FRAUD DETECTION IN FINANCIAL SYSTEMS USING EXPLAINABLE AI

Prykhodko Taisiia[1]

Received: 2026-02-01
Accepted: 2026-04-25
DOI: https://doi.org/10.5281/zenodo.20687156

Abstract. This article presents a systematic review of AI driven fraud detection methodologies, evaluating the efficacy and transparency limitations of supervised, unsupervised and deep learning paradigms. Central to this study is the examination of Explainable Artificial Intelligence (XAI) as a reconciliation layer, specifically analyzing feature attribution methods (SHAP, LIME) and counterfactual explanations. Drawing on empirical evidence from industrial case studies, including Danske Bank and JPMorgan Chase, the research validates that interpretability significantly mitigates false positive rates, but faces challenges regarding computational latency and cognitive load. Synthesizing these findings, the paper proposes a novel Human-in-the-Loop (HITL) architectural framework that integrates XAI tools directly into analyst workflows. This approach advocates for a synergistic socio-technical system, where machine scalability is augmented by human ethical judgment, ensuring a fraud detection ecosystem that is robust, legally compliant and operationally transparent.

Key words: explainable AI (XAI), human-in-the-loop (HITL), GDPR compliance, SHAP, fraud finance detection, black box.

[1] Bachelor's Degree in Information Systems and Technologies,
Igor Sikorsky Kyiv Polytechnic Institute,
Graduate Assistant, American University,
https://orcid.org/0009-0008-9374-2703

Introduction

The exponential digitalization of the global financial infrastructure has catalyzed a radical expansion in the velocity, volume and heterogeneity of transactional data streams. This structural evolution has rendered legacy fraud detection mechanisms - predicated on static, deterministic heuristics increasingly obsolete [1]. Designed for low-dimensional, stable environments, such rule-based systems exhibit systemic rigidity, failing to capture the non-linear and adaptive modalities of modern adversarial vectors.

To address the exigencies of real-time processing within high-frequency data environments, financial institutions have executed a paradigmatic transition toward stochastic learning architectures, specifically leveraging Deep Learning and ensemble-based gradient boosting methods [2, 3]. While these high-dimensional models excel at extracting latent feature interactions to maximize predictive precision, their deployment has engendered a critical methodological dichotomy - the accuracy-interpretability trade-off [4]. Consequently, state-of-the-art algorithms frequently operate as epistemologically opaque “black boxes”, yielding probabilistic risk assessments devoid of accessible causal reasoning [5].

Within the stringent regulatory frameworks of the financial sector, such opacity transcends technical limitation to constitute a severe governance and operational risk. Compliance mandates, notably the GDPR`s right to explanation and Basel III/IV operational risk standards, necessitate that automated decisioning systems possess transparency, auditability and contestability [6, 7]. The deployment of uninterpretable models impedes the identification of algorithmic bias, obstructs forensic audit trails and undermines legal defensibility in dispute resolution. Furthermore, the absence of explanatory logic precipitates a crisis of operational trust: analysts, confronted with unsubstantiated false positives, suffer from cognitive overload, which degrades validation efficiency and exacerbates friction in the customer experience.

To ameliorate these systemic vulnerabilities, the domain of Explainable Artificial Intelligence (XAI) has emerged as a critical reconciliation layer, aiming to bridge the gap between statistical optimization and human cognition [8]. By deploying feature attribution methodologies, such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) - alongside counterfactual analysis, XAI frameworks seek to deconstruct the decision boundaries of complex models [9, 10, 11]. These techniques transform abstract probability distributions into semantic, human-intelligible rationales, thereby enabling stakeholders to verify that model outputs are driven by legitimate risk indicators rather than spurious correlations or prohibited discriminatory proxies.

Materials and methods.

This article presents a comprehensive survey of contemporary fraud detection instruments, systematically analyzing the limitations of supervised, unsupervised and deep learning approaches through the lens of model transparency. The core focus of the study is the paradigm of Explainable Artificial Intelligence (XAI), positioned as a critical mechanism for bridging the gap between the statistical power of algorithmic models and the cognitive requirements of human decision-makers [12]. The paper provides an in-depth examination of feature attribution methods, including SHAP and LIME, as well as counterfactual analysis techniques, as tools for demystifying model decision logic. Drawing on an empirical analysis of real-world deployments within the banking sector and an assessment of existing technological constraints, the authors propose a Human-in-the-Loop (HITL) architectural framework [13]. This approach advocates for the direct integration of interpretability tools into analysts` operational workflows, enabling a synergistic fusion of machine-level precision and human expert judgment to establish an ethically robust, transparent and resilient financial security system [14].

The methodological framework is designed to deconstruct the technological landscape of financial security, proceeding from a rigorous data acquisition phase to a structured taxonomic evaluation and culminating in the development of a conceptual proposal. To ensure a comprehensive analysis of the state-of-the-art, a structured literature search was conducted across

major scientific databases, including IEEE Xplore, Scopus, Web of Science and the ACM Digital Library [15]. The search strategy prioritized peer-reviewed articles, conference proceedings and technical reports published between 2018 and 2024, a temporal constraint selected to reflect the rapid evolution of deep learning architectures and the concurrent emergence of explainable artificial intelligence. The query protocols utilized Boolean operators to synthesize keywords across three core semantic domains: financial fraud detection mechanisms, advanced machine learning architectures and governance-focused topics regarding interpretability and algorithmic fairness.

Following the acquisition of primary sources, the study applied strict inclusion criteria to filter the corpus, eliminating purely theoretical papers that lacked empirical validation or architectural implementation details. The remaining literature formed the basis for a comparative evaluation, wherein identified fraud detection methodologies were stratified into a three-tiered taxonomy comprising supervised learning, unsupervised or semi-supervised approaches, and deep learning architectures. This stratification facilitated a multi-dimensional analysis in which algorithms were not evaluated solely on standard predictive metrics such as precision, recall and AUC-ROC, but were simultaneously assessed against qualitative indicators of interpretability, computational latency and robustness to adversarial concept drift [16]. This dual-lens approach allowed for the systematic identification of the inverse relationship between model complexity and transparency, which serves as the central problem statement of this research.

To bridge the disconnect between theoretical efficacy and operational reality, the methodology incorporates a qualitative analysis of industrial use cases from major financial institutions, selected based on the availability of verifiable data regarding their deployment of explainable AI strategies. The analysis of these real-world implementations, such as those by Danske Bank and JPMorgan Chase, provided empirical evidence regarding the trade-offs between regulatory compliance and operational scalability [17, 18]. The synthesis of these empirical findings with the identified limitations of current algorithmic paradigms informed the final phase of the research: the theoretical design of a Human-in-the-Loop (HITL) framework. This conceptual model was developed by mapping the deficiencies of automated “black box” decision-making against the capabilities of modern feature attribution and counterfactual generation tools, thereby prioritizing the integration of algorithmic output into human cognitive workflows over fully autonomous processing.

Results

A review of AI-based fraud detection tools and their limitations in terms of transparency:

The rapid digitalization of the global financial ecosystem has resulted in an unprecedented escalation in transactional volume, velocity and heterogeneity, fundamentally undermining the effectiveness of legacy fraud detection systems grounded in static rule-based heuristics [19]. Such systems, designed for low-dimensional and weakly adversarial environments, lack the expressive capacity and adaptive flexibility required to counteract the continuously evolving strategies of sophisticated fraud actors. In response to this structural inadequacy, financial institutions have increasingly transitioned toward Artificial Intelligence (AI) and Machine Learning (ML) driven detection architectures capable of operating at scale and adapting to non-stationary threat landscapes in near real time [20]. However, this paradigm shift has surfaced a critical methodological and governance dilemma - the inverse relationship between predictive accuracy and algorithmic interpretability.

Contemporary algorithmic fraud detection methodologies can be systematically classified into three dominant learning paradigms, each characterized by distinct inductive biases and failure modes. Supervised learning approaches, typically instantiated via ensemble-based classifiers such as random forests and gradient-boosted decision trees, leverage historically labeled transactional corpora to achieve high discriminative performance against known fraud patterns [21]. Despite their operational maturity, these models are intrinsically constrained by their dependence on annotation fidelity, susceptibility to class imbalance and limited resilience to concept drift [22, 23,

24]. As a result, their effectiveness deteriorates when confronted with novel, adversarially adaptive fraud schemes that deviate from historical distributions.

Unsupervised learning techniques have been introduced to alleviate these limitations by modeling baseline transactional behavior and identifying deviations through anomaly detection mechanisms [25]. While such approaches enhance sensitivity to previously unseen attack vectors, they rely predominantly on statistical deviation metrics rather than semantic intent, leading to structurally elevated false-positive rates. This phenomenon imposes significant downstream costs, including analyst fatigue, reduced trust in automated alerts and diminished operational efficiency within financial intelligence units.

At the apex of antifraud system evolution lie deep learning architectures, encompassing multilayer neural networks and representation learning frameworks capable of extracting complex, high-dimensional, nonlinear feature interactions from large-scale transactional data streams. These models consistently outperform shallow learners in terms of raw predictive accuracy. Nevertheless, their internal decision-making processes remain largely opaque due to the entangled nature of learned latent representations. Consequently, deep models typically output probabilistic fraud scores without providing intelligible or causally grounded explanations, rendering their predictions epistemically inaccessible to human stakeholders [26].

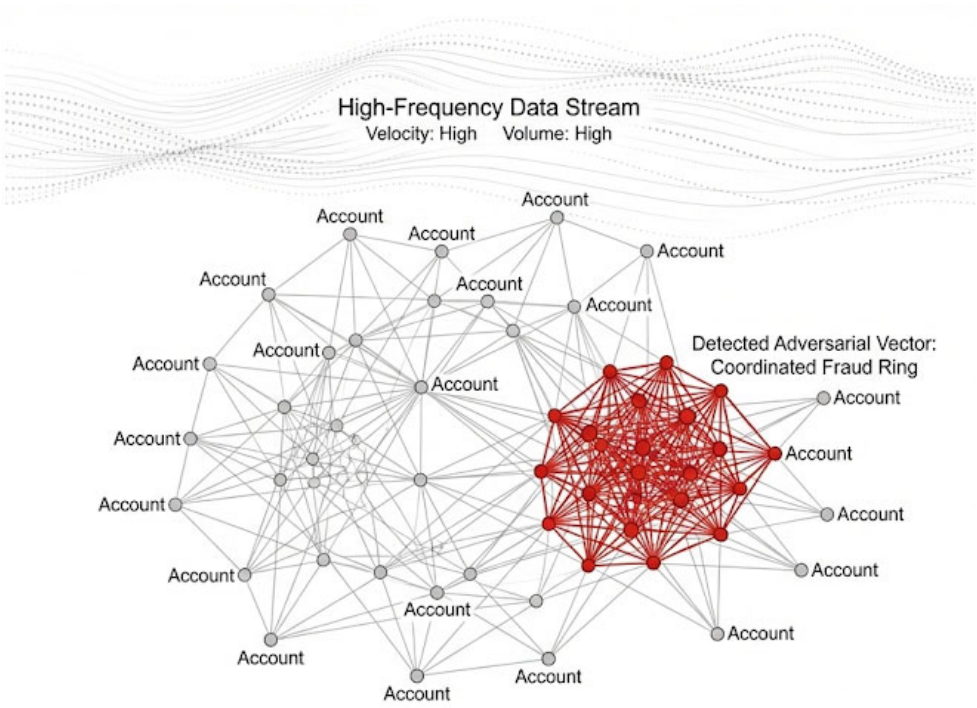

Figure 1. Visualization of high-dimensional transactional data relationships. The network graph reveals a coordinated adversarial cluster (highlighted in red), demonstrating the non-linear interactions and latent patterns detected by deep learning models that static heuristics fail to capture. Generative artificial intelligence was used to create the illustration

This algorithmic opacity introduces material risk within heavily regulated financial environments. Under stringent regulatory regimes, such as the General Data Protection Regulation (GDPR) and the Basel III/IV frameworks, and emerging supervisory mandates governing automated decision-making - financial institutions are required to ensure transparency, contestability and accountability of algorithmic decisions. Black-box models obstruct post hoc reasoning, inhibit the detection of latent representational biases, and compromise legal defensibility when automated decisions are challenged by customers or regulators [27].

Accordingly, the prevailing technological landscape necessitates a paradigmatic reorientation toward approaches that reconcile statistical learning performance with human-interpretable decision logic. This requirement has catalyzed the emergence of Explainable Artificial Intelligence

(XAI), a domain focused on augmenting predictive systems with interpretable, auditable and regulator-aligned explanatory mechanisms. By transforming opaque model outputs into structured, human-understandable rationales, XAI frameworks aim to restore epistemic visibility, enabling effective human oversight while preserving the operational advantages of advanced machine learning in fraud detection contexts.

Counterfactual explanations in synergy with explainable AI methods: role in elucidating model decision logic:

To operationalize principles of transparency in contemporary fraud detection architectures, it is imperative to deploy advanced analytical instruments capable of deconstructing the decision-making logic at the granularity of individual transactional events. At the vanguard of this paradigm shift are feature attribution frameworks, notably SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). These algorithmic methodologies mitigate the opacity inherent in "black box" machine learning models by quantifying the contribution of each input feature - ranging from transaction magnitude, geospatial coordinates, temporal patterns, to operation frequency to the model`s predictive output. This enables financial analysts to trace the specific features exerting maximal influence on the classification of a transaction as potentially fraudulent. Such a methodology translates abstract probabilistic outputs into rigorously justified, interpretable insights, underpinned by principles of cooperative game theory and local approximation techniques [28].

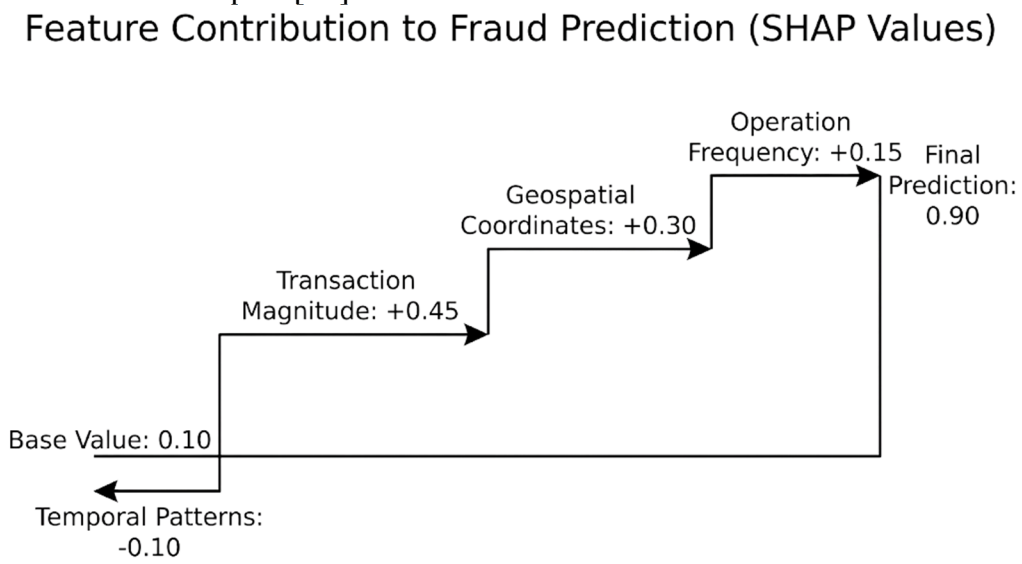

Figure 2. SHAP Waterfall plot illustrating local feature attribution for a high-risk transaction

Nevertheless, feature attribution approaches, while elucidating the rationale behind model predictions, often fall short of conveying the stability and sensitivity thresholds of these inferences. Counterfactual explanations address this limitation by providing a complementary interpretability paradigm through "what-if" scenario analysis. Unlike SHAP or LIME, which focus on the extant data configuration, counterfactual analysis identifies minimal perturbations in input variables necessary to invert the model`s classification outcome from "fraudulent" to "legitimate". For instance, a system may indicate that a transaction would have been approved if the inter-click latency were marginally increased or if the originating IP address aligned with the user`s residential region. This approach approximates human cognitive reasoning by emphasizing comparative and contrastive assessment of alternative decision pathways.

Moreover, the integration of feature attribution and counterfactual reasoning establishes a robust, multi-dimensional validation mechanism within mission-critical financial infrastructures. While SHAP and LIME illuminate anomalous patterns and direct expert scrutiny to high-risk regions of the feature space, counterfactual explanations provide sensitivity analysis, quantifying the confidence and resilience of the model`s classification under hypothetical perturbations [29]. This

synergistic deployment enhances the precision of investigative workflows by mitigating false positives, while simultaneously furnishing actionable intelligence for clients subjected to erroneous transaction blocks.

Consequently, this integrated framework transforms traditional fraud detection systems from opaque, procedural barriers into transparent, auditable and causally grounded decision-support instruments, wherein each predictive outcome is substantiated by both mechanistic attribution and scenario-based validation.

Trade-offs between model performance and interpretability - ethical and regulatory considerations with explainable AI positioned as a mechanism for bias mitigation, auditability and legal compliance:

The deployment of advanced machine learning methodologies in the financial sector inevitably confronts the classical accuracy-interpretability trade-off. Deep neural networks and ensemble learning methods, which exhibit superior performance in detecting sophisticated fraud schemes, operate as opaque "black boxes". In high-stakes fraud detection, where the cost of errors is substantial, financial institutions have historically prioritized predictive accuracy over model transparency, sacrificing insight into algorithmic reasoning to minimize direct financial losses.

However, in the contemporary regulatory environment, an exclusive focus on performance metrics constitutes an unacceptable strategic risk. Algorithmic efficacy can no longer justify opacity, particularly in the context of consumer protection and antidiscrimination legislation. The use of non-interpretable models fosters latent algorithmic bias, as systems may inadvertently learn from historical datasets containing social or demographic disparities [30]. Without Explainable AI (XAI) mechanisms, discriminatory outcomes, such as the blocking of transactions based on proxy variables correlated with ethnicity or socioeconomic status, remain invisible to system operators until a reputational or legal incident occurs.

In this context, XAI functions as a critical compliance enabler. European legislation, particularly the General Data Protection Regulation (GDPR), enshrines the "right to explanation", obligating organizations to provide stakeholders with meaningful information regarding automated decision logic. Explainable AI transforms fraud detection processes into auditable workflows: interpretability frameworks allow regulators and internal auditors to verify that high model performance is achieved through the identification of genuine fraud indicators rather than spurious correlations or prohibited attributes.

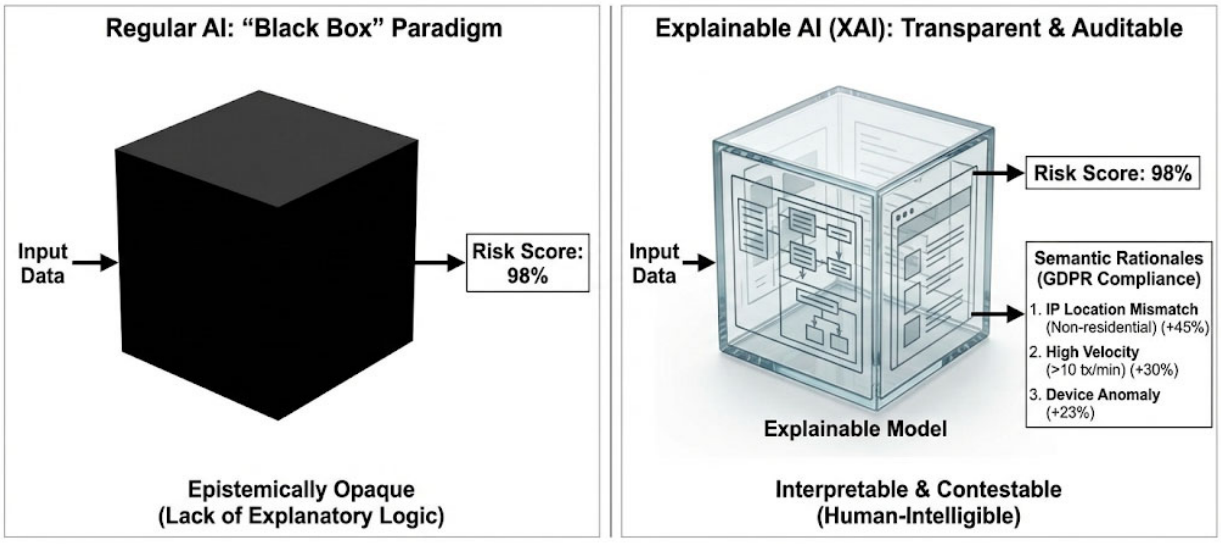

Figure 3. Comparison of traditional "black box" AI output vs Explainable AI. While the standard model provides only a probabilistic risk score, the XAI framework elucidates the semantic rationales driving the decision, thereby satisfying regulatory requirements for transparency and the right to explanation. Generative artificial intelligence was used to create illustration

Consequently, the integration of explainable models reconciles the tension between efficacy and transparency, enabling financial institutions to leverage the predictive power of modern AI technologies while maintaining adherence to legal and ethical standards.

Case studies of real-world XAI deployments in banking and fintech, highlighting both successful outcomes and existing limitations:

The transition from theoretical frameworks to the practical deployment of Explainable AI (XAI) in the financial sector has demonstrated tangible outcomes, corroborating the hypothesis that interpretability is a key driver of operational efficiency. A particularly illustrative case is the transformational experience of Danske Bank. Confronted with critically low fraud detection rates (~40%) and an overwhelming volume of false positives, reaching up to 1200 cases per day, the bank implemented a deep learning system augmented with behavioral analytics mechanisms. By moving away from rigid rule-based systems toward adaptive models complemented with anomaly interpretation tools, the institution reduced false alerts by 60% while simultaneously increasing the detection of genuine fraudulent activity by 50%. In this scenario, explainability played a pivotal role in fostering trust in the system: analysts could contextualize deviations, thereby accelerating the transaction validation process.

Similar successes have been observed among global financial leaders such as JPMorgan Chase and Mastercard, which leverage XAI to address scalability and compliance challenges [31]. JPMorgan Chase integrated Shapley value-based (SHAP) interpretability techniques into its transaction monitoring frameworks to ensure transparency for internal auditors and regulators. This integration also decreased the time required to detect security threats. Mastercard, deployed its Decision Intelligence solution, which utilizes explainable algorithms for real-time transaction analysis. By providing issuing banks with granular explanations of risk factors, such as atypical transaction timing or device mismatches, the system significantly enhanced the accuracy of legitimate transaction approvals while minimizing customer inconvenience.

Furthermore, despite these documented successes, large-scale XAI deployment encounters substantial technical and operational constraints. A fundamental challenge is the computational complexity of attribution methods: generating precise SHAP values for every transaction in high-throughput systems processing thousands of operations per second introduces unacceptable latency. Engineers are thus compelled to adopt approximated explanation methods, potentially sacrificing interpretability fidelity for processing speed. Additionally, there is the risk of explanation attacks, whereby detailed information regarding the rationale for transaction blocking may be exploited by adversaries to reverse-engineer models and adapt their attacks to algorithmic blind spots.

Moreover, cognitive limitations of end-users cannot be disregarded. Providing analysts with raw feature attribution data or complex visualizations can lead to information overload, paradoxically slowing decision-making rather than expediting it [32]. Research indicates that without interface adaptation to the operator`s expertise level, even mathematically precise explanations may be misinterpreted. Consequently, the current phase of technological evolution necessitates a shift in focus from developing novel explanation algorithms to integrating them into ergonomically optimized workflows, thereby bridging the gap between mathematical transparency and practical utility.

A proposed human-in-the-loop fraud detection framework in which interpretability tools are directly integrated into analysts` operational workflows:

Based on an in-depth analysis of technological capabilities and operational constraints, a conceptual framework for a fraud detection system is proposed, built on the principle of a Human-in-the-Loop (HITL) architecture. Unlike fully automated solutions, that exclude human operators from the decision-making process, or traditional systems in which analysts are overburdened with manual verification, the proposed framework positions AI algorithms as intelligent assistants, whose primary role is the augmentation of human expertise. The central element of this

architecture is the interaction interface, where machine predictions are intrinsically coupled with their interpretability.

From a technical standpoint, the proposed hybrid architecture envisions the integration of Explainable AI (XAI) modules directly into the monitoring dashboards of financial analysts. When the system flags a transaction as suspicious, the analyst is presented also with a visualized decision profile: key factors influencing the scoring (via SHAP value plots) and contextual cues elucidating deviations from the client`s normal behavioral patterns. This enables the expert to immediately assess the validity of the alert, distinguishing genuine threats from statistical anomalies, while substantially reducing cognitive load during incident processing. Here, interpretability functions as a filter, eliminating obvious model errors before initiating in-depth investigation.

A critically important component of the framework is the active learning mechanism facilitated through the feedback loop. Decisions made by the analyst, whether confirming fraudulent activity or marking a false positive, are continuously incorporated to retrain and refine the model [33]. In cases where the expert disagrees with the AI`s verdict and justifies their decision (for example labeling a transaction as legitimate despite an unusual geolocation), the system adjusts its internal weights, adapting to emerging patterns. This creates a synergistic effect: AI provides the speed and data processing scale beyond human capability, while the human operator contributes contextual understanding and ethical judgment unavailable to the machine.

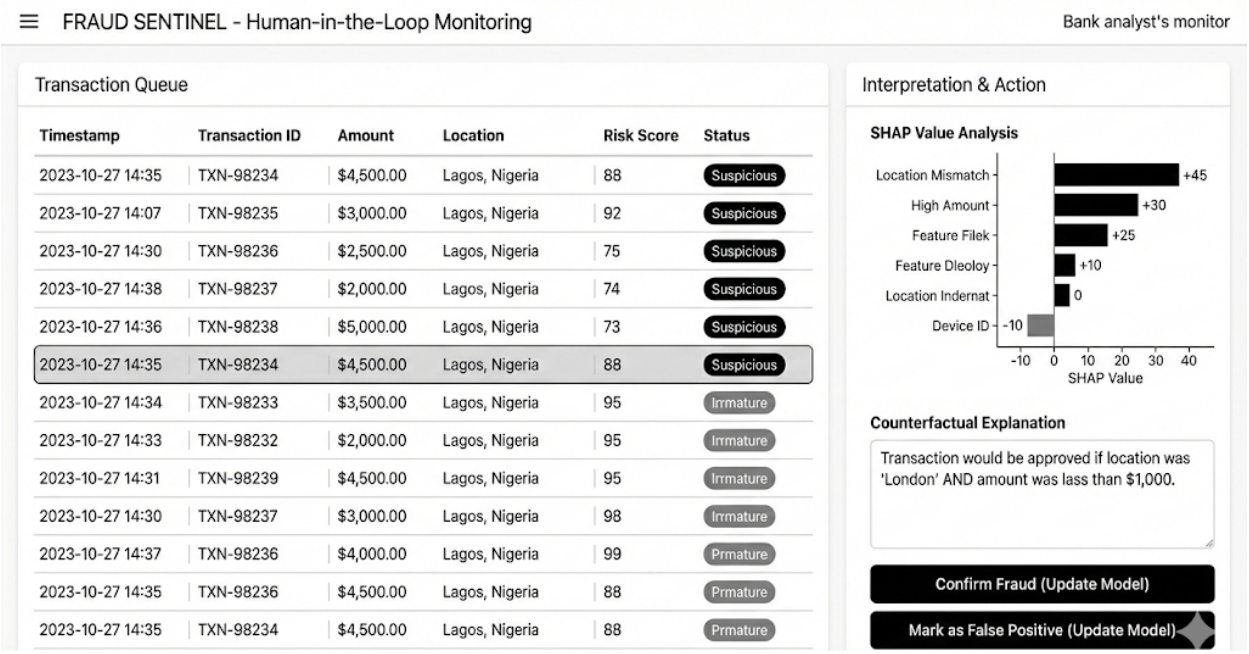

Figure 4. Dashboard mockup for an explainable fraud monitoring system. The interface of a system built on the principle of Human-in-the-Loop (HITL), where tools of interpretability are integrated into the analytical work process. Generative artificial intelligence was used to create illustration

Consequently, this approach ensures the resilience, transparency and continuous evolution of the financial security system.

Conclusions:

The rapid evolution of the global financial landscape has rendered the dichotomy between algorithmic precision and interpretability unsustainable. This study has demonstrated that while the transition from rule-based heuristics to deep learning architectures provides the necessary computational throughput to counter sophisticated fraud, the resulting epistemic opacity of 'black box" models introduces unacceptable systemic risks. The analysis confirms that in a regulated environment governed by frameworks, such as GDPR and Basel III, predictive accuracy alone is no longer a sufficient metric of viability. Operational resilience now demands that decision-making processes be transparent, auditable and devoid of latent algorithmic bias.

The critical review of Explainable Artificial Intelligence (XAI) methodologies highlights that tools like SHAP, LIME and counterfactual analysis are fundamental prerequisites for compliant fraud detection. By transforming abstract probabilistic scores into semantic rationales, these mechanisms bridge the cognitive gap between stochastic machine logic and human reasoning. The examination of industrial case studies validates this premise, evidencing that the integration of interpretability layers significantly reduces false positive rates and operational costs, as observed in the deployments by major financial institutions. However, the research also identifies persisting challenges, specifically the computational latency associated with real-time attribution and the cognitive overload risks for analysts, necessitating a careful calibration of explanation granularity.

In response to these findings, the proposed Human-in-the-Loop (HITL) framework synthesizes technical capability with operational necessity. By positioning the algorithm as an explainable assistant rather than an autonomous arbiter and by establishing an active learning feedback loop, this architecture ensures that the system evolves through human expert validation. Ultimately, the study concludes that the future of financial security lies not in the pursuit of full automation, but in the creation of synergistic socio-technical systems where machine scalability is governed by human ethical judgment and contextual understanding.

References:

1. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3784–3797.
2. Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3), 235–255. DOI: https://doi.org/10.1214/ss/1042727940
3. Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, 42(19), 6609–6619. DOI: https://doi.org/10.1016/j.eswa.2015.04.042
4. Kaggle & IEEE-CIS. (2019). *IEEE-CIS Fraud Detection*. Kaggle Competition. URL: https://www.kaggle.com/c/ieee-fraud-detection
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. DOI: https://doi.org/10.1145/2939672.2939785
6. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint*, arXiv:1702.08608. URL: https://arxiv.org/abs/1702.08608
7. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. DOI: https://doi.org/10.1038/s42256-019-0048-x
8. Basel Committee on Banking Supervision. (2017). *Basel III: Finalising post-crisis reforms*. Bank for International Settlements. URL: https://www.bis.org/bcbs/publ/d424.pdf
9. European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. *Official Journal of the European Union*. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj
10. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3), 50–57.
11. Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, 59(1), 5–27. DOI: https://doi.org/10.1177/0018720816681350

12. Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2), 44–58.

13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. URL: https://arxiv.org/abs/1705.07874

14. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778

15. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7), e1000097. DOI: https://doi.org/10.1371/journal.pmed.1000097

16. IEEE. (2023). *AI in financial services survey*.

17. Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *2015 IEEE Symposium Series on Computational Intelligence*, 159–166.

18. Sato, T., et al. (2020). Concept drift detection in finance. *ACM CIKM*.

19. Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. In N. Japkowicz & J. Stefanowski (Eds.), *Big Data Analysis: New Algorithms for a New Society* (pp. 91–114). Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-26989-4_4

20. Teradata. (2017). *Danske Bank Fights Fraud with Deep Learning and AI*. URL: https://assets.teradata.com/resourceCenter/downloads/CaseStudies/CaseStudy_EB9821_Danske_Bank_Saves_Millions_Fighting_Fraud_With_Deep_Learning_and_AI.pdf

21. JPMorgan Chase. (n.d.). *Explainable AI Center of Excellence*. URL: https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/explainable-ai-center-of-excellence

22. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans*, 30(3), 286–297.

23. Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, 18(1), 30–55. DOI: https://doi.org/10.1007/s10618-008-0116-z

24. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection and concept-drift adaptation with delayed supervised information. *2015 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

25. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), Article 15. DOI: https://doi.org/10.1145/1541880.1541882

26. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. DOI: https://doi.org/10.1038/nature14539

27. Financial Stability Board. (2017). *Artificial intelligence and machine learning in financial services: Market developments and financial stability implications*. URL: https://www.fsb.org/uploads/P011117.pdf

28. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012

29. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. DOI: https://doi.org/10.1038/s42256-019-0138-9

30. Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. URL: https://christophm.github.io/interpretable-ml-book

31. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31, 841–887. DOI: https://doi.org/10.2139/ssrn.3063289

32. Mastercard. (n.d.). *Decision Intelligence for Fraud and Risk Management*. URL: https://www.mastercard.com/ua/uk/business/cybersecurity-fraud-prevention/risk-decisioning/decision-intelligence.html

33. Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. URL: https://fairmlbook.org