Science Discovery Artificial Intelligence
2026, Vol. 1, No. 1, pp. 57–63
https://doi.org/10.11648/j.sdai.20260101.17

SciencePG
Science Publishing Group

Research Article

# Operational Security Assessment of Machine Learning Fraud Detection Systems: A Cybersecurity Perspective Using Stride, Explainability, and Anomaly Gating

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

[1]Department of Computer Science and Information Technology, The Cooperative University of Kenya, Nairobi, Kenya
[2]Department of Computer Science, St. Paul's University, Nairobi, Kenya

## Abstract

Machine learning (ML) enables large-scale analysis of transaction data and has become integral to financial fraud detection. Despite strong predictive performance, ML-based systems remain vulnerable to adversarial manipulation and are often insufficiently aligned with cybersecurity evaluation practices. This paper introduces an operational security assessment framework that shifts the focus from model optimisation toward holistic security evaluation. The framework combines STRIDE threat modelling to systematically identify vulnerabilities such as spoofing, tampering, and denial-of-service. It uses SHapley Additive exPlanations (SHAP) to embed contextual, SOC-ready alerts into Security Information and Event Management (SIEM) workflows and an anomaly-gating mechanism using Isolation Forest to assess resilience against adversarial and out-of-distribution samples. Using the IEEE-CIS dataset as a case study, the framework revealed susceptibility to identity spoofing, sensitivity to targeted feature perturbations and operational bottlenecks under simulated denial-of-service conditions. For Anomaly gating though it reduced false positives and captured adversarial manipulations it also imposed significant recall trade-offs, underscoring the challenge of balancing detection coverage with workload reduction. Embedding SHAP into structured alerts improved interpretability and supported drift-based anomaly identification. The study concludes that effective fraud detection requires moving beyond accuracy-centric evaluation toward integrated methodologies combining threat modelling, explainability and resilience testing. The proposed framework provides a structured blueprint for strengthening the operational security and trustworthiness of ML-driven fraud detection systems.

## Keywords

Cybersecurity, Machine Learning, Explainable AI, Threat Modelling, Fraud Detection, Operational Security, Anomaly Detection

## 1. Introduction

Machine learning (ML) offers powerful capabilities to an-alyse large-scale, high-dimensional data in real time. With all these capabilities it has become a cornerstone of fraud de-tection [6, 9]. This has led to the development of models

*Correspondence: Danson Gikonyo Mwarangu (dansongikonyo@gmail.com)

**Received:** 9 October 2025; **Accepted:** 28 February 2026; **Published:** 12 March 2026

that deliver high predictive accuracy. However, these models remain vulnerable to a series of security and operational challenges that compromise their reliability in adversarial environments. The concern lies in their susceptibility to ad-versarial manipulation where carefully engineered per-turba-tions to transaction features can cause misclassifications without being detectable by human analysts. This now raises questions about the robustness of these models and in this case fraud detection pipelines [7, 11]. Most of fraud detection models are ensemble and the opacity of complex ensemble models limits transparency. This transparency is critical for both analyst trust and regulatory compliance [1]. This is im-portant because most of these models are integrated in Se-curity operations systems and without clear interpretations alerts often overwhelm security operations centres (SOCs) with false positives or non-actionable outputs.

Structured threat modelling methodologies provide sys-tematic ways to identify vulnerabilities across a range of categories including spoofing, tampering, denial-of-service, and privilege escalation [22]. Similarly, explainable artificial intelligence (XAI) techniques have been proposed to enhance interpretability in these models [14]. Financial fraud detection requires more to this, they require integration of interpre-ta-bility into SOC workflows but studies rarely focus on these aspects [3]. Finally, though their trade-offs in fraud detection pipelines are poorly understood anomaly detection methods offer promising resilience against adversarial behaviours because they were originally developed for general cy-berse-curity applications [13].

This paper explores these gaps by applying an operational security assessment framework for ML-based fraud detection. Instead of focusing on model optimisation, the framework evaluates fraud detectors through the application of STRIDE threat modelling to identify systemic vulnerabilities, the gen-eration of explainable, SOC-ready alerts using SHAP values and finally, the integration of an anomaly-gating mechanism to test resilience against adversarial or out-of-distribution samples. While providing practical insights into analyst workflows and adversarial robustness the applied framework demonstrates how security evaluation can uncover weak-nesses overlooked by accuracy-focused benchmarks. The IEEE-CIS fraud detection dataset was used as a structured case study to operationalise and evaluate the framework.

This work contributes a pragmatic blueprint for financial in-stitutions seeking to evaluate and strengthen the operational security of their fraud detection pipelines. This is done by by situating ML-based fraud detection within the broader con-text of cybersecurity risk management. The objective here is to advance a methodology for assessing the resilience, transpar-ency, and usability of fraud detection systems in adversarial settings not just. another high-performing predic-tive model.

## 2. Background and Related Work

Research on fraud detection has long focused on improving the predictive performance of algorithms. A lot of focus is placed in achieving higher recall and precision in highly im-balanced datasets. This reinforced a narrow emphasis on metrics that do not adequately capture the operational re-ali-ties of deploying fraud detection systems in adversarial en-vironments [5, 9]. This brings the need of these models to be evaluated on ML-based security threats, interpretability re-quirements and resilience challenges that arise once they are integrated into financial institutions' workflows [5, 7]. Threat modelling, explainable artificial intelligence (XAI), and anomaly-based resilience are the central focus of the pro-posed framework.

### 2.1. STRIDE Threat Modelling and AI Security

Threat modelling provides structured methods for iden-ti-fying, categorising, and prioritising risks in complex sys-tems. Among the most widely applied threat modeling ap-proaches is STRIDE which wa developed by Microsoft. It classifies threats into six categories of spoofing, tampering, repudiation, information disclosure, denial of service and elevation of privilege [22]. its systematic coverage of both technical and operational vulnerabilities makes it a suitable approach. STRIDE has been extensively applied to software and net-work infrastructures [24] but its application to ma-chine learning pipelines appears promising. MITRE ATLAS are part Emerging frameworks that demonstrate attempts to map adversarial tactics to AI-specific vulnerabilities [15]. Our work extends this conversation by showing how STRIDE can be adapted to the evaluation of ML-based financial fraud de-tectors. The focus will be on transaction-level vulnerabili-ties such as identity spoofing and feature manipulation.

### 2.2. Explainable AI for Security Operations

Explainable artificial intelligence has emerged as a re-sponse to the opacity in machine learning models which a persistent obstacle in fraud detection. They offer methods such as Local Interpretable Model-Agnostic Explanations (LIME) [20] and SHapley Additive exPlanations (SHAP) [14] where analysts require justifications for alerts in order to make informed decisions. Studies in cybersecurity emphasise that explana-tions must be actionable and tailored to the workflows of human analysts rather than presented as raw feature weights [3]. Our study embedds SHAP-derived reason codes into alerts that are formatted for SIEM integration. This clearly demonstrates how explainability can support opera-tional decision-making and reduce the cognitive burden of fraud analysts.

### 2.3. Anomaly Detection and Adversarial Resilience

Anomaly detection has been explored as a means of iden-tifying outliers or unusual behaviours that may indicate

fraud or adversarial activity this is in parallel with predictive mod-elling, Recently Isolation Forest has gained particular traction due to its ability to isolate anomalies in high-dimensional data by recursively partitioning the feature space [13]. Anoma-ly-based methods have been used to detect network intrusions and insider threats [2, 8] but their utility in fraud detection has often been constrained by the trade-offs between false posi-tives, recall, and workload for human reviewers [11]. More re-search has suggested its potential role as secondary defences against adversarial manipulation [7, 17]. Our framework builds on this literature by positioning anomaly detection as a resilience mechanism that acts as a gate to flag suspicious transactions for manual review. This is to strengthen the system against out-of-distribution and ad-versarial samples.

While STRIDE provides a structured lens for analysing system vulnerabilities, explainable AI enables trust and usa-bility, and anomaly detection adds resilience, each has been treated largely in isolation within fraud detection studies. This fragmentation creates a gap for integrated methodologies that combine these elements into cohesive evaluation frameworks. Our contribution is to bridge this gap by pro-posing and testing such a framework in the context of finan-cial fraud detection.

## 3. Methodology

The proposed framework was implemented and evaluated using the IEEE-CIS Fraud Detection dataset. The dataset is a widely recognised benchmark containing anonymised trans-action records with diverse feature types. To replicate the deployment conditions of a real fraud detection system, we employed a temporal data split, allocating 80% of the trans-actions for training and the remaining 20% for validation. This was reducing the risk of leakage and improving the re-alism of evaluations by ensuring that future information was not inadvertently introduced into the training process. Pre-processing involved type downcasting to reduce memory overhead, median imputation for numerical features, and mode imputation for categorical variables. Frequency-based features, such as card usage counts, were engineered ex-clu-sively from the training set and then applied to the vali-dation data. This reflects a production pipeline where features must be derived without knowledge of future data. Additional transformations included ratio-based interactions between key attributes, temporal derivatives to capture behavioural pat-terns and log transformations of skewed distributions.

The detection layer was constructed as a stacked ensemble of Random Forest, XGBoost, and LightGBM classifiers. The stacking used an XGBoost meta-model to aggregate their outputs. This ensemble design reflected the diversity com-monly seen in operational fraud detection pipelines. This is where multiple algorithms are deployed in tandem to cap-ture different decision boundaries. Synthetic Minority Over-sampling Technique (SMOTE) was used during training within group-aware cross-validation folds to address severe class imbalance without contaminating the temporal split. The model's predictions were subsequently explained using SHapley Additive exPlanations (SHAP) values. For this study the values were computed primarily on the LightGBM base learner due to its balance of accuracy and computational ef-ficiency.

Rather than applying STRIDE generically, we adapted it to the specific context of a fraud detection pipeline. By giving attention to features such as card identifiers Spoofing threats were modelled by analysing synthetic or compromised iden-tities within the dataset. Tampering risks were evaluated by introducing adversarial perturbations to sensitive features such as reducing transaction amounts or modifying fre-quency-based variables, and observing the resulting changes in predictions. Denial-of-service was explored through simu-lated floods of borderline transactions to assess the capacity of the system and its impact on analyst workload. The re-maining categories such as repudiation, information disclo-sure, and privilege escalation were acknowledged but not empirically tested. This is because they depend on broader infrastructural elements beyond the scope of the dataset. This selective application of STRIDE allowed for a pragmatic, evidence-driven analysis while maintaining alignment with established cybersecurity practices see Figure 1.

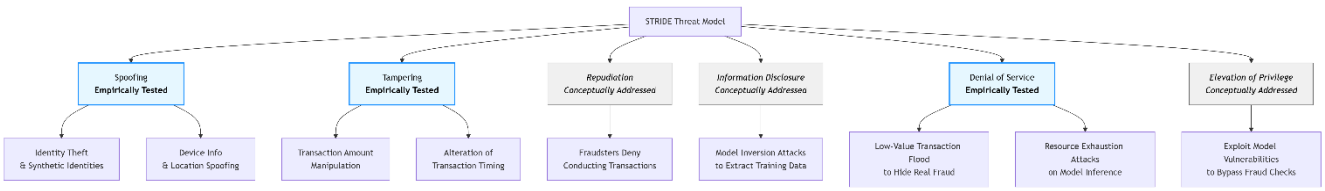


***Figure 1.*** *Stride Threat Model.*

To operationalise explainability, SHAP outputs were em-bedded into alerts formatted in JavaScript Object Notation (JSON) see Figure 2. This is a widely used format in Security Information and Event Management (SIEM) systems. Each alert contained transaction identifiers, the predicted fraud probability, the top contributing features with their SHAP values and a concise reason code translated into hu-man-readable language. For example, a transaction with

unusually high spending inconsistent with prior behaviour would generate an alert with the reason code "Transaction amount inconsistent with customer history." This trans-for-mation was intended to reduce the cognitive burden on an-alysts by contextualising model outputs within operational workflows.

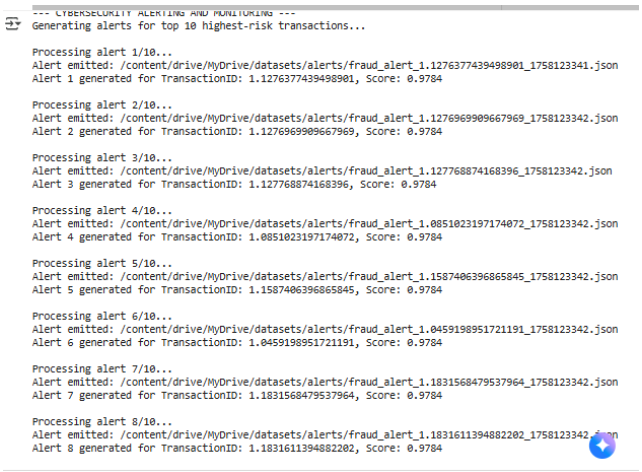


***Figure 2.*** *Alerts formatted in JSON.*

Anomaly-gating mechanism was implemented using an Isolation Forest model trained on the same transaction fea-tures. Transactions scoring above the 95th percentile in anomaly scores were withheld from automatic classification and instead flagged for manual review. This gating mecha-nism was designed to capture out-of-distribution cases and adversarial manipulations that might otherwise evade the predictive model. the framework created a two-tiered defence by layering anomaly detection on top of the ensemble pre-dictions. The first layer was to focus on predictive classifica-tion and the second acted as a safeguard against unusual or manipulated patterns. While acknowledging the trade-offs between recall and analyst workload, this design emphasised resilience and operational applicability.

## 4. Results

The stacked ensemble detector at the operating threshold optimised for the F1 score (0.2661), the model achieved an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.9040, with a 95% confidence interval ranging from 0.8985 to 0.9095. Precision was measured at 0.6360 (95% CI: 0.6173–0.6520), while recall reached 0.4446 (95% CI: 0.4277–0.4599), yielding an overall F1-score of 0.5234. Out of 118,108 evaluated transactions 1,807 fraudulent cases were correctly identified. This was accompanied by 1,034 false positive alerts. The overall accuracy remained high at 0.9721 this was due to the overwhelming majority of legitimate cases in the dataset. Cross-validation confirmed the stability of these performance levels because only minor fluctuations observed across folds.

The application of STRIDE to the pipeline produced clear evidence of specific vulnerabilities. Spoofing risks were apparent in the form of synthetic or compromised identities, revealed through frequency analyses of identity-linked. Tampering was confirmed by adversarial perturbations this was because when transaction amounts were manipulated, the model's predictions shifted significantly. It demonstrated sensitivity to small but targeted feature changes. Denial-of-service threats were simulated by introducing high volumes of borderline transactions. It did not affect predictive metrics directly but resulted in inflated queues of cases for review. The result was pointing to potential operational bottlenecks if exploited at scale. The remaining categories were not empirically testable within the dataset but were retained in the framework for conceptual completeness.

Interpretability was evaluated by embedding SHAP values into SIEM-compatible alerts. Feature attribution consistently highlighted the ratio of key behavioural variables, the transaction amount and card-related identifiers as the strongest predictors of fraud. These outputs were presented as structured JSON objects containing transaction identifiers, predicted risk, top contributing features and reason codes written in analyst-friendly term. Alerts citing "Unusual ratio between spending behaviour features C1 and C14" or "High transaction amount inconsistent with user profile" were generated automatically. Analyst can be able to triage based on context rather than binary classification alone. This is because Across the validation set 1,034 alerts previously categorised as false positives were enriched with explanatory details. This was found to streamline review processes by highlighting underlying drivers of the flagged predictions.

The anomaly-gating defence flagged approximately 7% of validation cases as anomalous. Of these, 63% were confirmed fraudulent and its ability to identify out-of-distribution cases. However, the gating also removed 3% of legitimate cases from automatic classification that resulted to a drop in recall to 0.2527 while precision declined slightly to 0.6091. This meant that while 404 false positives were eliminated, 732 true frauds were also missed. Despite these trade-offs, the approach was effective in mitigating adversarial perturbations this was because the manipulated inputs in particular tampering of transaction amounts were disproportionately captured within the anomalous group, reducing their success in bypassing the primary model. This quantitative contrast illustrates the central operational dilemma where resilience mechanisms can reduce analyst workload and mitigate adversarial risk, but at the cost of detection coverage. Rather than optimising purely for predictive gain, the framework makes this trade-off explicit, enabling informed deployment decisions depending on institutional risk tolerance.

```
SHAP Drift Detection Performance:
  Detected perturbations (Spearman): 0.2490
  Detected perturbations (L2): 0.2500
  Detected perturbations (Combined): 0.4052
```

***Figure 3.*** *SHAP drift metrics.*

SHAP drift metrics provided an additional perspective on resilience. Approximately 40.7% of adversarial manipulations were detected as abnormal. This was done by comparing SHAP value distributions across clean and perturbed samples as shown below in Figure 3.

## 5. Discussion

The results confirmed that the proposed framework not only matched expected predictive baselines but also produced actionable insights into threats, interpretability and opera-tional resilience. They demonstrate the necessity of evaluating financial fraud detection systems through a cybersecurity lens rather than solely through predictive metrics. While the stacked ensemble achieved an AUC of 0.9040, its perfor-mance did not surpass that of simpler baselines reported in literature. This reinforces the observation that increased model complexity does not always translate to superior dis-crimination in highly imbalanced domains such as fraud detection [16], particularly when adversarial robustness and operational integration are considered. It clearly explains the limited value of pursuing marginal accuracy gains without addressing systemic vul-nerabilities and operational challenges.

The selective application of STRIDE provided a structured means of uncovering threats that are rarely considered in conventional ML evaluations. Spoofing risks that are evident in the dataset align with real-world scenarios where fraudsters employ synthetic or stolen identities [4]. The successful pertur-bation of transaction features further confirmed tampering vulnerabilities. This supports recent findings that adversarial manipulation of financial features can significantly alter classifier behaviour [12]. From the simulated denial-of-service conditions it was revealed that even models with stable pre-dictive performance are susceptible to operational overload, a factor that directly affects the efficiency of Security Opera-tions Centers (SOCs). These insights position STRIDE not merely as a conceptual tool but as a practical framework for highlighting the systemic risks of ML-driven fraud detection.

Embedding SHAP-based explanations into SIEM alerts illustrated the potential for bridging technical predictions with analyst workflows. False positives are inevitable in imbal-anced settings but this bridge become more manageable when accompanied by contextual explanations. This is consistent with prior research emphasising that explainability enhances analyst trust and reduces the cognitive burden associated with reviewing large volumes of alerts [18, 23]. The framework con-tributes to ongoing efforts to operationalise explainable AI within financial cybersecurity environments by demonstrat-ing that alerts can be transformed into actionable intelligence.

The anomaly-gating experiments highlighted a critical trade-off between reducing analyst workload and maintain-ing fraud detection coverage. While the gate successfully reduced false positives and captured adversarial manipula-tions, it did so at the expense of recall, eliminating a substan-tial number of genuine fraud cases. This finding suggests that static thresholds are inadequate for real-world deployment and that adaptive anomaly-based mechanisms are required to balance precision and coverage dynamically [19, 21]. Rather than a failure, this outcome is an important empirical con-tribution since it demonstrates that naïve implementations of anomaly filters may be counterproductive serving as a warning for practitioners seeking to integrate such defences into fraud detection pipelines. All in all these demonstrated the defensive potential of anomaly gating, albeit with a clear cost to detection coverage.

SHAP drift analysis revealed its dual role in both interpretability and resilience. The detection of 40.7% of pertur-bations, although modest, illustrates that explainability methods can serve as lightweight adversarial detectors. This complements findings from adversarial ML research where model explanations expose inconsistencies caused by manipulations [10]. While insufficient as a standalone defence it provided a foundation for multi-layered detection strategies consistent with defence-in-depth principles thus supporting the notion that interpretability methods can serve both ex-planatory and defensive purposes.

From a scalability perspective, the framework introduces additional computational layers beyond conventional fraud detection pipelines. The integration of SHAP explanations and anomaly gating increases inference overhead compared to standalone classifiers. In high-throughput financial environments processing millions of transactions daily, this may necessitate distributed computation or asynchronous explanation generation. Furthermore, embedding SHAP outputs into SIEM-compatible alerts requires structured logging pipelines and governance controls to prevent information leakage. While these considerations do not invalidate the framework, they highlight that operational adoption requires infrastructure alignment and careful threshold tuning to balance latency, workload and security objectives.

Overall, this discussion affirms that the strength of the proposed framework lies not in delivering higher accuracy but in offering a structured cybersecurity evaluation of fraud detection pipelines. This shift in focus from accuracy opti-misation to holistic evaluation represents a meaningful con-tribution to the ongoing discourse on secure and trustworthy AI in financial cybersecurity.

## 6. Conclusion

The key contribution of this work is not the proposal of a

new model but the operationalisation of a structured evalua-tion methodology that connects machine learning outputs with cybersecurity principles. This provides a practical blue-print for shifting fraud detection research from a narrow emphasis on predictive accuracy to a broader concern with resilience, transparency and system-level robustness.

The evaluation was conducted using the IEEE-CIS dataset as a structured benchmark. While this dataset provides sufficient complexity for methodological validation, further empirical validation across multiple institutional datasets would strengthen generalisability. Similarly, adversarial testing focused on controlled feature perturbations and simulated workload stress; broader adversarial scenarios, including adaptive multi-step attacks, remain outside the scope of this study. These extensions represent important avenues for future research rather than omissions in the present framework.

Adaptive anomaly detection strategies represent a critical next step to mitigate the severe trade-offs observed in this study. Equally important is the need to explore priva-cy-preserving explainability techniques which can provide transparency for analysts without exposing sensitive model information to adversaries. Finally, expanding STRIDE analysis to empirically test unerexplored in this study would enhance the completeness of the security evaluations.

This framework contributes to building fraud detection systems that are not only accurate but also secure, inter-pretable and operationally viable. It tries to advance the integration of threat modeling, explainability and anoma-ly-based defences.

## Abbreviations

| | |
|---|---|
| ML | Machine Learning |
| AI | Artificial Intelligence |
| XAI | Explainable Artificial Intelligence |
| SOC | Security Operations Center |
| SIEM | Security Information and Event Management |
| SHAP | SHapley Additive exPlanations |
| STRIDE | Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege |
| IEEE-CIS | Institute of Electrical and Electronics Engineers – Computational Intelligence Society |
| DoS | Denial of Service |

## Author Contributions

**Danson Gikonyo Mwarangu:** Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing

**Shem Mbandu Angolo:** Methodology, Supervision

**Boniface Mwirigi Kiula:** Supervision, Validation

## Data Availability Statement

The data that support the findings of this study can be found at: https://www.kaggle.com/search (a publicly available repository url).

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

[1] Adadi, A. & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31. https://doi.org/10.1016/j.jnca.2015.11.016

[3] Aldeen, Y. A. A., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. SpringerPlus, 4, 694. https://doi.org/10.1186/s40064-015-1481-x

[4] Alhashmi, A. A., Alhashmi, S. S., & Al-Mekhlafi, A. M. (2023). Hybrid ensemble learning approach for fraud detection in financial transactions. Engineering, Technology & Applied Science Research, 13(6), 6401-6407. https://doi.org/10.48084/etasr.6401

[5] Bauder, R. A., Khoshgoftaar, T. M., Kemp, C., & Seliya, N. (2018). A survey on the state of healthcare upcoding fraud analysis and detection. Health Services and Outcomes Research Methodology, 18, 62-83. https://doi.org/10.1007/s10742-017-0180-1

[6] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602-613. https://doi.org/10.1016/j.dss.2010.08.008

[7] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331. https://doi.org/10.1007/s10742-016-0154-8

[8] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1-58. https://doi.org/10.1145/1541880.1541882

[9] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection and concept-drift adaptation with delayed supervised information. 2017 International Joint Conference on Neural Networks (IJCNN), 1-8. https://doi.org/10.1109/TNNLS.2017.2736643

[10] Fidel, G., Bitton, R., & Shabtai, A. (2019). When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures. arXiv preprint. https://arxiv.org/abs/1909.03418

[11] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6572

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

[13] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. 2008 Eighth IEEE International Conference on Data Mining, 413-422. https://doi.org/10.1109/ICDM.2008.17

[14] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765-4774. https://arxiv.org/abs/1705.07874

[15] MITRE. (2023). MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. Retrieved from https://atlas.mitre.org

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

[17] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 372-387. https://doi.org/10.1109/EuroSP.2016.36

[18] Radha, R., Singh, R., Agarwal, S., & Bafna, R. (2024). Explainable machine learning approaches in cybersecurity defense systems. In M. Kumar & A. Gupta (Eds.), Handbook of Artificial Intelligence in Cybersecurity (pp. 215-229). Springer. https://doi.org/10.1007/978-981-96-3358-6_14

[19] Raza, M., Ali, F., & Hussain, M. (2024). Machine learning-based anomaly detection for cybersecurity defense. Security and Safety, 3(2), 102-118. https://publishing.emanresearch.org/Journal/FullText/5955

[20] 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 (KDD), 1135-1144. https://doi.org/10.1145/2939672.2939778

[21] Sarker, Iqbal H. AI-Driven Cybersecurity and Threat Intelligence. Springer Nature Switzerland, 2024. https://doi.org/10.1007/978-3-031-54497-2

[22] Shostack, A. (2014). Threat Modeling: Designing for Security. Wiley.

[23] Vijayanand, D., & Smrithy, G. S. (2025). Explainable AI-enhanced ensemble learning for financial fraud detection in mobile money transactions. Intelligent Decision Technologies, 19(1), 52-67. https://doi.org/10.1177/18724981241289751

[24] Xiong, W., & Lagerström, R. (2019). Threat modeling - A systematic literature review. Computers & Security, 84, 53-69. https://doi.org/10.1016/j.cose.2019.03.010

## Biography

**Danson Gikonyo Mwarangu** is a cybersecurity researcher, technology trainer, and a developer who has recently worked on machine learning security in fraud detection model. He recently completed his Master of Science in Cyber Security in 2025 at The Co-operative University of Kenya, where his research focused on developing a leakage-safe hybrid fraud detection model for financial institutions by integrating supervised ensemble learning, anomaly detection, explainable artificial intelligence, and structured threat modeling. Beyond research, he is actively involved in technology training and practical learning.

## Research Field

**Danson Gikonyo Mwarangu:** machine learning, machine learning security, AI, adversarial machine learning, explainable artificial intelligence, threat modeling for machine learning, data leakage prevention techniques.

**Shem Mbandu Angolo:** wireless sensor networks, cybersecurity and cryptography, internet of things security, machine learning applications, data science and analytics, digital forensic readiness, electronic learning systems, health informatics systems, ubiquitous healthcare security.

**Boniface Mwirigi Kiula:** information and communication technology policy, digital leadership and governance, quality management systems, organizational performance management, strategic human capital development, electronic readiness in higher education, institutional digital transformation, health informatics systems, data analytics and business intelligence.