. The review also prioritizes methodological transparency over performance comparison, and findings should be interpreted within this scope.

4. Discussion of Findings

Many fraud detection studies rely on anonymized or synthetic datasets that lack behavioral richness and fail to capture real-world adversarial patterns. Evaluation often focuses narrowly on metrics like accuracy and AUC, neglecting explanation quality, interpretability, and forensic value. This creates a gap: models may perform well statistically but remain unsuitable for cyber defense, where transparency and accountability are crucial. To address this, we propose the Three-Pillar Framework for Explainable AI, which provides a structured approach to building transparent, operationally trustworthy fraud detection systems. These pillars are:

- *Algorithmic Transparency*: the capacity of ML/DL models to expose decision logic through interpretable architectures or post-hoc explanation tools such as SHAP, LIME, or attention mechanisms.
- *Evaluation Accountability*: the use of multi-metric performance assessment that aligns quantitative accuracy with qualitative interpretability, ensuring fairness, fidelity, and stakeholder trust.
- *Data Diversity and Traceability*: the degree to which model explainability reflects the structure, representativeness, and provenance of underlying datasets, enabling reproducibility and auditability.

Together, these pillars define a socio-technical ecosystem in which transparent machine intelligence supports regulatory oversight, ethical governance, and alignment between model behavior and human understanding. Theoretical trends indicate a shift from treating explainability as an external diagnostic layer to integrating it directly into model design, as evidenced by attention-based models and game-theoretic interaction analyses. This evolution positions interpretability as a design objective rather than a post-hoc requirement and suggests future research should optimize for transparency alongside accuracy and robustness.

At the applied level, financial institutions favor risk-averse, legally defensible methods, such as tree-based ensembles paired with SHAP, that provide clear, auditable explanations. Operational constraints, including data privacy, limited cross-institutional datasets, and regulatory inconsistencies, restrict the deployment of more advanced intrinsic-XAI models, widening the gap between exploratory academic work on Transformers, Graph Neural Networks (GNNs), and attention mechanisms and the practical need for simplicity and transparency in regulated environments. Overall, explainability in financial fraud detection is context-dependent and stakeholder-driven: the usefulness of an explanation is shaped not by abstract completeness but by its relevance to regulators, fraud analysts, and customers. Consequently, XAI evaluation in finance should adopt multidimensional criteria that combine technical performance with human-centered qualities such as comprehensibility, actionability, and regulatory compliance.

4.1 Positioning the Three-Pillar Framework within Existing XAI Taxonomies

The proposed Three-Pillar Framework for Explainable Fraud Detection extends existing XAI taxonomies by integrating algorithmic transparency, evaluation accountability, and data traceability into a unified socio-technical model. While rooted in the DARPA-XAI taxonomy (Gunning and Aha, 2019), which focused primarily on model transparency and user-oriented explanations, the framework advances this view by elevating evaluation and dataset provenance as coequal pillars. This reframes explainability from a technical interface problem into a systemic feedback loop that links model logic, performance validation, and data foundations within a single evaluative ecosystem.

The framework also aligns with and expands the XAI 2.0 perspective by Longo et al. (2024), which emphasizes epistemic transparency, ethical accountability, and stakeholder-centric design. Within this lens, Algorithmic Transparency serves as the epistemic layer, Evaluation Accountability as the ethical–regulatory layer, and Data Diversity and Traceability as the empirical foundation for representative and auditable explanations. Synthesizing DARPA-XAI's technical foundations with XAI 2.0's ethical paradigm, the Three-Pillar model operationalizes both in a form tailored to financial AI systems, where interpretability, compliance, and auditability must converge.

4.2 Cross-Domain Applicability of the Three-Pillar Framework

Although developed for financial fraud detection, the Three-Pillar Framework applies broadly to domains involving high-risk, regulated decision-making. In insider threat detection, Algorithmic Transparency supports interpretable behavioral anomaly models by enabling analysts to justify flagged actions. In healthcare fraud and diagnostic prediction, Evaluation Accountability ensures that explanations meet ethical and clinical standards. Likewise, Data Diversity and Traceability are crucial in cybersecurity and critical infrastructure monitoring, where verifiable data provenance underpins auditing, forensics, and incident response.

Collectively, these mappings demonstrate that the framework offers a scalable, domain-agnostic architecture for trustworthy AI in sectors where transparency, accountability, and auditability are central to operational integrity. Treating explainability as a measurable, reproducible property of model design rather than a purely qualitative add-on, the framework advances the goal of developing human-centered, reliable metrics for transparent intelligent systems.

5. Key Findings and Conclusion

This study identifies prevailing trends, research gaps, and prospective directions to advance transparency, interpretability, and trust in AI-driven decision systems.

5.1 Research Gaps

- Existing explainable fraud and insider-threat detection models rarely use recurrent architectures or latent behavioral representations. As a result, they struggle to model evolving attack behavior.
- The field exhibits heavy reliance on post-hoc, model-agnostic tools such as SHAP and LIME, with minimal exploration of causal, counterfactual, or visual explanation methods like Gradient-weighted Class Activation Mapping (Grad-CAM) and Anchors. This methodological homogeneity restricts interpretability to feature attribution rather than causal reasoning or decision transparency, an essential requirement in regulated financial environments.
- Evaluation practices focus predominantly on Accuracy and ROC–AUC, with limited consideration of domain-sensitive and explanation-centric metrics such as Recall, Precision, False Positive Rate, explanation fidelity, or user satisfaction. This imbalance risks overstating performance and obscuring the operational reliability of models deployed in real-world financial systems.
- Research is heavily concentrated on a few structured and publicly available datasets (IEEE-CIS, UCI Repository, and Credit Card datasets). Emerging ecosystems, blockchain transactions, smart contracts, mobile money, and insurance claims remain underrepresented, limiting model generalizability and contextual validity.

5.2 Conclusions

This paper reviewed the role of Explainable Artificial Intelligence (XAI) in financial and insider fraud detection within cyber warfare and security contexts. The review showed that most existing systems prioritize predictive accuracy over transparency. This imbalance reduces analyst trust, weakens forensic investigation, and complicates regulatory compliance. In high-risk security environments, black-box models hinder effective, accountable decision-making.

To address these challenges, the paper proposed a Three-Pillar Framework for Explainable Fraud Detection. The framework consists of Algorithmic Transparency, Evaluation Accountability, and Data Diversity and Traceability. It treats explainability as a core design requirement rather than a post-hoc feature. The framework provides a structured approach for aligning AI models with operational, regulatory, and human-centered needs in cyber defense settings.

Algorithmic Transparency focuses on making model decisions understandable to analysts. This can be achieved through interpretable models, attention mechanisms, or post-hoc tools such as SHAP and LIME. Evaluation Accountability extends performance assessment beyond accuracy and ROC–AUC. It includes operational metrics and explanation quality to support trust, auditability, and responsible use. Data Diversity and Traceability ensure that datasets are representative, well-documented, and auditable. This supports reproducibility and regulatory review.

A key gap identified in this review is the limited use of temporal and behavioral modeling in explainable fraud detection. Such models are essential for capturing how fraud evolves. Variational Autoencoders (VAEs) can represent latent behavioral patterns, while Bidirectional LSTMs (BiLSTMs) can model sequential dependencies. When applied within the proposed Three-Pillar Framework, these approaches can guide future research on insider financial fraud detection while maintaining interpretability and auditability.

Ethics Declaration: The authors declare that ethical clearance was not required for the research. No human subjects or sensitive data used

AI Declaration: The authors declare that AI tools were not used to create this paper. Grammarly was employed to ensure clarity, grammar, and professional tone in the documentation.

References

Abbas, Z.A., Hilal, Z.M., Jabbar, H.G., 2025. Click Fraud Detection in Online Advertising: A Comparative Study of Machine Learning Models. International Journal of Safety and Security Engineering 15, pp 427–441. https://doi.org/10.18280/IJSSE.150303

Adetumi Adewumi, Somto Emmanuel Ewim, Ngodoo Joy Sam-Bulya, Olajumoke Bolatito Ajani, 2024. Enhancing financial fraud detection using adaptive machine learning models and business analytics. International Journal of Scientific Research Updates 8, pp 012–021. https://doi.org/10.53430/IJSRU.2024.8.2.0054

Agomuo, O.C., Uzoma, A.K., Khan, Z., Otuomasirichi, A.I., Muzamal, J.H., 2025. Transparent AI for Adaptive Fraud Detection. Proceedings of the 2025 19th International Conference on Ubiquitous Information Management and Communication, IMCOM 2025. https://doi.org/10.1109/IMCOM64595.2025.10857433

Ahmad, W., Vashist, A., Sinha, N., Prasad, M., Shrivastava, V., Muzamal, J.H., 2025. Enhancing Transparency and Privacy in Financial Fraud Detection: The Integration of Explainable AI and Federated Learning. Communications in Computer and Information Science 2244 CCIS, pp 139–156. https://doi.org/10.1007/978-3-031-75201-8_10

Al-hchaimi, A.A.J., Alomari, M.F., Muhsen, Y.R., Sulaiman, N. Bin, Ali, S.H., 2024. Explainable Machine Learning for Real-Time Payment Fraud Detection: Building Trustworthy Models to Protect Financial Transactions. Lecture Notes in Networks and Systems 1033 LNNS, pp 1–25. https://doi.org/10.1007/978-3-031-63717-9_1

Aljunaid, S.K., Almheiri, S.J., Dawood, H., Khan, M.A., 2025. Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection. Journal of Risk and Financial Management 18. https://doi.org/10.3390/JRFM18040179

Ezzeddine, F., Saad, M., Ayoub, O., Andreoletti, D., Gjoreski, M., Sbeity, I., Langheinrich, M., Giordano, S., 2024. Differential Privacy for Anomaly Detection: Analyzing the Trade-Off Between Privacy and Explainability. Communications in Computer and Information Science 2155 CCIS, pp 294–318. https://doi.org/10.1007/978-3-031-63800-8_15

Financial Stability Institute, 2011. Principles for the Sound Management of Operational Risk (PSMOR). https://www.bis.org/publ/bcbs195.pdf.

Gunning, D., Aha, D.W., 2019. DARPA's explainable artificial intelligence program. AI Mag 40, pp 44–58. https://doi.org/10.1609/AIMAG.V40I2.2850

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. Del, Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., Stumpf, S., 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion 106, 102301. https://doi.org/10.1016/J.INFFUS.2024.102301

Mozolewski, M., Bobek, S., Nalepa, G.J., 2024. Visual Explanations and Perturbation-Based Fidelity Metrics for Feature-Based Models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 14835 LNCS, pp 294–309. https://doi.org/10.1007/978-3-031-63772-8_27

Nha HONG, Q., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., Rousseau, M.-C., Vedel, I., n.d. MIXED METHODS APPRAISAL TOOL (MMAT) VERSION 2018 User guide.

Ofori, H.K., Bell-Dzide, K., Brown-Acquaye, W.L., Lempogo, F., Frimpong, S.O., Agbehadji, I.E., Millham, R.C., 2025. Application of Machine Learning and Deep Learning Techniques for Enhanced Insider Threat Detection in Cybersecurity: Bibliometric Review. Symmetry 2025, Vol. 17, Page 1704 17, 1704. https://doi.org/10.3390/SYM17101704

Pennella, L., Pinelli, F., Galletta, L., 2025. X-SPIDE: An eXplainable Machine Learning Pipeline for Detecting Smart Ponzi Contracts in Ethereum. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3569565

Praveenraj, D.D.W., Victor, M., Vennila, C., Alawadi, A.H., Diyora, P., Vasudevan, N., Avudaiappan, T., 2023. Exploring Explainable Artificial Intelligence for Transparent Decision Making. E3S Web of Conferences 399. https://doi.org/10.1051/E3SCONF/202339904030

Qin, Z., Liu, Y., He, Q., Ao, X., 2022. Explainable Graph-based Fraud Detection via Neural Meta-graph Search. International Conference on Information and Knowledge Management, Proceedings pp 4414–4418. https://doi.org/10.1145/3511808.3557598

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-August-2016, pp 1135–1144. https://doi.org/10.1145/2939672.2939778

Roshan, K., Zafar, A., 2021. Utilizing Xai Technique to Improve Autoencoder Based Model for Computer Network Anomaly Detection with Shapley Additive Explanation(SHAP). International Journal of Computer Networks and Communications 13, pp 109–128. https://doi.org/10.5121/IJCNC.2021.13607

Sakil, M.B.H., Hasan, M.A., Mozumder, M.S.A., Hasan, M.R., Opee, S.A., Mridha, M.F., Aung, Z., 2025. Enhancing Medicare Fraud Detection with a CNN-Transformer-XGBoost Framework and Explainable AI. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3562577

The European Parliament, 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence. Office of the European Union L-, Publications.

The European Parliament, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

Tian, Y., 2025. Financial Fraud Detection Based on an XGBoost and LSTM Fusion Model: A Comparative Study on the Enhancement of Time-Series Features. Advances in Economics, Management and Political Sciences 170, pp 101–111. https://doi.org/10.54254/2754-1169/2025.LH23993

Vivek, Y., Ravi, V., Mane, A., Naidu, L.R., 2025. Explainable One Class Classification for ATM Fraud Detection. International Conference on Communication Systems and Networks, COMSNETS pp 114–119. https://doi.org/10.1109/COMSNETS63942.2025.10885737

Zhang, K., Wang, H., Chen, M., Chen, X., Liu, L., Geng, Q., Zhou, Y., 2025. Leveraging machine learning to proactively identify phishing campaigns before they strike. J Big Data 12. https://doi.org/10.1186/S40537-025-01174-X

Zhu, W., Chen, Z., 2024. An Intelligent Financial Fraud Detection Model Using Knowledge Graph-Integrated Deep Neural Network. Journal of Circuits, Systems and Computers 33. https://doi.org/10.1142/S0218126624502670

Zhu, W., Zhang, C., Li, J., Wang, Z., 2025. An Intelligent Financial Fraud Detection Model Based on Dilated Convolution and Generative Adversarial Network. Journal of Circuits, Systems and Computers 34. https://doi.org/10.1142/S0218126625501701