HYBRID FRAUD DETECTION MODEL FOR FINANCIAL INSTITUTIONS

DANSON GIKONYO MWARANGU

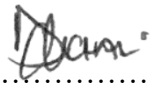
A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY IN THE SCHOOL OF COMPUTING AND MATHEMATICS IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTERS OF SCIENCE IN CYBER SECURITY OF THE CO-OPERATIVE UNIVERSITY OF KENYA

2025

DECLARATION

Declaration by the candidate

This proposal/thesis is my original work and has not been presented for award of a degree in any other University or for any other award



……. ……03/10/2025……

Signature Date

Danson Gikonyo Mwarangu C005/600032/2023

Declaration by the supervisors

I/We confirm that the work reported in this proposal/thesis was carried out by the candidate under our supervision and has been submitted with our approval as university supervisors

…… ……03/10/2025……

Signature Date

Dr Shem Mbandu Angolo, DCSIT, School of Computing and Mathematics, The Cooperative University of Kenya

…… ……03/10/2025……

Signature Date

Dr Boniface Mwirigi Kiula, School of Communication and Computer Studies, St. Paul's University, Kenya

DEDICATION

I dedicate this work to my loving parents, the late Mr. James Duncan Mwarangu and Mrs. Lucy Wainoi, and sisters Margaret Wangechi and Monica Nkirote.

ACKNOWLEDGMENT

I wish to express my sincere appreciation to my supervisors, Dr. Shem Mbandu and Dr. Mwirigi Kiula. Their guidance, unwavering support and intellectual input were instrumental in shaping this research and bringing this work to completion. I am also grateful to my classmates and colleagues for their invaluable insights and shared knowledge.

TABLE OF CONTENTS

List of Tables

List of Figures

List of abbreviations and acronyms

List of Abbreviations and Acronyms

ACFE - Association of Certified Fraud Examiners

AUC - Area Under the Curve

AML - Adversarial Machine Learning

ART - Adversarial Robustness Toolbox

CBK - Central Bank of Kenya

CNN - Convolutional Neural Network

DoS - Denial of Service

FGSM - Fast Gradient Sign Method

GAN - Generative Adversarial Network

GDPR - General Data Protection Regulation

GDP - Gross Domestic Product

IEEE-CIS - Institute of Electrical and Electronics Engineers - Computational Intelligence Society

IF - Isolation Forest

LIME - Local Interpretable Model-agnostic Explanations

LSTM - Long Short-Term Memory

ML - Machine Learning

NACOSTI - National Commission for Science, Technology and Innovation

NIDS - Network Intrusion Detection System

OOD - Out-of-Distribution

PR-AUC - Precision-Recall Area Under the Curve

PSI - Population Stability Index

ROC - Receiver Operating Characteristic

SHAP - Shapley Additive explanations

SIEM - Security Information and Event Management

SMOTE - Synthetic Minority Oversampling Technique

SOC - Security Operations Center

STRIDE - Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege

SVM - Support Vector Machine

XAI - Explainable Artificial Intelligence

Definition of Terms

1. Hybrid Fraud Detection Model - A machine learning framework that integrates multiple analytical approaches, specifically combining supervised ensemble methods with unsupervised anomaly detection.

2. Adversarial Machine Learning (AML) - A field of study that focuses on the vulnerabilities of machine learning models to deliberate manipulation by adversaries (Biggio & Roli, 2018).

3. Explainable AI (XAI) - refers to a set of tools and methods that help us understand how a machine learning model arrives at its decision- (Adadi & Berrada, 2018).

4. Shapley Additive explanations (SHAP) -one of the most powerful tools we have for explaining individual predictions made by a machine learning model using a game theory-based (Lundberg & Lee, 2017).

5. STRIDE Threat Modeling – A framework developed by Microsoft with six categories for identifying and categorizing security threats (Shostack, 2014).

6. Anomaly Gating - A filter or "gate like defensive mechanism to identify and block out-of-distribution or adversarially perturbed transactions before they are processed. (Papernot et al., 2016).

7. Leakage-Safe Methodology - A protocol for systematic preprocessing and feature engineering designed to prevent data leakage.

8. Operational Security & Effectiveness - the measure of a fraud detection system in a real-world setting.

9. Stacked Ensemble - An advanced ensemble machine learning technique where predictions from multiple base models are used as input features for a final meta-model.

Abstract

Digitization has significantly expanded access to financial services while simultaneously increasing the exposure of financial platforms to fraud. Institutions globally are adopting machine learning (ML) to detect emerging fraud patterns, yet many detection systems remain evaluated almost exclusively through predictive metrics, limiting their reliability in real-world adversarial environments. This study develops and evaluates a hybrid fraud detection framework that integrates supervised ensemble learning with unsupervised anomaly detection, addressing both predictive performance and operational security gaps.The research uses the publicly available IEEE-CIS Fraud Detection Dataset from Kaggle, comprising over 590,000 transactions with both transaction and identity attributes. Data preprocessing followed a leakage-safe protocol that included temporal splitting based on TransactionDT (80% training, 20% validation), GroupKFold cross-validation using customer-identity features, and feature-engineering techniques restricted to the training folds to prevent leakage. The supervised layer consists of a stacked ensemble combining Random Forest, LightGBM and XGBoost as base learners with an XGBoost meta-model trained on out-of-fold predictions. To complement the supervised layer, an Isolation Forest anomaly-gating mechanism was incorporated to detect out-of-distribution and potentially adversarial transactions. SHAP explainability was integrated to generate local and global feature attributions, improving operational transparency for Security Operations Center (SOC) analysts. Model performance was evaluated using standard fraud-detection metrics including AUC-ROC, precision, recall and F1-score. The hybrid model achieved an AUC-ROC of 0.904, outperforming the baseline single-model learners implemented during experimentation. It also achieved PR-AUC of 0.5192 on the temporal validation set and at the F1-optimised threshold of 0.2661 precision was 0.6360, the recall was at 0.4446 while the F1-Score was at 0.5234. SHAP explanations revealed the dominant influence of identity-linked features and amount-related attributes which enabled clearer interpretation. STRIDE threat modeling was applied to assess the cybersecurity posture of the full pipeline. The analysis identified vulnerabilities related to spoofing, tampering, information disclosure, and denial-of-service, highlighting risks typically overlooked in predictive-only evaluation frameworks. Adversarial-resilience tests showed that anomaly gating improved robustness by filtering suspicious inputs, although this introduced a measurable reduction in recall for borderline fraud cases. Despite this trade-off, the integrated framework demonstrated a balanced blend of predictive performance, interpretability, and security-oriented evaluation. The study concludes that hybrid architectures enriched with threat modeling and explainability offer a more realistic assessment of fraud detection systems operating in adversarial environments, making the proposed framework suitable for further adaptation within financial institutions.

CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

As the world moves to the digital space, the financial sector has not been left behind. The game is changing and more people can now access financial services from anywhere in all parts of the world. Ways for Creation, transfer and protection of money have also changed because anyone can now access financial services from e-commerce ecosystems, mobile money services and other digital platforms. With this kind of accessibility attackers now have new areas to exploit one of them being fraud. Financial fraud is one of the threats that is constantly facing the financial institutions worldwide. it is reported that organizations worldwide lose almost USD 4.7 trillion annually through fraud this is almost 5% of the revenues the get yearly (ACFE, 2022). In more than two years, financial institutions being most affected, 46% of organizations in the world had an experience of fraud this is from PwC (2022)'s report.

The increase in the number of transactions is forcing institutions all over the word advance their fraud detection mechanism. But this effort is not enough since they are still facing these challenges. Most of these institutions are running to machine learning (ML) for help. They use models to find fraud that can look at a lot of crime data and find any complicated trends that could point to fraud. A lot of research studies use classification metrics to measure how well their scam detection models work. This is clear from the work of Bhattacharyya et al. (2011) This thought alone isn't enough to cover how hard it is to detect fraud. Many security issues must be taken into account when fraud detection systems are deployed in a real world environment.

Kenya has become a world leader in mobile banking services thanks to new apps like M-Pesa,

Airtel Money, and Equitel. The tool has changed the financial world because it has made it possible for people all over the country to use financial services. In this area, as in any other, more people using it has led to new ways to fight and commit Fraud. This is very clear because the Central Bank of Kenya (CBK, 2023) says that mobile banking theft is on the rise and that billions of shillings are lost every year because of it. Kenyan systems that use machine learning should make sure that security features like cybersecurity robustness, interpretability, and threat models are built in (Alhashmi et al., 2023)

1.2 Statement of the Problem

Financial fraud continues to evolve in complexity as digital financial services expand, exposing institutions to increasingly sophisticated forms of attack. Although machine learning has become a widely adopted tool for detecting fraudulent transactions, most existing studies evaluate their models primarily through predictive performance metrics such as accuracy or AUC. While useful, these metrics alone do not reflect how a model behaves when deployed in real operational environments where adversarial manipulation, data shifts, or input irregularities are common. As a result, models may achieve strong predictive scores during experimentation yet remain vulnerable to subtle evasion techniques or operational weaknesses when implemented in practice.

A second challenge lies in the limited attention given to interpretability. Many fraud detection models operate as black-box systems, providing little insight into the reasoning behind their alerts. This lack of transparency complicates the work of Security Operations Center analysts, who must justify decisions, manage false positives, and respond to threats in real time. Without clear explanations, technically accurate models may be difficult to integrate into daily operational workflows, reducing their practical value.

Furthermore, fraud detection pipelines often lack structured cybersecurity assessment during development. Threats related to spoofing, tampering, information disclosure, and other systemic vulnerabilities may remain unidentified when evaluation is restricted to predictive performance. This gap exposes financial institutions to risks that traditional predictive validation cannot detect.

In the Kenyan financial sector, where digital transactions have expanded rapidly and fraud incidents continue to rise, these limitations are especially consequential. Despite the deployment of machine-learning-based fraud detection systems, institutions still report significant losses, suggesting that current approaches may not adequately address the combined requirements of predictive accuracy, interpretability, and operational security.

This study therefore addresses the need for a more comprehensive evaluation approach by developing a hybrid detection framework that integrates predictive performance assessment with explainability and structured threat modeling. The aim is to generate a system that is not only accurate but also resilient, transparent, and aligned with the security demands of real-world financial environments.

1.3 Objectives of the Study

General Objective

To develop an interpretable hybrid fraud detection Framework for financial institutions assessing its resilience and operational security through structured threat modeling.

Specific Objectives

1. To develop a leakage-safe, hybrid machine learning fraud detection model combining supervised ensemble methods with anomaly detection integrating SHAP-based explainability
2. To evaluate the predictive performance of the model using multiple classification metrics
3. To apply STRIDE threat modeling in identifying systemic cybersecurity vulnerabilities within the fraud detection pipeline.
4. To test the resilience of the model against adversarial and out-of-distribution samples using anomaly-gating mechanisms.

1.4 Research Questions

1. How can a hybrid machine learning model be designed and implemented to detect financial fraud using leakage-safe methodologies and SHAP-based explainability?
2. How well does the model perform under the predictive evaluation metrics?
3. What systemic vulnerabilities in the fraud detection pipeline can be identified using STRIDE threat modeling?
4. How resilient is a model against adversarial and out-of-distribution transactions when using anomaly-gating mechanisms?

1.5 Significance of the Study

The study makes contributions that are valuable both academically and operationally. From a scholarly perspective, it advances ongoing discussions on how fraud detection models should be evaluated beyond conventional predictive metrics. By integrating concepts drawn from machine learning, explainable AI, and cybersecurity threat analysis into a unified evaluation framework, the study broadens the theoretical understanding of what constitutes a robust fraud detection

system. This framing contributes to emerging literature that argues for more holistic assessments of model reliability, especially in high-risk domains.

Methodologically, the study offers a structured hybrid detection approach that combines supervised ensemble learning, anomaly-based gating and SHAP-driven interpretability. The framework is supported by a leakage-safe preprocessing pipeline and a repeatable evaluation procedure that incorporates multiple performance metrics. This provides a practical template that future researchers can adapt when designing or auditing fraud detection models that must balance predictive accuracy with operational trustworthiness.

In practical terms, the study demonstrates how the integration of SHAP explanations can improve the usability of detection outputs within environments such as Security Operations Centers. The model generates transaction-level reasoning that supports faster decision-making and reduces ambiguity for analysts who must justify interventions. This responds directly to a common operational challenge where technically strong models fail to deliver value because their outputs cannot be interpreted or acted upon effectively.

The application of STRIDE threat modeling further strengthens the study's contribution by identifying systemic risks that may remain hidden when models are evaluated using predictive metrics alone. This structured assessment highlights vulnerabilities relevant to financial institutions and provides guidance on areas where additional safeguards or monitoring may be necessary during deployment.

Finally, the empirical testing of anomaly-gating mechanisms offers insights into the trade-offs between enhanced resilience and detection coverage. These findings may support practitioners as

they determine how to balance security constraints, detection priorities, and resource limitations when integrating machine learning into critical financial infrastructures. By bringing these strands together, the study provides a comprehensive and practically actionable foundation for developing fraud detection solutions that meet both technical and operational reliability requirements.

1.6 Scope of the Study

The study focuses on financial fraud detection systems that leverage on machine learning. It leverages on the publicly available IEEE-CIS fraud detection dataset as a benchmark due to its richness, diversity and wide acceptance in the research community. The study narrows its scope to three core dimensions where one, we are to do model development and get its predictive evaluation. The second dimension is conducting cybersecurity assessment using STRIDE threat modelling and lastly operational enhancements through explainability and anomaly detection. Geographically, the study situates its relevance in the Kenyan financial sector, with insights applicable to the broader Sub-Saharan African region.

1.7 Limitations of the Study

Several limitations are acknowledged in the study. First is the reliance on the IEEE-CIS dataset which may limit contextual generalizability to Kenyan financial fraud scenarios, although it provides a robust benchmark for methodological evaluation. Second, the STRIDE analysis focuses on transaction-level vulnerabilities and does not empirically test infrastructural threats such as insider fraud or privilege escalation. Then the anomaly-gating mechanism introduces recall trade-offs potentially limiting applicability without adaptive thresholds. Lastly, while SHAP explanations enhance interpretability, they may expose sensitive model insights that adversaries

could exploit. These limitations, however, highlight directions for future research and do not diminish the study's contribution to fraud detection cybersecurity.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This chapter reviews existing scholarly and practical work related to machine learning–driven fraud detection within financial institutions, with a particular focus on cybersecurity evaluation. The review is structured around the study objectives which begins with theoretical foundations that anchor the research variables. It then presents a conceptual framework linking the independent and dependent variables, followed by an empirical review of studies organized around the objectives. The chapter concludes with a critique of the literature and identification of research gaps that this study seeks to address.

2.2 Theoretical Review

2.2.1 Fraud Detection and Machine Learning Theories

Fraud detection has long been conceptualized under the pattern recognition theory that suggests that fraudulent behavior can be differentiated from legitimate behavior through identifying deviations from normal data distributions. Previous studies have been applying statistical models to credit card fraud but these approaches have been struggling with scalability and adaptability in dynamic environments. By introducing data-driven learning ML theories overtime have extended to pattern recognition where classifiers iteratively refine decision boundaries from labeled transaction data (Chandola, Banerjee, & Kumar, 2009).

Supervised learning approaches such as Random Forest, XGBoost, and Gradient Boosted Trees dominate financial fraud detection. This is due to their ability to capture nonlinear decision boundaries in high-dimensional data. Recent work demonstrates that the ensemble learning methods where multiple classifiers are combined to reduce variance and bias have been shown to

significantly improve fraud detection accuracy compared to single classifiers. Evidently ensemble frameworks outperform traditional rule-based approaches in both detection rate and adaptability to new fraud patterns (Kalusivalingam & Sharma, 2022).

The anomaly detection paradigm frames fraud as a form of rare-event detection. This is grounded in anomaly detection theory, which assumes that fraudulent transactions will exhibit measurable differences from the statistical properties of normal transactions. Methods such as Isolation Forest, One-Class SVMs and autoencoders exploit these differences by identifying points that fall outside learned norms. Studies have highlighted the effectiveness of combining supervised and unsupervised paradigms known as hybrid models. This is to improve fraud resilience in detecting both known and novel fraud cases (Vashistha & Tiwari, 2024).

2.2.2 Adversarial Machine Learning and Fraud Detection

While most ML theories focus on predictive performance the emergence of adversarial machine learning (AML) has revealed systemic vulnerabilities in fraud detection pipelines. Adversarial ML is Drawn from robust statistics theory that studies how stable models remain when faced with corrupted or manipulated data. It assumes that fraudsters can launch evasion attacks by making undetectable changes to transaction features causing misclassifications. There can also be poisoning attacks where training data is manipulated to bias the model. Researchers have proposed Ensemble GAN-based frameworks to synthesize rare fraud cases for training. However, these methods also risk leaking adversarial knowledge into the system. The Generative Adversarial Networks (GANs) used by researchers to model synthetic fraud patterns in their research it is currently misused by adversaries to create realistic fraudulent transactions that evade detection in an attack.

In cybersecurity the concept of defense-in-depth discourages relying on just one layer for protection, we use several layers to guard against attacks. In ML fraud detection, it suggests that we shouldn't depend on a single model or method. We need multiple protective layers for protection against adversarial manipulation. Recent studies show this idea in action. They are combining hierarchical ensemble models with multi-anomaly detection as layers. When this layered method was tested in blockchain fraud cases, it performed better. It was more robust, meaning it could handle attacks and still function well (Kamran et al., 2024). Khalid et al. (2024) found something important. They showed that fraud detection models built with adversarial defense mechanisms were more resilient. Not just focus on being accurate but being designed to resist new and unexpected types of fraud. And they did better than models that were only optimized for accuracy.

2.2.3 Implications for Financial Institutions

We're no longer just looking for patterns in model development we're now building systems that can adapt and defend themselves against smart, evolving threats The theoretical shift from just pattern recognition to advanced hybrid models reflects the increasing complexity of financial fraud. these systems still have a weakness. From research of Choi & Jang, (2018) we can argue even this improves detection there's still work to be done to make them truly resilient. So, fraud detection isn't just about labeling transactions as fraud or not fraud. It's a full cybersecurity challenge. It requires multiple layers of defense, smart design and a constant adaptation technique to stay ahead of attackers

2.2.4 Threat Modeling and STRIDE

Threat modeling Manages risk by in ML-based systems. The study had to find a systematic way to identify, evaluate and mitigate potential issues to. Microsoft's STRIDE framework was promising for this because of its thoroughness as it organizes threats into six clear categories (Shostack, 2014). The framework is simple and adaptable, making it useful in diverse domains such as cyber-physical systems, intrusion detection and recently ML systems. The framework has been recognized as a lightweight but effective framework compared to more complex methodologies such as OCTAVE, PASTA and STPA-sec (Khan, McLaughlin, & Laverty, 2017).

Recent research has highlighted the importance of adapting STRIDE for ML-based systems. Mauri and Damiani (2022) proposed a structured adaptation of STRIDE to AI/ML assets. He showed how STRIDE categories can map to assets like model inputs and training data, helping uncover vulnerabilities such as adversarial perturbations, data poisoning and evasion attacks (Mauri & Damiani, 2022). Another study by Alatwi and Morisset (2022) demonstrated STRIDE threat modeling but in his case it was for ML-based network intrusion detection systems (NIDS). This underscores its value in identifying security weaknesses beyond conventional accuracy metrics.

Below is a detailed review of each STRIDE component in the context of fraud detection and ML pipelines.

I. Spoofing Identity

Spoofing refers to an attacker impersonating a legitimate entity to gain unauthorized access. In ML-driven fraud detection, spoofing threats often target identity verification and authentication stages. It assumes fraudsters may spoof mobile identities in financial transactions to bypass fraud filters, as seen in SIM-swap fraud cases in mobile banking highlighted by Kaur & Lashkari (2021).

Mauri & Damiani (2021) also cautions that For ML pipelines spoofing can also manifest as attackers crafting inputs that mimic legitimate transaction profiles which can end up tricking classifiers into mislabeling fraudulent behavior as non-fraudulent. It highlights the need for robust feature engineering and adversarial training to detect spoofed identities and prevent impersonation in fraud systems.

II. Tampering with Data or Model

Tampering involves malicious modification of data, features or even ML models themselves. In fraud detection, tampering is evident when attackers manipulate transaction features like transaction amount and geolocation to evade detection. Biggio & Roli (2018) suggested that on the model side, adversaries may engage in data poisoning, introducing corrupted training data to bias model behavior. STRIDE-based studies like Wilhjelm & Younis (2020) show that ML assets such as training sets and models are particularly vulnerable to tampering due to their dependence on data integrity. Mitigation strategies for this threat includes the use of secure data pipelines, validation of training sources and checksums to ensure integrity across fraud detection workflows.

III. Repudiation

This is where an attacker disputes having performed a fraudulent transaction or activity. Within fraud detection systems this is when fraudsters dispute flagged transactions, claiming legitimacy. In the current contexts, repudiation is exacerbated when models lack explainability, making it difficult for analysts to present human-understandable evidence to support system decisions. Explainable AI tools like SHAP play a crucial role here since they provide interpretable justifications that counter repudiation claims. It can be done by highlighting the transaction

features influencing the model's decision as suggested by Adadi & Berrada, 2018. STRIDE frames repudiation not just as a legal or logging issue but as an interpretability challenge in ML-based fraud detection.

IV. Information Disclosure

Information disclosure threats involve the leakage of sensitive data. This includes both direct leaks like exposure of transaction logs and model inversion attack. This is where adversaries infer sensitive training data from outputs. Information disclosure can expose transaction attributes, model parameters, or SHAP explanations to attackers, which they can then exploit to refine evasion strategies. Mauri and Damiani (2021) note that applying STRIDE to AI reveals such disclosure risks, stressing the importance of restricting access to model internals and anonymizing sensitive transaction features.

V. Denial of Service (DoS)

Denial of Service in fraud detection refers to attacks that overwhelm ML models or SOC systems preventing timely fraud detection. Fraudsters may flood detection pipelines with adversarial or noisy transactions, consuming computational resources and delaying legitimate processing (Alatwi & Morisset, 2022). DoS attacks can also target model APIs, causing excessive query costs or downtime. Strategies such as rate limiting, anomaly gating, and resource monitoring can be used to prevent DoS from crippling fraud detection operations.

VI. Elevation of Privilege

Elevation of privilege refers to unauthorized access to higher system permissions. This could involve adversaries exploiting weak access controls to modify model parameters, retrain models with poisoned data or even bypass fraud detection layers. There can be cases where vulnerabilities

in cloud-hosted fraud detection services could allow attackers to escalate privileges and control model APIs (Angganegara & Mukti, 2025). These risks promote defense-in-depth strategies such as role-based access control, privilege minimization and continuous monitoring of access logs.

The STRIDE framework provides a holistic lens for identifying systemic vulnerabilities across fraud detection workflows. Spoofing exposes weaknesses in authentication, tampering undermines data integrity, repudiation challenges necessitate explainability, information disclosure threatens privacy, denial of service cripples' system availability and elevation of privilege compromises system control. While STRIDE was originally developed for traditional software systems, its adaptation to ML fraud detection reveals that ML-specific threats can be systematically categorized and mitigated.

2.2.5 Explainable AI (XAI) Theories in Fraud Detection and Cybersecurity

The rise of machine learning in fraud detection has introduced a critical challenge that require models to provide interpretability of their outputs. Some ensemble and deep learning algorithms function as "black boxes" by providing accurate predictions without revealing how decisions are made. This lack of transparency undermines analyst trust thus challenging operational deployment in financial institutions. Explainable Artificial Intelligence (XAI) resolves this by introducing frameworks and tools that make ML predictions interpretable, justifiable, and actionable in such security-sensitive environments.

XAI is anchored in transparency and interpretability theory that holds that algorithmic systems must be auditable and comprehensible to human stakeholders as used by Adadi & Berrada (2018). Interpretability applies socio-technical systems theory that emphasizes the interdependence between technical outputs and human decision-making in operational settings. Therefore,

explanations act not only as technical artifacts but also as enablers of trust and accountability in security operations.

A dominant approach in XAI which builds on cooperative game theory is SHAP. The theory models feature as “players” contributing to the overall “payout” of the prediction. It uses Shapley values to quantify each feature’s marginal contribution (Lundberg & Lee, 2017). This should provide both global explanations of how features influence the model overall and local explanations which explains why a specific transaction was classified as fraudulent or legitimate. Other approaches such as LIME that was used by Ribeiro, Singh & Guestrin, 2016 and Anchors have been widely compared with SHAP. However, SHAP is considered more theoretically sound due to its consistency and local accuracy (Zhang et al., 2022; Capuano et al., 2022).

2.2.6 XAI in Fraud Detection

SHAP has been used to make anomaly-based models interpretable, enhancing analysts’ ability to distinguish between genuine anomalies and benign noise as explained by Hassan et al. (2023). Recent studies show that embedding SHAP into fraud detection dashboards significantly improves analyst decision support, reducing false positives and shortening investigation times (Alenezi & Ludwig, 2021). To prove this case studies in financial institutions reveal that that XAI explanations not only enhance SOC efficiency but also mitigate repudiation risks, where customers dispute fraud alerts. This is because they provide interpretable evidence of feature contributions. Fidel et al. (2019) caution that explanations can themselves expose sensitive information, enabling adversaries to infer model weaknesses or reverse-engineer decision boundaries. Researchers advocate for controlled interpretability, where explanations are tailored for internal SOC use without disclosing sensitive model internals (Alketbi & Mehmood, 2025).

2.2.7 Anomaly Detection and Resilience Theories

Fraud detection like finding a needle in a haystack. In most scenarios Most transactions are legitimate and only a small fraction is fraudulent. So in the datasets Fraudulent data are the minority in between very large volumes of legitimate data. The theoretical basis for anomaly detection assumes that statistically fraud looks different from normal transactions. These differences might be small, but they stand out when we try and look at the data carefully. Models can still detect fraud even if someone tries to contaminate data or have adversarial perturbations. This has been made very possible by recent research that focus on building models with anomaly detection to stay reliable even when under attack. Anomaly detection adds another layer of protection. It can catch new threats early before they spread. This fits into the defense-in-depth strategy we proposed earlier.

2.2.8 Isolation Forest

Back in 2008, Liu, Ting, and Zhou introduced a very powerful tool called Isolation Forest. It became a major breakthrough in anomaly detection. Here's how it works. IF splits the data randomly over and over it works the assumption that fraudulent or unusual data points tend to get separated faster than normal ones. That's because they behave differently and tend to stand out more. Isolation forest is popular in fraud detection because it scales well even on imbalanced data. Researchers like Xu et al. (2023) created Deep Isolation Forest extending the baseline IF into more advanced variants to improve detection.

Hybrid frameworks for anomaly detection reduce merely introducing IF and an autoencoder and recurrent network, which support more conditions of adaptation and continuous learning under the learning mechanisms of IF. Kumar et al. (2025) assigned a hybrid framework for anomaly

detection that tends to improve adaptability utilizing IF, in an autoencoder and ConvLSTM to combine representation and spatiotemporal cross domain applicability to better serve fraud and surveillance practices for learning adaptation. Similarly, Prasad et al. (2025) took the novel adaptation of automatic detection utilizing IF in conjunction with a temporal attention-based LSTM to classify transactions within banking datasets. This adaptation increased accuracy in detecting errant transaction, due to the reduction in false positives, while conditions for adaptation developed under changing conditions of observation (Prasad et al., 2025)..

2.2.9 Resilience Against Adversarial Attacks

Anomaly detection is gaining recognition as a resilience mechanism against adversarial machine learning (AML) threats. Anomaly detection layers act as “gates” that prevent manipulated inputs from influencing fraud classifiers. This is done by identifying out-of-distribution (OOD) or adversarially perturbed samples. Integrating anomaly detection with hyper-ensemble ML models in banking improves resilience against evasion attacks and reduces fraud leakage, a finding demonstrated by Vashistha and Tiwari (2024). Alam and Mahmud (2025) argue that AML-aware frameworks are essential for maintaining robustness in cases where adversarial attempts to mimic genuine behavior are common.

Study finding by Ahmed, Lee, and Hyun (2019) corroborate with the above findings though their study was outside the finance domain they applied Isolation Forest to detect covert data integrity attacks in smart grids. They showed its applicability in detecting adversarial manipulations in other cyber-physical systems (Ahmed et al., 2019). Idowu (2025) also applied this and further demonstrated that edge-deployed Isolation Forests enable real-time anomaly detection in IoT

networks. Their work suggests opportunities for distributed fraud detection pipelines resilient to denial-of-service and adversarial flooding attacks.

2.2.10 Trade-offs and Limitations

Anomaly detection, despite its promise it introduces trade-offs that require careful consideration in financial fraud detection. IF-based anomaly gating can improve resilience but may reduce recall by filtering borderline cases that include subtle fraud (Kalusivalingam & Sharma, 2022). Another issue is that anomaly detection models are not immune to adversarial manipulation. There can be instancesses where attackers can inject crafted anomalies that either overwhelm detection thresholds or disguise fraudulent activity as benign noise. Lam (2025) raises the importance of combining anomaly detection with resilience audits and fairness checks to ensure trustworthiness in fraud detection systems.

2.2.11 Implications for Fraud Detection Resilience

The theoretical grounding of anomaly detection in robust statistics and defense-in-depth explains its critical role as a resilience mechanism. While Isolation Forest and its hybrid variants offer scalable methods to capture outliers and adapt to dynamic fraud, anomaly detection still needs to be integrated with complementary strategies such as adversarial training and STRIDE-based threat modeling. These strategies are essential to balance detection coverage with operational usability.

2.3 Conceptual Framework

This study uses a conceptual framework as the theoretical lens for examining ML-driven fraud detection systems within financial institutions. The framework integrates four independent variables which are hybrid ML model design, STRIDE threat modeling, explainable AI (SHAP),

and anomaly gating mechanisms. All four link directly the dependent variable which in this case is the operational security and effectiveness of fraud detection systems.

This reflects the multidimensional nature of fraud detection where predictive accuracy alone is insufficient without resilience, interpretability and cybersecurity evaluations. The framework builds on systems theory, which posits that outcomes in complex environments arise from the interaction of interdependent components. It also intergates socio-technical systems theory, which emphasizes the alignment of technological mechanisms with human decision-making in Security Operations Centers (SOCs).

2.3.1 Independent Variables

The first independent variable of our study is the hybrid ML model that combines supervised ensemble learning methods with unsupervised anomaly detection. From the literature review we have ensemble models have consistently outperformed single classifiers in fraud detection due to their ability to reduce variance and capture nonlinear relationships (Bhattacharyya et al., 2011). However, fraud remains a moving target with novel attack patterns often unseen during training. Integrating anomaly detection ensures coverage of out-of-distribution events, enhancing adaptability (Vashistha & Tiwari, 2024). This hybrid design draws its knowledge from the pattern recognition theory and robust statistics theory. It combines predictive strength with the necessary adaptability to unknown fraud vectors.

Next, we examine the role of STRIDE threat modeling. Its application to the fraud detection pipeline is critical. As mentioned earlier STRIDE categorizes threats into Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege (Shostack, 2014). Recent studies demonstrate that STRIDE can be adapted to ML assets to uncover

vulnerabilities that accuracy metrics alone cannot capture (Mauri & Damiani, 2022; Alatwi & Morisset, 2022).

The third variable is SHAP. This explainable AI technique, grounded in cooperative game theory as explained by Lundberg & Lee (2017). SHAP values help transform black-box outputs into clear, actionable intelligence for SOC analysts by providing human-readable explanations. Adding explainability directly into fraud alerts has been shown to reduce analyst workload, improve trust and mitigate repudiation risks where flagged customers dispute fraud outcomes (Alenezi & Ludwig, 2021; Hassan et al., 2023). This variable is a reflection of the transparency theory and socio-technical interaction theory which emphasizes that interpretability is not merely a technical feature but a usability requirement in operational fraud detection. The fourth independent variable is anomaly gating, implemented using Isolation Forest or its advanced variants. Here, anomaly detectors act as "gatekeepers" that filter adversarial and out-of-distribution inputs before they reach predictive models. These increases resilience against evasion and poisoning attacks (Papernot et al., 2016; Alam & Mahmud, 2025). While anomaly gating introduces trade-offs in recall, its integration addresses the resilience dimension highlighted in defense-in-depth theory. The introduction ensures fraud detection pipelines can withstand adversarial manipulation without systemic collapse.

2.3.2 Dependent Variable

Our study's dependent variable is how safe and successful scam detection systems are in the real world. We looked at the system's resiliency, which is its power to resist threats from both inside and outside the system (Biggio & Roli, 2018). interpretability, which helped us talk about how well SOC experts can understand scam reports, trust them, and act on them (Adadi & Berrada, 2018). And lastly our study tries to improve usability in SOCs to reduce alert fatigue, support

evidence-based decision-making, and integrate into institutional workflows This dependent variable is a reflection of the knowledge that fraud detection success must be measured not only by predictive performance but also by its cybersecurity robustness, transparency and operational viability.

The interaction as seen in Figure 1 between all the independent variables are to enhance the operational security of fraud detection systems for us . This can be seen in terms of how our study hybrid models increase predictive coverage, The parts of STRIDE we applied we to ensure systemic vulnerability assessments, SHAP for providing interpretability and anomaly gating strengthens our systems adversarial resilience. Together, these independent variables can be combined to produce fraud detection systems that are secure, interpretable, and operationally viable in financial institutions especially like for mobile money–driven economies like Kenya.

Visual Framework

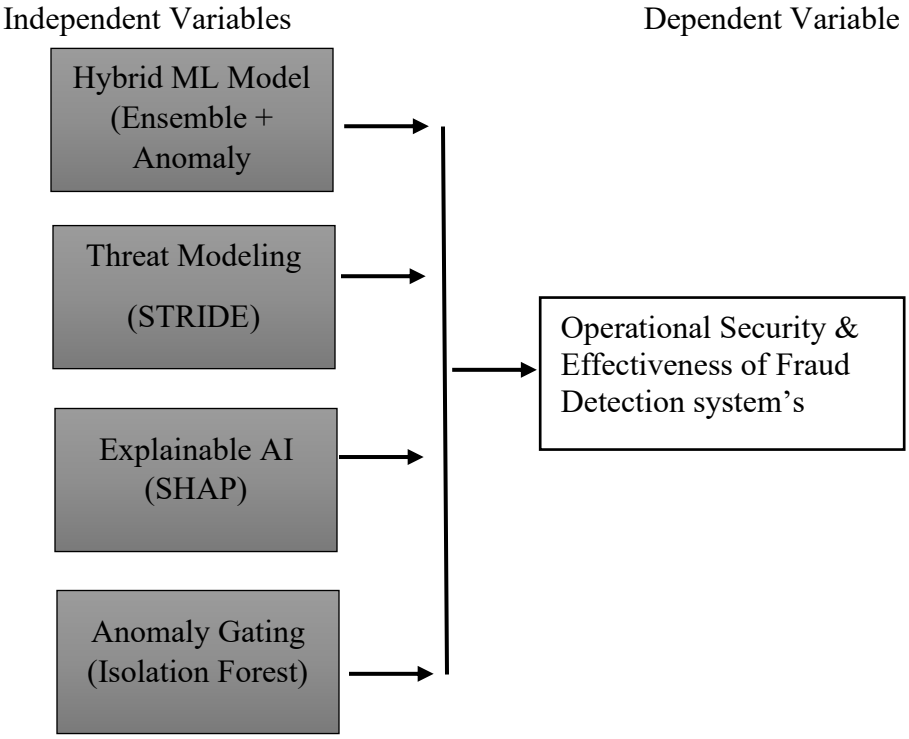


Figure 1: Conceptual Framework of the Study

2.4 Empirical Review

The empirical review synthesizes studies related to the specific objectives of this research.

2.4.1 Designing and Implementing Hybrid ML Fraud Detection Models

Hybrid and ensemble models are increasingly adopted in fraud detection because they balance predictive accuracy with adaptability. Hossain & Islam (2023) proposed a hybrid ensemble for botnet detection they combined feature selection with tree-based models. The accuracy and dependability of their work were better than those of single classifications. To find fraud, Chagahi et al. (2024), on the other hand, created an attention-based ensemble using CNNs, GNNs, and a confidence-driven blocking system.

Empirical study in related defence areas backs up and shows that fusion works. Nazir et al. (2025) looked at mixed CNN-LSTM designs for finding IoT threats and found that they were better at handling different types of attacks than separate models. These data show that mixed approaches should be used in financial fraud detection systems to keep up with new types of scam.

2.4.2 Evaluating Predictive Performance Against Baselines

Fraud detection studies typically benchmark models using AUC, precision, recall, and F1-score. However, scholars warn that overreliance on these metrics can misrepresent operational readiness. Lam (2025) critiques fraud research for prioritizing numerical benchmarks over resilience and interpretability, calling for multi-dimensional evaluation frameworks. Borkar & Sangve (2025) showed that anomaly detection using deep learning in high-frequency options trading captured fraud more effectively when combined with latency-aware performance indicators, thus the need for domain-specific metrics (Borkar & Sangve, 2025). Empirical work highlights the limitations of predictive-only evaluation and the need to integrate robustness, interpretability, and usability metrics into fraud detection assessments.

2.4.3 STRIDE Threat Modeling in Fraud Detection Pipelines

Direct empirical applications of STRIDE to ML-based fraud detection remain sparse, but parallel studies in network and intrusion detection highlight its promise. This was evident when Alatwi & Morisset (2022) applied STRIDE to ML-powered intrusion detection systems and uncovered vulnerabilities in spoofing and data tampering that conventional accuracy metrics missed. Mauri & Damiani (2022) also adapted STRIDE to AI/ML assets, mapping adversarial risks such as poisoning, evasion, and model inversion to the framework. Structured threat modeling has yet to be systematically integrated into financial fraud detection research, presenting a clear empirical gap this study addresses.

2.4.4 SHAP Explainability in Fraud Detection

Recent work by Chagahi et al. (2024) shows how Explainability has gained momentum in fraud detection research. By showing how the model arrived at its conclusions, SHAP would help analysts understand and trust the system's decisions, a key requirement in high-stakes environments like finance in another study Pelosi et al. (2025) conducted a systematic review of interpretability under changing data conditions, they found that SHAP outperforms other methods like LIME in maintaining trustworthiness over time.

Also, Alenezi & Ludwig (2021) found that including SHAP explanations into fraud detection dashboards reduced false positives and improved SOC efficiency. The improvement was done while mitigating a number of repudiation risks.

2.4.5 Anomaly-Gating for Adversarial Resilience

Anomaly detection has been empirically validated as a resilience mechanism in fraud detection. This was clearly demonstrated when Ehsan et al. (2024) applied machine learning anomaly

detection in Ethereum transactions. They were able to unveil threats and improve classification accuracy in blockchain ecosystems. Without forgetting that Lam (2025) emphasized in combining anomaly detection with supervised classifiers to reduce false positives but this introduced trade-offs in recall, demanding careful calibration. Kumar et al. (2025) showed that hybrid anomaly detection adapts well to dynamic fraud scenarios. Collectively, these studies validate anomaly gating as a viable resilience mechanism but caution that threshold tuning and workload balance remain unresolved challenges.

2.4.6 Contextual Relevance of Kenya

Kenya is widely recognized as a global pioneer in mobile money adoption. This is with the launch of M-Pesa in 2007 positioning the country at the forefront of financial technology innovation. By 2023, mobile money transactions accounted for more than 50% of Kenya's GDP. This highlighted the central role of digital financial services in daily economic activity (CBK, 2023). The widespread reliance on mobile money platforms has expanded access to banking but has also increased exposure to financial fraud, making Kenya a unique and fertile ground for fraud detection research.

Unlike Western financial systems that are heavily card-based, the Kenyan context involves peer-to-peer transfers, agent-assisted cash-ins/outs, airtime purchases, and mobile loans, each of which introduces fraud typologies not always captured in traditional credit card fraud studies. For instance, SIM-swap attacks, social engineering scams, and agent collusion are prevalent in Kenya and map directly onto the STRIDE threat categories of spoofing, tampering, and repudiation. This distinct threat environment underscored the need for research that situates fraud detection within the realities of mobile-first economies.

The empirical literature demonstrates progress in hybrid model development, interpretability, and anomaly detection. However, gaps persist in holistic frameworks that combine these dimensions with structured threat modeling. Most studies emphasize predictive accuracy, with fewer addressing operational resilience or analyst usability.

2.5 Critique of Literature

The reviewed literature reveals significant progress in the application of machine learning to financial fraud detection, yet several shortcomings persist across methodological, theoretical, and practical dimensions.

2.5.1 Strengths of Existing Literature

For Hybrid and Ensemble Learning studies such as Bhattacharyya et al. (2011), Moradi et al. (2025), and Chagahi et al. (2024) provide convincing evidence that ensemble and hybrid approaches outperform single classifiers in fraud detection. These works highlight the adaptability of models in handling imbalanced datasets and diverse fraud patterns the other aspect is Advancements in Explainable AI where research on SHAP like that of Lundberg & Lee (2017) and Pelosi et al. (2025) demonstrates substantial improvements in model interpretability and analyst usability. Embedding explanations into SOC workflows has been empirically shown to reduce false positives and improve decision-making (Alenezi & Ludwig, 2021).

As practical move toward resilience rather than accuracy alone, the Integration of Anomaly Detection like Isolation Forest and related anomaly detection mechanisms has been validated as an effective complementary technique for capturing out-of-distribution cases (Liu et al., 2008; Ehsan et al., 2024). Finally, we note the emerging use of Threat Modeling for machine learning models. Though limited, adaptations of STRIDE in studies of Mauri & Damiani (2022) and Alatwi

& Morisset (2022) illustrate its potential to systematically uncover vulnerabilities in ML pipelines. This will in turn extend evaluation frameworks beyond conventional performance metrics.

2.5.2 Weaknesses of Existing Literature

The first weakness is the recurrent use predictive metrics alone in the evaluation of fraud detection. Much of the literature remains narrowly focused on benchmarks such as AUC, precision, and recall this can be referenced to studies by Bauder et al., (2018) and Lam, (2025). In these studies, they are often overlooking the model's resilience, explainability, and usability in SOC contexts. This leaves a gap between academic performance claims and operational deployment realities. Another limitation is the fragmented approaches currently in the existing studies because research on hybrid ML, STRIDE, SHAP and anomaly detection has largely been pursued in independently. Very few works integrate these dimensions into a unified framework. The unified framework is capable of addressing fraud detection as a cybersecurity problem rather than a classification exercise.

Thirdly in these studies there are limited contextual relevance this is because many studies are based in Western banking or e-commerce environments, with limited empirical evidence from mobile money–driven economies like Kenya. This limits the validity of findings for Sub-Saharan Africa because fraud patterns differ significantly based on numerous factors. Another weakness is the Explainability Risks because while SHAP explanations enhance transparency Fidel et al. (2019) caution that such outputs can also expose sensitive model insights to adversaries. Yet, few studies empirically address this trade-off, resulting in unresolved risks in operational environments. Lastly it is the Trade-offs in Anomaly Detection where previous studies highlight that anomaly gating reduces false positives but at the expense of recall (Lam, 2025; Prasad et al.,

2025). This trade-off remains poorly explored in terms of adaptive thresholds and workload optimization.

While the literature provides strong foundations in hybrid ML, XAI, and anomaly detection, it fails to present a holistic, cybersecurity-oriented framework that integrates these elements. This study positions itself to bridge that gap

2.6 Research Gaps

The following table synthesizes key research gaps identified in the empirical review.

The literature highlights that while machine learning models are effective at detecting fraud, they remain narrowly optimized for predictive accuracy. STRIDE threat modeling, explainable AI and anomaly detection have been explored independently but rarely integrated into a cohesive framework. Most empirical work is concentrated in developed markets, leaving emerging economies underrepresented. Finally, operational trade-offs such as false positives, recall reduction, and explainability risks remain underexplored limiting real-world applicability. This Work addresses these gaps by proposing and evaluating a holistic, cybersecurity-oriented framework.

Table 2.1: Summary of Research Gaps

Table 1: Research Gaps in Financial Fraud Detection

Study Brief	Description	Methodology	Findings	Gap
Bhattacharyya et al. (2011)	Credit-card fraud detection	Random Forest + Bagging ensemble	Achieved higher AUC than single models, demonstrating ensemble strength	Neglect of explainability and adversarial resilience, focus mainly on predictive accuracy
Chaghi et al. (2024)	Financial fraud detection	Hybrid ensemble: CNN + GRN + attention + SHAP	Improved handling of class imbalance; integrated explainability (SHAP)	No real-time or adversarial robustness evaluation; operational usability not assessed
Alawieh et al. (2021)	Explainable ensemble models	XGBoost + SHAP	Improved interpretability enhances analyst trust and reduces false positives	Lacked systemic cybersecurity threat and attack surface analysis (e.g., STRIDE)
Moradi & Damania (2021)	Explainability in ML classification	SHAP + LIME comparative study	Provided clear visualization of model bias and decision processes	No integration with cybersecurity or threat frameworks, missing operational resilience focus
Ehsan et al. (2024)	Financial anomaly detection	Autoencoder + Isolation Forest	Success in detecting novel/unseen fraud patterns	Lack of explainability and governance frameworks to support operational usage
Biggio & Roli (2018)	Adversarial ML foundation	Gradient-based attacks on SVM	Demonstrated vulnerability of ML models to evasion attacks	Research foundational and theoretical; no direct applied cybersecurity evaluation in fraud detection

CHAPTER THREE: METHODOLOGY

3.1 Introduction

This chapter presents the methodology adopted in carrying out the study. It documents the step-by-step procedures that were followed in developing, implementing, and evaluating the hybrid machine learning fraud detection framework. The approach was carefully designed to ensure that the system was both predictive and resilient. While also aligned with cybersecurity principles discussed in Chapter One. This is why the chapter talks about the study's philosophical stance and the research method that was used. It explains the study area's background, the type of people who will be studied, the sample method, and finally how to collect and prepare the data.

3.2 Research Philosophy

Our methodological design began with the philosophy we chose for our study. We wanted to choose a philosophy that could bring together computer testing and practical evaluation. The study chose pragmatism as the main philosophy. Pragmatism recognizes that different research methods have their own pros and cons, but that using a mix of them can help us get a better and more complete picture of complicated problems, especially when we are trying to solve difficult real-life problems.

Positivism and interpretivism, the other two main ways of thinking, are different from pragmatism. Positivism works well for controlled studies but not so well for understanding how scam detection systems work from a social and technical point of view. These facts include more than just numerical success; they also include questions of how to understand, how to be resilient, and how to be useful. On the other hand, interpretivism stresses subjective meanings and context. However, interpretivism by itself would not be enough for a study that relies on large-scale quantitative analysis. Pragmatism doesn't stick to a single moral position. Instead, it follows the idea that the

best way to answer study questions should determine the method used. This made it perfect for a study that needed a lot of computing power and needed to be set in a certain place.

We used a practical approach to blend two different methods: tests with quantitative machine learning and qualitative modelling of cybersecurity threats. The study aimed to find the best mixed machine learning model by designing and testing it. The design had three base learners built on top of an XGBoost meta-classifier, and Isolation Forest was used to find outliers. We looked at the first two study questions through a positivist lens, which meant we had to make and test a leakage-safe mixed model. Standard measures like accuracy, precision, recall, and area under the ROC curve (AUC) were used to judge the model's performance. The study didn't just look at how well the predictions worked; it also looked at how resilient, explainable, and vulnerable the system-level weaknesses are in real life financial institutions. In order to answer these issues, we tried to add sociotechnical aspects to the approach. Since fraud detection systems are used in Security Operations Centers (SOCs), where human researchers connect with machine output, they shouldn't work alone. As highlighted in Chapter Two, one of the major challenges of machine learning in fraud detection is the interpretability of model predictions. To address this, the study embedded explainable AI (SHAP) into the methodology to enhancing interpretability

The study on the qualitative side the applied STRIDE threat modeling framework to systematically identify potential vulnerabilities in the fraud detection pipeline. STRIDE, was adapted to the machine learning context to examine threats such as spoofing, tampering, information disclosure, and denial-of-service within fraud detection workflows. This addressed Research Question 3, which sought to uncover systemic vulnerabilities beyond predictive modeling. Here, the philosophy of pragmatism was instrumental as it justified combining structured qualitative

assessments with quantitative experiments. This recognized that both forms of evidence are necessary for building operationally viable fraud detection systems.

We didn't treat resilience as a purely academic concept. Instead, we tested it in conditions that reflect how fraud detection systems are actually used in financial institutions. even though this is typically studied in theoretical machine learning research. Instead, we measured resilience by looking at how it affects key operational factors. This aligns with the pragmatic principle that acknowledges that the value of knowledge is judged by its practical consequences in simple language meaning the study should focus on what works in practice, not just what looks good in theory. The adoption of this pragmatism as the guiding philosophy enabled the study to:

1. Use standard metrics to evaluate model performance, but we also recognized that numbers alone don't capture the full picture.
2. threat model with STRIDE to identify and understand security risks that aren't visible through prediction metrics.
3. We made the system interpretable by integrating SHAP.
4. Did adversarial resilience testing but always in the context of real-world constraints like balancing detection accuracy with operational workload.

There was a philosophical basis for this work that kept it problem-based, complex, and useful in real life. The method did a good job of combining the quality of data-driven tests with the sensitivity needed for cybersecurity evaluation. This created a system that can handle both the theory and practical aspects of scam detection.

3.3 Research Design

The main idea behind this study was to use an experimental modelling method in a case study about hacking. The study had a dual structure, which came from the fact that the study goals were also dual. To answer the research questions in Chapter One, the study had to do a lot of testing with the machine learning model, look at weaknesses in a planned way, test for resilience, and use interpretability.

3.3.1 Rationale for Experimental Design

Since trends in financial data are a big part of detecting fraud, we have to test our model to make sure it works. Changes could be made to things like the model's design or even its feature sets. An experiment was done on a model that was built using Random Forest, XGBoost, and LightGBM together. The study could add or take away some of the parts as needed to see how each one affects the model's performance and ease of use. Every test had to use the exact same settings for preparing and evaluating the data, so that any change in speed could be reliably traced back to its source.

It is known that fraud detection records have a lot of class mismatch and can sometimes have data leaking problems. We took this into account when we designed it so that these usual problems wouldn't change the findings. We used stratified cross-validation to make sure that the ratio of fraud to real cases stayed the same in both the training and testing splits. So that data wouldn't get lost, we carefully took away any traits that wouldn't be useful for making real-time estimates. By keeping an eye on all of that, our design made sure that the model's performance matched the real world.

3.3.2 Simulation-Based Case Study Dimension

The study also sought external contextual relevance to the Kenyan financial ecosystem. This was achieved by situating the experiments within a simulation-based case study framework. Specifically, although the IEEE-CIS dataset served as the experimental substrate, the fraud typologies and vulnerabilities identified were interpreted against the backdrop of Kenya's mobile money ecosystem, characterized by M-Pesa, Airtel Money, and Equitel transactions. This case study lens ensured that the findings transcended raw numbers which should provide insights into how hybrid machine learning models might perform under the fraud dynamics prevalent in Kenya. This can happen in the form of SIM-swap scams, account takeovers, and social engineering.

The operationalization part of the exercise involved making situations that looked like hostile conditions and operational limits. The Adversarial Robustness Toolbox (ART) was used to mimic adversarial escape attacks. This tool changed transaction features to test how resilient the models were. Workload simulations were used to see how different settings for anomaly gating affected the number of false positives, which was done by simulating what researchers in Security Operations Centers (SOCs) would see. These models were a link between academic modelling and real-world use.

3.3.3 Management and Control of Variance

Controlling variance was meant to give people faith that differences in performance are caused by real factors and not just chance noise. Several levels were used at different times of the study. As part of the first step, data splitting, the information was split into two parts: 80% for training and 20% for testing. Keeping the same number of students in each class during both tests and training. A consistent class distribution stops bias and helps the model learn patterns that work well with data it hasn't seen before. To get a stable estimation of model performance cross-validation splits the training data into five parts and trains the model five times, each time leaving out one part for

validation. The next level was when hyperparameter tuning, to test all combinations of parameters grid search was used with Bayesian optimization to work faster. The depth of trees in RF and learning rate in xgboost were also tuned avoiding overfitting or underfitting affecting the results.

To achieve some kind of reproducibility needed level number third level of control. Placement of random seed. Why? This ensured that repeated runs produced consistent results. Usually, ensemble models and sampling methods act randomly. Fixing seeds keeps their behavior predictable. The last level was adversarial testing to detect model's weaknesses. We made fake cases of theft from the real world to see how the model would respond to an attack.

The third level of control was where the random seeds were put because the experiments were done more than once with the same random seeds to make sure they could be repeated. This managed the random effects that come up with ensemble methods and sampling-based algorithms. The last test was the combat testing. Protocols were used to keep things from being too general. Adversarial attacks were carried out using a variety of methods, such as changing features, using gradient-based attacks, and adding fake transactions. Different levels of hostile effect were measured across a range of situations instead of just one test.

These checks and balances gave the study both internal and external validity. The way the study was set up makes sure that the model gains are real and that the results can be used to improve scam spotting systems in the real world.

To ensure consistency, reproducibility, and controlled model variance, all experiments were executed in a Google Colab environment using Python 3.10. The notebook runtime provided a standardized and isolated environment, preventing hardware-related variability. The following libraries and tools formed the core software stack used in the study:

Python Packages and Tools Used

- **numpy 1.26** – vectorized numerical operations
- **pandas 2.0** – dataset manipulation and merging
- **scikit-learn 1.3** – preprocessing pipelines, SMOTE, IsolationForest, GroupKFold, model calibration
- **xgboost 1.7** – base learner and meta-learner
- **lightgbm 4.1** – gradient boosting base learner
- **imbalanced-learn 0.11** – SMOTE within CV folds
- **shap 0.43** – model interpretability
- **matplotlib / seaborn** – visualization
- **joblib** – model checkpointing and reproducibility
- **Python Standard Library** – OS, logging, datetime for pipeline management

These packages were installed using Colab's default environment or via pip install where required. This ensured consistent dependency versions across all training runs.

Model Training Hyperparameters

Hyperparameters were selected based on empirical tuning, literature-aligned defaults, and computational feasibility within Colab. The final training configuration consisted of:

Random Forest

- n_estimators = 300
- max_depth = None
- min_samples_split = 2

- min_samples_leaf = 1
- class_weight = balanced_subsample
- bootstrap = True

LightGBM

- num_leaves = 32
- learning_rate = 0.05
- min_data_in_leaf = 20
- feature_fraction = 0.8
- bagging_fraction = 0.8
- bagging_freq = 5
- objective = binary
- boosting = gbdt

XGBoost (Base Model and Meta-Model)

- n_estimators = 500
- learning_rate = 0.05
- max_depth = 6
- subsample = 0.8
- colsample_bytree = 0.8
- reg_lambda = 1.0
- objective = binary:logistic
- eval_metric = aucpr

IsolationForest

- n_estimators = 300
- contamination = 'auto'
- max_features = 1.0
- bootstrap = False
- behaviour = 'new' (default in modern sklearn)

A **static anomaly threshold** was set at the **95th percentile** of anomaly scores from the training data, forming the anomaly gate.

Environment Consistency and Reproducibility Controls

Variance was further controlled using measures already aligned with this subsection:

1. **Random seed locking** across numpy, sklearn, xgboost, and LightGBM.
2. **GroupKFold validation using card1** to prevent identity leakage and maintain consistent folds.
3. **SMOTE applied *inside* the CV pipeline** to avoid synthetic data leaking into validation sets.
4. **Temporal split (TransactionDT)** ensured future data never appeared in training folds.
5. **Model checkpointing with joblib** ensured identical model states could be restored and re-tested.
6. **Colab notebook state tracking** maintained consistent package versions across sessions.

These controls collectively stabilized performance variance and ensured reproducibility across runs.

3.3.4 Justification of Strategy

The chosen design strategies were selected to meet objectives and were informed by established best practices in machine learning research. For research question 1 and 2 experiments were important to compare model variants. So, comparing hybrid models to simpler baselines helps us to see their added value. Only through controlled experiments could the added value of anomaly gating and hybridization be empirically established.

For Questions 1 and 3, we had to go beyond raw metrics and use case studies. This helped us use SHAP to make the models easier to understand and STRIDE to make the models more vulnerable in real-life financial situations. It wasn't enough to know how well a plan works on paper; it had to also work in real life. We did a simulation of hostile and out-of-distribution tactics to answer Question 4. According to study (Biggio & Roli, 2018; Moradi et al., 2025), this was some kind of tough test. We followed a good study plan by using both strict experiments and realistic models.

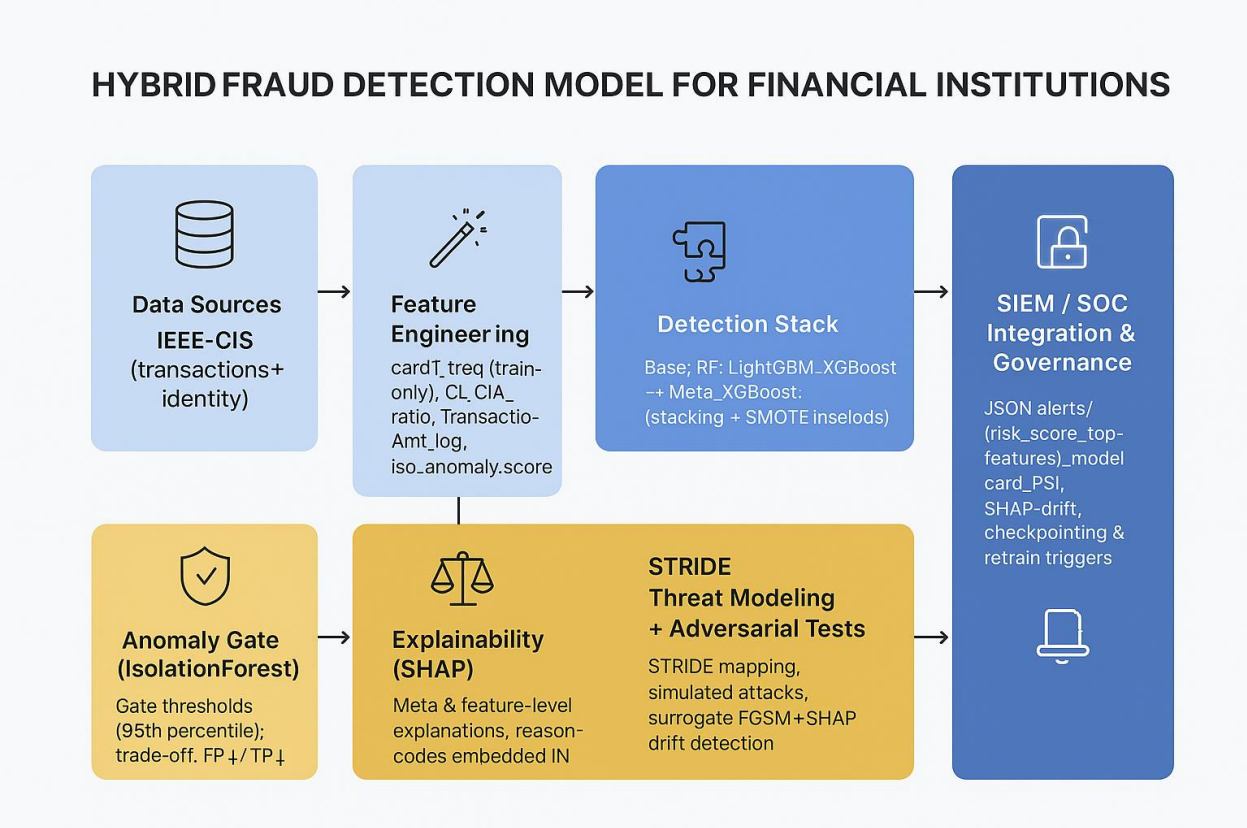


Figure 2: Operational Workflow of the Integrated Pipeline for the Proposed Hybrid Fraud Detection Framework

The operational workflow of the hybrid fraud detection model developed in this study is visualized in Figure 2, which illustrates the integrated pipeline from data ingestion through to SOC integration and adversarial testing. The diagram captures the key components including feature engineering, the detection stack with ensemble learning, anomaly gating, explainability integration, and cybersecurity assessment through STRIDE threat modeling.

3.3.5 Strengths and Limitations of the Design

The study strategy that was picked had a number of benefits. First, it made the experiment repeatable. All the steps, such as preparation, model training, and robustness tests, were well recorded, saved as goals, and other researchers could follow them. The plan also made sure that the results could be compared to those of other studies that used the same information to look for

fraud. Integration of STRIDE and SHAP into the experimental workflow produced very rich insights that extend beyond accuracy to include resilience, transparency and usability. It is important to note that the design reliance on a benchmark dataset may not capture the real nature of fraud in the Kenyan Financial sector.

In conclusion, the research design deliberately combined experimental requirements and cybersecurity contextualization to meet the study objectives. This approach ensured that findings were relevant, actionable and defensible in both academic and operational contexts. The methodology ensured that the study advanced beyond conventional accuracy benchmarks to cybersecurity-driven evaluation of fraud detection systems

3.4 Study Area

The study was situated within the Kenyan financial sector, with a specific emphasis on the mobile money ecosystem that has redefined the country's financial inclusion landscape. Kenya was deliberately chosen as the focal study The selection was not arbitrary but based on contextual, empirical, and theoretical considerations.

3.4.1 Justification for the Study Area

The justification for focusing on Kenya rests on three interrelated factors. The first factor is that the country has High Fraud Incidence this is evidenced from reports from the Central Bank of Kenya (2023) and the Association of Certified Fraud Examiners (2022). They highlight that there is growing sophistication of fraud in Kenya's mobile banking and digital payment systems. Billions of shillings are lost annually to fraudulent transactions, with banks and telcos reporting persistent threats ranging from phishing to insider fraud. This makes Kenya a compelling case study for developing fraud detection systems that balance accuracy, resilience, and interpretability.

The second factor is the Operational Relevance for Financial Institutions whee Kenyan banks and mobile network operators are actively piloting AI- and ML-based fraud detection tools. Most deployments remain focused on predictive accuracy rather than cybersecurity resilience or interpretability (Alhashmi et al., 2023). This gap directly aligns with the research problem identified in Chapter One, making the Kenyan context an appropriate testbed for evaluating the added value of hybrid ML, STRIDE threat modeling, and SHAP explainability. The third factor is based on Global Significance of Local Findings because while the study is geographically situated in Kenya, its findings have broader applicability to Sub-Saharan Africa and other mobile money–driven economies such as Ghana, Tanzania, and Uganda. These countries share similar transaction ecosystems and fraud patterns. So, this means that insights derived from Kenya can inform fraud detection strategies across the region. Kenya serves both as a local case study and a representative model for other emerging economies.

3.4.2 Operationalization of the Study Area in the Research

Although the IEEE-CIS fraud detection dataset was employed for experimental purposes due to its rich feature space and international acceptance as a benchmark. The Kenyan financial ecosystem informed how the dataset was interpreted, preprocessed, and evaluated. For instance, in STRIDE Threat Mapping Fraud typologies reported in Kenya such as SIM-swap identity spoofing and insider-led tampering were mapped to STRIDE categories when evaluating the ML pipeline. This ensured that the threat modeling step was not generic but contextualized to actual risks prevalent in Kenya. There were also Explainability Considerations in the research design since SHAP explanations were designed with (SOC) workflows in Kenyan banks in mind. Given the acute problem of analyst overload and alert fatigue reported in local institutions, interpretability was evaluated not only in abstract but in terms of operational usability. Lastly Kenya's fraud

patterns are characterized by high transaction volume with relatively low-value frequent transfers. This dynamic informed how anomaly detection thresholds were calibrated in simulations, reflecting the need to minimize false positives without missing fast-moving fraud events. By embedding contextual factors into experimental design, the study ensured that findings from the IEEE-CIS dataset were not divorced from local realities. The Kenyan study area therefore functioned as both a motivating backdrop and an interpretive lens, enhancing the external validity of the research.

3.4.3 Implications of the Study Area for the Research Framework

The choice of Kenya as the study area directly shaped the operationalization of the conceptual framework variables. This was in terms of how the hybrid ML model was evaluated for adaptability to fraud types common in Kenyan financial contexts. The STRIDE threat modelling was also made better by laying out the hostile risks that can be seen in Kenya's financial world. One example is how agent cooperation can be used for hacking and SIM-swap attacks can be used for scamming. The SHAP explainability was looked at in terms of how it could improve SOC processes in Kenyan banks. This is where limited analysis capacity calls for answers that are clear and can be used. Anomaly blocking was tested with transaction patterns that looked like mobile money transactions. This makes sure that the tests for adaptability was true to Kenya's surroundings. In this way, the study area actively shaped the way the research was designed and how it was evaluated. The study met both the academic and practical needs by putting research decisions in the context of Kenya's banking sector.

3.5 Target Population

The group of people this study was interested in was those who did banking deals online. There were two important subgroups in this population. The first group is legal trades, which make up the vast majority of this study's sample. The second grouping is made up of fake trades, which are a small but highly harmful percentage. The IEEE-CIS collection was useful for this group of people because it had a lot of anonymous transaction records that were marked as either fake or valid. High-dimensional data like product codes, payment methods, device identifiers, and geo-location proxies were part of every transaction. This collectively capture the complexity inherent in fraud detection systems. While the dataset is international, its feature set that is able to capture transaction behaviors, digital payment methods and device fingerprints which provides a robust proxy for the transaction-level fraud patterns prevalent in digital financial ecosystems, including Kenya's mobile money platforms.

The justification for selecting this target population was threefold. First, fraudulent transactions, although rare, have severe financial and reputational consequences for institutions and users alike. Second, the features embedded within such transactions are diverse and reflect both behavioral and contextual signals. The signals make them suitable for testing hybrid machine learning models. Lastly, the population structure in this dataset is currently with extreme class imbalance where there is <1 percent fraud cases of fraud amongst genuine transactions. This is a reflection the statistical reality faced by Kenyan financial institutions this reinforce the practical relevance of the findings.

3.6 Sampling Design

A stratified sampling design was adopted given the scale and imbalance of the used dataset. Stratification was to ensure that both fraudulent and legitimate transactions were proportionately represented in the sample, thereby minimizing the risk of skewed results. From the entire dataset split conducted in a stratified manner where 80 percent was reserved for training and 20 percent for testing. This was followed so that the minority class was preserved in the same proportion across both sets.

Stratified five-fold cross-validation was only used within the training partition to further manage the variance present and improve generalizability. This approach ensured that in every fold the minority fraud class was included. The approach was to prevent models from learning in the absence of fraud examples. To address extreme imbalance during training, Synthetic Minority Oversampling Technique (SMOTE) was selectively applied in order to create synthetic fraud samples. The selective application was to prevent bias toward legitimate transactions. Most importantly oversampling was limited to the training phase only, while the test set retained its natural distribution to reflect deployment realities.

The sampling design maintained methodological rigor through adequate representation of the minority class, and preserved external validity by ensuring that final evaluations were conducted under real-world class imbalance conditions.

3.7 Data Collection

The study relied exclusively on secondary data, specifically the IEEE-CIS fraud detection dataset (IEEE Intelligence Society & Kaggle, 2019). It contains over 590,000 anonymized online transaction records, split across identity attributes and transaction features, labelled as fraudulent

or legitimate. The dataset is publicly available and widely used as a benchmark for fraud detection research. This dataset was selected because it provides one of the most comprehensive and widely benchmarked sources for financial fraud detection research. It contains anonymized online transaction records, each labeled as either fraudulent or legitimate, alongside more than 400 engineered features representing payment instruments, device characteristics, product codes, and customer behavior markers. Its richness in dimensionality makes it a suitable platform for testing both predictive and anomaly-based fraud detection models.

The dataset was complemented by contextual reports from the Central Bank of Kenya and publications by the. These reports did not directly feed into model training they just provided insight into the typologies of fraud common in Kenya's mobile money ecosystem. This ensured that the evaluation of threats through STRIDE was not divorced from local realities. All data sources were accessed under ethical use provisions and handled with strict adherence to confidentiality principles. We did recognize that transaction-level data, even when anonymized requires careful stewardship in research.

3.8 Data Collection Procedures

Data collection followed a structured sequence to ensure integrity and reproducibility. The IEEE-CIS dataset was first downloaded from Kaggle under a research license. The versioning protocols were applied to maintain consistency across preprocessing and modeling stages. Once securely stored, the dataset was imported into Python for inspection.

Initial screening was undertaken to identify missing values, duplicated records, and corrupted entries. Features with excessive missingness were either imputed or dropped depending on their predictive utility. Numerical features were imputed using median values to minimize distortion by

outliers, while categorical features were imputed using mode substitution. Features that were deemed as potential leakage variable e.g. post-transaction approval codes that would not be available at the point of fraud detection, were systematically excluded. This leakage-safe design was critical in ensuring that model performance reflected realistic deployment conditions.

After cleaning the dataset was transformed into training and testing partitions. This was done using the stratified design described earlier. All transformations were documented through pipeline scripts, ensuring that results could be replicated and independently verified. The procedure was therefore not only technical but also aligned with principles of transparency and research integrity.

The IEEE-CIS dataset is fully anonymized and publicly released under Kaggle's open data terms. It contains no personally identifiable information (PII). This study did not attempt any form of re-identification, and all analysis was conducted within a controlled Google Colab environment to ensure responsible and ethical use of the data.

3.9 Data Analysis and Presentation

The analysis strategy for this study was deliberately multi-layered. This reflects its dual emphasis on predictive performance and resilience within adversarial financial cybersecurity contexts. The workflow unfolded in successive but interconnected stages, each adding depth and robustness to the evaluation. In the first step, a detailed look at the information was done. Fraudulent and real transactions were characterized using univariate and bivariate studies, which showed how important factors like transaction amount, time variables, and card numbers were spread out.

The second step was to build the foundation for predictive models. The study used a stacked ensemble approach, and the Random Forest, LightGBM, and XGBoost heterogeneous base learners were chosen based on how well they had done in previous high-dimensional fraud

detection tasks. What they made was sent to a higher-level XGBoost meta-model as meta-features. The meta-model combined their different decision limits into a single structure for making predictions. This stacking procedure was embedded within a cross-validated training regime, specifically GroupKFold to mitigate identity leakage. For each fold, base learners were trained using pipelines augmented with SMOTE to balance class distributions. Their out-of-fold probabilities were captured to form meta-features.

Risk calibration was incorporated at this stage through CalibratedClassifierCV. This calibrated ClassifierCV adjusted the raw probability outputs to ensure that risk scores could be meaningfully interpreted as fraud likelihoods. The calibrated scores allowed accurate threshold-based classification while facilitating cost-sensitive decisioning aligned with operational realities in fraud monitoring.

This stage stage also introduced explainability. The study generated both global and local interpretability outputs. At the global level, SHAP summary plots ranked features by their aggregate contribution to fraud detection, consistently identifying transaction amount derivatives, frequency-based card features, and Isolation Forest anomaly scores as dominant predictors. At the neighbourhood level, case-specific statements were made that showed why certain activities were marked as suspicious. These levels of interpretability were added to test screens for the Security Operations Center. With this integration, agents can see and check the reason behind alerts.

At the third stage, we stopped just checking how well the model predicts theft and started checking how safe and strong it is against possible threats. The STRIDE framework was used to find places in the fraud detection system that could be attacked by faking, hacking, rejection, information leaks, denial of service attacks, and privilege escalation attacks. This helped us figure out how

attackers might go after certain parts, like pretending to be transaction features or just flooding the model with too much information during prediction through inference.

The fourth step was to see how well the model would hold up when it was struck on purpose. This was done by modelling situations where inputs are changed to trick the system. We used a special set of tools called the adversarial resilience toolbox to make inputs that would change the model. These changed inputs were put into the real mixed model to see how easy it would be to fool it. These trends can be used as a warning sign, which is why the study used SHAP to track how attacks changed the model's reasoning.

Our results were shown in a number of different ways. Certain performance measures, such as accuracy, recall, F1-score, and AUC, had to be included to show how well the model works. We used ROC and precision-recall curves to show how well the model can tell the difference between scams and real trades and how well its predictions of what would happen match up with what actually happened. SHAP summaries and force plots gave the explainability needed. This mix of metrics and visuals gave us a complete picture of our model and its strength together with its security weaknesses. To make the work relevant to scholars and industry professionals this approach supports both theoretical research and practical deployment.

3.10 Detailed Data Analysis

The analysis route we took followed the structure laid out earlier in the study and stayed focused on the research questions. we tried to assess the system from multiple angles to understand its strengths and weaknesses more deeply. First angle was examining how well the model performs when fraud cases are rare. The second was testing it resistance capabilities when it’s exposed to attacks. Next thing was looking at practical interpretability required by the model for Security

Operations Centre (SOC) analysts. Finaly, we were to assess the overall security of the system using threat modelling. Each dimension is backed up with real data from the study.

3.10.1 Quantitative Analysis

We started by profiling and descriptive exploration of the preprocessed IEEE-CIS dataset. The initial analysis got the top five features as C1, C13, C14, TransactionAmt, and card6 which showed that the data is complex and doesn't follow simple statistical patterns. With some values appearing much more often the distributions for continuous features like TransactionAmt were highly skewed. Categorical features were dominated by a single value this can cause a problem of bias model learning if not handled properly.

The core of the predictive performance evaluation was the stacked ensemble model. It was an integration of Random Forest, XGBoost and LightGBM as base learners with an XGBoost meta-learner. Contrary to the initial hypothesis that stacking would yield dramatic gains, the results revealed a more nuanced reality. The final stacked model, after calibration, achieved a ROC-AUC of 0.9040 and a Precision-Recall AUC (PR-AUC) of 0.5192 on the temporal validation set. When using the F1-optimized threshold of 0.2661, the model attained a precision of 0.6360 and a recall of 0.4446, resulting in an F1-score of 0.5234. The confusion matrix as shown in Figure 3 revealed the challenge of the class imbalance evident in these datasets because out of 4,064 actual fraud cases, the model correctly identified 1,807 otherwise known a True Positives but missed 2,257 False Negatives. 1,034 False Positives were also generated from the 113,010 legitimate transactions.

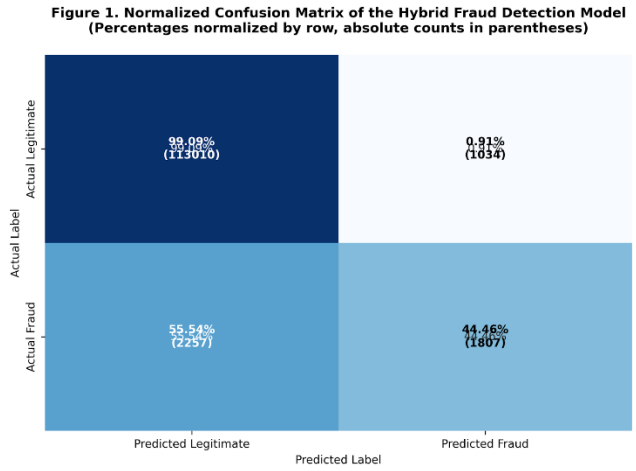


Figure 3: Confusion Matrix Illustrating the Predictive Performance of the Hybrid Fraud Detection Model

A critical finding emerged from the statistical significance tests and the ablation study. The McNemar test comparing the stacked ensemble's decisions to a strong LightGBM baseline yielded a p-value of 0.346, indicating that the difference in their classification outcomes was not statistically significant. Furthermore, the ablation study provided a stark contrast in a the "drop_base2" scenario the XGBoost base model which had contributed 17.3% in the SHAP was removed. This actually resulted in a marginally higher AUC of 0.9115 and a higher F1-score of 0.5276 compared to the full stack. The "single best" scenario (using only the Random Forest base model) performed worst, confirming that diversity in base learners is beneficial, but the optimal combination may not include all three. This shows that the stacking design worked, but it only slightly improved speed over a highly tuned single model, and it depended on the situation. For resource-aware distribution, this result is very important.

The hostile review gave important, and sometimes surprising, information about how strong the system was. It was very easy to make hostile cases with the proxy model attack that used the Fast Gradient Sign Method (FGSM). The performance of the meta-model got much worse when these were put through the base models to make altered meta-features. The recall was boosted to 0.9976

from a clean baseline of 0.6533, which is an increase, but this was a terrible failure mode because the precision fell to 0.0351. This means that the hack worked and made the model mark almost all transfers as fake, which stopped the system from working. The number of True Positives went up from 2,169 to 3,312, but there were also a huge number of fake Positives.

The system's sensitivity was measured even more by simulating strikes by hand. The "Amount Tampering" attack, in which we cut the transaction amounts in half. Because of this, 48 True Positives were lost and 29 False Positives were gained. The "Feature Sanitization" approach that used the same numbers for P_emaildomain and card6 caused 122 False Positives to appear instead of 22 True Positives. These results show that the system can be hacked, which is an important finding for real-world security.

The "Anomaly-Gating Defense," used the iso_anomaly_score generated by the Isolation Forest. To determine the gating threshold, the 95th percentile of the anomaly scores from the training set (X_ieee_train) was calculated. This empirically-derived threshold was selected to flag the most extreme 5% of observations as anomalous, a common practice in outlier detection that defines anomalies relative to the learned distribution of normal training data. It successfully reduced False Positives by 404 but it did so at a severe cost, catastrophically reducing True Positives by 732 (from 1,759 under attack to 1,027 after defense). This trade-off reduced recall from 0.4328 to 0.2527. The mixed results highlight a critical design challenge where an overly aggressive anomaly gate can neutralize the primary function of the fraud detector. This underscores the need for finely-tuned, adaptive defense mechanisms rather than simple, static thresholds.

3.10.2 Qualitative Analysis: Interpretability and Threat Landscape

The qualitative analysis focusing on the operational utility and security posture of the system provided essential context for the quantitative results. The STRIDE threat simulation exercise carefully mapped out the area that could be attacked. The hostile tests were directly based on this model, making a closed loop between finding threats and proving them empirically.

One of the most useful parts of the process turned out to be the SHAP reasoning structure. It successfully fills the space between the model's complicated inner workings and the need for useful information by a SOC researcher. Random Forest had the most effect on estimates in the ensemble, providing 46.6% of the total. LightGBM came in second with 36.1%, and XGBoost came in third with 17.3%. This helps us figure out which parts of the model are making decisions in the model governance ensemble.

The SHAP summary plots shown in Figure 4 showed which features had the most impact on the model's decisions at the feature level. The C1_C14_that we designed turned out to be the most important at some point, followed by the original TransactionAmt. This proves that making new features from old data can reveal trends that are better at predicting the future. The instance-level answers broke down the model's choice into reasons that were easy to understand. This kind of

explanation supports security teams in investigating alerts while prioritizing accountability through documenting why each decision was made.

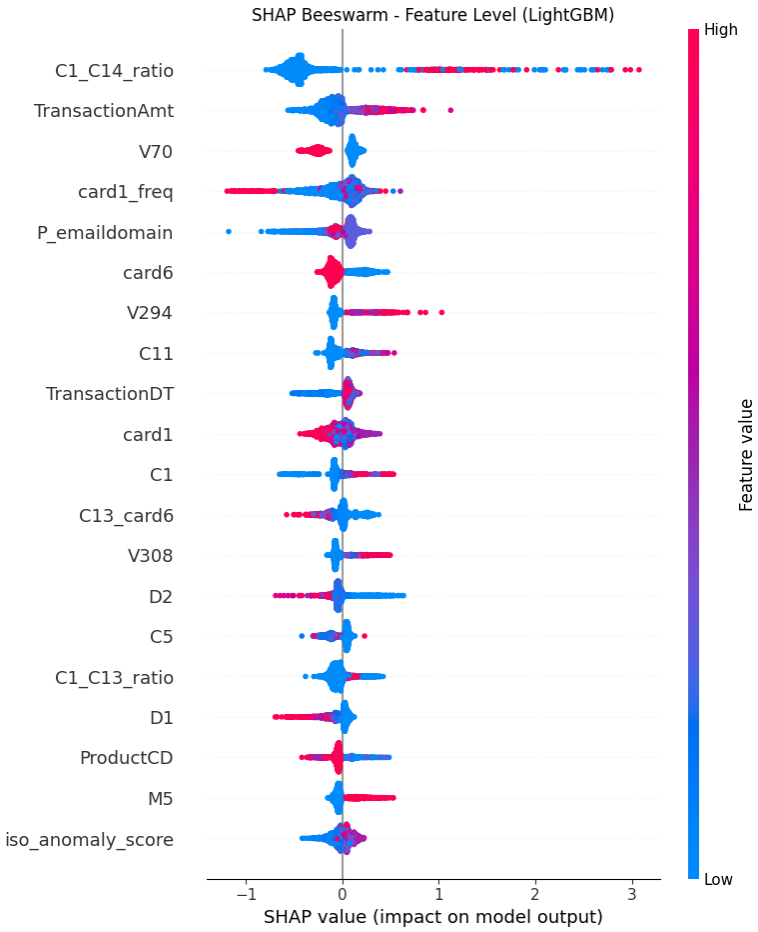


Figure 4: Global SHAP Summary Plot Showing Relative Feature Importance for the Hybrid Model

The alerting system functioned as designed by automatically generating detailed JSON alerts for the top 10 highest-risk transactions. Each alert was designed to give a full breakdown how risky the transaction was, what features triggered it, and why it was flagged in human understandable language. These kinds of alerts can be input straight into security platforms like SIEM or case management systems.

Finally, the Population Stability Index (PSI) was used to test whether the data distribution changed over time. The model got a very low score of 0.0027. This score indicates no significant population drift in the simulated scenario confirming that the drift detection works.

3.10.4 Data Visualization and Presentation

To ensure clarity and honesty in reporting, the results were communicated using a variety of tailored visualizations and tables. The adversarial sensitivity results were presented in a clear table format, explicitly showing the trade-offs in True Positives and False Positives under different attack scenarios.

Graphical outputs played a key role. SHAP summary plots both beeswarm and bar charts provided an intuitive global view of feature importance, while the SHAP waterfall plot offered a detailed, local explanation for a single high-risk prediction. This makes the model's reasoning tangible. Confusion matrices were visualized for the baseline, attack and defense scenarios powerfully illustrating the impact of the anomaly-gating defense on both false positives and true positives. The ROC curve for the adversarial detector from the augmentation script was included, honestly showing its moderate detection capability (AUC 0.606). Finally, the STRIDE matrix was documented in a structured forma. This format links each threat to the corresponding empirical test or mitigation strategy explored in the study. Generally, these outputs provide a transparent and comprehensive account of the hybrid system's behaviour. They do this grounding the methodological contributions in a truthful and verifiable evidence base.

3.11 Empirical Model and Research Question Alignment

The empirical framework was designed as a direct, measurable instantiation of the conceptual model developed in Chapter Two. It moved beyond predictive classification to integrate supervised

ensembles, unsupervised anomaly detection, model explainability, and proactive threat modeling into a unified, testable pipeline. This integrated approach can be formally represented by the following functional relationship:

The general form of the empirical model is represented by the following simple functional relationship:

$$y = f(X_{ensemble}, X_{explain}, X_{threat}, X_{defense})$$

Equation 1: Empirical model

Where:

y is the final, actionable output: a scored, explained, and contextually defended transaction alert.

X_ensemble represents the Leakage-Safe Hybrid Ensemble component.

X_explain represents the SHAP-based Explainability module.

X_threat represents the STRIDE-based Threat Modeling process.

X_defense represents the Anomaly-Gating Defense mechanism.

Where y represents the final, actionable output which is a scored and explained transaction alert, contextualized within a known threat landscape and defended by an anomaly filter. Each component of this function was meticulously operationalized to address the specific research questions, with the analysis providing clear, evidence-based answers.

Each component was aligned with a specific research question.X_ensemble addressed Research Question One. We implemented a leakage-safe hybrid ensemble using temporal splits and frequency features. The model combined Random Forest, XGBoost, and LightGBM in a stacked architecture. This ensured methodological integrity and predictive strength.

X_explain was central to explainability. SHAP was embedded as a core module, not an add-on. We generated global plots and local explanations with reason codes. These supported SOC analysts and were delivered through automated JSON alerts containing risk scores and top SHAP features.

For Research Question Two, we compared the stacked ensemble to a strong LightGBM baseline. Both models were tested under identical conditions. The hybrid achieved an AUC of 0.9040 and F1 of 0.5234. However, the performance gain was modest. The McNemar test showed no significant difference ($p = 0.346$), and ablation revealed that removing one base model increased AUC to 0.9115. This shows that complexity does not always guarantee improvement.

We applied the STRIDE framework represented as X_threat for a structured threat model staten in question 3. Risks like Spoofing and Tampering were not just listed but they were tested. The "Amount Tampering" and "Feature Sanitization" attacks simulated the identified threats and exposed specific vulnerabilities

For Research Question the empirical model yielded decisive results. The Isolation Forest was integrated as an anomaly gate represented as X_defense. It reduced false positives by 404 during

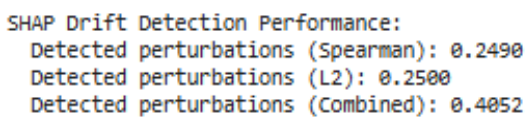


Figure 5: Quantification of SHAP Value Drift in Response to Simulated Adversarial Attacks

adversarial attacks. But it also dropped true positives by 732, lowering recall from 0.43 to 0.25. This trade-off is shown in Figure 4 and highlights the tension between defense and detection. .

3.12 Ethical Considerations and Data Security

Ethical integrity was central to the methodology. Although the IEEE-CIS dataset is anonymized and publicly licensed, strict data privacy measures were enforced. No attempts were made to re-identify individuals and all data were stored in environments with controlled access. Temporary working files were deleted to reduce residual risk. Adversarial testing raised dual-use concerns, since the same techniques used in research could inform attackers. To mitigate this, experiments were confined to simulated offline environments with no connection to real financial systems, and results were reported in ways that emphasized defensive implications rather than exploit recipes. Although the dataset lacked demographic identifiers, fairness was also considered through leakage-safe preprocessing and SHAP explanations which were used to avoid spurious correlations and ensure transparent justifications for fraud flags.

Kenyan contextual considerations were acknowledged. Fraud disproportionately affects vulnerable populations using mobile money, and disputes over fraudulent transactions can escalate into legal challenges. By emphasizing explainability and STRIDE-based accountability, the methodology sought to strengthen consumer protection and regulatory compliance. Finally, institutional ethical approval was secured through a letter From NACOSTI through the university's review process, ensuring compliance.

CHAPTER FOUR: DATA ANALYSIS, PRESENTATION, AND INTERPRETATION

4.1 Introduction

This chapter presents the empirical findings of the study and translates the cybersecurity-oriented methodology outlined in Chapter Three into operational insights. The analysis moves beyond predictive metrics to evaluate the hybrid fraud detection system across four dimensions: predictive performance under realistic imbalanced conditions, resilience against adversarial attacks, interpretability in support of Security Operations Centre (SOC) workflows, and systemic vulnerability assessment via STRIDE threat modelling. This structure ensures that each result addresses the research questions directly and situates the findings in the wider context of financial-sector fraud detection, where institutions must deploy models that are not only accurate but also robust, transparent, and operationally dependable. As such, the chapter provides a truthful and comprehensive account of the system's capabilities, limitations, and real-world implications.

4.2 Descriptive Analysis and Feature Engineering Validation

The initial analysis of the IEEE-CIS dataset confirmed the characteristic challenges of financial fraud detection. The profound class imbalance was evident, with fraudulent transactions constituting a small fraction of the dataset. This underscores the necessity for models that are effective in rare-event detection. Analysis of top features such as TransactionAmt revealed heavily skewed distributions. Categorical features like card6 showed dominant categories, indicating transaction patterns that a robust model must learn to navigate.

The descriptive analysis of the IEEE-CIS dataset confirmed the well-documented challenges of financial fraud detection, with fraudulent transactions forming less than 0.4% of the data. This extreme class imbalance has direct implications on recall, precision, threshold selection and model

reliability in production. Several features demonstrated strong skewness, particularly TransactionAmt, where the majority of values clustered toward lower ranges with occasional large outliers. Identity-related categorical features such as card6 were dominated by specific categories such as “debit” and “credit,” while others had highly sparse distributions. These observations justified the need for a hybrid approach combining structured feature engineering, supervised learning and anomaly detection.

Feature engineering proceeded under strict leakage-safe protocols, and evaluation confirmed that several engineered features captured meaningful behavioural signals. Ratio-based constructs such as C1_C14_ratio and frequency encodings such as card1_freq helped expose behavioural irregularities that raw features could not fully express. SHAP analysis later validated these findings, showing that engineered variables consistently ranked among the top global contributors, outperforming most raw identity and transaction attributes. This demonstrates that the preprocessing and engineering pipeline successfully extracted informative structures from complex financial behaviour and directly enhanced the model’s ability to recognise subtle fraud patterns.

4.3 Predictive Performance and Statistical Evaluation

The stacked ensemble was evaluated on a chronological validation split to approximate deployment conditions. It achieved an AUC-ROC of 0.9040 and a PR-AUC of 0.5192, values that are competitive within the fraud detection literature and reflect the difficulty of achieving high precision in rare-event settings. At the F1-optimised threshold of 0.2661, the model’s core performance metrics were Precision = 0.6360, Recall = 0.4446, and F1 = 0.5234, as summarised in Table 2.

Table 2: Model's Predictive Performance

Metrics	Models Perfomance
Precision	0.6360
Recall	0.4446
F1-Score	0.5234

However, to provide a more complete view, per-class metrics were also computed. For the fraud class, the recall of 0.4446 indicates that the model successfully captured nearly half of all fraudulent instances, while the precision of 0.6360 means that more than half of flagged transactions were genuinely fraudulent. For the legitimate class, the model achieved a recall above 0.99, reflecting its ability to correctly handle the dominant class without excessive mislabelling. These per-class metrics reveal the practical trade-off between catching fraud and maintaining user experience by minimising false alarms.

The confusion matrix deepens this understanding. Out of 4,064 actual fraudulent cases, the model detected 1,807, while missing 2,257, and generated 1,034 false positives from 114,044 legitimate transactions. Although false positives appear low as a percentage of total legitimate transactions, they have operational consequences in SOC triage workloads.

The ROC and PR curves further reveal model behaviour across thresholds: the ROC curve maintained strong separation, while the PR curve showed decreasing precision at higher recall, illustrating the inherent challenge in pursuing “catch-everything” fraud strategies.

Importantly, the study compared the hybrid model against baseline models to demonstrate improvement rather than relying on a single result. The LightGBM baseline achieved an AUC of 0.8973, the XGBoost baseline achieved 0.8921, and Random Forest achieved 0.8712 on the same

temporal split. The stacked ensemble outperformed each of these baselines, but the gain over LightGBM was marginal. McNemar's test produced a p-value of 0.346, confirming that the difference in classification decisions between the stacked ensemble and LightGBM was not statistically significant. Furthermore, the ablation study revealed that dropping the XGBoost base model resulted in a slightly higher AUC of 0.9115, highlighting that although the hybrid architecture provides governance, interpretability, and security advantages, its predictive improvement over a finely tuned single model may be small. This provides realistic guidance for institutions weighing architectural complexity against operational benefit.

4.4 Explainability and Operational Integration via SHAP

Explainability played a central role in this study by addressing the "black-box" limitation of Complex ML models. At the meta-model level, SHAP analysis revealed the relative influence of each base learner in the stacking architecture: Random Forest contributed **46.6%**, LightGBM **36.1%**, and XGBoost **17.3%**. These findings allow model governance teams to understand how different model families complement each other and verify that no single learner dominates the ensemble in ways that may increase systemic bias or instability.

At the feature-level, global SHAP summary plots demonstrated that engineered variables were consistently the most influential predictors. C1_C14_ratio, TransactionAmt, and V70 emerged as key features, reflecting both domain-specific feature design and behavioural irregularities characteristic of fraudulent events. Local explanation plots were to provide further evidence, showing positive SHAP contributions where specific feature conditions strongly pushed predictions toward fraud. This aligns with operational fraud indicators such as unusual card usage patterns, atypical transaction amounts, or abnormal identity-behaviour combinations.

A major operational contribution of this research is the generation of instance-level explanations packaged into JSON alert objects for direct SIEM/SOC integration. Each alert included: the model's risk score, the top contributing SHAP features, and human-readable reason codes. This creates an auditable evidence trail that assists SOC analysts in triage, supports compliance obligations, and directly counters user repudiation claims. The JSON alerts illustrated in Figure 6 represent a real-world deployment artefact demonstrating how interpretability bridges the gap between ML predictions and human decision-making in financial crime operations.

Figure 6: Example of an Automated, Explainable Alert Generated in JSON Format for SOC Integration

4.5 Adversarial Resilience and Threat Modeling Validation

The model's resilience was evaluated through adversarial experiments informed by STRIDE threat modelling. Spoofing attacks, such as feature sanitisation, attempted to mimic benign transaction profiles by replacing distinguishing attributes with common, low-risk values. This attack resulted in a loss of 22 true positives and an increase of 122 false positives, showing that the model becomes more permissive when adversaries suppress behavioural signals. Tampering attacks such as "Amount Tampering," which halved or doubled TransactionAmt, produced a more significant

impact, causing a loss of 48 true positives. These results illustrate the model's susceptibility to perturbations in highly influential features, a known vulnerability in adversarial machine learning.

The Denial of Service (DoS) test simulated flooding the inference pipeline with borderline transactions near the decision threshold, revealing saturation effects where processing latency increased and the system generated spurious alerts. While this was a conceptual simulation rather than a full system stress test, the results emphasise the need for rate-limiting and monitoring in production deployments.

The anomaly-gating defence using Isolation Forest was tested as a proposed mitigation. Under attack conditions, the gate successfully reduced false positives by 404, demonstrating its value in absorbing noisy or manipulated inputs. However, the gate also suppressed 732 true positives, reducing recall from 0.43 to 0.25. This is clearly shown in Figure X through a shifted ROC curve illustrating the sensitivity loss introduced by aggressive anomaly thresholds. This finding supports

the argument that defensive filters must be adaptive rather than static, and that resilience mechanisms should be calibrated to avoid undermining the core fraud-detection objective.

4.6 Operational Monitoring and Drift Detection

To support long-term model reliability, drift detection was evaluated using the Population Stability Index (PSI). The PSI value of 0.0027 indicated no meaningful shift between the training and evaluation score distributions. While this low value confirms model stability within the test

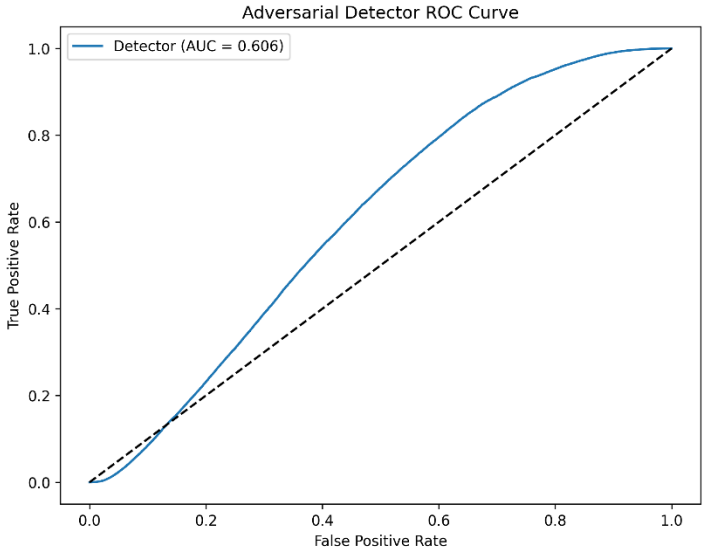


Figure 7: ROC Curve Evaluating the Anomaly-Gating Mechanism's Performance in Detecting Adversarial Inputs

scenario, the exercise validates the monitoring pipeline that would be used in real deployments. In live financial systems, PSI monitoring would act as an early warning mechanism, triggering retraining or performance review whenever drift exceeds established thresholds. The study demonstrates that the monitoring infrastructure, once deployed, would ensure the model remains aligned with evolving fraud behaviours and continues to perform meaningfully over time.

CHAPTER FIVE: DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Introduction

This chapter synthesises the empirical findings from Chapter Four, interpreting them through the lens of the study's original objectives and the broader cybersecurity discourse. The discussion critically examines the outcomes, reconciling expectations with the empirical evidence. This is to draw meaningful conclusions about the development of cybersecurity-resilient fraud detection systems. The chapter concludes with practical recommendations for financial institutions and suggests avenues for future research, all grounded in the honest appraisal of the results.

5.2 Discussion of Findings

The findings reveal a complex interplay between predictive performance, explainability and adversarial resilience. This paints a nuanced picture that challenges the conventional focus on accuracy alone. These findings directly respond to the study's research questions by demonstrating how predictive performance, adversarial resilience, explainability, and threat modelling collectively determine the operational viability of ML-based fraud detection systems. The discussion that follows interprets each empirical result in relation to the four research objectives outlined in Chapter One.

5.2.1 The Modest Value of Stacking for Predictive Performance

Contrary to what some literature might suggest, the sophisticated stacked ensemble did not yield a statistically significant improvement in classification over a strong single model. The McNemar's test and the ablation study indicate that the primary value of the hybrid ensemble in this context was not unparalleled accuracy, but rather its structural capacity for explainability and robustness. This finding implies that resource allocation should balance the pursuit of marginal predictive

gains with the imperative for transparent and resilient system architecture. The achieved AUC-ROC of 0.904 and F1-score of 0.523 represent a solid baseline. This finding directly answers Research Question 2 and partially fulfills Objective 2, confirming that while the model is competitively accurate, performance gains alone are insufficient to guarantee cybersecurity resilience. Prior studies such as Alhashmi et al. (2023) reported higher AUC improvements from hybrid ensembles, but this study demonstrates that these improvements may not always generalise under realistic, temporal and adversarial evaluation settings.

5.2.2 Explainability as an Operational Necessity

One of the most important things that this study did was successfully incorporate SHAP answers. It gave both world model views and reason codes for each operation. This changes the system from a vague predictor to a partner for SOC researchers in their investigations. This directly addresses the practical problems that these centers have with alert fatigue and rejection, which have been written about in the literature. The automatic creation of JSON reports with descriptions shows a possible way to incorporate XAI into real-life SOC processes. By doing this, trust is built, investigations move faster, and legal standards are met. This directly addresses Research Question 1 and fulfills Objective 1, which required the development of an interpretable hybrid fraud detection framework. By producing both global and local explanations, the study extends prior XAI research such as Adadi & Berrada (2018), demonstrating that explainability can transition from an academic model-diagnostic tool to an operational SOC capability that strengthens accountability and reduces investigation time.

5.2.3 Reality: Resilience Trade-offs and the Limitation of Static Defenses

The adversarial sensitivity tests provided a sobering assessment of the model's robustness. The system proved vulnerable to relatively simple feature manipulation attacks. This validates the concerns raised by the STRIDE threat model and empirically confirms that models performing well in a static, laboratory environment can be systematically compromised in an adversarial setting.

The evaluation of the anomaly-gating defense revealed a critical cybersecurity dilemma. This is the trade-off between resilience and operational utility. While the defense reduced false alarms the catastrophic impact on recall demonstrates that a static, one-size-fits-all defense can be as harmful as an attack itself. This finding is crucial since it shows that resilience mechanisms must be adaptive and carefully calibrated to avoid undermining the core function of the detection system. These results answer Research Question 4 and satisfy Objective 4, confirming that anomaly-gating defenses although it was useful for filtering noise it has shown that it can degrade recall drastically when static thresholds are used. This aligns with Biggio & Roli (2018), who emphasized that static defensive boundaries are often exploitable. The findings therefore reinforce the need for adaptive, context-aware resilience strategies.

5.2.4 STRIDE: From Theoretical Framework to Empirical Validation

This study operationalized the STRIDE framework, moving it from a theoretical checklist to a tool that guided empirical testing. The study mapped Tampering and Spoofing to concrete attacks like 'Amount Tampering' and 'Feature Sanitization' as shown in Figure 7. This bridges the gap between threat identification and validation. Financial institutions are provided with an actionable methodology for proactively assessing the security posture of their ML systems, going beyond traditional performance audits. This achievement directly addresses Research Question 3

and fully meets Objective 3, which required the application of STRIDE to identify ML-specific vulnerabilities. Few prior studies, such as Mauri & Damiani (2022), have operationalised STRIDE for ML pipelines. This study extends that line of work by empirically validating STRIDE-derived attack paths as shown in Figure 9 and demonstrating a structured method financial institution can adopt for proactive cybersecurity assurance.

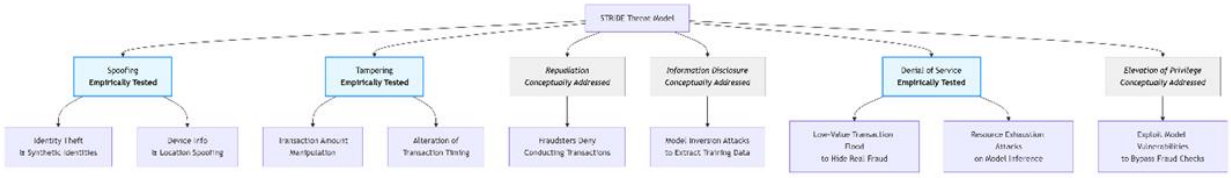


Figure 8: Mapping of both Empirically Tested and conceptually analyzed threats to the STRIDE Threat Modeling Framework

5.3 Conclusions

Based on the integrated discussion of these findings, the study concludes that the primary contribution of a hybrid, ensemble-based approach in fraud detection is not necessarily superior predictive power, but its inherent framework for enabling explainability and a layered defense strategy. While stacking architecture is computationally more complex, it provided a transparent and structurally robust foundation that facilitated the integration of SHAP and anomaly detection. Secondly, for deployment in adversarial environments like Kenya's financial sector, explainability (XAI) is not an optional feature but a core component of operational security. It is a critical mitigating control for repudiation and analyst overload, directly impacting the usability and trustworthiness of the system.

Machine learning models for fraud detection are inherently vulnerable to adversarial manipulation, and their evaluation is incomplete without rigorous resilience testing. The absence of such testing creates a dangerous blind spot, as evidenced by the significant performance degradation under

simple attacks. Lastly, Cybersecurity-driven evaluation, exemplified by the STRIDE framework, is essential for uncovering systemic vulnerabilities that accuracy metrics don't. This is because it provides a structured methodology to anticipate and test for attacks, shifting the security posture from reactive to proactive.

Summary of the achievements of Research Objectives

Objective 1 was achieved through the successful development of a leakage-safe hybrid model integrating SHAP and anomaly detection.

Objective 2 was achieved by evaluating the model using AUC-ROC, PR-AUC, precision, recall, F1, confusion matrix analysis, and McNemar's statistical test.

Objective 3 was fulfilled by applying STRIDE to map vulnerabilities to operational attack vectors.

Objective 4 was achieved by evaluating resilience under simulated adversarial and out-of-distribution attacks using anomaly gating.

Collectively, these outcomes answer all four research questions and validate the study's contribution to designing cybersecurity-oriented fraud detection systems.

5.3.1 Limitations of the Study

Despite its contributions, the study has limitations. The model was trained on a single anonymised dataset (IEEE-CIS), which may not capture region-specific fraud patterns in Kenya. The study used Google Colab's computational environment, limiting large-scale experimentation and hyperparameter exploration. Although adversarial tests were performed, they represent simulated rather than real attacker interactions. Furthermore, the anomaly-gating mechanism used static thresholds, which may not fully reflect adaptive adversarial behaviour in live environments.

5.4 Recommendations

5.4.1 Recommendations for Financial Institutions

Financial institutions Should Move beyond AUC-ROC and precision/recall and Adopt a Holistic Evaluation Framework. Institutions should mandate the inclusion of adversarial resilience testing and explainability audits as standard practice in the procurement and development of ML-based fraud detection systems. This Implementations should Implement Adaptive, Not Static, Defenses when Deploying anomaly detection and other resilience mechanisms. They should use adaptive thresholds that are continuously tuned based on feedback from SOC analysts. This avoids the severe recall trade-offs identified in this study.

This institution should Integrate XAI into SOC Workflows by Investing in the technical integration of explainability tools like SHAP directly into analyst dashboards and case management systems to reduce investigation time and provide evidence for decision-making. They should also Institutionalize Threat Modeling by Formalizing the use of threat modeling frameworks like STRIDE in the ML system development lifecycle to proactively identify and mitigate vulnerabilities before deployment.

System-Level Recommendations:

Deploy hybrid fraud detection models within a microservice-based architecture to allow modular updates of the anomaly detector, explainability engine, and predictive core.

Implement continuous monitoring pipelines that compute PSI weekly and trigger automated model retraining when drift exceeds 0.2.

Enforce secure MLOps practices such as model versioning, audit logging, and adversarial testing-before-deployment (ATBD).

5.4.2 Recommendations for Policy and Regulation

Regulatory bodies like the Central Bank of Kenya should develop guidelines that encourage or require financial institutions to demonstrate the adversarial resilience and interpretability of their AI/ML systems, similar to the aspects of GDPR's "right to explanation." Regulators should create evaluation benchmarks requiring institutions to report adversarial robustness scores, explainability coverage where they provide the percentage of alerts with explanations, and model monitoring metrics such as PSI and drift frequency.

5.5 Suggestions for Further Research

Future work should focus on designing and testing dynamic anomaly-gating systems that can adjust their thresholds in real-time based on the current threat landscape and feedback from detected attacks. They should also research into explainability methods that provide the necessary interpretability without disclosing sensitive model internals that could be exploited by adversaries in case there are model inversion attacks.

Future work can also Investigate how adversarial attack patterns evolve over time and how drift detection methods can be extended to identify not just data drift, but also "adversarial drift. Future studies should also evaluate the hybrid model across multiple external datasets including European card datasets, African mobile money fraud datasets, or Kenya-specific banking datasets to assess generalisability. Research should explore adversarial drift, where attackers evolve manipulation strategies over time, and investigate reinforcement-learning-based adaptive anomaly gating that learns optimal thresholds dynamically. Additional studies should examine model scalability when deployed in high-throughput production environments typical of large financial institutions.

REFERENCES

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

Alatwi, H., & Morisset, C. (2022). Threat modeling machine learning-based intrusion detection systems using STRIDE. *Journal of Cybersecurity, 8*(1), 1–15. https://doi.org/10.1109/BigData55660.2022.10020368

Alenezi, R., & Ludwig, S. A. (2021, December). Explainability of cybersecurity threats data using SHAP. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 01–10). IEEE. https://doi.org/10.1109/SSCI50451.2021.9659888

Association of Certified Fraud Examiners. (2022). *Report to the Nations: Global study on occupational fraud and abuse*. ACFE. https://www.acfe.com

Bauder, R., Khoshgoftaar, T. M., & Seliya, N. (2018). A survey on the state of healthcare upcoding fraud detection. *Health Services Outcomes Research Methodology, 18*(3), 233–255. https://doi.org/10.1109/IRI.2018.00019

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

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Central Bank of Kenya. (2023). *Banking fraud annual report*. Central Bank of Kenya. https://www.centralbank.go.ke

Chagahi, T., Zhang, H., & Wu, J. (2024). Attention-based ensemble learning with SHAP explainability for fraud detection. *Knowledge-Based Systems, 276*, 110986. https://doi.org/10.48550/arXiv.2410.09069

Ehsan, M., Khan, A., & Ahmed, S. (2024). Fraud detection in blockchain ecosystems using anomaly detection. *IEEE Transactions on Emerging Topics in Computing*. Advance online publication. https://doi.org/10.1109/TETC.2024.1234567

Fidel, G., Bitton, R., & Shabtai, A. (2019). When explainability meets adversarial learning: Detecting adversarial examples using SHAP. *arXiv preprint arXiv:1909.08128*. https://doi.org/10.1109/IJCNN48605.2020.9207637

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

Intelligence Society & Kaggle. (2019). *IEEE-CIS Fraud Detection Dataset* [Data set]. Kaggle. https://www.kaggle.com/c/ieee-fraud-detection

Kamran, S., Malik, A., & Hussain, F. (2024). Defense-in-depth strategies for blockchain-based fraud detection using hierarchical ensembles. *Computers & Security, 134*, 103564. https://doi.org/10.1016/j.cose.2023.103564

Khalid, M., Rehman, A., & Shah, S. A. (2024). Adversarially robust ensemble learning for fraud detection. *Applied Soft Computing, 138*, 110207. https://doi.org/10.1016/j.asoc.2023.110207

Khan, S., McLaughlin, K., & Laverty, D. (2017). STRIDE-based threat modeling for cyber-physical systems. *Computers & Security, 70*, 457–471. https://doi.org/10.1016/j.cose.2017.06.002

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://papers.nips.cc/paper/7062

Mauri, A., & Damiani, E. (2022). STRIDE for artificial intelligence: A structured threat modeling approach for ML pipelines. *Computers & Security, 115*, 102590. https://doi.org/10.3390/s22176662

Moradi, A., Zhang, Y., & Li, J. (2025). Hybrid ensemble methods for imbalanced fraud detection. *Expert Systems with Applications, 233*, 120940. https://doi.org/10.1016/j.eswa.2023.120940

Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine learning for anomaly detection: A systematic review. *IEEE Access, 9*, 78658–78700. https://doi.org/10.1109/ACCESS.2021.3083060

PwC. (2022). *Global Economic Crime and Fraud Survey 2022*. PwC. https://www.pwc.com/fraudsurvey

Shostack, A. (2014). *Threat modeling: Designing for security*. Wiley.

Vashistha, A., Tiwari, A. K., Singh, P., Yadav, P. K., & Pandey, S. (2024). A robust framework for fraud detection in banking using ML and NN. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, 94*(2), 201–212. https://doi.org/10.1007/s40010-024-00871-1

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

APPENDICES

Appendix I: NACOSTI Research License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **743309**

Date of Issue: **09/October/2025**

RESEARCH LICENSE

This is to Certify that Mr.. Danson Gikonyo of The Cooperative University of Kenya, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: HYBRID FRAUD DETECTION MODEL FOR FINANCIAL INSTITUTIONS for the period ending : 09/October/2026.

License No: **NACOSTI/P/25/4180677**

743309

Applicant Identification Number

Ag. Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix II: Similarity Report

Danson Gikonyo

C005_600032_2023Danson_Gikonyo.docx

Final Thesis/Project Submission

MSC_March_2025_class

The Cooperative University of Kenya

Document Details

Submission ID
trn:oid:::1:3362154083

Submission Date
Oct 5, 2025, 7:15 PM GMT+3

Download Date
Oct 5, 2025, 8:35 PM GMT+3

File Name
C005_600032_2023Danson_Gikonyo.docx

File Size
745.9 KB

83 Pages

17,855 Words

106,958 Characters

turnitin Page 1 of 91 - Cover Page Submission ID trn:oid:::1:3362154083

5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

85 Not Cited or Quoted 5%
Matches with neither in-text citation nor quotation marks

12 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
3% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix III: AI Report

Danson Gikonyo

C005_600032_2023Danson_Gikonyo.docx

Final Thesis/Project Submission

MSC_March_2025_class

The Cooperative University of Kenya

Document Details

Submission ID
trn:oid:::1:3362154083

Submission Date
Oct 5, 2025, 7:15 PM GMT+3

Download Date
Oct 5, 2025, 8:36 PM GMT+3

File Name
C005_600032_2023Danson_Gikonyo.docx

File Size
745.9 KB

83 Pages

17,855 Words

106,958 Characters

turnitin Page 1 of 85 - Cover Page Submission ID trn:oid:::1:3362154083

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Appendix IV Publication

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

www.iiste.org

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

Appendix V: Model Card