Application of Explainable AI in Identifying Fraudulent Transactions

Guanming Shi
Max Planck Institute of Security and Privacy
Bochum, Germany

Abstract

The proliferation of digital financial transactions has expanded opportunities for fraud, prompting substantial investment in AI-driven detection systems by technology firms, financial institutions, and regulators. Yet these systems continue to exhibit high false-positive rates, and their outputs often lack the interpretability that operational stakeholders require to act on them confidently. This paper argues that Explainable AI (XAI) that could satisfy analysts' goals and needs in fraud detection could better support their decision-making in the high-stakes scenarios of fraud detection. Drawing on ongoing semi-structured interviews with 21 fraud detection professionals, we identify key workflow pain points, characterize the forms of explanation that practitioners find actionable, and assess the risks of poorly implemented XAI. We conclude with a discussion of implications for the design and evaluation of XAI tools in fraud detection, grounded in the operational realities practitioners described.

1 Introduction

Global losses attributable to fraud amount to billions of dollars annually, e.g., USD 21 billion in the United States [5] and €7.4 billion in the European Union [7] just in 2024. Beyond direct monetary losses, fraud fuels broader societal harms, including organized crime and drug trafficking [20]. Monetary and psychological costs of fraud are often borne by companies and private individuals, many of whom are unable to fully recover their losses. Even when reimbursement is provided, victims may face significant administrative burdens, reputation losses, and psychological stress following victimization [2]. Together, these impacts underscore financial crimes as a persistent and high-stakes societal challenge.

In response, governments, financial institutions, and technology companies have turned to Artificial Intelligence (AI) systems to support the detection of fraudulent transactions. These systems offer powerful capabilities for identifying anomalous transaction patterns and predicting risky behavior. However, their practical deployment is often constrained by high false-positive rates and limited transparency. Wedge et al. [19] quantitatively demonstrated that one in five blocked financial transactions was fraudulent and that every sixth transaction was mistakenly flagged. When system outputs cannot be meaningfully interpreted, particularly in the presence of frequent errors, trust in automated decisions erodes, and human operators struggle to act on model recommendations.

This challenge has brought renewed attention to XAI in supporting analysts in identifying fraudulent transactions. Yet, we argue that existing approaches in this space often prioritize technical transparency over the operational and cognitive needs of the diverse stakeholders who rely on these systems. As a result, explanations risk overwhelming users without supporting effective judgment or action in practice.

2 Related Work

2.1 AI in Fraud Detection

Three trends have led banks, law enforcement, and technology companies to move away from traditional rules-based fraud detection toward AI-powered tools [9]. First, the growing volume of digital transactions renders conventional methods insufficient to handle modern financial system complexities. Second, fraud is highly adversarial — fraudsters continuously adapt to circumvent static rule sets, whereas ML models can ingest new patterns with fewer manual updates. Third, severe class imbalance in transaction data, where fraudulent cases are rare by nature, causes rules-based systems to either miss fraud or over-generate false positives; ML approaches have

Workshop on Designing Human–AI Teaming for Security (HATS), co-located with the Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23, 2026, Hannover, Germany.

shown greater robustness to this through techniques such as cost-sensitive learning [13]. Together, these pressures have driven widespread adoption of AI-based detection across the financial sector.

2.2 Explainable AI Approaches

There is no established consensus on the complete set of techniques for XAI, prior work shared the goal of making an AI system's functioning or decisions easy to understand by people [6]. Following the seminal paper by Ali et al. [3], XAI techniques can be divided by three broad categories: data-based, complexity-based and model-based.

- **Data**: Techniques that calculate a weight for each input variable, revealing how much influence that input has on a model's predictions. They can be local (explain a single prediction) or global (explain model behavior). LIME [16] is a seminal example of a local explanation, whereas SHAP [12] can provide by local and global explanations.
- **Complexity**: Simpler models like Linear Regression and Random Forest are inherently explainable [1]. However, previous studies across cyber security, healthcare and fraud detection have shown that these models underperform "black-box" Neural Networks and Deep Learning models in anomalous detection, as these data are highly non-linear and complex [18]. To overcome this trade-off, post-hoc XAI techniques are implemented to provide users with explanations.
- **Model**: Methods are either model-specific (tied to a particular algorithm) or model-agnostic (applicable across ML models). In fraud detection, model-agnostic XAI methods are preferred as they provide flexibility across the heterogeneous ensemble of models typically deployed in financial institutions [10].

Most prior work in XAI evaluates explanation effectiveness through traditional AI performance metrics such as accuracy, precision, and ROC-AUC, rather than how intended users perceive or act on them [14]. To address this gap, we conducted semi-structured interviews with fraud detection and investigation professionals to understand how AI and explainability could be integrated into their workflows. Specifically, we would like to address three questions:

- **RQ1**: How are AI systems currently deployed in anti-fraud operations and to what extent is explainability integrated into these systems?
- **RQ2**: Why do anti-fraud teams want explainability for their AI tools? What are these explanations used for?
- **RQ3**: Which forms of AI-generated explanations (e.g. graphical, risk scores, text) are perceived as most effective by anti-fraud teams?

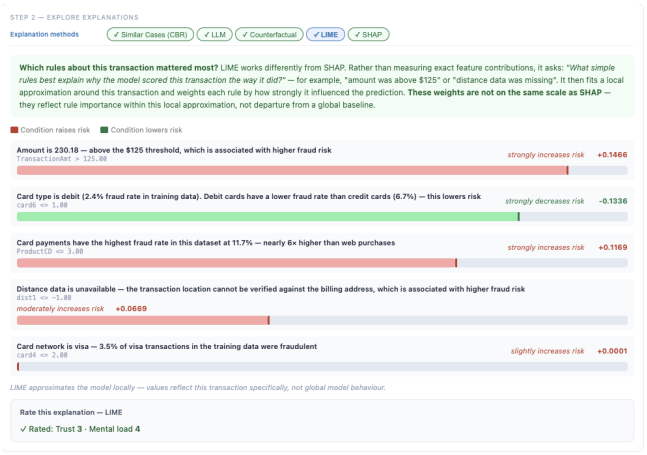

Figure 1: The fraud detection dashboard used as part of the interview. Participants classified transactions, then explored explanations (LIME shown) across five methods, rating trust and mental load for each before a final reflection step.

3 Method

The following section describes the recruitment, interview procedure and qualitative data analysis.

3.1 Recruitment

A total of 21 experts across 16 countries in Europe, North America, Asia and Africa were recruited for this study, with eight interviews completed as of the submission of this paper. Participant demographics are summarized in Table 1. The experts had an average of 7.26 years of experience in fraud detection or investigation (SD = 3.13). Experts were recruited through two channels - (1) snowball sampling of professional contacts of the first author and (2) freelancers from a platform called Upwork. Participants were included if they had at least two years of experience in fraud detection, investigation, or a related role; those without direct operational experience in this domain were excluded. Each expert recruited through Upwork were screened by a ten-minute introductory session with the first author, where they would answer their experience in fraud detection and knowledge of the workflow. Knowledge of AI or XAI was not necessary for our study. Interviews took place between May and August 2026. All experts were compensated 30 Euros for their participation, with experts recruited through Upwork compensated in accordance to the platform's Terms of Services.

3.2 Interview Protocol

Participants were informed about the data collection, storage and analysis through a consent form prior to the interview. The interviews were conducted remotely and recorded via Zoom and transcribed using MAXQDA. All participants went

Table 1: Demographics of experts recruited for interviews (N=21). Yrs. = years of experience in anti-fraud operations.

ID	Gender	Yrs.	Industry	Position
P1	F	6.92	FinTech	Product Manager
P2	F	8.75	FinTech	Head, Anti-Money Laundering and Fraud
P3	F	10.58	Banking	Risk Consultant
P4	M	14.25	Consulting	Fraud/AML Investigator
P5	M	3.92	Consulting	Senior Financial Crimes Consultant
P6	M	4.58	Government	Lead Data Scientist
P7	M	6.83	FinTech	Data Scientist
P8	F	12.92	Social Media	Payment Risk/Fraud Senior Specialist
P9	M	5.00	Banking	Senior Consultant
P10	M	10.00	Banking	Senior Manager, Risk Management
P11	F	6.33	Government	Assistant Director, Investigations
P12	M	3.50	Banking	Data Scientist
P13	M	6.50	Banking	Group Compliance Director
P14	M	9.00	Banking	Lead Architect
P15	M	9.91	FinTech	Chief Anti-Money Laundering and Fraud Officer
P16	F	2.83	Government	Compliance Officer
P17	F	4.00	FinTech	Compliance Analyst
P18	F	6.25	Government	Head of Fraud Reporting Operations
P19	F	4.00	Social Media	Trust and Safety Specialist
P20	M	7.83	Consulting	Consultant - Fraud Investigations and Chargebacks
P21	M	8.50	FinTech	Payment Disputes and Chargeback Specialist

through three parts for the interview - (1) Fraud Detection Workflow, (2) Demands for Explainability and (3) Interaction with Fraud Dashboard. To address RQ3, participants interacted with a fraud detection dashboard created by the first author (Figure 1), where financial transactions were explained using SHAP, LIME, Counterfactuals, LLMs and Case-based Reasoning, and rated which they thought was the best and how much they trusted or understood the explanations. Rating questions were adapted from prior interview-based studies on XAI [4].

3.3 Data Analysis

For data analysis, the first author employed a hybrid inductive-deductive thematic analysis [8], a choice suited to our exploratory research questions and the lack of practitioner input in existing fraud-detection XAI frameworks. This dual approach integrated both top-down and bottom-up coding [15]. In the top-down deductive phase, initial codes were derived from existing XAI literature in other high-stakes fields such as cybersecurity [17] and healthcare [11]. To complement these ideas and account for our limited prior experience in fraud detection, we integrated a bottom-up inductive approach. This allowed new codes and themes to emerge directly from the participants' accounts, bridging existing literature with context-specific empirical insights. As of submission, all eight completed interviews have undergone this coding process, and the themes reported in Section 4 reflect this initial analysis; coding will be revisited iteratively as remaining interviews are completed. As data collection is ongoing, Section 4 presents preliminary findings from the eight interviews conducted to date.

4 Findings and Discussion

Preliminary findings from our ongoing interviews suggest a nuanced view of how fraud detection professionals conceptualize the role of AI in their work. Several participants expressed reluctance toward full reliance on AI, raising concerns such as hallucinations, imbalanced training data, and the adversarial adaptability of fraudsters who may learn to circumvent automated detection. Accountability pressures from external stakeholders were also noted as a reason to keep human judgment central to consequential decisions, reflecting broader tensions around trust calibration in high-stakes security contexts.

However, participants did not dismiss AI outright. Some articulated a complementary role for AI as a productivity aid, one that could handle repetitive investigative tasks and routine data retrieval, freeing analysts for higher-order reasoning. Notably, explanations emerged as a potential enabler of this vision: participants suggested that AI could become more useful if it could synthesize information across disparate tools and sources into a coherent investigative narrative, rather than surfacing isolated signals. This points to properly designed explanations being an opportunity to support Human-AI teaming in fraud detection.

4.1 Integrating AI into Fraud Detection Workflows

Participants' use of AI varied considerably by organizational context. When applying risk scores to transactions, AI models such as XGBoost operated in near real-time with minimal human involvement. P3 described a fully automated pipeline where deep neural networks generate risk scores and trigger tiered responses (e.g., auto-block, customer alert, analyst review), with humans entering only to handle edge cases, customer callbacks, data annotation, and model monitoring. For investigations, participants described a more manual process in which AI served primarily as a productivity aid. P4 said that he used AI to filter large volumes of transactions, flagging surface-level anomalies, and compiling background information, with human analysts retaining the responsibility in making the final decision. Several participants noted using commercial Large Language Models (LLMs) for what P2 called "lazy work" - automating routine searches and document summaries, but still requiring human review to catch gaps and verify accuracy. Across participants, a shared pattern was that AI was most accepted when it acted as a starting point for the fraud detection workflow, and not a substitute for human decision-making.

4.2 Challenges with AI-assisted Fraud Detection

4.2.1 Hallucinations of AI Models

Concerns about AI inaccuracies surfaced across interviews. Participants described LLMs conflating information from different entities, generating plausible but unverifiable narratives, and producing outputs that required rereading multiple times before the intended meaning was clear. P2 noted that even a model with 75% accuracy rate will be well below what would be tolerated from a human analyst and could translate into significant compliance failures given the volume of cases processed daily. P6 said that the hallucinations from LLM outputs mean that they still required heavy human verification, limiting their time-saving benefit.

4.2.2 Fraudsters adapting faster than AI

Several participants (P1, P2, P4 and P7) expressed concerns that adversarial actors actively and rapidly learn to circumvent automated detection. P7 described fraudsters as consistently "two steps ahead," exploiting the gap between when new patterns emerge and when models can be retrained. A related concern was that rule-based AI systems become less effective over time as fraudsters learn which thresholds and behaviors trigger alerts, and adapt their transaction amount to circumvent these rules. P7 also noted that this dynamic made model monitoring and regular rule effectiveness reviews a necessary and ongoing human responsibility.

4.2.3 Regulatory pressure

A recurring challenge was the expectation that fraud decisions remain explainable and defensible to regulators and senior management. P2 described that under the Revised EU Payment Services Directive (PSD2), it was a legal requirement to not only flag fraud correctly but to keep full documentations of identified and missed cases, supported by an auditable trail. Hence, AI tools that improve detection speed but operate as black boxes are difficult to deploy in regulated environments where human accountability is legally mandated. Several participants described human-in-the-loop requirements not as a preference but as a compliance obligation, with senior management sign-off required for high-risk decisions.

4.2.4 Unrealistic organizational expectations

A related tension emerged around the gap between what organizations expect AI to deliver and what it can currently do. P3 noted that some companies are already reducing compliance headcount on the assumption that AI will replace analyst functions — a view she considered premature given correction rates she estimated at around 70%. P6 described a frustration common among newer analysts who expect to upload a bulk of documents and receive a fraud verdict, when in reality AI tools are only reliably useful for narrow, well-specified tasks. This disconnect, several participants suggested, risks both over-reliance on AI outputs and disillusionment when those outputs require as much human verification as manual work would have.

4.3 Views on Explainable AI

Participants' preferences for explanation type were shaped largely by their role and the time pressure of their workflow. Analysts operating under high case volumes favored concise, visual explanations that allowed fast triage. Several participants (P3, P4 and P7) preferred LIME-style feature importance displays in the fraud detection dashboard, describing them as quick to scan and effective at highlighting the key drivers of a risk score without requiring deep interpretation. One noted that color-coded feature bars allowed experienced analysts to pattern-match rapidly against known fraud typologies. SHAP outputs, while informative, were considered too technical for front-line use without additional training. LLM-generated narratives, though appreciated for their comprehensiveness, were seen as too verbose for workflows where analysts might handle hundreds of cases per day. For investigators handling complex or high-value cases, the priorities shifted. P1 valued similar-case comparisons, noting that seeing past confirmed fraud cases with matching patterns helped build confidence in borderline decisions — and could serve as a useful learning tool for less experienced analysts. This divergence in requirements for details suggests that explana-

tion design may need to be role-sensitive rather than uniform across a deployment.

A recurring desideratum across participants was for explanations to synthesize signals across multiple data sources into a coherent narrative, rather than surfacing isolated feature scores. Critically, several participants (P2, P3 and P4) emphasized that explanations are of limited value without broader transactional context: knowing that an amount exceeded a threshold means little without knowing the customer's payment history, typical spending behavior, the identity of the recipient, and whether the destination is a known high-risk jurisdiction. P6 described reviewing a transaction as inherently requiring a longitudinal view — a single flagged event only becomes interpretable against the backdrop of what the customer normally does and who they normally transact with. Several noted that their current tools do not provide this context automatically, leaving analysts to manually reconstruct it before a decision can be made.

5 Recommendations

Our findings suggest that while AI offers meaningful productivity gains, particularly in automating routine investigative tasks, practitioners remain cautious about full reliance on AI-driven decisions, citing concerns about hallucinations, adversarial adaptation, and regulatory accountability. Critically, explanations are not uniformly useful: their value depends heavily on the role of the analyst, the time available for review, and whether they incorporate sufficient contextual information such as transaction history and recipient identity.

These findings point to several implications for organizations evaluating XAI investments. First, explanation design should be role-sensitive: front-line analysts handling high case volumes have different needs than investigators working complex, high-value cases. Second, explanations that surface isolated feature scores without broader transactional context risk being ignored or misinterpreted in practice. Third, the gap between organizational expectations of AI and its current capabilities represents a deployment risk in itself: premature reduction of human oversight, driven by overestimated AI reliability, may undermine rather than support fraud detection effectiveness. We encourage future work to evaluate XAI tools against concrete practitioner needs through application-grounded user studies, rather than relying solely on computational performance metrics.

References

[1] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). 6:52138–52160.

[2] Eman Alashwali, Ragashree Mysuru Chandrashekar, Mandy Lanyon, and Lorrie Faith Cranor. Detection and Impact of Debit/Credit Card Fraud: Victims' Experiences. In *Proceedings of the 2024 European Symposium on Usable Security*, EuroUSEC '24, pages 235–260. Association for Computing Machinery.

[3] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. 99:101805.

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. 5:1–21.

[5] Federal Trade Commission. New FTC Data Show a Big Jump in Reported Losses to Fraud to $12.5 Billion in 2024.

[6] Upol Ehsan and Mark O. Riedl. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach.

[7] European Central Bank. Joint EBA-ECB report on payment fraud: Strong authentication remains effective but fraudsters are adapting.

[8] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. 5(1):80–92.

[9] Waleed Hilal, S. Andrew Gadsden, and John Yawney. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. 193:116429.

[10] Farhina Sardar Khan, Syed Shahid Mazhar, Kashif Mazhar, Dhoha A. AlSaleh, and Amir Mazhar. Model-agnostic explainable artificial intelligence methods in finance: A systematic review, recent developments, limitations, challenges and future directions. 58(8):232.

[11] Minjung Kim, Saebyeol Kim, Jinwoo Kim, Tae-Jin Song, and Yuyoung Kim. Do stakeholder needs differ? - Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces. 181:103160.

[12] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777. Curran Associates Inc.

[13] Sara Makki, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand-Saïd Hacid, and Hassan Zeineddine. An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. 7:93010–93022.

[14] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, prefix=van useprefix=false family=Keulen, given=Maurice, and Christin Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. 55:1–42.

[15] Kevin Proudfoot. Inductive/Deductive Hybrid Thematic Analysis in Mixed Methods Research. 17(3):308–326.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

[17] Neele Roch, Hannah Sievers, Noé Zufferey, and Verena Zimmermann. You know why, but still rely: The impact of explainable ai on trust, task load, and performance in cybersecurity decision-making. In *Proceedings of the 35th USENIX Security Symposium (USENIX Security 26)*. USENIX Association, 2026. To appear.

[18] Iqbal H. Sarker. Machine Learning: Algorithms, Real-World Applications and Research Directions. 2(3):160.

[19] Roy Wedge, James Max Kanter, Santiago Moral Rubio, Sergio Iglesias Perez, and Kalyan Veeramachaneni. Solving the "false positives" problem in fraud prediction.

[20] Jarrod West and Maumita Bhattacharya. Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57:47–66.