Innovative Systems Design and Engineering
ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online)
Vol.15, No.1, 2025

A Cybersecurity Evaluation Framework for Fraud Detection: Integrating STRIDE Threat Modelling, Explainable Alerts and Anomaly Gating

Danson Gikonyo Mwarangu[1*], Shem Mbandu Angolo[2] Boniface Mwirigi Kiula[3]

1. School of Computing and Mathematics, The Cooperative University of Kenya, 24814 – 00502, Karen, Nairobi, Kenya
2. School of Computing and Mathematics, The Cooperative University of Kenya, 24814 – 00502, Karen, Nairobi, Kenya
3. School of Communication and Computer Studies, St. Paul's University, Private Bag, Limuru, Kenya

* E-mail of the corresponding author: dansongikonyo@gmail.com

Abstract

Financial fraud continues to evolve in complexity, challenging traditional detection methods. Machine learning has provided powerful tools but it remains vulnerable to adversarial manipulation, requires transparency and may operate disconnected from established cybersecurity frameworks. This study proposes a hybrid evaluation framework which combines selective STRIDE threat analysis, SHAP-based explainable alerts and an anomaly-gating mechanism that leverages on Isolation Forest scores. The study uses IEEE-CIS dataset to uncover critical vulnerabilities in financial detection such as identity spoofing and feature tampering. The Model that was used in this study integrated explainable alerts to improve analyst decision-making and operational transparency. Despite severe recall trade-offs anomaly gating effectively reduces false positives and workload demonstrating the practical difficulty of balancing precision and resilience. The results of this study highlight that effective fraud detection requires moving beyond accuracy-focused models by integrating frameworks that embed explainability, threat modeling and cybersecurity principles. This work contributes a realistic blueprint for moving fraud detection research beyond narrow accuracy metrics toward integrated, security-aware frameworks that prioritize explainability, resilience and operational integration.

Keywords: financial fraud detection, STRIDE threat modelling, explainable AI, SHAP explanations, anomaly detection, Isolation Forest, adversarial robustness, cybersecurity resilience, Security Operations Center (SOC), SIEM integration

DOI: 10.7176/ISDE/15-06
Publication date: October 31st 2025

1. Introduction

Financial fraud has evolved into a complex cybersecurity challenge for the traditional rule-based or statistical detection systems. This increasing sophistication in exploiting vulnerabilities in large-scale digital payment ecosystems require counter measures that are both adaptive and explainable for defense (Narender & Anand, 2025). Currently machine learning (ML) has been widely adopted in the banking and financial sectors for fraud detection However this deployment introduces significant challenges like adversarial manipulation of inputs, limited interpretability of model decisions and weak integration with established cybersecurity frameworks.

Adversarial attacks highlight a critical vulnerability in ML-based systems where small, carefully crafted perturbations to input features can alter predictions without detection by human analysts (Ijiga, Idoko, Ebiega, & Olajide, 2024). Such attacks have been demonstrated across various domains including this specific area of financial fraud, raising concerns about the robustness of AI defenses (Gupta, Jain, Agarwal, & Modake, 2025). Equally pressing in these ML models is the lack of transparency in complex ensemble models. This is important because it impedes regulatory compliance and might end up reducing analyst trust (Radha, Singh, Agarwal, & Bafna, 2024; Vijayanand & Smrithy, 2025). Alerts may overwhelm security operations centers (SOCs) with false positives or unexplainable outputs without interpretable justifications.

Cybersecurity research has begun to explore structured methodologies such as STRIDE to cover Spoofing, Tampering, Repudiation, Information disclosure, Denial of Service and Elevation of Privilege as systematic approaches to threat modeling in AI systems (Sharif, 2023; Demyanchuk & Yashchuk, 2025). The application of

STRIDE to financial fraud detection tries to bridge the gap between AI-driven risk scoring with established cybersecurity frameworks. At the same time studies on solation Forests, have shown promise in identifying abnormal or adversarial behaviors when integrated as a second line of defense (Raza, Ali, & Hussain, 2024; Sarker, 2024).

This research builds on these Knowledge by proposing a hybrid cybersecurity evaluation framework for financial fraud detection systems. It will not focus solely on predictive accuracy but will try to give emphasizes to systematic threat modeling of ML fraud detection systems using STRIDE, explainable alert generation through SHAP-based outputs formatted for SIEM and SOC integration, anomaly-gating mechanisms that enhance resilience against adversarial perturbations. The primary objective is to demonstrate a cohesive framework that embeds cybersecurity principles, explainable AI, and anomaly-based defenses into fraud detection. We evaluate this framework by its ability to provide operational transparency, identify systemic vulnerabilities, and enhance resilience, thereby bridging the gap between ML performance and SOC operational needs.

2. Materials and Methods

The proposed framework was evaluated using the IEEE-CIS Fraud Detection dataset, a widely recognized benchmark for financial transaction classification. Transactions were split temporally: 80% for training and 20% for validation to reflect real-world deployment. Preprocessing included type downcasting for memory optimization, median imputation for numeric features and mode imputation for categorical variables. To prevent leakage, frequency-based features (e.g., card usage counts) were derived from the training partition before being applied to validation data. Additional engineered features included transaction time derivatives, log transformations of transaction amounts, and ratio-based interaction terms between high-importance variables.

The fraud detection system was designed as a stacked ensemble comprising three tree-based classifiers which are Random Forest, XGBoost, and LightGBM. Their outputs were aggregated by an XGBoost meta-model. This reflects the diversity typically found in production fraud detection pipelines. These pipelines include multiple algorithms that contribute as complementary perspectives rather than optimizing for marginal performance gains. Synthetic Minority Oversampling Technique (SMOTE) was applied within group-aware cross-validation folds to mitigate class imbalance. SHapley Additive exPlanations (SHAP) were computed primarily on the LightGBM learner to provide global and local interpretability of predictions (Lundberg & Lee, 2017).

The STRIDE methodology was applied selectively to the fraud detection pipeline because the study followed a pragmatic approach dictated by dataset constraints and the specific threat profile of a batch-processing ML model. The Empirical analysis that was done focused on spoofing, tampering, and denial-of-service vulnerabilities, as these represent the most immediate and empirically testable threats within the transaction data. Spoofing threats were modeled by examining synthetic and compromised identities within the dataset. On the other hand tampering risks were evaluated via adversarial perturbations of transaction attributes. Finally, Denial-of-service risks were explored by simulating high volumes of borderline transactions to assess analyst workload inflation.

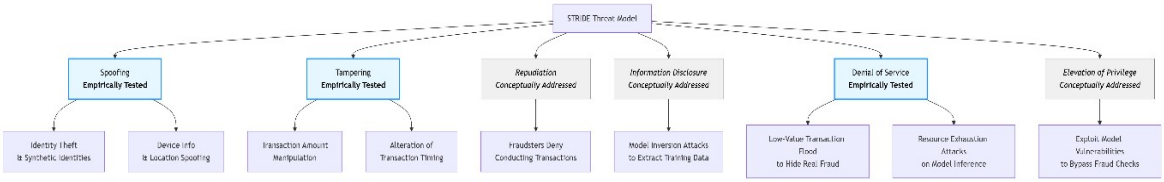

Figure 21 Stride Threat Model

The other categories repudiation, information disclosure, and privilege escalation were acknowledged but not empirically tested as shorn in Figure 1, reflecting their reliance on broader system components beyond the scope of an isolated model evaluation.

SHAP-based explanations were embedded into SIEM-compatible JSON alerts. Each alert included transaction identifiers, predicted risk scores, SHAP-based top contributing features, and reason codes written in analyst-friendly language. Such as“High transaction amount inconsistent with historical profile”. This aimed to reduce

cognitive load for SOC analysts and bridge the gap between predictive outputs and operational decision-making (Vijayanand & Smrithy, 2025).

Lastly An Isolation Forest anomaly detector was trained on transaction features to provide a secondary filter. Transactions with anomaly scores exceeding the 95th percentile were withheld from automatic classification and escalated for manual review. This anomaly-gating mechanism was intended to capture out-of-distribution or adversarial examples. However, it was recognized that this kind of defenses introduce trade-offs between precision, recall, and operational workload (Raza et al., 2024).

3. Results

The hybrid stacked ensemble model that comprised of Random Forest, XGBoost, and LightGBM base classifiers stacked via an XGBoost meta-model, was evaluated on validation set using a temporally split dataset to mimic real-world deployment. Performance metrics at the F1-optimized operating threshold (0.2661) showed an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.9040, with a 95% confidence interval between 0.8985 and 0.9095, reflecting robust discrimination capability despite the dataset's severe class imbalance. The model achieved a precision of 0.6360 (95% CI: 0.6173–0.6520) and a recall of 0.4446 (95% CI: 0.4277–0.4599), translating to an F1-score of 0.5234.

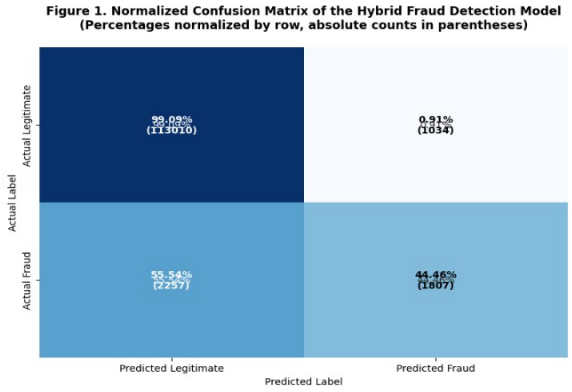

Figure 22 Confusion Matrix

As detailed in Figure 2, the confusion matrix analysis revealed that out of 118,108 transactions evaluated, the model detected 1,807 true fraud cases correctly while generating 1,034 false positive alerts., Due to the dominant presence of legitimate transactions the overall accuracy was high at 0.9721. Cross-validation analyses that was used in this work confirmed stability and reliability of these performance metrics, reinforcing confidence in the generalizability of the hybrid approach.

SHapley Additive exPlanations (SHAP) were computed on the LightGBM base learner for interpretability analyses. Feature importance rankings identified key predictors of fraud, prominently including the ratio between features C1 and C14 (contributing 53.8%), transaction amount (17.2%), and additional behavioral and card-related features. Visualizations of t SHAP values facilitated both global and local explanations. This enables transparency into individual transaction risk scores and their driving factors. Alerts generated through SHAP-based feature attributions were delivered in SIEM-compatible formats for operational consumption.

Adversarial testing clearly eposes the system's sensitivity to perturbations. With corresponding modest declines in precision, recall rates also experienced minor reductions under amount tampering (from 0.4446 to 0.4328) and feature sanitization attacks (to 0.4392). These experiments highlight the ongoing vulnerabilities tolerable within designed operational margins.

The Isolation Forest gate used in this work flagged approximately 7% of transactions as anomalous. Of these, 63% were fraudulent. This mechanism came at a cost where it was found to also withhold legitimate cases. Precision slightly dropped to 0.6091, while recall collapsed to 0.2527. Although false positives were reduced by 404, the trade-off was a loss of 732 true positives. This finding illustrates the fragility of anomaly gating where thresholds are rigid and uncalibrated.

SHAP drift metrics provided an additional layer of adversarial detection. It demonstrates the method's utility in operational threat monitoring by combining methods which detected 40.7% of the perturbed cases.

4. Discussion

These results demonstrate that while predictive performance is important it can be insufficient for evaluating fraud detection systems. The stacked ensemble provided reliable discrimination but failed to outperform a LightGBM baseline. The perfomance echoed literature that highlights diminishing returns from model complexity (Moradi et al., 2025). it is clear that the unique contribution of the framework lies not in accuracy but in its operational design, where explainability, threat modeling, and anomaly defenses are explicitly integrated.

The selective application of STRIDE confirmed that ML-based systems and in this case fraud detection on financial systems are exposed to spoofing, tampering and denial-of-service vulnerabilities. These findings support the call for embedding AI-driven fraud detection into structured cybersecurity risk models (Sharif, 2023; Demyanchuk & Yashchuk, 2025). The explainable alerting impressively translated black-box outputs into actionable intelligence thus addressing regulatory compliance and trust challenges emphasized in recent work (Vijayanand & Smrithy, 2025).

The anomaly-gating experiment revealed the steep trade-offs inherent in defensive filtering. The drastic recall drop evidently supports that a static, one-size-fits-all anomaly threshold is operationally untenable. Because while it reduced false positives the collapse in recall makes clear that the used anomaly gating may be counterproductive. Rather than a failure this outcome represents a critical empirical finding in this area where effective anomaly defenses require adaptive strategies or integration as a continuous feature, not static thresholds

5. Limitations and Future Work

Several limitations qualify the contributions of this study. As real-world fraud varies across geographies and transaction streams the exclusive reliance on the IEEE-CIS dataset restricts generalizability of this work.

The stacked ensemble developed did not outperform a LightGBM baseline as expected which reflects the bounded predictive gains of complex ensembles in imbalanced datasets. It is also important to note that STRIDE was applied selectively, focusing on spoofing, tampering, and denial-of-service, while other categories were considered only conceptually. While reducing false positives the anomaly-gating defense imposed a severe recall penalty which highlights the need for adaptive thresholds or hybrid defenses. Finally, SHAP-based explanations faces the risk of information leakage which is a dual-use problem that shows the inherent tension between explainability and security if exploited by adversaries. This requires that future work must prioritize privacy-preserving interpretability techniques. (Aljunaid et al., 2025).

Future research should extend this work by validating the framework across multiple datasets, develop anomaly defenses that are adaptive, apply STRIDE holistically to deployed systems and integrating explainability safeguards. Such extensions will be critical for advancing fraud detection frameworks from proof-of-concept to robust, real-world deployments.

6. Conclusions

This study proposed a hybrid evaluation framework for fraud detection that integrates STRIDE threat modeling, SHAP-based explainable alerts, and anomaly-gating defenses. While the ensemble classifier's performance was comparable to a strong baseline, the framework demonstrated that resilience and operational transparency of a fraud detection cannot be captured by statistical performance metrics alone. STRIDE analysis widely exposes overlooked vulnerabilities in ML based fraud detection, explainable alerts on the other hand showed potential in improving analyst workflows and anomaly gating revealed the challenging trade-offs present between workload reduction and detection coverage.

The study recommends a paradigm shift in fraud detection research. Moving away from isolated model optimization toward integrated, security-aware systems that address adversarial risks and operational realities by embedding explainability and resilience alongside machine learning. The proposed framework offers a practical blueprint for aligning fraud detection research with the complex demands of financial cybersecurity operations for effective implementation and integrations.

References

Alhashmi, A.A., Alhashmi, S.S. and Al-Mekhlafi, A.M. (2023) 'Hybrid ensemble learning approach for fraud

detection in financial transactions', *Engineering, Technology & Applied Science Research*, 13(6), pp. 6401–6407. Available at: https://doi.org/10.48084/etasr.6401

Aljunaid, M.F., Alenezi, M., Alzain, H., Alghamdi, A.S. and Oussalah, M. (2025) 'Explainable AI-driven federated learning model for financial fraud detection', *Journal of Risk and Financial Management*, 18(4), p. 179. Available at: https://doi.org/10.3390/jrfm18040179

Almalki, F. and Masud, M. (2025) *Financial fraud detection using explainable AI and stacking ensemble methods*. *arXiv preprint*. Available at: https://arxiv.org/abs/2505.10050

Demyanchuk, Y. and Yashchuk, T. (2025) 'Threat modeling of information and communication systems based on the STRIDE methodology', *Scientific Bulletin of Lviv State University of Life Safety*, 23(1), pp. 55–63. Available at: https://sci.ldubgd.edu.ua/jspui/handle/123456789/16141

Fidel, G., Bitton, R. and Shabtai, A. (2019) *When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures*. *arXiv preprint*. Available at: https://arxiv.org/abs/1909.03418

Gupta, R., Jain, J., Agarwal, A. and Modake, P. (2025) *Adversarial attacks and fraud defenses: Leveraging data engineering to secure AI models in the digital age*. *ResearchGate preprint*. Available at: https://www.researchgate.net/publication/388469709

Ijiga, S.A., Idoko, V.E., Ebiega, O.S. and Olajide, O. (2024) 'Adversarial machine learning in cybersecurity: A survey of threats and defenses', *International Journal of Information Security Science*, 13(2), pp. 55–72. Available at: https://ijirt.org/publishedpaper/IJIRT169990_PAPER.pdf

Johnson, M. (2025) 'Artificial intelligence in financial crime compliance: Balancing efficiency, explainability, and regulatory expectations', *Journal of Financial Crime*, 32(1), pp. 145–162. Available at: https://doi.org/10.1108/JFC-09-2024-0205

Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, 30, pp. 4765–4774. Available at: https://arxiv.org/abs/1705.07874

Moradi, M., Tarif, K. and Homaei, H. (2025) *Robust fraud detection with ensemble learning: A case study on the IEEE-CIS dataset*. *Preprints*. Available at: https://www.preprints.org/manuscript/202507.1711/v1

Narender, M. and Anand, A.J. (2025) 'Artificial Intelligence in Financial Fraud Detection', in *Handbook of AI-Driven Threat Detection and Prevention*. CRC Press, pp. 193–207. Available at: https://doi.org/10.1201/9781003521020

Radha, R., Singh, R., Agarwal, S. and Bafna, R. (2024) 'Explainable machine learning approaches in cybersecurity defense systems', in Kumar, M. and Gupta, A. (eds.) *Handbook of Artificial Intelligence in Cybersecurity*. Springer, pp. 215–229. Available at: https://doi.org/10.1007/978-981-96-3358-6_14

Raza, M., Ali, F. and Hussain, M. (2024) 'Machine learning-based anomaly detection for cybersecurity defense', *Security and Safety*, 3(2), pp. 102–118. Available at: https://www.researchgate.net/publication/393440976_Machine_Learning-Based_Anomaly_Detection_for_Cyber_Threat_Prevention

Sarker, I.H. (2024) 'Machine learning for cybersecurity: Threat detection, adversarial defense and future directions', *Journal of Network and Computer Applications*, 236, p. 103683. Available at: https://doi.org/10.1016/j.jnca.2024.103683

Sharif, A. (2023) *Threat modeling for AI-based systems: Applying STRIDE methodology*. Bachelor thesis, University of Zurich. Available at: https://files.ifi.uzh.ch/CSG/staff/vonderassen/extern/theses/ba-sharif.pdf

Vijayanand, K. and Smrithy, M. (2025) 'Explainable AI-enhanced ensemble learning for mobile financial fraud detection', *International Journal of Distributed Sensor Networks*, 21(3), pp. 1–15. Available at: https://doi.org/10.1177/18724981241289751