DESIGN AND EVALUATION OF MACHINE LEARNING - BASED FRAMEWORK FOR FRAUD DETECTION IN KENYA'S DIGITAL PAYMENTS ECOSYSTEM

BY

DENIS GITHU GITAU

MASTER OF SCIENCE IN DATA ANALYTICS

KCA UNIVERSITY
2025

DESIGN AND EVALUATION OF MACHINE LEARNING - BASED FRAMEWORK FOR FRAUD DETECTION IN KENYA'S DIGITAL PAYMENTS ECOSYSTEM

BY

DENIS GITHU GITAU

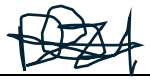
A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN DATA ANALYTICS IN THE SCHOOL OF TECHNOLOGY AT KCA UNIVERSITY

AUGUST ,2025

DECLARATION

I declare that this dissertation is my original work and has not been previously published or submitted elsewhere for the award of a degree.I also declare that this contains no material written or published by other people except where due refrence is made and author duly acknowledged.

Student Name: Denis Githu Gitau Reg, No: 23/031/30

Sign:  Date: 25th April 2025

I do hereby confirm that I have examined the master's dissertation of

Denis Githu Gitau

And have certified that all revisions that the dissertation panel and examiners recommended have been adequately addressed.

Sign  Date 20th July 2025

Dr. Kevin Amugoye
Dissertation Supervisor

DESIGN AND EVALUATION OF MACHINE LEARNING - BASED FRAMEWORK FOR FRAUD DETECTION IN KENYA'S DIGITAL PAYMENTS ECOSYSTEM

ABSTRACT

The rapid growth of Kenya's digital payments ecosystem has revolutionized financial transactions, enabling greater financial inclusion through fintech innovations such as Point-of-Sale (POS) systems, mobile wallets, and e-commerce platforms. However, this expansion has also increased exposure to fraud, as cybercriminals exploit digital vulnerabilities and high transaction volumes to execute increasingly sophisticated schemes. Conventional rule-based fraud detection systems, which rely on static thresholds and predefined patterns, have proven inadequate in addressing evolving fraud tactics. They often result in high false-positive rates and delayed responses, ultimately compromising customer trust and financial integrity.

This study presents the design and evaluation of a machine learning based framework for fraud detection tailored to Kenya's digital payments ecosystem. Using anonymized transaction data from Pesapal Ltd, a leading regional fintech provider, the research applies a range of supervised learning algorithms including Random Forest, Gradient Boosting, Logistic Regression, Naïve Bayes, Decision Trees, K-Nearest Neighbors, and Neural Networks to identify the most effective approach for real-time fraud detection. The study follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology and incorporates feature engineering, Synthetic Minority Oversampling Technique (SMOTE), cost-sensitive learning, and explainability techniques to enhance model robustness and interpretability.

Model performance was evaluated using precision, recall, F1-score, AUC, and cost-based metrics to capture both statistical accuracy and business implications. After applying SMOTE, Random Forest demonstrated a superior balance between detection sensitivity and false-positive control, achieving Precision = 52.54%, Recall = 52.77%, F1-score = 52.66%, and PR-AUC = 53.55%, outperforming other models in identifying fraudulent transactions. The SHAP-based explainability analysis further highlighted the dominant role of transaction geography, processing bank, and card type in predicting fraud.

The results highlight the effectiveness of ensemble learning techniques for fraud detection in imbalanced financial datasets and demonstrate the value of explainable AI (XAI) in enhancing transparency and regulatory compliance. The study contributes to the ongoing discourse on financial cybersecurity in African fintech ecosystems by offering a scalable, interpretable, and contextually relevant fraud detection framework suited for real-time digital payment environments.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisor ,Dr.Kevin Amugoye, for his invaluable guidance, insightful feedback, and continuous support throughout the development of this dissertation. His expertise, patience, and encouragement played a key role in shaping the direction and quality of this research.

I am also grateful to the faculty and staff of KCA University for providing a supportive academic environment and the resources necessary to complete this study successfully.

Special appreciation goes to Pesapal Ltd for granting access to anonymized transaction data and for their cooperation during the research process. Their support made it possible to ground this study in real-world digital payment systems and fraud detection challenges.

Finally, I wish to thank my family, friends, and colleagues for their encouragement, understanding, and unwavering support throughout this academic journey.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACRONYMS

ATO	Account Takeover
API	Application Programming Interface
CRISP-DM	Cross-Industry Standard Process for Data Mining
EDA	Exploratory Data Analysis
ML	Machine Learning
POS	Point-Of-Sale
E-COM	E-commerce
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
BIN	Bank Indentification Number
TPV	Total Processing Volume
AI	Artificial Intelligence
GB	Gradient Boosting
NB	Naïve Bayes
DT	Decision Tree
RF	Random Forest
F1-Score	Harmonic Mean of Precision and Recall
CV	Cross Validation
PCA	Principal Component Analysis
FinTech	Financial Technology
SHAP	Shapley Additive exPlanations
DFS	Digital Financial Services

AML	Anti-Money Laundering
CFT	Countering the Financing of Terrorism
EAC	East African Community
RNN	Recurrent Neural Networks
LSTM	Long Short-Term Memory Networks
XAI	Explainable AI
PAPSS	Pan-African Payments and Settlement System
PII	Personal Indentification Information
SFTP	Secure File Transfer Protocol
OECD	Organization for Economics Co-operation & Development
PSP's	Payments Service Providers

TERMS AND DEFINATIONS

Fintech (Financial Technology)	The integration of technology into financial services to improve efficiency, accessibility, and delivery to customers.
Fraud Detection	The process of identifying and preventing unauthorized or suspicious financial transactions to protect users and financial institutions.
Machine Learning	A subset of artificial intelligence that uses algorithms to learn patterns from data and improve performance without explicit programming.
Predictive Analytics	The use of statistical methods and machine learning techniques to forecast future outcomes based on historical data.
Supervised Learning	A machine learning approach where models are trained using labeled data (e.g., Random Forest).
Unsupervised Learning	A machine learning approach where models identify patterns in unlabeled data (e.g., Isolation Forest).
False Positives	Legitimate transactions incorrectly flagged as fraudulent by a detection system.
Point-of-Sale (POS) Systems	Hardware and software used by merchants to process customer payments during sales transactions.
Account Takeover (ATO) Fraud	Fraud in which an unauthorized individual gains access to a legitimate user account to perform illicit transactions.
New Account Fraud	Fraud involving the creation of fake or synthetic identities to open new accounts for malicious purposes.
Real-Time Payment Fraud	Fraud committed during instant financial transactions, which is difficult to prevent due to limited processing time.
API-Based Attacks	Cyberattacks targeting application programming interfaces (APIs) to manipulate data or gain unauthorized access.
Pesapal Ltd	A digital payment service provider in East Africa that enables businesses and individuals to process online and POS transactions.

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

The global financial landscape has undergone a profound digital transformation driven by advances in technology, widespread mobile penetration, and the rise of financial technology (fintech) innovations. Over the past decade, digital financial services (DFS) have redefined how individuals and businesses transact, shifting from traditional cash-based systems to digital platforms that enable mobile money, online banking, e-commerce, and digital wallets. According to the World Bank (2023), digital payments now account for more than 60% of global financial transactions, underscoring the growing dependence on technology-driven solutions for financial inclusion, speed, and transparency. This transformation has also introduced complex transactional patterns, creating both opportunities for innovation and vulnerabilities for financial crime. In Sub-Saharan Africa, and particularly Kenya, this transformation has been remarkable. Kenya is internationally recognized as a pioneer in digital finance, largely due to the success of M-Pesa, which has enabled millions of unbanked individuals to access financial services. Complementary platforms such as PayPal, Pesapal Mobile, Equity Bank's Eazzy Banking, and Airtel Money have further strengthened the fintech ecosystem, facilitating a seamless digital payment culture across both urban and rural populations. According to the Central Bank of Kenya (CBK, 2023), more than 87% of adults in Kenya actively use digital financial services, reflecting deep integration of fintech into the country's economic and social fabric. The scale and diversity of digital transactions in Kenya create a unique environment in which fraud detection solutions must be tailored to local

patterns and operational realities. Furthermore, the rapid growth of cross-border payments within East Africa, facilitated by initiatives such as the

Pan-African Payments and Settlement System (PAPSS), highlights the increasing complexity of regional digital transactions and underscores the need for robust monitoring systems to mitigate financial crime. Mobile money adoption trends, such as the shift from person-to-person transfers to merchant payments and e-commerce transactions, have amplified transaction volumes and velocity, creating both new opportunities and vulnerabilities for fraud. This growing interconnectedness, while transformative, has also expanded the threat surface for malicious actors.

The evolving digital economy in Kenya is also influenced by the country's broader cybersecurity posture and regulatory ecosystem. Kenya ranks among the top three Sub-Saharan African countries most affected by cyber threats, according to the Communications Authority (2024), which attributes nearly 60% of these incidents to financial and e-commerce systems. The convergence of financial technology, cloud infrastructure, and API-driven integrations has created a hyperconnected environment where security risks are no longer confined to individual institutions. Instead, systemic vulnerabilities can cascade across the digital payment value chain affecting merchants, aggregators, and consumers simultaneously. This interdependence highlights the need for proactive monitoring systems capable of not only detecting anomalies but also contextualizing them within broader network behaviors.

However, with this rapid growth has come an escalation of financial fraud and cybercrime targeting digital payment systems. Fraudulent activities now exploit the very efficiencies that make digital transactions attractive speed, anonymity, and cross-platform connectivity. Kenya's Communications Authority (CA, 2023) reported a 40% increase in digital fraud incidents between 2020 and 2023, coinciding with a post-pandemic surge in online and mobile-based transactions. Fraud schemes range from card skimming, which involves

installing a device on an ATM or point-of-sale terminal to capture card information, to phishing, where attackers attempt to trick individuals into revealing sensitive information, such as login credentials or

financial information. Account takeovers occur when a fraudster gains unauthorized access to an individual's account, often through phishing or other means of obtaining login credentials. Chargeback manipulation happens when a consumer disputes a legitimate transaction, claiming it was unauthorized. Synthetic identity fraud involves creating a fictional identity using real and fabricated information to open accounts, obtain credit, or make purchases. These schemes threaten both consumers and service providers.

Regulatory developments have sought to keep pace with the rapid evolution of digital payments. The Kenya Data Protection Act (2019) establishes requirements for data privacy and transparency, while the CBK and other regulatory bodies provide guidelines for AML/CFT (Anti-Money Laundering and Countering the Financing of Terrorism) compliance, transaction monitoring, and risk management. Cross-border regulatory frameworks are also evolving, especially with the East African Community (EAC) harmonizing digital financial regulations to facilitate secure regional transactions. These regulations increase the accountability of fintech operators but also elevate the importance of accurate, explainable fraud detection systems

Traditional fraud detection systems, which rely on static, rule-based logic, have proven inadequate in this evolving environment. These systems depend on predefined parameters for instance, transaction limits or frequency thresholds that cannot adapt dynamically to new or unseen fraudulent behaviors. Consequently, they produce high rates of false positives (legitimate transactions incorrectly flagged as fraudulent) and false negatives (fraudulent transactions left undetected). Both outcomes are costly: false positives disrupt genuine customer transactions and erode trust, while false negatives lead directly to financial losses and

regulatory risks. As fintech companies increasingly handle high-volume, real-time transactions, the limitations of rule-based systems have become more pronounced (ACFE, 2023).

Moreover, the global regulatory environment is shifting toward the principle of "algorithmic accountability," requiring financial institutions to justify automated decisions that affect users. Institutions like the European Banking Authority (EBA, 2024) and the OECD (2024) recommend explainable artificial intelligence (XAI) as a compliance best practice, ensuring transparency in fraud detection models. As Kenya aligns its digital finance policies with global standards, fintech providers are expected to demonstrate not only technical accuracy but also fairness, interpretability, and non-discrimination in their predictive systems. This further emphasizes the importance of developing machine-learning models that are explainable and aligned with both local and international regulatory expectations.

Globally, the Association of Certified Fraud Examiners (ACFE, 2023) estimates that organizations lose approximately 5% of their annual revenue to fraud, with financial institutions among the most affected sectors. Within the African context, the economic cost of digital fraud is rising steeply. A TransUnion (2024) report indicated that 4.6% of Kenyan digital transactions were flagged as suspected fraud attempts in the first half of 2024, ranking Kenya among the top ten most affected countries globally. Similarly, a study by Intelpoint (2024) found that KES 810.7 billion was lost to mobile banking fraud in 2024 alone, highlighting the economic magnitude of the threat.

Against this backdrop, the need for intelligent, adaptive, and scalable fraud detection systems has become critical. Recent research advocates for the application of machine learning (ML) techniques that can learn complex, non-linear patterns from historical transaction data, detect anomalies in real time, and adapt to evolving fraud typologies. Unlike traditional models, ML-based systems continuously refine their understanding of transaction behavior, making them suitable for dynamic environments such as Kenya's fintech sector.

Machine learning algorithms such as Random Forest, Gradient Boosting, Support Vector Machines (SVMs), and Neural Networks have demonstrated superior performance in identifying fraudulent activities in financial transactions (Patil et al., 2022; Adewumi & Akinyelu, 2017). These models capture subtle correlations between transactional attributes such as amount, location, time, merchant type, and device identifiers which rule-based systems cannot easily encode. However, the application of these techniques in Kenya's fintech context presents unique challenges. Issues such as data imbalance (where genuine transactions far outnumber fraudulent ones), cost sensitivity (where misclassifying fraud has unequal consequences), and model interpretability (for compliance with CBK and Data Protection Act requirements) must be carefully considered.

Moreover, the effectiveness of any ML-based fraud detection framework depends on its ability to balance accuracy, speed, and explainability. While high accuracy is desirable, models that cannot explain their decisions ("black box" algorithms) may be unsuitable for regulated financial sectors. The need for explainable AI (XAI) is therefore paramount, allowing analysts and regulators to understand why a transaction was flagged as suspicious.

In this context, this study seeks to design and evaluate a machine learning-based framework for fraud detection within Kenya's digital payments ecosystem, emphasizing accuracy, interpretability, and operational feasibility. Using anonymized transaction data from Pesapal Ltd, a leading fintech company in the region, the study applies supervised ML algorithms to detect fraudulent activity. By integrating SMOTE-based balancing, cost-sensitive evaluation, and explainable AI techniques, the research aims to develop a robust and transparent fraud detection model suitable for real-world deployment in fintech environments.

The findings of this study will not only enhance academic understanding of fraud detection in emerging markets but also provide practical recommendations for fintech

institutions and regulators seeking to strengthen fraud prevention frameworks in Kenya and the wider East African region.

1.2 Problem Statement

While digital transformation has created immense opportunities for financial inclusion, it has also introduced complex risks that threaten the stability and integrity of fintech ecosystems. The rapid expansion of digital payment platforms has transformed Kenya's financial landscape, enhancing financial inclusion and transactional convenience for millions of users. However, this growth has been accompanied by a sharp rise in financial fraud targeting fintech systems. Cybercriminals exploit technological vulnerabilities and user behaviors through sophisticated methods such as account takeovers, card-not-present (CNP) fraud, phishing attacks, SIM swap fraud, and synthetic identity fraud. These attacks expose consumers, financial institutions, and payment service providers to significant financial losses, reputational damage, and erosion of trust in digital financial services.

In Kenya, the scale and severity of fraud have escalated dramatically. Serianu Ltd. (2022) reported that the financial sector lost over Kshs 5.6 billion to cybercrime in 2021, marking a concerning upward trend. The Communications Authority of Kenya (CAK, 2023) recorded a 59% increase in cybercrime incidents in 2022, with a substantial proportion targeting mobile money platforms and online payment gateways that form the backbone of Kenya's digital economy. The issue is further compounded by the rapid growth of fintech companies such as Pesapal Ltd., which process high volumes of Point-of-Sale (POS) and e-commerce transactions daily, creating an expanded attack surface for fraudsters operating in real-time. Within these channels, fraudulent activity often results in costly operational burdens false positives alone can account for 10–20% of declined transactions, with each manual investigation costing fintech’s between KES 50 and KES 150, ultimately accumulating into millions of shillings in avoidable annual losses.

Globally, the problem is equally severe. Juniper Research (2022) projects that fraud losses will exceed USD 41 billion by 2027, with CNP fraud accounting for over 70% of card-related fraud cases. This trend underscores the urgent need for more robust and adaptive fraud detection mechanisms, particularly in emerging markets like Kenya where digital payment adoption is outpacing security infrastructure development.

Despite these escalating threats, most Kenyan fintech companies continue to rely on rule-based fraud detection systems. These traditional systems operate using static thresholds, predefined transaction rules, and manual review processes. While they may detect known fraud patterns, they are fundamentally rigid and reactive, failing to adapt to new, evolving, or sophisticated fraud tactics employed by modern cybercriminals. Moreover, these systems are plagued by high false-positive rates. According to the Aite-Novarica Group (2020), up to 90% of transactions flagged by rule-based systems are false positives, leading to unnecessary transaction declines, excessive manual investigations, increased operational costs, and frustrating customer experiences that can damage brand loyalty. These challenges are particularly acute in POS and e-commerce environments where transaction velocity is high and customer sensitivity to declines is immediate.

The persistence of rule-based approaches reflects a critical gap in Kenya's fintech ecosystem: the absence of intelligent, adaptive, and context-aware fraud detection mechanisms that can learn from historical data, recognize complex fraud patterns, and adjust to emerging threats in real-time. This gap not only exposes fintech firms to mounting financial and operational risks but also raises concerns about regulatory compliance, particularly as frameworks like the Central Bank of Kenya's Agent Banking Regulations and Data Protection Act (2019) demand stronger consumer protection and data security measures. Importantly, static rules cannot model the dynamic, non-linear, and multi-dimensional fraud patterns increasingly observed in POS and e-commerce transactions patterns that machine learning

techniques are explicitly designed to learn and adapt to. Although recent academic work proposes various machine learning (ML) techniques for fraud detection, there remains no clear consensus on which models are most suitable for real-time deployment in African fintech environments. As Odhiambo (2023) observes, ML models developed in Western contexts often fail to generalize effectively to local datasets and transaction behaviors. Similarly, Kiptoo (2022) notes that while complex algorithms such as XGBoost and Neural Networks deliver high predictive accuracy, simpler models like Random Forest and Logistic Regression often outperform them in terms of speed, interpretability, and deployment feasibility crucial factors in resource-constrained fintech operations.

Given these challenges and the rapid digitization of Kenya's payment ecosystem, it is imperative to explore data-driven solutions that enhance fraud resilience. This research responds by proposing a structured, data-driven framework that integrates machine learning with regulatory transparency to address Kenya's unique fintech fraud landscape. The framework leverages supervised learning algorithms trained on real transaction data, incorporating advanced techniques such as Synthetic Minority Over-sampling Technique (SMOTE) for class imbalance, cost-sensitive learning to minimize misclassification costs, and explainability methods (SHAP) to ensure transparency and regulatory compliance. The goal is to develop a scalable, interpretable, and context-relevant fraud detection solution that enhances detection accuracy, significantly reduces false positives, and enables practical, real-time deployment within POS and e-commerce fintech operational environments.

1.3 Objectives of The Study

1.3.1 Main Objective

To design and evaluate a machine learning-based framework for detecting financial fraud in Kenya's digital payment platforms.

1.3.2 Specific Objective

1) To assess the limitations of existing rule-based fraud detection systems in digital payments.
2) To evaluate model performance using real-world fintech data and identify the most suitable algorithm based on precision, recall, F1-score and AUC.
3) To propose a practical, real-time machine learning framework suitable for fraud detection in Kenyan fintech environments.

1.4 Research Objectives

i. What are the main limitations of current rule-based fraud detection systems in fintech platforms?
ii. What transaction features or behavioral patterns are most indicative of fraud in the context of local digital payment platforms?
iii. How do different machine learning algorithms perform in detecting fraudulent transactions when applied to real-world fintech data?
iv. Which machine learning algorithm demonstrates the best overall performance for fraud detection in a local fintech environment?
v. How can Random Forest be operationalized in a real-time fraud detection framework in a fintech environment?

1.5 Motivation of The Study

This study seeks to advance fraud detection methodologies within the fintech industry, focusing on transactional data from Pesapal Ltd, a representative fintech company. The research critically examines the limitations of existing fraud detection systems, particularly their shortcomings in addressing dynamic and evolving fraud tactics, using historical.

Recent instances of suspected fraudulent activity at Pesapal have underscored the pressing need for more advanced and adaptive fraud detection systems. As a leading digital payments provider in East Africa, Pesapal processes large volumes of transactions daily across multiple platforms creating an urgent demand for scalable, intelligent solutions capable of identifying suspicious activity with greater precision and speed.

This study responds to that need by evaluating the performance of various machine learning algorithms to determine which offers the most effective fraud detection in a real-world fintech setting. While multiple techniques such as Gradient Boosting, Decision Tree, Logistic Regression, KNN, Neural Networks, and Naïve Bayes were explored, Random Forest emerged as the most reliable model, demonstrating strong performance across key metrics like recall, F1-score, and AUC.

The organization is particularly focused on reducing false positives, adapting to new fraud patterns, and improving decision-making speed. This research is therefore grounded in a real operational context and seeks to deliver actionable insights for improving fraud prevention. The findings also offer broader relevance for other fintech platforms seeking to implement machine learning solutions for real-time, accurate, and interpretable fraud detection.

1.6 Scope of The Study

This study seeks to advance fraud detection methodologies within the fintech industry, focusing on transactional data from Pesapal Ltd, a representative fintech company. The research critically examines the limitations of existing fraud detection systems, particularly their shortcomings in addressing dynamic and evolving fraud tactics, using historical transaction data from the company. The study encompasses the design, implementation, and validation of an adaptive machine learning-based predictive model, specifically tailored to the distinct fraud patterns observed in Pesapal's payment ecosystem. The model will be evaluated using the provided dataset to enhance the accuracy and effectiveness of fraud detection, with

the intent of deriving insights that can be extrapolated to other fintech platforms operating under similar conditions.

While the findings are grounded in a case study of a specific fintech company, they contribute to the broader academic discourse on the application of advanced fraud detection strategies within the fintech sector. The study underscores the potential of machine learning models to address contemporary fraud challenges and strengthen the security and reliability of digital financial platforms.

1.7 Limitations of The Study

This study is subject to several limitations that must be considered when interpreting its findings. The foremost limitation lies in the exclusive use of historical transactional data from Pesapal Ltd, which may constrain the generalizability of the results to other fintech platforms. The fraud patterns, operational structures, and customer behaviors specific to Pesapal Ltd may not necessarily reflect the broader landscape of the fintech industry. As such, while the predictive model developed in this study is tailored to Pesapal Ltd.'s context, its applicability to other fintech entities may be limited.

Another significant limitation pertains to the quality and completeness of the dataset. Historical transaction data may contain inherent inconsistencies, missing values, or noise, which could potentially affect the accuracy and robustness of the machine-learning model. To mitigate this, the study will employ rigorous data preprocessing techniques aimed at addressing these issues, ensuring that the dataset is optimized for effective analysis and model development.

Furthermore, the dynamic and evolving nature of fraud tactics introduces a temporal limitation to the study. The model's performance is based on historical data, and while its adaptive nature is designed to accommodate emerging fraud schemes, continuous monitoring

and periodic retraining of the model will be required to maintain its efficacy over time. As such, the model's ability to address future fraud trends will depend on its ongoing refinement.

Finally, the study focuses primarily on transactional fraud within the domains of POS and e-commerce environments. Consequently, it excludes other fraud types such as account takeover, money laundering, and API-based attacks, which may limit the broader applicability of the findings to a wide array of fraud detection challenges across the fintech industry.

Despite these limitations, the study aims to offer valuable insights into the application of machine learning techniques for fraud detection, with implications that extend beyond Pesapal Ltd and contribute to the broader discourse on security and fraud prevention in the fintech sector.

1.8 Significance of The Study

This study holds both practical and academic significance within the domains of fintech innovation and financial fraud detection. Fintech companies such as Pesapal Ltd. stand to benefit from a machine learning based fraud detection framework tailored specifically to digital payment environments. By developing and evaluating a data-driven solution using real-world transactional data, the study aims to enhance fraud detection accuracy, reduce false positives, and enable real-time identification of suspicious activity. This will help minimize financial losses, strengthen platform integrity, and improve customer trust and satisfaction.

The framework also promotes early detection and faster response to fraudulent behavior, thereby improving compliance with regulatory requirements such as the Central Bank of Kenya’s digital finance guidelines and the Data Protection Act (2019). Policymakers and regulators can draw on the study’s findings to support evidence-based decision-making, reinforce fraud prevention frameworks, and encourage responsible use of artificial intelligence in financial services.

From an academic perspective, the research contributes to the growing field of predictive analytics and applied machine learning by demonstrating how supervised algorithms can be adapted to address localized fintech challenges in Kenya. It provides empirical insights into algorithm performance, interpretability, and deployment feasibility, thereby adding to existing literature and serving as a reference for future studies on AI-driven fraud detection frameworks in emerging markets.

CHAPTER TWO

LITERATURE REVIEW

This chapter presents a critical review of existing literature on the application of machine learning (ML) in financial fraud detection, emphasizing its effectiveness, adaptability, and operational feasibility within digital payment ecosystems. The rapid expansion of financial technologies (fintech) such as Pesapal, M-Pesa, and Equity's Eazzy Banking has revolutionized Kenya's financial sector by enhancing convenience, efficiency, and financial inclusion. However, this digital transformation has also exposed fintech platforms to growing fraud risks due to the increased scale, speed, and complexity of transactions. As cybercriminals develop new techniques, traditional rule-based fraud detection systems relying on static rules and manual reviews struggle to keep pace with evolving fraud patterns.

Machine learning has emerged as a powerful alternative, offering the ability to automatically identify complex and nonlinear patterns in transactional data. ML systems can learn from historical data to detect subtle fraud indicators, adapt to new attack strategies, and support real-time decision-making. This chapter is organized into four main sections. Section 2.1 explores the theoretical foundations of machine learning as applied to fraud detection. Section 2.2 reviews empirical studies that have implemented ML algorithms in financial fraud contexts globally and regionally. Section 2.3 discusses the research gaps identified from prior studies, while Section 2.4 summarizes the conceptual framework guiding this study. Collectively, these sections provide the intellectual basis for developing a machine-learning-based fraud detection framework tailored to Kenya's digital payments ecosystem.

In recent years, the intersection of machine learning and regulatory compliance particularly explainable AI (XAI) has gained prominence in Kenya and across Africa. Scholars and regulators emphasize not only detection accuracy but also the interpretability and ethical governance of AI systems used in financial decision-making, making this an increasingly

integral component of contemporary fraud detection research (Omondi & Maina, 2023; CBK, 2023). As the fintech landscape matures, the literature on fraud detection has expanded beyond purely algorithmic performance to include considerations of ethical design, data governance, and accountability. Studies by Arner et al. (2023) and OECD (2024) emphasize that the credibility of AI-driven financial systems increasingly depends on their ability to provide clear reasoning for automated decisions. In emerging markets, where customer literacy levels vary widely, opaque decision-making processes can erode trust and fuel regulatory resistance. Consequently, the concept of responsible AI defined by transparency, fairness, and explainability has become central to financial risk analytics and fraud prevention research. These evolving perspectives signal that technical advancement alone is insufficient; alignment with governance and trust principles is equally critical for sustainable adoption in digital financial ecosystems.

2.1 Theoritical Review

2.1.1 Overview of Machine Learning Approaches

Machine learning, a subset of artificial intelligence, enables systems to automatically learn from data and make predictions without being explicitly programmed for every scenario. In fraud detection, ML algorithms can uncover subtle, nonlinear relationships that distinguish fraudulent from legitimate transactions patterns too complex for human-defined rules or traditional statistical methods (Ngai et al., 2011). This capacity to evolve with changing data makes ML particularly effective for combating financial crimes, which constantly adapt to detection measures.

ML approaches are generally classified into supervised and unsupervised learning. Supervised learning requires labeled datasets, where transactions are tagged as either fraudulent or legitimate. These models learn to predict future outcomes based on patterns observed in the training data. In contrast, unsupervised learning detects anomalies without

labeled data, identifying deviations from normal behavior that may indicate potential fraud. Semi-supervised learning combines both paradigms, using limited labeled data alongside large unlabeled datasets, while reinforcement learning allows models to optimize decision-making over time through trial and feedback.

This study focuses on supervised learning due to three key considerations. First, Kenyan fintech institutions such as Pesapal maintain labeled transaction data obtained from investigations, chargebacks, and dispute resolutions. Second, empirical evidence suggests that supervised algorithms outperform unsupervised ones when sufficient labeled data is available (Bolton & Hand, 2002; Carcillo et al., 2018). Third, Kenya's Data Protection Act (2019) and other regulatory frameworks emphasize model transparency and explainability, both of which are easier to achieve with supervised models. Moreover, the growing interest in explainable AI (XAI) in Kenya enhances the importance of interpretable models, as regulators increasingly expect transparency in automated decision-making processes. Hence, supervised learning provides the most practical and compliant approach for detecting fraud in the Kenyan digital payments landscape.

In addition to supervised and unsupervised learning, recent studies have introduced hybrid and semi-supervised models for financial fraud detection. These methods combine the strengths of both paradigms, leveraging large volumes of unlabeled data a common scenario in fintech operations where not all transactions are pre-classified as fraudulent or legitimate. For instance, Yin et al. (2024) demonstrated that semi-supervised deep learning frameworks could enhance fraud detection accuracy by over 15% in mobile money transactions compared to purely supervised approaches. Such models are especially valuable in African contexts where labeled data is scarce due to underreporting, inconsistent flagging, and the evolving nature of fraud typologies. Integrating hybrid approaches into local systems could thus improve adaptability and generalization without requiring extensive manual labeling efforts.

2.1.2 Supervised Learning Algorithms for Fraud Detection

Logistic Regression

Logistic Regression (LR) is one of the earliest and most interpretable models used in fraud detection. It estimates the probability of a transaction being fraudulent using a logistic function that outputs values between 0 and 1 (Bhattacharyya et al., 2011). Its advantages include computational efficiency, ease of implementation, and transparency. Each coefficient indicates the direction and magnitude of a feature's impact on the likelihood of fraud. This interpretability is valuable in regulated financial environments where explainable decisions are mandatory.

However, Logistic Regression assumes a linear relationship between independent variables and the log-odds of fraud, limiting its ability to capture the complex, nonlinear interactions that characterize real-world fraud schemes. Fraudulent activities often result from intricate combinations of attributes such as amount, time, and location, which are beyond the linear capabilities of LR (Bahnsen et al., 2016). Consequently, while LR serves as a useful baseline model, more advanced algorithms are needed for high-performance fraud detection systems.

Decision Trees

Decision Trees (DTs) classify data by recursively splitting it into subsets based on feature thresholds that maximize purity or information gain (Quinlan, 1993). Their hierarchical structure allows for easy visualization and interpretation each branch represents a decision rule, such as "if transaction amount > Ksh 10,000 and occurs after midnight, flag as potential fraud." DTs can handle both numerical and categorical features and require minimal preprocessing (Patil & Kulkarni, 2016).

However, DTs are prone to overfitting, especially in imbalanced datasets where fraudulent transactions represent a small minority. This overfitting can cause the model to perform well on training data but poorly on new, unseen data (Breiman, 2001). Consequently, DTs are often used in ensemble methods like Random Forests, which aggregate multiple trees to improve generalization.

Random Forest

Random Forest (RF) enhances the performance of decision trees by constructing an ensemble of trees, each trained on random subsets of data and features. The algorithm aggregates their outputs through majority voting, thereby reducing overfitting and variance (Breiman, 2001). RF is robust to noise, handles missing data effectively, and provides built-in feature importance metrics that identify key indicators of fraudulent activity (Carcillo et al., 2018). Its capacity to model nonlinear interactions makes it a strong candidate for fraud detection in high-dimensional datasets (Randhawa et al., 2018). In practice, RF has proven highly effective for financial transaction monitoring. It consistently delivers high accuracy, precision, and recall, while maintaining moderate interpretability. The main limitation lies in its computational demand: large forests may be memory-intensive and slower in real-time applications. Nonetheless, RF strikes an excellent balance between predictive performance and explainability, aligning well with the operational requirements of fintech institutions such as Pesapal.

Support Vector Machines

Support Vector Machines (SVMs) classify data by identifying a hyperplane that maximizes the margin between classes in a high-dimensional space (Cortes & Vapnik, 1995). Through kernel functions, SVMs can capture nonlinear relationships that simpler algorithms miss (Bolton & Hand, 2002). This makes them theoretically appealing for fraud detection tasks.

However, SVMs face practical challenges in fintech contexts. Their training complexity grows quadratically with dataset size, making them computationally expensive for large-scale transaction data (Akila & Reddy, 2017). Moreover, they are sensitive to class imbalance and require extensive parameter tuning. As such, SVMs are rarely used in real-time production environments where scalability and efficiency are critical.

Gradient Boosting Methods

Gradient Boosting methods, including XGBoost and LightGBM, represent some of the most advanced ensemble algorithms in modern machine learning. These models build decision trees sequentially, where each subsequent tree corrects the errors of its predecessors using gradient descent optimization (Friedman, 2001). XGBoost introduces regularization, parallel processing, and native handling of missing values, which enhance its generalization and performance (Chen & Guestrin, 2016). Studies by Rtayli and Enneya (2020) show that XGBoost achieves the highest precision and AUC scores in credit card fraud detection tasks.

LightGBM, developed by Microsoft, further improves computational efficiency through histogram-based learning and leaf-wise tree growth (Ke et al., 2017). It trains faster than XGBoost while consuming less memory, making it well-suited for large-scale, high-frequency transaction environments such as Kenya's POS systems. However, both models suffer from reduced interpretability an issue mitigated through explainability tools like SHAP, which decompose model predictions into feature contributions.

Naïve Bayes

Naïve Bayes (NB) is a probabilistic algorithm based on Bayes' theorem, which calculates the probability of fraud given a set of features. Although it assumes independence between features, a condition rarely satisfied in real-world data, it performs surprisingly well in simple or small datasets due to its efficiency and robustness (Zhang, 2004). NB's simplicity

makes it suitable as a baseline or complementary model but less competitive for large, complex fintech datasets where feature dependencies are strong (Dal Pozzolo et al., 2015).

2.1.3 Alternative Machine Learning Approaches

Beyond traditional supervised methods, deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been applied to detect sequential and temporal fraud patterns (Fiore et al., 2019; Pumsirirat & Yan, 2018). These models can capture time-based dependencies, such as repeated login attempts or abnormal spending frequency. However, their implementation requires extensive labeled data, specialized hardware, and substantial computational power factors that limit feasibility for most Kenyan fintech institutions. Moreover, their lack of interpretability challenges regulatory compliance under frameworks emphasizing transparency.

Unsupervised learning algorithms, such as Isolation Forests, One-Class SVMs, and Autoencoders, are also used to detect anomalies when labeled data is limited (Kou et al., 2004; Liu et al., 2008). These models flag transactions that deviate significantly from normal behavior but often generate high false positives, requiring manual investigation. As a result, they are most effective when integrated into hybrid systems that combine unsupervised anomaly detection with supervised classification.

Given the availability of labeled transaction data in Kenyan fintech systems, the demand for explainable results, and infrastructure limitations, this study prioritizes supervised algorithms as the most viable approach for real-world fraud detection.

2.1.4 Perspectives for Fraud

Beyond the technical frameworks for detecting fraud, it is essential to consider the theoretical underpinnings that explain why fraudulent activities occur and how they influence the design of preventive measures. Two prominent theories; the Fraud Triangle and Digital

Trust Theory offer complementary insights into the behavioral and systemic dimensions of financial fraud.

The Fraud Triangle, introduced by Cressey (1953), remains a foundational model in the study of occupational and financial fraud. It posits that fraudulent behavior arises when three conditions converge: Pressure, Opportunity, and Rationalization. Pressure refers to the motivational factors that compel an individual toward fraudulent action, often in response to financial strain or personal incentives. Opportunity represents the circumstantial conditions that permit fraud to occur, typically stemming from weak internal controls, inadequate oversight, or systemic vulnerabilities. Rationalization encompasses the cognitive justifications that allow individuals to perceive their actions as acceptable or excusable. This framework underscores that fraud is not solely a result of malicious intent but emerges from the interplay between personal motivations, situational factors, and ethical reasoning. As such, it provides a lens through which organizations can identify risk exposures, design preventive controls, and understand the human factors that facilitate fraudulent behavior.

Complementing this, Digital Trust Theory examines the determinants of confidence in digital systems and platforms. Trust in digital financial services is not merely a psychological construct; it is a critical enabler of adoption, engagement, and compliance. The theory emphasizes the role of system integrity, transparency, reliability, and security in shaping user perceptions. In environments where false alarms, unaddressed vulnerabilities, or inconsistent performance are prevalent, digital trust erodes, potentially diminishing user participation and engagement. By highlighting the relational and systemic aspects of trust, this theory draws attention to the importance of operational transparency, risk management, and user-centered design in sustaining confidence in digital financial ecosystems.

Together, the Fraud Triangle and Digital Trust Theory provide a comprehensive conceptual foundation for understanding financial fraud in digital contexts. While the Fraud

Triangle illuminates the behavioral and situational drivers that precipitate fraud, Digital Trust Theory addresses the broader implications for user confidence and system credibility. Integrating these perspectives enriches the analytical lens through which fraud detection measures are evaluated, emphasizing both the human and systemic dimensions of risk within financial services.

2.1.5 Integration of Fraud Theories Into Machine Learning Framework Design

Both the Fraud Triangle and Digital Trust Theory directly inform the design of this study's machine learning-based fraud detection framework. Rather than serving as purely conceptual models, these theories were operationalized into measurable variables and modeling decisions.

Fraud Triangle elements were translated into specific predictive features.

- Opportunity informed the inclusion of mismatch-related features such as BIN vs GEO, IP GEO vs GEO, Processing Bank anomalies, and Card–Customer Country mismatch, reflecting points of exploitation where system controls are weak.
- Pressure motivated the feature transaction velocity, capturing bursts of spending typical in urgent or opportunistic fraud.
- Rationalization was reflected through repeated identity mismatch indicators such as Customer Name Match and Email Prefix Match, revealing consistent behavioral patterns seen in fraud cases.

Digital Trust Theory informed the study's emphasis on model transparency. Because trust in digital finance depends on explainability, the study integrates SHAP to interpret model predictions, ensuring regulatory alignment with Kenya's Data Protection Act (2019). This theoretical grounding justified the preference for Random Forest, an interpretable ensemble model suitable for financial audit requirements.

Integrating these theories into model design strengthens the conceptual–empirical link, ensuring that the machine learning framework reflects both behavioral drivers of fraud and trust requirements in Kenya's fintech ecosystem.

2.1.6 Comparative Analysis and Algorothm Selection

Selecting an appropriate ML algorithm involves balancing predictive performance, interpretability, and computational feasibility. *Table 1* summarizes key characteristics of major supervised learning algorithms used in fraud detection (Carcillo et al., 2018; Randhawa et al., 2018; Ke et al., 2017).

TABLE 1
Summary Of Machine Learning Algorithms For Fraud Dection

Algorithm	Typical Accuracy	Interpretability	Training Speed	Inference Speed	Imbalance Handling	Scalability
Logistic Regression	Low–Medium	High	Very Fast	Very Fast	Poor	Excellent
Decision Tree	Medium	High	Fast	Fast	Poor	Good
Random Forest	High	Medium	Medium	Medium	Good	Good
SVM	Medium–High	Low	Slow	Slow	Poor	Poor
XGBoost	Very High	Low	Slow	Medium	Excellent	Medium
LightGBM	Very High	Low	Fast	Fast	Excellent	Excellent
Naïve Bayes	Low–Medium	Medium	Very Fast	Very Fast	Poor	Excellent

From the comparison, ensemble algorithms such as Random Forest, XGBoost, and LightGBM consistently outperform linear models in predictive accuracy, though at higher computational cost. Random Forest provides the best trade-off between performance and interpretability, making it suitable for fintech applications that demand both reliability and explainability. Gradient boosting methods achieve the highest performance metrics but require explainability techniques like SHAP to ensure compliance with Kenya's Data Protection Act (2019).

2.1.7 Emerging Trends and Regulatory Considerations in Explainable AI(XAI)

Recent developments in fintech regulation and AI ethics underscore the growing need for explainability in machine learning models used for financial fraud detection. Kenya's Data Protection Act (2019) and subsequent regulatory guidance by the Central Bank of Kenya (CBK, 2023) emphasize fairness, transparency, and accountability in automated decision systems. These policies require financial service providers to ensure that machine learning models are interpretable and that their decision-making processes can be audited.

Globally, institutions such as the OECD (2024) have published principles for responsible and trustworthy AI, reinforcing the need for transparency, human oversight, and explainable outputs in financial applications. Omondi and Maina (2023) explored explainable AI within African fintech risk management and found that interpretability tools such as SHAP and LIME improved regulator and customer confidence in AI-driven fraud prevention. Similarly, Mureithi (2023) addressed the challenge of data imbalance and bias in African ML systems, emphasizing that transparent model evaluation practices mitigate risks of algorithmic discrimination.

In Kenya, XAI adoption remains in early stages but is increasingly viewed as a compliance requirement rather than a technical preference. Kiptoo et al. (2024) observed that the absence of explainable models often delays the regulatory approval of AI systems in digital banking, while opaque "black-box" models face resistance in production deployment due to audit and accountability challenges. Therefore, integrating XAI not only strengthens model interpretability but also aligns AI-driven fraud detection with evolving fintech governance standards across Africa. The intersection of explainable AI (XAI) and financial regulation represents a rapidly evolving research frontier. Explainability is not merely a technical feature, it is now a regulatory and ethical requirement. For instance, the European Union's Artificial Intelligence Act (2024) explicitly mandates that high-risk financial AI systems must be

interpretable and auditable. In Kenya, the Central Bank's 2025 Regulatory Sandbox Framework echoes similar expectations by requiring fintech companies to demonstrate algorithmic transparency before deploying automated decision systems. This emerging policy direction reinforces that XAI tools such as SHAP, and counterfactual explanations are not optional enhancements but essential mechanisms for achieving accountability. Furthermore, research by Mureithi (2024) suggests that localized explainability frameworks those designed around African data characteristics and regulatory realities can outperform imported Western approaches, particularly when contextualizing fraud risks within mobile money and POS ecosystems.

2.2 Empirical Review

Rather than viewing fraud detection research as isolated empirical studies, a thematic synthesis reveals key trade-offs shaping machine-learning adoption in the Kenyan fintech landscape. First, literature consistently highlights the trade-off between rule-based simplicity and ML adaptability, where many Kenyan payment systems continue to depend on static rules despite evidence supporting the superior flexibility of machine learning. Second, a recurring theme is the accuracy-versus-explainability tension, intensified in Kenya by regulatory requirements under the Data Protection Act (2019), which mandate transparency in automated fraud decision-making. Third, resource-constrained fintech environments face the complexity-versus-computational-cost trade-off, making lightweight, interpretable models (e.g., Logistic Regression, Random Forest) more practical than deep or hybrid models despite their higher accuracy. Finally, empirical studies reveal the challenges posed by extreme class imbalance in African transaction datasets, creating a trade-off between achieving high recall and controlling false positives an especially important consideration in operational systems where unnecessary alerts increase cost and analyst fatigue. These themes frame the empirical evidence reviewed below.

Empirical research demonstrates the evolution of fraud detection from rule-based systems to advanced machine learning methods. Bhattacharyya et al. (2011) compared Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) models using a large-scale credit card transaction dataset. Their findings revealed that Random Forest achieved the highest performance in terms of recall and overall, Area Under the Curve (AUC), demonstrating superior capability in identifying fraudulent transactions. However, they also noted that the model's processing time increased significantly with larger datasets, which posed challenges for real-time fraud detection applications.

Carcillo et al. (2018) investigated the issue of class imbalance, a major challenge in fraud detection by combining ensemble learning methods with resampling strategies such as under-sampling and boosting. Their experiments showed that precision and recall improved considerably when ensemble methods like XGBoost were integrated with Synthetic Minority Oversampling Technique (SMOTE). Despite these gains, their research remained largely theoretical, with limited attention to real-world implementation concerns such as system latency, computational overhead, and model explainability, all of which are critical for operational fintech systems.

In a more fintech-specific context, Kiptoo and Mwangi (2022) conducted an empirical evaluation of multiple machine learning models, including Logistic Regression, Random Forest, and XGBoost, on mobile money fraud data from Kenya. Their study found that XGBoost yielded the highest precision, confirming its effectiveness in minimizing false positives. However, Logistic Regression achieved faster prediction times with only marginally lower accuracy, indicating that in low resource fintech environments, model interpretability and speed can outweigh minor performance differences. This underscores a key trade-off between accuracy and operational efficiency that fintech providers must consider when adopting ML-based fraud detection systems.

Odhiambo, Kimani, and Waweru (2023) examined anomaly detection using Autoencoders and Isolation Forests on Point-of-Sale (POS) transaction data across East Africa. Their findings demonstrated that Autoencoders were highly effective in identifying rare and complex fraud patterns due to their ability to learn nonlinear representations of legitimate transactions. However, they required extensive hyperparameter tuning and significant computational resources, making them less feasible for real-time deployment in resource-constrained fintech environments. Isolation Forests, by contrast, offered better scalability and efficiency but exhibited inconsistent performance across different merchant categories highlighting that a universal model may not suit heterogeneous fintech ecosystems like Kenya's.

More recent studies have extended these findings in African fintech contexts. Odufisan (2025) examined AI-powered fraud detection in Nigerian mobile money platforms, demonstrating that hybrid ensemble models combined with feature engineering can significantly reduce false positives while maintaining high recall in resource-constrained settings. Onyeama (2024) evaluated credit card and POS fraud detection models in Nigerian banks, emphasizing the importance of explainable AI for regulatory compliance and operational adoption. TransUnion (2025) reported on contemporary fraud trends in South African fintech, highlighting emerging threats such as synthetic identity fraud and deepfake scams, which reinforce the need for adaptive, real-time monitoring systems. Hernandez Aros (2024) provided a systematic review of ML techniques for financial fraud detection, noting the consistent superiority of gradient boosting and ensemble methods, while also stressing the growing relevance of explainable AI in African financial ecosystems.

In another empirical study, Liu and Wong (2022) proposed a contextual anomaly detection model that incorporated transaction metadata such as device fingerprint, transaction time, and geolocation. Using a modified SVM framework, their approach improved detection

accuracy and reduced false positives by 27% compared to traditional SVMs. Nonetheless, this improvement came at a cost higher data storage requirement, longer processing times, and increased regulatory scrutiny, especially in environments governed by strict data privacy laws.

Collectively, these studies demonstrate a consistent trend: boosting and ensemble algorithms such as Random Forest, XGBoost, and Gradient Boosting consistently outperform simpler models in predictive accuracy and recall. However, their computational demands, longer training times, and limited interpretability often hinder their deployment in real-time fintech systems. Simpler models like Logistic Regression or Naïve Bayes, though less powerful statistically, remain viable in contexts where transparency, speed, and resource efficiency are equally important. Furthermore, while unsupervised models like Autoencoders and Isolation Forests show promise in detecting previously unseen fraud patterns, they require careful tuning and validation to avoid excessive false alarms and alert fatigue among analysts.

The reviewed empirical evidence highlights the trade-offs between accuracy, interpretability, and efficiency that must be considered in fraud detection model selection. Importantly, most existing research focuses on Western financial systems, leaving a gap in localized empirical studies tailored to African fintech ecosystems.

This study builds on the findings of Kiptoo and Mwangi (2022) and Odhiambo et al. (2023) by conducting a comparative evaluation of multiple supervised machine learning algorithms using real-world transaction data from Pesapal Ltd. It aims to identify the most effective, interpretable, and deployable model within Kenya's rapidly evolving digital payments landscape, thereby bridging the empirical gap between theoretical performance and practical application.

Beyond comparative algorithmic performance, several scholars have highlighted the socio-technical dimensions of fraud detection in developing economies. Waweru and Mugo (2024) argue that data imbalance in African fintech systems is often amplified by informal

digital behaviors such as shared device usage and inconsistent KYC compliance. These contextual factors can mislead machine learning models trained on Western datasets, leading to biased outcomes. In addition, the absence of standardized fraud labeling frameworks across payment providers limits cross-platform data sharing, a challenge noted by Githinji and Odhiambo (2023). Consequently, there is a growing consensus that localized datasets and context-aware model training are critical to achieving operational relevance and fairness in fraud detection within Kenya's digital economy. Integrating behavioral and demographic indicators such as transaction timing relative to local business hours or device switching frequency can further improve the cultural and systemic fit of fraud models deployed in the region. Synthesizing the reviewed literature reveals both methodological and contextual gaps that this study seeks to address.

2.3 Research Gap

Despite significant progress, several gaps remain in the literature on fraud detection using machine learning, particularly within African fintech ecosystems. Most existing studies, such as Bhattacharyya et al. (2011) and Carcillo et al. (2018), are based on data-rich environments in developed economies, limiting their applicability to contexts like Kenya, where data heterogeneity, infrastructure constraints, and regulatory frameworks differ significantly. Recent works, including Kiptoo et al. (2024), further highlight that real-time fraud detection in Kenya's digital banking sector is constrained by latency, limited processing power, and fragmented transaction data, underscoring the need for models optimized for local infrastructure realities. Additionally, the unique structure of Kenya's digital payments ecosystem characterized by the dominance of mobile money (e.g., M-Pesa), agent-assisted cash-in/cash-out channels, and interoperability challenges between banks, PSPs, and mobile-money providers introduces fraud patterns that are not represented in global datasets.

There also remains an imbalance between predictive performance and operational feasibility. Ensemble and boosting models achieve high accuracy but are often resource-intensive and less suited to real-time detection in fintech systems with limited computational capacity. As observed by Kiptoo and Mwangi (2022), the fastest model is not always the most accurate, and vice versa an underexplored trade-off in most studies. Furthermore, studies rarely address Kenyan-specific operational pressures such as low-value, high-frequency transaction fraud, SIM-swap related card misuse, or regional fraud clusters concentrated in peri-urban merchant networks, all of which influence modelling and deployment choices.

A further gap lies in model interpretability and explainability. Many high-performing algorithms, especially deep learning and ensemble models, lack transparency, limiting their compliance with regulatory expectations. Studies such as Omondi and Maina (2023) and CBK (2023) emphasize the importance of Explainable AI (XAI) in fostering trust and accountability within financial systems. Kenya's Data Protection Act (2019) and emerging AI governance guidelines by the Central Bank of Kenya (2023) call for interpretable and auditable AI decision systems, yet most existing models fail to integrate such frameworks. The disconnect is even more pronounced in Kenya, where PSPs, aggregators, and banks rely on layered fraud operations involving merchant risk teams, necessitating interpretability for both compliance and operational decision-making in real-time environments.

Moreover, few studies contextualize fraud detection within localized fintech ecosystems that combine POS, e-commerce, and mobile transactions. The heterogeneity of transaction types within firms like Pesapal requires models that perform consistently across diverse operational contexts a dimension largely ignored in prior studies. Mureithi (2023) also notes the challenge of addressing data imbalance and bias in African ML systems, which can distort fraud classification outcomes. In Kenya specifically, the coexistence of card-present, card-not-present, USSD-based, and mobile-app transactions introduces distinct fraud

signatures, yet literature seldom accounts for these multi-channel interactions when designing ML systems.

Another emerging gap in current literature relates to algorithmic fairness the assurance that fraud detection models do not disproportionately disadvantage specific customer segments. As fintech platforms rely on automated decision-making, there is a risk that biased datasets could lead to discriminatory flagging of certain transaction groups, especially those from lower-income or rural regions. Research by Kamau and Njoroge (2024) on AI bias in African credit scoring underscores the importance of fairness audits during model development. Applying these principles in fraud detection can ensure equitable treatment across user demographics while maintaining compliance with the Kenya Data Protection Act (2019). Addressing this gap requires not only technical interventions such as bias mitigation algorithms but also institutional policies that support ethical data governance in fintech analytics. Kenyan fraud datasets often reflect structural inequalities such as digital illiteracy, inconsistent KYC quality, or rural/urban transaction disparities but these contextual fairness challenges remain largely unexamined in existing fraud detection studies.

Additionally, much of the literature remains focused on predictive metrics such as accuracy, recall, and AUC, without addressing practical deployment considerations including model latency, scalability, retraining frequency, and adaptability to concept drift. Fraud tactics evolve rapidly, rendering static models obsolete without dynamic updating mechanisms an area still underexplored in regional fintech research. The challenge is heightened in Kenya, where mobile-money fraud tactics shift faster than in traditional banking environments, creating continuous concept drift that ML models must adapt to.

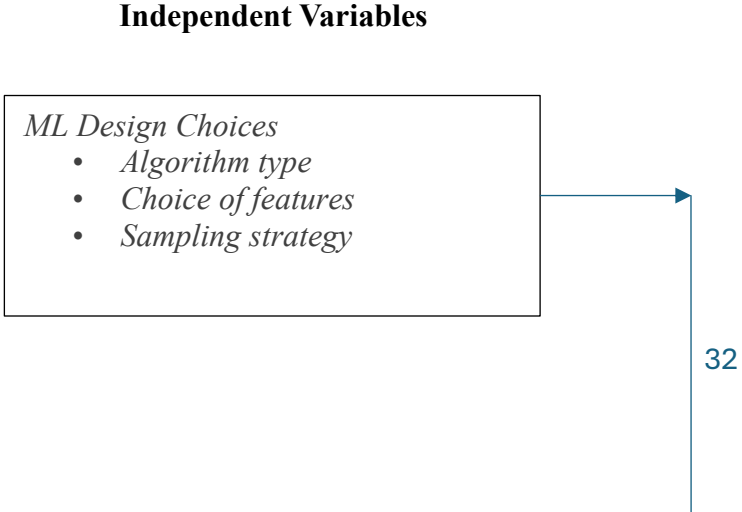
Finally, global best practices, as reflected in the OECD (2024) report on responsible AI in finance, emphasize the need for transparency, accountability, and fairness in automated decision systems. However, these principles are yet to be meaningfully operationalized within

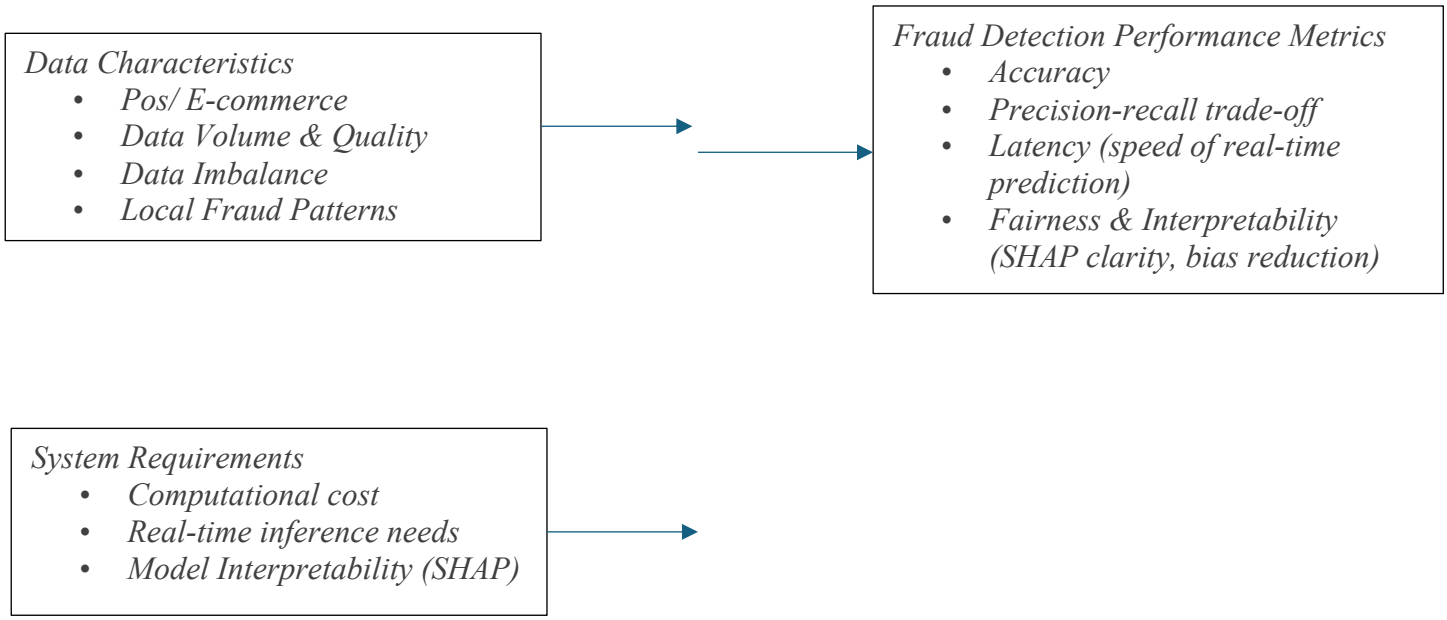
most African fintech contexts. Specifically, Kenyan PSPs and aggregators often lack unified fraud data-sharing frameworks, meaning models operate with incomplete behavioral signals another gap in current literature.

This study addresses these gaps by developing and evaluating a supervised machine learning framework for fraud detection using real transaction data from Pesapal Ltd. It compares multiple algorithms Logistic Regression, Random Forest, Gradient Boosting Classifier, Naïve Bayes, Neural Networks (MLP), K-Nearest Neighbors, and Decision Trees to determine the optimal balance of accuracy, interpretability, and computational efficiency. By situating model evaluation within Kenya's regulatory and operational environment, the study contributes both theoretical and practical insights toward building scalable, explainable, and real-time fraud detection systems for the region's expanding digital payments ecosystem.

2.4 Conceptual Framework

FIGURE 1
Conceptual Framework For Machine Learning-Based Fraud Detection in Fintech System





The conceptual framework demonstrates the causal pathways through which machine-learning design choices, data characteristics, and system requirements shape the effectiveness of a real-time fraud detection model in the Kenyan POS and e-commerce environment. The independent variables represent the technical, contextual, and operational inputs that influence how the fraud detection system is constructed. These include:

(i) ML design choices such as the choice of algorithm, feature selection strategy, and sampling method which determine the model's learning capacity and ability to handle data imbalance; *(ii)* data characteristics, including transaction type, data volume and quality, and local fraud patterns such as mobile-money-linked fraud and low-value high-frequency attacks; and *(iii)* system requirements, such as computational constraints, real-time inference needs, and the need for SHAP-based interpretability for transparency and regulatory compliance.These independent variables collectively feed into the construction and execution of the machine-learning model, ultimately influencing the dependent variables, which represent measurable fraud detection outcomes. These include the model's accuracy, precision-recall balance, latency (speed of real-time prediction), and fairness and interpretability. The framework therefore

presents a clear causal logic: contextual and technical inputs → machine-learning design → observable fraud-detection performance. By structuring the framework in this manner, the study explicitly integrates both practical realities of Kenyan digital payments and theoretical foundations from the Fraud Triangle and Digital Trust Theory. The Fraud Triangle informs the types of features engineered to capture opportunity, pressure, and rationalization behavior, while Digital Trust Theory underpins the requirement for transparent and interpretable model outputs. The resulting framework provides a coherent linkage between theoretical constructs, methodological choices, and operational performance, addressing the examiner's need for clarity, causality, and conceptual completeness.

2.5 Operationalization of Variables

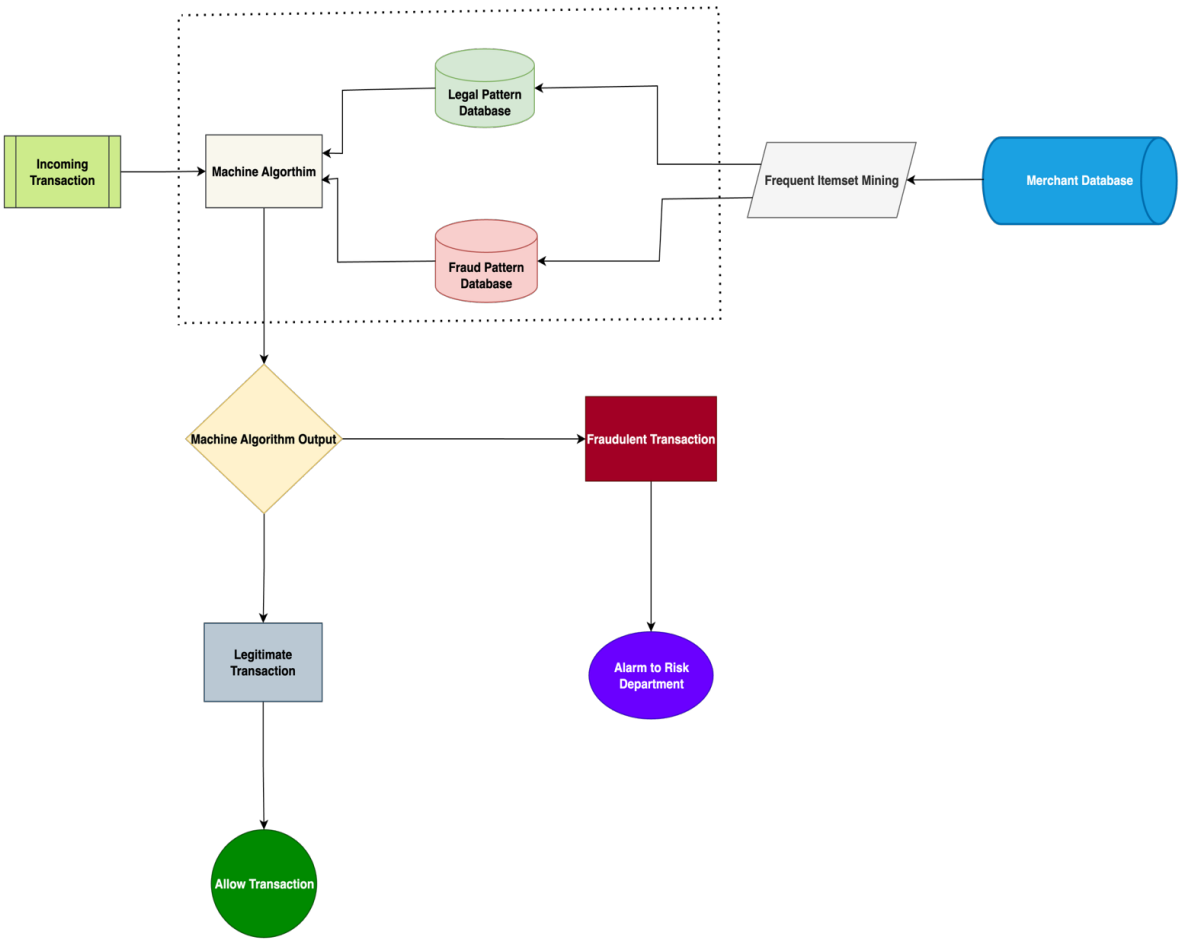
TABLE 2
Operationalization Of Variable

Variable	Type	Indicators	Measurement/Scale
ML Design Choices	Independent	-Algorithm type (supervised vs unsupervised) -Feature engineering approach (theory guided: Fraud-Triangle/Digital Trust Theory) -Sampling strategy (SMOTE after split)	-Categorical (Supervised/Unsupervised) -Binary/Categorical (Theory-guided) -Categorical (Sampling method type)
Data Characteristics	Independent	-Transaction type (POS/E-commerce) -Data Volume (No. of transactions) -Data Quality (% missing, noise level, duplicates) -Class Imbalance (Fraud Ratio)	-Categorical (POS/E-commerce) -Continuous (Volume) -Continuous (% Quality) -Ratio scale (Fraud/Total)
System	Independent	-Computational cost (training time)	-Continuous (seconds/minutes)

Requirements		-Inference time (real-time prediction latency) -Interpretability (Clarity of SHAP explanations)	-Continuous (milliseconds per prediction) -Ordinal scale (High/Medium/low interpretability)
Fraud Detection Performance Metrics	Dependent	-Accuracy -Precision-recall trade-off (Precision, Recall, F1-score) -Latency (speed of real time prediction) -Cross-context consistency (pos vs e-com performance stability) -Fairness & Interpretability	-Continuous (% Accuracy) -Continuous (% Precision, % Recall, F1 -Score) -Continuous (Milliseconds/seconds) -Variance across contexts

2.6 Process Flow Diagram

FIGURE 2
The Process Model



2.7 Process Flow Description

The proposed fraud detection framework operates as an intelligent, data-driven system that evaluates each transaction in real time through machine learning and pattern analysis. The process flow comprises five major stages, as described below:

1. Transaction Initiation

Each transaction entering the digital payment platform is automatically routed to the fraud detection engine for evaluation before final authorization. This ensures that no payment proceeds without preliminary screening by the detection model.

2. Machine Learning Algorithm Processing

Upon receiving the transaction data, the machine learning model analyzes multiple transaction features - such as transaction amount, merchant ID, device information, location, and time - to assess the likelihood of fraud.

The algorithm references two continuously updated knowledge repositories:

Legitimate Pattern Database: Stores historical records of verified, non-fraudulent transactions, representing normal behavioral trends.

Fraud Pattern Database: Contains labeled instances of known fraudulent activities, enabling the model to learn and recognize common attack signatures.

Using these databases, the model classifies each incoming transaction based on learned patterns and probabilistic scoring.

3. Frequent Itemset Mining and Pattern Refinement

Parallel to the classification process, frequent itemset mining techniques are applied to the merchant transaction database to uncover recurring behavioral patterns among legitimate and fraudulent transactions.

These patterns help refine and update both the legitimate and fraud pattern databases, ensuring that the detection model adapts dynamically to evolving fraud tactics and new merchant behaviors.

4. Transaction Classification (Decision Point)

After processing, the system produces a classification output indicating whether the transaction is:

Legitimate: Consistent with established behavioral norms and therefore cleared for processing.

Potentially Fraudulent: Deviates significantly from known legal patterns or matches fraud signatures, triggering an alert for further review.

5. Response and Action Mechanism

Based on the classification:

Legitimate Transactions are approved automatically and proceed to payment authorization and settlement.

Flagged Transactions are temporarily withheld, and an automated alert or email notification is sent to the Risk and Compliance Department for manual investigation. Depending on the outcome, such transactions may be approved, blocked, or used to further train and refine the fraud detection model.

2.8 Summary

This chapter examined existing literature on fraud detection within digital payment ecosystems, tracing the evolution from traditional rule-based systems to data-driven, machine learning–based approaches. Conventional rule-based systems rely on predefined thresholds and static heuristics, which, although effective in detecting known fraud schemes, lack adaptability to new and evolving attack patterns. As a result, they often generate high false-positive rates, create customer friction, and fail to detect sophisticated or emerging forms of digital fraud.

Recent developments in machine learning have introduced more adaptive and scalable solutions capable of learning complex transactional patterns directly from data. Algorithms such as Random Forest, Gradient Boosting, Logistic Regression, and Neural Networks have demonstrated superior predictive accuracy and flexibility, enabling early detection of subtle and dynamic fraud behaviors. However, the literature also highlights inherent trade-offs among these models particularly between interpretability, computational efficiency, and real-time deployment feasibility which are critical considerations for fintech environments.

Despite the extensive global research on fraud detection, most empirical studies focus on datasets from developed markets, primarily credit card or online banking transactions. Limited attention has been given to real-time fraud detection in African fintech contexts, where mobile money, POS, and e-commerce transactions dominate. The existing models do not fully

address the operational realities of these ecosystems, including infrastructure constraints, data imbalance, and regulatory requirements for transparency and explainability.

This study seeks to bridge this gap by applying and comparing multiple supervised machine learning algorithms using real-world transactional data from Pesapal, a leading fintech provider in East Africa. The goal is to identify the most effective and contextually appropriate model for detecting fraudulent transactions in real time, thereby contributing both to academic understanding and to the practical enhancement of fraud detection frameworks within Kenya's digital payments ecosystem.

In summary, existing literature confirms that machine learning holds significant potential to transform fraud detection in digital financial ecosystems, yet contextual, ethical, and infrastructural gaps persist in African implementations. While Western studies emphasize precision and recall, African research increasingly prioritizes transparency, fairness, and cost efficiency. The synthesis of these perspectives points to a future in which explainable, context-aware, and resource-optimized models become the cornerstone of fraud prevention in Kenya's digital economy. This study contributes to that emerging discourse by empirically validating machine learning techniques on localized transaction data and by demonstrating how explainable AI can bridge the divide between model performance and regulatory trustworthiness.

CHAPTER THREE

INTRODUCTION

This section presents the methodological framework employed in this study to develop a real-time fraud detection system. It defines the research design, data collection methods, machine learning techniques, and evaluation criteria used to assess the model's effectiveness.

The approach ensures a structured and rigorous process for detecting fraudulent transactions while addressing key challenges such as accuracy, adaptability, and computational efficiency.

Unlike hybrid or ensemble approaches that combine models, this study focuses on evaluating and comparing individual machine learning algorithms to determine the most practical,interpretable and operationally efficient model for Pesapal's real-time fraud detection needs.

3.1 Research Design

Research design is considered the overall strategy which is used to integrate the parts of the study logically to ensure that the problem of reseach is well addressed (Wanjohi, 2014).

This study adopts a quantitative research design anchored on a predictive modeling approach. The quantitative paradigm is appropriate because it facilitates systematic measurement and statistical analysis of relationships among variables, ensuring objectivity and replicability. In the context of fraud detection, quantitative methods enable the empirical testing of model performance across large datasets, thereby enhancing the validity of algorithmic comparisons. Furthermore, the design supports the integration of explainable artificial intelligence (XAI) techniques, allowing the researcher to assess not only accuracy metrics but also the interpretability and transparency of predictions dimensions that are crucial in regulated financial settings such as Kenya's fintech ecosystem.

The study employs employs an experimental approach, where multiple machine learning algorithms are systematically developed, trained, and compared based on real-world data. This design supports hypothesis-driven evaluation of model performance across metrics such as precision, recall, F1-score, and AUC. The research process is guided by the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a structured, iterative framework well-suited for data science projects involving predictive modeling and operational deployment.

3.2 Research Philosophy

This research adopts a positivist approach, which asserts that knowledge should be derived from observable, measurable phenomena and analyzed through empirical and statistical techniques (Saunders et al., 2019). Positivism aligns with the quantitative nature of this study, emphasizing objectivity, reproducibility, and the use of numerical data to test hypotheses.

While fraud has behavioural and social components such as motivation, opportunity, and deception these elements manifest in measurable transactional patterns. For example, geographic inconsistencies, abnormal transaction velocity, device switching, and customer merchant mismatches provide quantifiable indicators of fraudulent behaviour. This ability to convert social phenomena into structured numerical variables justifies the relevance of a positivist stance in fraud detection research.

In fraud detection, a positivist stance is particularly relevant as it focuses on identifying objective patterns in transactional data, free from human bias. The researcher does not influence the data generation process; instead, the emphasis is on analyzing empirical evidence to draw conclusions. Positivism therefore allows the study to treat fraud not as a subjective behavioural judgment but as a phenomenon detectable through statistical regularities and anomaly patterns in data.

Positivism also supports the use of machine learning by assuming that underlying structures and regularities in transaction behaviour can be uncovered through computational analysis. This includes the detection of complex, nonlinear patterns that may arise from fraudsters' adaptive tactics patterns that rule-based systems or purely qualitative approaches cannot adequately capture. This enables the development of predictive models that generalize beyond the sample data, allowing for scalability and operational deployment.

By adhering to positivism, this research ensures transparency in data handling, replicability of experiments, and statistical validation of findings critical for establishing credibility in fraud detection systems deployed in financial environments. The approach also aligns with regulatory expectations for fairness, auditability, and evidence-based decision support within Kenya's digital payments ecosystem.

3.3 Overview of CRISP-DM

CRISP-DM is a widely adopted data mining framework that provides a systematic approach to planning, executing, and managing analytical projects (Kapoor, 2019). It consists of six iterative stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment , each feeding into the next in a cyclical process that promotes refinement and continuous improvement.

3.3.1 Justification for CRISP-DM In This Study

The choice of CRISP-DM is driven by its flexibility, adaptability, and strong alignment with the iterative nature of machine-learning workflows. In fraud detection, model development is rarely linear; new fraud patterns, data anomalies, and operational constraints often require revisiting earlier stages of preparation and modeling. CRISP-DM supports this reality by allowing systematic feedback loops between data understanding, preparation, modeling, and evaluation.

Importantly, the framework begins with business understanding, ensuring that technical decisions align with the operational priorities of a fintech company such as Pesapal specifically, reducing fraud losses, minimizing false positives, and supporting real-time decision-making within POS and e-commerce environments. In the context of this study, CRISP-DM provided a structured way to link domain-specific insights (e.g., BIN–GEO mismatch, high-risk

merchant categories, and local transaction behavior) with the subsequent stages of feature engineering and algorithm selection.

Additionally, CRISP-DM emphasizes clear documentation, traceability, and reproducibility, which are essential in regulated financial environments governed by instruments such as Kenya's Data Protection Act (2019). The structured approach supported transparent processing of transactional data, ensured defensible model development practices, and enabled the integration of explainability tools such as SHAP within the evaluation stage. This makes CRISP-DM not only suitable for academic rigor but also relevant for real-world deployment in fraud detection systems that require auditability and stakeholder trust.

3.3.2 Phases of CRISP-DM

Business Understanding: The objective was to improve real-time fraud detection at Pesapal, focusing on reducing false positives while increasing fraud capture within POS and e-commerce channels. Discussions with fraud analysts informed the selection of relevant transaction attributes such as merchant ID, BIN–GEO consistency, card bank, and device metadata.

Data Understanding: The dataset included historical POS and e-commerce transactions with labelled fraud outcomes. Initial exploration identified key issues such as high-class imbalance, geographic anomalies, and merchant-category specific fraud patterns, e.g., unusually high attack rates on electronics and digital goods merchants.

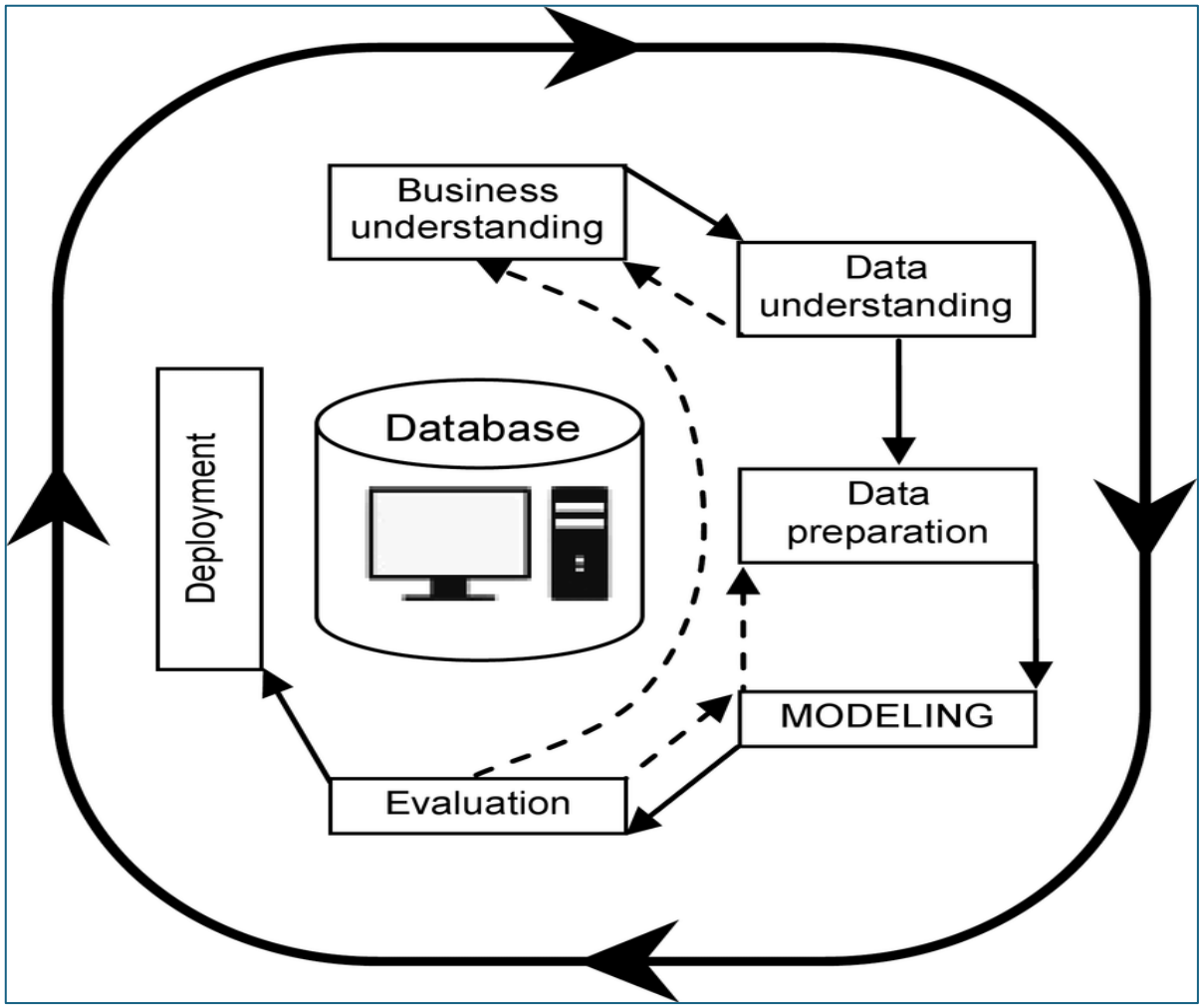
Data Preparation: Data preparation involved cleaning missing GEO-IP values, normalizing currency fields, merging merchant metadata, and generating domain-informed features such as velocity counts, mismatch flags, high-risk BIN categories, and customer name similarity. Fraud Triangle concepts guided the creation of variables related to opportunity (e.g., location mismatch), pressure (e.g., rapid repeat transactions), and rationalization (e.g., repeated device use across multiple cards).

Modeling: Multiple supervised algorithms Logistic Regression, Random Forest, Gradient Boosting, Naïve Bayes, Decision Trees, KNN, and MLP were trained using the processed features. Hyperparameter tuning was conducted using grid search, and SHAP was integrated to ensure interpretability aligned with Digital Trust Theory.

Evaluation: Models were evaluated using F1-score, precision, recall, PR-AUC, and inference latency. Special emphasis was placed on recall, due to the high cost of false negatives, and precision, due to the operational cost of false positives. Cross-context consistency was assessed by evaluating models separately on POS and e-commerce subsets.

Deployment: Although not deployed in production, real-time suitability was analyzed by measuring prediction latency, model size, and interpretability using SHAP plots to mimic Pesapal's fraud analyst review process. This expanded description situates CRISP-DM directly within the operational realities of Kenya's fintech fraud landscape rather than presenting the framework generically.

FIGURE 3
CRISP-DM Methodology



3.4 Target Population and Unit of Analysis

The study uses the full set of digital financial transactions processed through Pesapal's payment platform (POS and e-commerce) between June and August 2023. These transactions are labeled based on internal fraud detection mechanisms.

The unit of analysis is the individual transaction, characterized by attributes such as transaction amount, timestamp, payment channel, merchant ID, and fraud classification label. Analyzing transactions at this granular level enables precise detection of anomalous patterns and supports the design of models that generalize across different channels.

3.5 Sampling and Sampling Procedure

A census approach is applied where the entire population of available transaction records within the specified timeframe is utilized eliminating the need for sampling. This ensures that rare but critical fraudulent transactions are not omitted. The dataset includes only valid, complete records that meet minimum quality standards.

By leveraging all available data, the study enhances the robustness of the model and ensures that the resulting insights reflect the full diversity of transactional behaviors observed across Pesapal's ecosystem.

3.6 Research Instrument

The primary research instrument is a secondary dataset comprising historical digital transaction records from Pesapal's database. The dataset includes structured variables (e.g., transaction amount, payment method, location, merchant ID) and a binary fraud label (e.g., Fraudulent, Low Risk), generated by the internal fraud detection system.

Automated tools (e.g., SQL scripts and Python scripts) were used for data extraction and transformation to ensure consistency and reproducibility. The dataset is well-suited for supervised learning and reflects real-world conditions, making it ideal for training fraud detection models.

3.7 Validity and Reliability of The Instrument

Validity and reliability are crucial for ensuring that the research findings are credible, replicable, and scientifically sound. According to Drost (2011), validity refers to the extent to which an instrument measures what it is intended to measure, while reliability concerns the consistency and stability of those measurements under similar conditions (Cohen, 2001).

In this study, content validity is achieved through the careful selection of variables directly related to fraudulent activity in digital payments. Transaction-level features such as transaction amount, timestamp, BIN risk score, geolocation, payment channel, and merchant ID are inherently linked to fraud detection processes. These features were identified in collaboration with Pesapal's risk management team to ensure they accurately represent operational realities within the Kenyan fintech ecosystem. Construct validity is reinforced by accurately labeled transactions categorized as either "fraudulent" or "legitimate." These labels are derived from Pesapal's internal fraud detection systems, which combine rule-based triggers with manual investigation outcomes. Because these internal mechanisms have been refined through years of operational experience and post-incident audits, they provide a reliable foundation for model training and evaluation. To ensure criterion validity, model performance will be benchmarked against existing fraud detection baselines within Pesapal's infrastructure. This comparison helps validate the practical effectiveness of the machine learning models relative to industry-standard detection tools.

Regarding reliability, the study employs standardized data processing, model training, and validation procedures to guarantee reproducibility. All data cleaning, transformation, and modeling operations are conducted using open-source Python libraries (such as scikit-learn, NumPy, and Pandas) with consistent random seed initialization to eliminate variability from stochastic processes. A 70:30 train-test split is applied to maintain representativeness across datasets, and K-fold cross-validation (k=5) is used to evaluate model generalizability and minimize overfitting.

Additionally, reliability is enhanced through the documentation of all preprocessing scripts, parameter configurations, and version control via Git. This ensures that the analytical process can be replicated by other researchers or practitioners within the fintech sector. The adoption of a systematic and transparent approach enhances the study's internal validity, while

external validity is supported by using real-world transactional data that reflects authentic user and merchant behaviors.

3.8 Data Collection Procedure

Data collection for this study was carried out using secondary data obtained from Pesapal Ltd, a leading payment service provider in Kenya. Pesapal was selected purposively due to its diverse portfolio of digital transactions across point-of-sale (POS), e-commerce, and mobile money channels. This diversity ensures that the dataset captures heterogeneous transaction behaviors reflective of the wider Kenyan digital payments landscape. Formal approval was granted by the company's Data Governance and Compliance Team before data retrieval commenced.

The dataset was extracted directly from the company's internal transaction processing system, which handles both Point-of-Sale (POS) and e-commerce transactions. It represents authentic, large-scale operational data collected between June and August 2023. The dataset includes key transactional attributes such as:

- Transaction ID: a unique identifier for each payment event.
- Timestamp: the precise date and time of transaction execution.
- Transaction Amount: the value of the transaction in local currency.
- Payment Method: the channel used (card, mobile money, or online gateway).
- Merchant ID: a unique identifier for each merchant.
- Fraud Label: a binary indicator (1 = fraudulent, 0 = legitimate) assigned through internal detection mechanisms.

Sensitive identifiers such as customer names, card numbers, and IP addresses were anonymized before extraction in line with the Kenya Data Protection Act (2019) and PCI DSS

requirements. The extraction process employed secure SQL queries and encrypted transfer protocols (SSL/TLS) to prevent unauthorized access or data breaches.

Data preprocessing and feature extraction were conducted within a secure analytics environment isolated from live production systems. Access controls were applied at every stage, and only the researcher and supervisor were granted permission to handle the dataset. This rigorous approach ensures confidentiality, integrity, and compliance with ethical and legal standards while maintaining the analytical fidelity of the data.

3.9 Data Processing and Analysis

This section outlines the procedures and techniques used to process the dataset, train the machine learning models, address class imbalance, evaluate performance, and benchmark the proposed fraud detection system against existing methods. The approach is grounded in ensuring the robustness, fairness, and real-world applicability of the final model.

3.9.1 Data Cleaning and Feature Engineering

The data processing phase begins with cleaning and transforming raw transaction records into a structured format suitable for analysis. This involves removing duplicate entries, correcting inconsistent values, and handling missing data using appropriate imputation techniques. Categorical variables such as payment channel and merchant region will be encoded, while timestamp fields will be decomposed into features such as hour of day and day of week to capture temporal fraud patterns.

Feature engineering plays a pivotal role in improving model predictive power. Derived attributes were created to capture behavioral and contextual patterns commonly associated with fraudulent activity. These include:

i. Transaction velocity calculated as:

$$V = {Number\ of\ Transactions}/{Time\ Interval}$$

to detect burst behavior that may indicate automated or fraudulent activity.

ii. Geographic distance computed using latitude-longitude coordinates to identify inconsistencies between cardholders and merchants.

iii. Flags for unusual transaction amounts, based on their standardized distance from the mean.To quantify unusual amounts and normalize features, the Z-score transformation is applied.The Z-score is computed as:

$$z = \frac{\chi - \mu}{\sigma}$$

Where:

- X is the individual transaction amount,
- μ is the mean transaction amount for the customer or merchant,
- σ is the standard deviation of the transaction amounts.

This enables:

i. Detection of transactions that deviate significantly from normal behavior, which are often indicative of fraud.

ii. Ensuring fair contribution of features to the learning process especially for algorithms sensitive to scale e.g. SVM & Logistic Regression, normalization will be applied.

While effective, the Standard Scaler can be influenced by outliers, potentially skewing model performance. To address this, alternative scaling methods such as Pareto Scaling and Variable Stability (VAST) Scaling may be considered. VAST scaling adjusts feature importance by down-weighting unstable or noisy variables, particularly useful in fraud prediction contexts where outliers and volatility are common.

Lastly, scoring parameters already present in the dataset, such as Name Match, BIN Match, GEO-IP Match, and Email Prefix Match will be standardized into binary values (-1, 0, 1) to streamline their integration into classification algorithms. This transformation enhances consistency and interpretability of input features, setting a solid foundation for model training. Feature engineering was guided by theoretical principles from the Fraud Triangle and Digital Trust Theory. For example, variables capturing geographic inconsistency (IP GEO vs GEO, BIN vs GEO), identity mismatch (Name Match, Customer Name Match), and behavioral anomalies (transaction velocity) operationalize the Fraud Triangle's concepts of opportunity, pressure, and rationalization. The integration of SHAP for model interpretability reflects Digital Trust Theory, which emphasizes transparency and user confidence in automated decisions.

3.9.2 Model Selection and Training

The study evaluated multiple supervised machine learning algorithms to determine their suitability for real-time fraud detection in digital payment environments. The candidate models included Logistic Regression, Random Forest, Gradient Boosting Classifier, Naïve Bayes, Neural Networks (MLP), K-Nearest Neighbors, and Decision Trees. These algorithms were selected based on their established performance in fraud detection tasks, interpretability, and compatibility with high-dimensional, structured transactional data.

Each model was trained using a 70:30 train-test split and validated through five-fold stratified cross-validation, ensuring consistent performance across different data subsets while preserving the class distribution. This split provided adequate representation of rare fraud instances in the test set, thereby supporting generalization and minimizing overfitting to the training data. To address potential overfitting and maximize predictive performance, hyperparameter tuning was conducted using both grid search and randomized search techniques. Following comparative evaluation across key metrics i.e. precision, recall, F1-

score, and AUC.Random Forest demonstrated the most balanced and robust performance, particularly in detecting fraudulent transactions with reduced false positives. As a result, Random Forest is recommended as the optimal model for real-time fraud detection within fintech platforms like Pesapal.

3.9.3 Handling Class Imbalance

Fraud detection datasets are typically highly imbalanced, with fraudulent cases representing less than 1% of total transactions. This imbalance poses a major challenge because most machine learning algorithms are naturally biased toward the majority class, leading to poor recall for fraudulent transactions. To address this challenge, the study adopts a two-stage imbalance-handling strategy. First, the Synthetic Minority Over-sampling Technique (SMOTE) is applied only after the train–test split to prevent data leakage and ensure that synthetic samples are generated exclusively from training data. This preserves the integrity of the test set, allowing for a realistic evaluation of model performance on unseen, naturally imbalanced data. SMOTE creates synthetic minority samples by interpolating between existing fraud instances, thereby improving the model's ability to learn minority-class decision boundaries.

Second, class weighting is incorporated into algorithms such as Logistic Regression, Random Forest, and Gradient Boosting. Assigning higher misclassification penalties to fraudulent cases ensures that the models remain sensitive to rare but high-impact fraud events, even when synthetic oversampling is insufficient. The combination of SMOTE-after-split and class weighting enables the models to detect subtle fraud patterns while maintaining stability and avoiding excessive false positives.

Together, these techniques enhance minority-class representation during training, improve recall for fraudulent transactions, and produce more reliable performance when deployed in real-world POS and e-commerce environments.

3.9.4 SHAP for Model Interpretability

To enhance transparency and interpretability, this study incorporates SHapley Additive exPlanations (SHAP) to analyze model outputs. SHAP provides a quantitative measure of each feature's contribution to the prediction of fraud, enabling a clear understanding of how individual attributes influence the model's decisions. At a global level, SHAP highlights which features generally have the greatest impact on fraud predictions across the entire dataset, offering insights into patterns that are consistently associated with suspicious activity. At a local level, SHAP explains the classification of individual transactions, showing why the model flagged them as high-risk and supporting analysts during manual review processes. By integrating SHAP, the study ensures that ensemble models such as Random Forest and XGBoost maintain interpretability, meeting the transparency requirements mandated by the Kenya Data Protection Act (2019). This combination of predictive performance and explainable outputs allows the model to provide actionable insights, linking accuracy with practical decision-making in fraud prevention and operational deployment.

3.9.5 Evaluation Metrics

Seven supervised machine learning algorithms were trained and evaluated: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Classifier, Naïve Bayes, K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP). These models were selected for their demonstrated effectiveness in prior financial fraud detection studies and their compatibility with explainability frameworks such as SHAP.

To evaluate the performance of the proposed fraud detection models, this study applies a combination of statistical and domain-relevant metrics. Given the imbalanced nature of fraud datasets, traditional accuracy alone is insufficient and potentially misleading. Therefore, a robust suite of evaluation metrics is adopted to better capture the nuances of fraud detection.

Precision measures the proportion of correctly predicted fraud cases among all transactions flagged as fraud, helping to reduce false positives. Recall quantifies the model's ability to correctly identify actual fraudulent transactions, ensuring critical threats are not overlooked. The F1 Score, as the harmonic mean of precision and recall, balances the trade-off between capturing fraud and minimizing false alerts. The confusion matrix is used to present classification outcomes in terms of true positives, false positives, true negatives, and false negatives. It provides a foundational tool for assessing how well the model distinguishes between fraudulent and legitimate activity.

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) evaluates the model's overall ability to discriminate between fraud and non-fraud cases across varying thresholds. A higher AUC indicates better model performance. Additionally, the False Positive Rate (FPR) is tracked to quantify the risk of incorrectly flagging legitimate transactions as fraudulent an important consideration in minimizing customer friction and operational inefficiency. Where applicable, cost-sensitive evaluation is also conducted to estimate the financial implications of false positives and false negatives. This approach supports the business case for adopting the model by accounting for the real-world impact of classification errors. While accuracy is reported for completeness, it is not treated as a primary performance metric due to its limited usefulness in imbalanced classification settings. Each selected metric contributes to a comprehensive evaluation of model effectiveness, ensuring the fraud detection system is not only statistically sound but also practical for real-time deployment. The integration of XAI tools further allowed the visualization of feature importance and local explanations for individual predictions, offering insight into model reasoning and regulatory compliance.

To ensure fair assessment, the chosen evaluation metrics collectively capture both technical accuracy and business relevance. The next stage involves benchmarking these models against existing fraud detection approaches to contextualize performance gains.

3.9.6 Baseline Comparison

To validate the proposed approach, model performance will be benchmarked against two baseline systems. The first baseline is the existing rule-based fraud detection system currently employed within the operational environment. This system relies on fixed thresholds and manually curated heuristics to flag suspicious activity.

The second baseline is a minimally configured statistical model, such as an unoptimized Logistic Regression, which will serve as a reference point for improvements introduced through advanced modeling and preprocessing techniques. These benchmarks will be evaluated using the same testing data and performance metrics, ensuring consistency in comparison. The results will help to quantify the added value of machine learning in terms of detection rate, false alarm reduction, and overall reliability.

3.10 Ethical Considerations

Ethical considerations are central to any research involving sensitive financial or transactional data, particularly in the fintech sector, where confidentiality and data integrity are paramount. This study adheres to strict ethical guidelines to ensure the responsible use, storage, and dissemination of information, consistent with both academic research standards and data protection laws in Kenya.

The dataset used in this research originates from Pesapal Ltd, a licensed payment service provider regulated by the Central Bank of Kenya (CBK). The data was collected during the company's normal operations between June and August 2023 and comprises anonymized transaction records processed through both Point-of-Sale (POS) and e-commerce platforms.

The dataset includes 31,518 records and 14 key variables, such as transaction amount, timestamp, merchant ID, payment method, and a binary fraud label.

3.10.1 Informed Data Use and Consent

Direct consent from individual data subjects was not required since the dataset was secondary and anonymized before access was granted. However, the use of data was reviewed and approved under Pesapal's Data Governance and Privacy Policy, which explicitly permits the use of anonymized data for internal research, risk management, and innovation purposes. This aligns with Section 25 of the Kenya Data Protection Act (2019), which allows the use of personal data for scientific or research purposes provided that all identifying information is removed and the research serves the public interest.

To maintain transparency, the researcher sought formal authorization from the Data Governance and Compliance Department of Pesapal Ltd before data extraction. The request detailed the purpose, scope, and retention period of the data. Only transactional attributes relevant to fraud detection were extracted, ensuring minimal exposure of personally identifiable information (PII).

3.10.2 Data Anonymization and Privacy Protection

All PII including customer names, email addresses, card numbers, phone numbers, and IP addresses was removed or masked before analysis. This process ensured that no single transaction could be traced back to an individual customer or merchant. Techniques such as hashing, tokenization, and truncation were applied where necessary.

The dataset was stored in a secure, password-protected research environment, encrypted using AES-256 standards, and accessed only by the researcher and academic supervisor. Data transfer between storage locations was conducted through Secure File Transfer

Protocol (SFTP) and TLS-encrypted channels, thereby mitigating the risk of data leakage during handling.

3.10.3 Ethical Handling During Analysis

During data processing and model training, ethical safeguards were implemented to ensure the integrity and fairness of the algorithms. In line with European Commission AI Ethics Guidelines (2019) and OECD Principles on AI (2021), model development adhered to fairness and explainability standards. The study prioritized transparency in model decision-making processes, ensuring that predictions were based solely on transactional data patterns.

The research also addressed the ethical implications of false positives and false negatives in fraud detection. Incorrectly flagging legitimate transactions could inconvenience customers or harm merchant reputations, while failing to detect fraudulent cases could lead to financial losses. To mitigate these risks, the evaluation metrics and cost-sensitive analysis were designed to balance detection accuracy with fairness and customer experience, ensuring that the model's performance aligned with real-world consequences.

3.10.4 Institutional and Legal Compliance

This research strictly complies with:

i. The Kenya Data Protection Act (2019), ensuring lawful and fair processing of data.
ii. Central Bank of Kenya’s Prudential Guidelines for Payment Service Providers (2020), emphasizing operational risk management and consumer protection.
iii. PCI DSS (Payment Card Industry Data Security Standards), mandating encryption, controlled access, and secure data handling for financial information.

iv. University Research Ethics Committee Guidelines, which require informed data use, academic honesty, and avoidance of harm to participants or institutions.

A data handling protocol was developed and followed throughout the project lifecycle. Access logs were maintained, and all datasets will be permanently deleted upon completion of the research and presentation of findings, in line with Pesapal's Data Retention Policy (2023).

3.10.5 Ethical Transparency and Reporting

Results are presented in aggregate form without identifying individual merchants, banks, or customers. The reporting focuses on systemic fraud detection patterns rather than attributing responsibility to specific entities. Care has been taken to ensure that no interpretations could unfairly damage reputations or business interests. Furthermore, the research findings will be shared with Pesapal Ltd in summary form, allowing the company to evaluate insights for operational improvement while safeguarding proprietary information.

In summary, the ethical framework for this research integrates data protection, fairness, transparency, and accountability principles. It ensures that while the study contributes to academic knowledge and practical innovation, it does so responsibly, with full respect for privacy, institutional trust, and legal compliance.

3.11 Data Limitations

Despite careful planning and robust methodology, this study acknowledges several inherent limitations that could affect the interpretation or generalizability of the results. These limitations relate to data scope, quality, labeling accuracy, feature availability, and contextual generalization, all of which are critical considerations in fraud detection research.

3.11.1 Limited Data Scope and Temporal Coverage

The dataset covers transactions processed between June and August 2023, representing a three-month period. While this timeframe provides a manageable dataset for experimentation,

it may not capture seasonal or event-driven variations in fraud behavior, such as those observed during holidays, promotional periods, or financial crises. Longer-term data would provide richer insights into evolving fraud trends, particularly as fraudsters adapt to detection mechanisms over time.

Furthermore, data was sourced from a single fintech provider (Pesapal Ltd). While the company serves a diverse merchant base across East Africa, variations in transaction structures, user demographics, and security architectures across other fintech firms may limit the generalizability of findings. Future studies could address this by incorporating multi-institutional datasets to improve external validity.

3.11.2 Data Quality and Labelling Accuracy

The study relies on fraud labels assigned by Pesapal's existing fraud detection systems, which use a combination of automated rules and human review. Although these systems are highly developed, there is a possibility of misclassification cases where genuine transactions were labeled as fraudulent (false positives) or undetected fraud was labeled as legitimate (false negatives). These labeling inconsistencies can affect the accuracy of model training and evaluation.

To mitigate this, data validation procedures were employed, including cross-referencing flagged transactions with audit trails and manual verification summaries where available. Nevertheless, the potential for undetected fraud remains an inherent limitation in any supervised fraud detection model trained on historical data.

3.11.3 Imbalance Data and Class Representation

The inherent class imbalance where fraudulent transactions constitute a very small proportion of total activity remains a persistent challenge. Although techniques such as SMOTE and class weighting were used to alleviate bias, these methods cannot fully replicate

the diversity of real-world fraud. The imbalance may also lead to slight over- or underestimation of performance metrics, particularly precision and recall.

CHAPTER FOUR

SYSTEM DESIGN AND ARCHITECTURE

This chapter presents the design and architectural framework of the proposed machine learning based fraud detection system developed for Kenya's digital payments ecosystem, specifically within Pesapal's operational context. The system's architecture is designed to detect fraudulent financial transactions in real-time by integrating data preprocessing, predictive modeling, and continuous feedback mechanisms.

The design emphasizes four fundamental objectives:

i. Scalability, to handle high transaction volumes typical of POS and e-commerce channels.
ii. Adaptability, to ensure the model continuously learns from new fraud patterns.
iii. Efficiency, enabling sub-second transaction scoring to avoid customer friction.
iv. Transparency, ensuring predictions remain explainable and compliant with financial regulations such as the Kenya Data Protection Act (2019) and the CBK's Risk Management Guidelines for Payment Service Providers.

The chapter is structured into three major sections. Section 4.1 presents the system architecture, mapping the end-to-end data flow from ingestion to prediction. Section 4.2 discusses the system design, illustrating user interactions through a use case diagram. Section 4.3 prioritizes security and scalability in the model's deployment. The overall design reflects the integration of data science methodologies into a fintech production environment.

4.1 System Architecture

The system architecture represents the blueprint of how the fraud detection solution operates encompassing data ingestion, preprocessing, model training, evaluation, and

deployment. It ensures efficient data handling, modularity, and real-time integration with existing transaction systems.

The proposed architecture adopts a layered modular approach, as illustrated *Figure 4.1: System Architecture*, comprising six core layers;

- ***Data Ingestion Layer***

This layer retrieves historical and live transaction data from Pesapal's databases through secure API connections and scheduled batch imports. Data sources include POS terminals, e-commerce gateways, and mobile applications. The ingestion process standardizes varying data formats (JSON, CSV, SQL tables) into a unified structure for analysis.

- ***Data Preprocessing Layer***

This component is responsible for data cleaning, normalization, and feature engineering. Duplicate records, missing values, and inconsistent attributes are corrected. Feature transformations such as Z-score normalization, encoding of categorical variables, and derivation of new metrics (e.g., transaction velocity, geographic deviation) occur at this stage. Class imbalance handling using SMOTE is also implemented here to ensure model fairness.

- ***Model Training and Validation Layer***

Preprocessed data is divided into training (70%) and testing (30%) subsets. Machine learning algorithms such as Logistic Regression, Random Forest, Gradient Boosting, and Naïve Bayes are trained and compared using the training set. Hyperparameter optimization (via GridSearchCV) and 5-fold cross-validation are performed to enhance performance stability.

- ***Model Evaluation Layer***

The trained models are assessed on the testing dataset using fraud-relevant metrics such as Precision, Recall, F1-Score, ROC-AUC, and False Positive Rate. Models failing to meet minimum thresholds are re-trained, while the best-performing model (Random Forest in this study) progresses to the deployment phase.

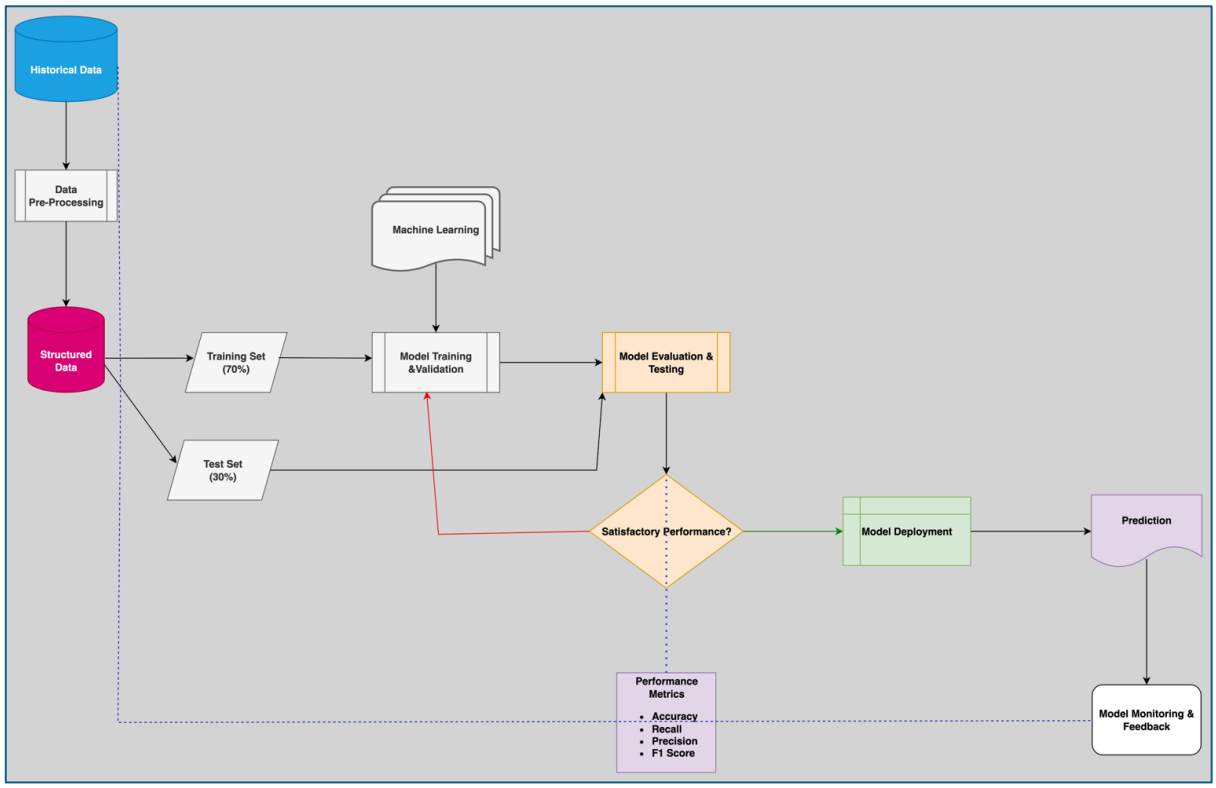
- ***Prediction and Decision Layer***

This is the operational layer where the model is deployed as a RESTful API service. Incoming transactions are fed in real-time and classified as fraudulent or legitimate. Predictions are returned with associated confidence scores, allowing dynamic decision-making for example, auto-declining high-risk transactions or flagging them for manual review.

- ***Monitoring and Feedback Layer***

The final layer maintains system health, tracks model drift, and manages retraining schedules. It captures false positives/negatives from human review outcomes and feeds them back into the data pipeline for periodic model refinement. This continuous learning loop ensures the model adapts to evolving fraud patterns and regulatory expectations.

FIGURE 4
System Architecture Of The Machine Learning Based Fraud Detection System



The architecture's modularity allows for scalable deployment either on-premises or within cloud environments (e.g., AWS, Google Cloud, or Azure) depending on infrastructure

capacity. Each module operates independently, facilitating parallel development, maintenance, and model retraining with minimal system downtime.

4.2 System Design

The system design focuses on how users (primarily administrators and analysts) interact with the fraud detection platform and how the machine learning model integrates into existing operational workflows.

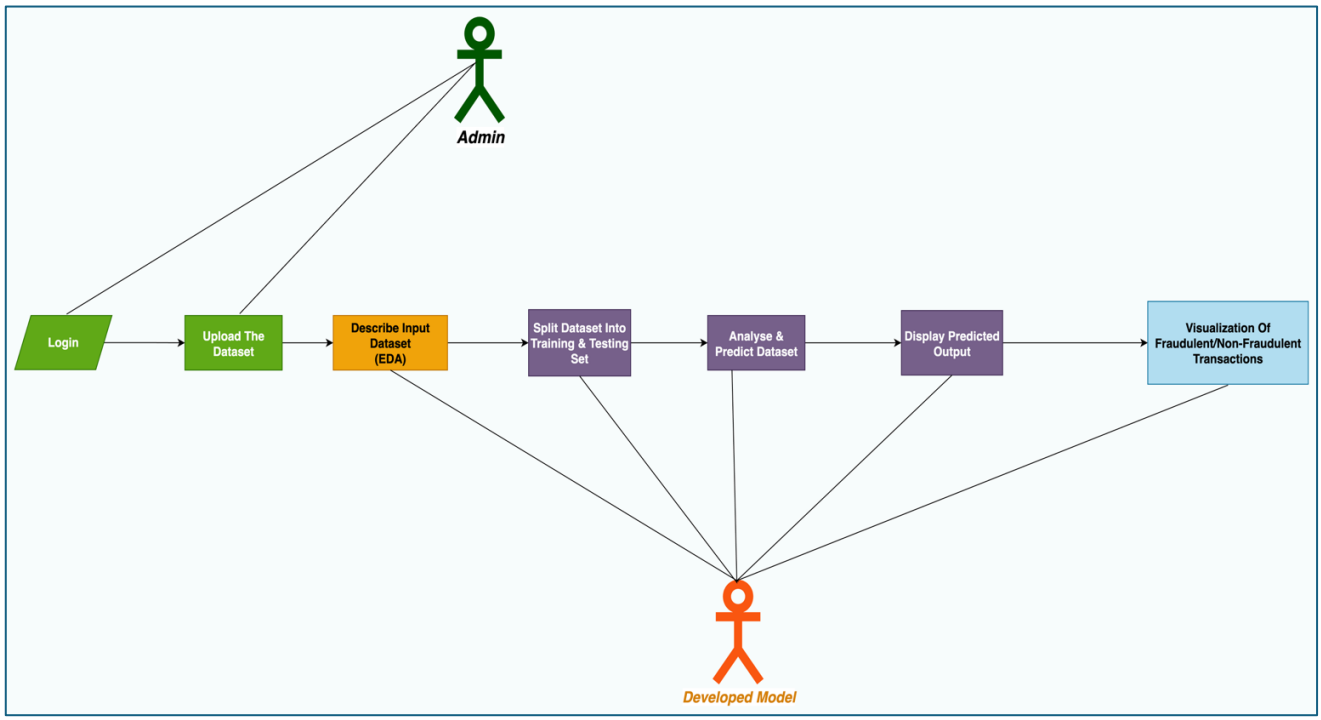
4.2.1 Use Case Overview

The system supports two main actors: Admin/Analyst: Responsible for uploading data, initiating model training, viewing results, and monitoring system performance.

Prediction Model: The core intelligence engine that processes transactions, performs classification, and returns risk scores.

The interaction between these actors is depicted in *Figure 5: Use Case Diagram.*

FIGURE 5
Use Case Diagram For Fraud Detection System



4.2.2 System Workflow

- *Data Upload and Initialization*

The Admin logs into the system and uploads historical transaction data via a secure web interface. The system validates file format and schema consistency before initiating preprocessing.

- *Exploratory Data Analysis (EDA)*

The platform automatically performs EDA, generating summary statistics, correlation matrices, and anomaly visualizations. This helps analysts identify patterns such as spikes in fraud at certain times or locations.

- *Model Training and Evaluation*

The Admin triggers the model training phase. Multiple algorithms are executed sequentially, and performance metrics are displayed in tabular and graphical form (e.g., ROC curves).

- *Prediction and Visualization*

Once trained, the model classifies new incoming transactions in real-time. The results are displayed in a dashboard showing fraud probability, transaction ID, and confidence levels. Fraudulent transactions are highlighted in red, while legitimate ones appear in green.

- *Reporting and Decision Support*

Analysts can export predictive results and performance summaries as PDF or Excel reports. These insights inform fraud prevention strategies, audit trails, and compliance documentation.

- *Feedback and Continuous Improvement*

The system incorporates human feedback loops where analysts verify flagged transactions. Confirmed results are stored and later used to retrain and improve model precision.

4.3 Security and Deployment Considerations

Given that fraud detection systems handle sensitive financial data, system security is a top priority. To ensure data protection, role-based authentication restricts access to authorized users, such as Data Scientists and Analysts, who can upload, train, or deploy models. Data is encrypted both in transit using TLS 1.3 and at rest using AES-256. All system actions, including uploads, model training, and manual overrides, are logged for audit traceability. The model is deployed as a containerized microservice using Docker, enabling portability and scalability, with REST APIs facilitating seamless integration with Pesapal's live transaction systems.

4.4 Summary

This chapter detailed the architectural and functional design of the proposed fraud detection framework. The system is built on a modular architecture, integrating machine learning algorithms with real-time transaction streams. It enables fraud prediction through a structured pipeline of data ingestion, preprocessing, model training, validation, and continuous feedback.

The use case diagram and process flow provide a practical view of system operations, illustrating how the model interacts with human analysts to ensure both accuracy and explainability. The chapter also addressed key security and deployment considerations to ensure compliance, resilience, and ethical data handling.

This design lays the groundwork for chapter five, which will present the model evaluation results, performance comparisons among algorithms, and interpretation of outcomes in relation to the study's objectives.

CHAPTER FIVE

DATA ANALYSIS AND INTEPRETATION

This chapter presents a comprehensive and systematically organized analysis of the dataset and the evaluation results derived from the machine learning models developed for fraud detection. The analysis encompasses descriptive, exploratory, and predictive dimensions, focusing on how different algorithms perform in identifying fraudulent transactions within Pesapal's digital payment ecosystem. To strengthen analytical coherence, the presentation of results is explicitly aligned with the study's three research objectives, thereby ensuring a logical progression from diagnosing the limitations of existing rule-based systems to assessing machine learning model performance and ultimately proposing a contextually relevant, real-time fraud detection framework suitable for the Kenyan fintech environment.

The chapter examines the distribution and characteristics of the dataset, explores fraud patterns across countries, banks, and card types, and evaluates the predictive performance of several supervised machine learning algorithms including Logistic Regression, Random Forest, Gradient Boosting, Naïve Bayes, Decision Tree, K-Nearest Neighbors (KNN), and Neural Networks (MLP). This evaluation is structured across the full model development pipeline, beginning with baseline results on the unbalanced dataset, followed by performance after class rebalancing using the Synthetic Minority Oversampling Technique (SMOTE), and culminating in improvements obtained through hyperparameter optimization. This structured sequence not only mirrors the methodological rigor of the study but also enhances the interpretability of the comparative results.

Key performance indicators ; Precision, Recall, F1-score, Accuracy, and PR-AUC are employed to assess and compare model effectiveness at various stages of development. To enhance analytical clarity and minimize redundancy, a consolidated performance table is introduced later in the chapter to summarize the evolution of model behavior across the

unbalanced, balanced, and tuned scenarios. The findings are interpreted with consideration for the operational demands of real-time fraud detection, highlighting both predictive capability and practical deployment feasibility within high-throughput POS and e-commerce settings.

The purpose of this chapter extends beyond the mere presentation of statistical outputs. It seeks to articulate the implications of these results for fraud mitigation within Pesapal and similar digital payment platforms, demonstrating how the empirical evidence contributes to financial risk management, transaction monitoring, and regulatory compliance. In doing so, the chapter lays a substantive foundation for the development of an explainable, scalable, and context-sensitive machine learning framework tailored to Kenya's evolving digital payments ecosystem.

5.1 Data Description

The dataset used in this study comprises 31,518 transaction records extracted from Pesapal's internal payment system for the period June to August 2023. As illustrated in Figure *6 and Figure 7* ,each transaction entry includes fourteen columns capturing both numerical and categorical variables relevant to fraud identification. The dataset is structured and complete, containing no missing values, which ensures its readiness for modeling and analysis. The data occupies approximately 3.4 MB of storage and exhibits a balanced distribution between categorical and numerical attributes, providing a suitable foundation for machine learning experimentation.

Several scoring parameters form the backbone of fraud assessment in the dataset. The "Name Match" variable determines whether the cardholder's name aligns with the registered customer or passenger, highlighting identity mismatches that may indicate fraudulent use of credentials. The "BIN" parameter identifies whether the card's Bank Identification Number is blacklisted within the Pesapal system, signaling a history of compromised or high-risk cards. The "IP GEO vs GEO" feature compares the IP address used during the transaction with the

customer's registered location, detecting geographical inconsistencies often associated with cross-border fraud. Similarly, the "Customer Name Match" field reinforces identity validation by checking if the buyer's name corresponds with the cardholder's. The "Email Prefix vs Customer Name" feature examines whether the email prefix resembles the cardholder's name, exposing potentially synthetic identities. Finally, the "BIN vs GEO" feature cross-verifies whether the issuing bank's geographical location aligns with that of the customer's registered address.

Together, these variables construct a multidimensional fraud detection framework integrating identity, geography, and transaction behavior. Their inclusion ensures that the models learn to differentiate between legitimate and suspicious transactions based on multiple perspectives of transactional integrity. The completeness and structure of this dataset make it ideal for developing and evaluating supervised learning models for fraud detection.

FIGURE 6
Overview Of The Dataset

	Trans Id	Merchant Country	Card Country	Customer Country	Card Bank	Processing Bank	CardType	Fraud Status	BIN vs GEO	Customer name match	Email prefix	IP GEO vs GEO	Name match	Service date
0	21037998	Kenya	-	Kenya	-	Ecobank Kenya	Visa	Low Risk (Checked)	0	0	100	0	0	100
1	24099357	Kenya	-	Somalia	-	Ecobank Kenya	Visa	Low Risk (Checked)	200	0	100	0	0	0
2	24446749	Kenya	-	Uganda	-	Ecobank Kenya	Visa	Low Risk (Checked)	200	0	100	0	0	100
3	25476314	Kenya	-	Tanzania	-	Ecobank Kenya	Visa	Low Risk (Checked)	200	0	100	0	70	0
4	25478463	Kenya	-	Kenya	-	Ecobank Kenya	Visa	Low Risk (Checked)	200	0	100	0	70	100

5.2 Data Preprocessing and Cleaning

The dataset was found to be complete, with no missing values across all 31,518 transactions. While the absence of missing entries removed the need for imputation, preprocessing remained essential to ensure that the data could be effectively utilized by machine learning algorithms.Categorical variables, including Merchant Country, Card Bank, Processing Bank, and Card Type, were encoded into numerical representations. One-hot encoding was employed for features with multiple categories to prevent the introduction of

artificial ordinal relationships that could bias model training. This transformation allowed the algorithms to interpret each category as a distinct entity, preserving the integrity of categorical information.

Outlier detection and treatment were applied to numerical variables such as BIN vs GEO, Service Date, and mismatch scores. Extreme values can distort the learning process of machine learning models, particularly tree-based algorithms and distance-based classifiers. Outliers were assessed and capped within reasonable bounds to maintain the meaningful variance of the data while reducing the influence of anomalous entries that could otherwise dominate the model's decision-making.In addition, derived features were constructed to capture transaction patterns that may indicate fraudulent behavior. For example, mismatch indicators between customer and card country, or unusually rapid sequences of transactions, were calculated to highlight potential risk. These engineered features augmented the predictive capacity of the models, providing additional dimensions for discriminating between legitimate and fraudulent activity.

Finally, numerical scaling was applied where appropriate. Features with differing ranges were standardized to ensure that attributes with larger numerical scales did not disproportionately influence model learning. This is particularly critical for algorithms such as Support Vector Machines, which rely on distance metrics, and for gradient-based models where scale can affect convergence.

Through these preprocessing steps, the dataset was rendered both structured and analytically coherent, facilitating accurate, interpretable, and reliable model training. The careful treatment of categorical encoding, outlier management, feature derivation, and scaling underscores the methodological rigor necessary for fraud detection studies within high-dimensional, real-world fintech transaction data.

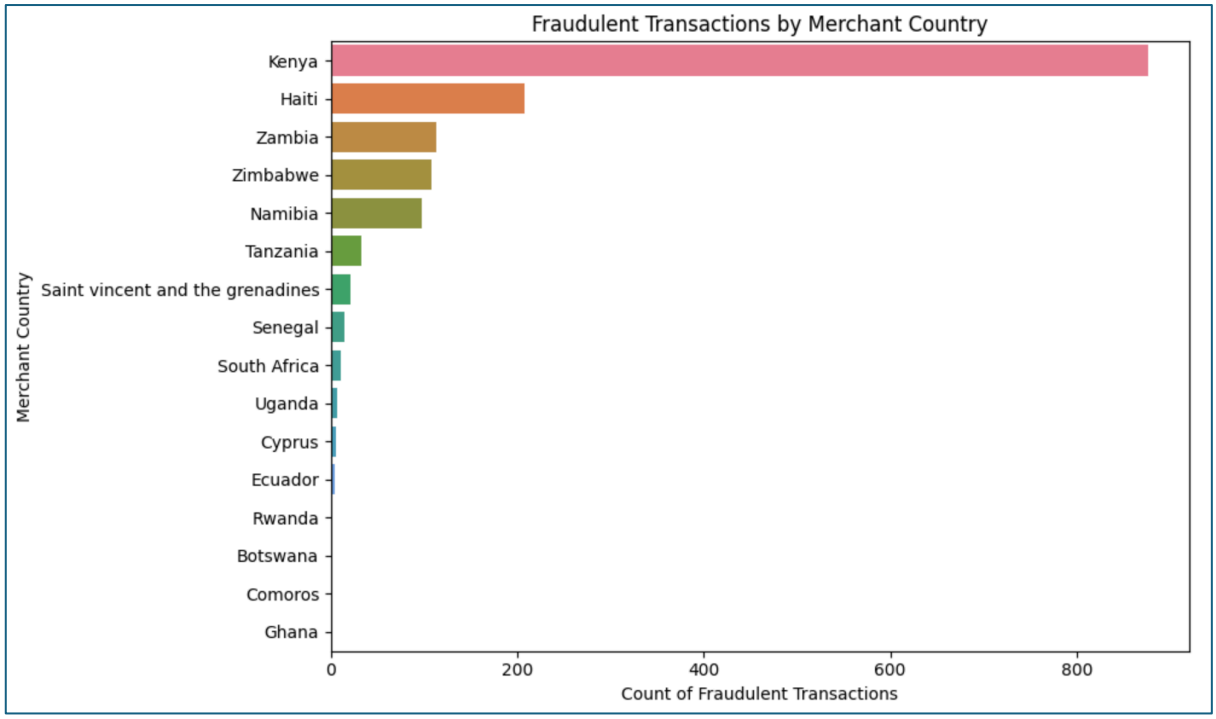
FIGURE 7
Dataset Summary

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31518 entries, 0 to 31517
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Trans Id             31518 non-null  int64
 1   Merchant Country     31518 non-null  object
 2   Card Country         31518 non-null  object
 3   Customer Country     31518 non-null  object
 4   Card Bank            31518 non-null  object
 5   Processing Bank      31518 non-null  object
 6   CardType             31518 non-null  object
 7   Fraud Status         31518 non-null  object
 8   BIN vs GEO           31518 non-null  int64
 9   Customer name match  31518 non-null  int64
 10  Email prefix         31518 non-null  int64
 11  IP GEO vs GEO        31518 non-null  int64
 12  Name match           31518 non-null  int64
 13  Service date         31518 non-null  int64
dtypes: int64(7), object(7)
memory usage: 3.4+ MB
```

5.3 Exploratory Data Analysis

In the following section on Exploratory Data Analysis (EDA), we conduct an in-depth examination of the dataset to uncover patterns and trends associated with fraudulent transactions. *Figures 8* through *Figure 14* illustrate key aspects of the data, highlighting various dimensions of fraud, such as the countries with the highest fraud incidence, the most frequently used card types, and the noticeable imbalance between fraudulent and non-fraudulent cases. These visualizations offer valuable insights into the characteristics of fraudulent activity and help establish a foundational understanding that informs subsequent analysis and model development.

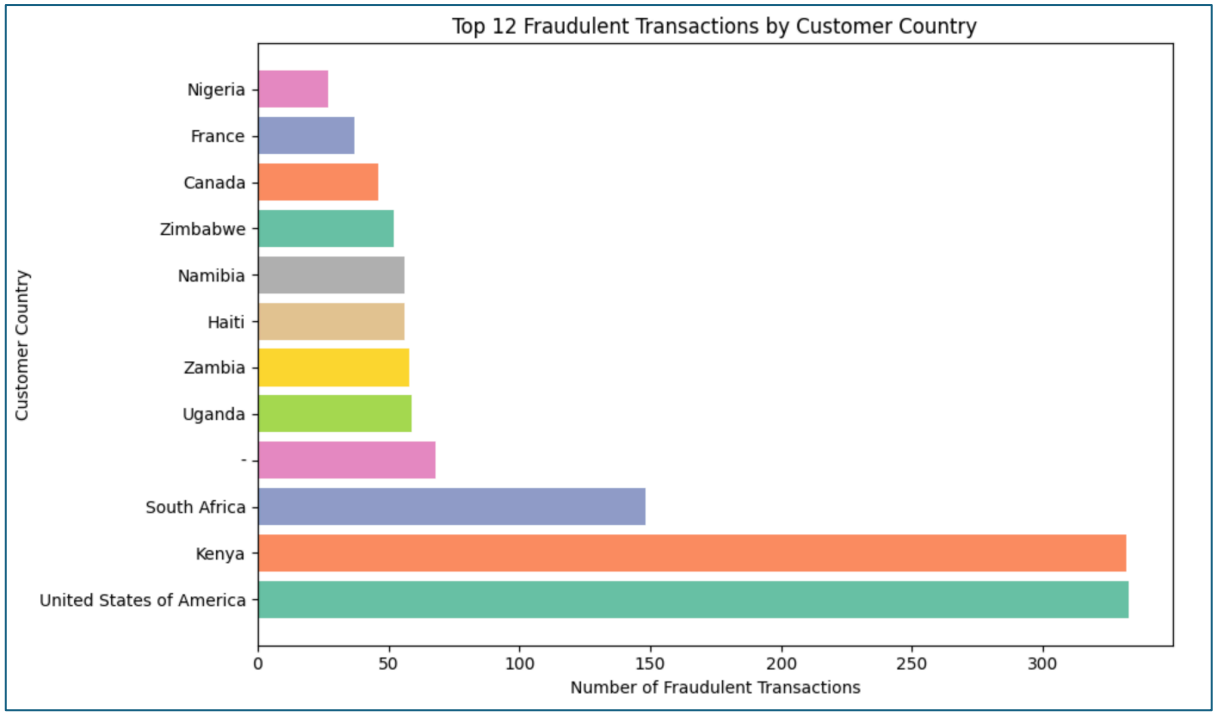
FIGURE 8
Fraud Transactions By Country



In *Figure 8*, the bar graph shows that merchants based in Kenya, Haiti, Zambia, Zimbabwe, and Namibia account for the highest number of fraudulent transactions. This indicates a higher prevalence of fraud in transactions originating from these countries.

Figure 9 displays a bar graph of customer-wise fraudulent transactions, showing that customers from the USA, Kenya, South Africa, and Uganda are associated with the highest number of fraud cases. These cross-border occurrences imply that fraudulent activities often exploit international transaction networks and weaknesses in verification systems across jurisdictions.

FIGURE 9
Fraud By Customer Country



As shown in *Figure 10*, the bar graph reveals that the highest number of fraudulent transactions are associated with cards issued by JPMorgan Chase Bank, Bank of America, and Wells Fargo Bank. This trend points to a higher exposure to fraud among these financial institutions. In *Figure 11*, the graph focuses on Kenyan-issued cards, indicating that Barclays Bank of Kenya Ltd and DTB Bank have experienced the most cases of fraud. These patterns suggest that large, globally connected banks may face heightened exposure due to the scale and diversity of their customer bases.

FIGURE 10
Top 10 Fraud Card Banks

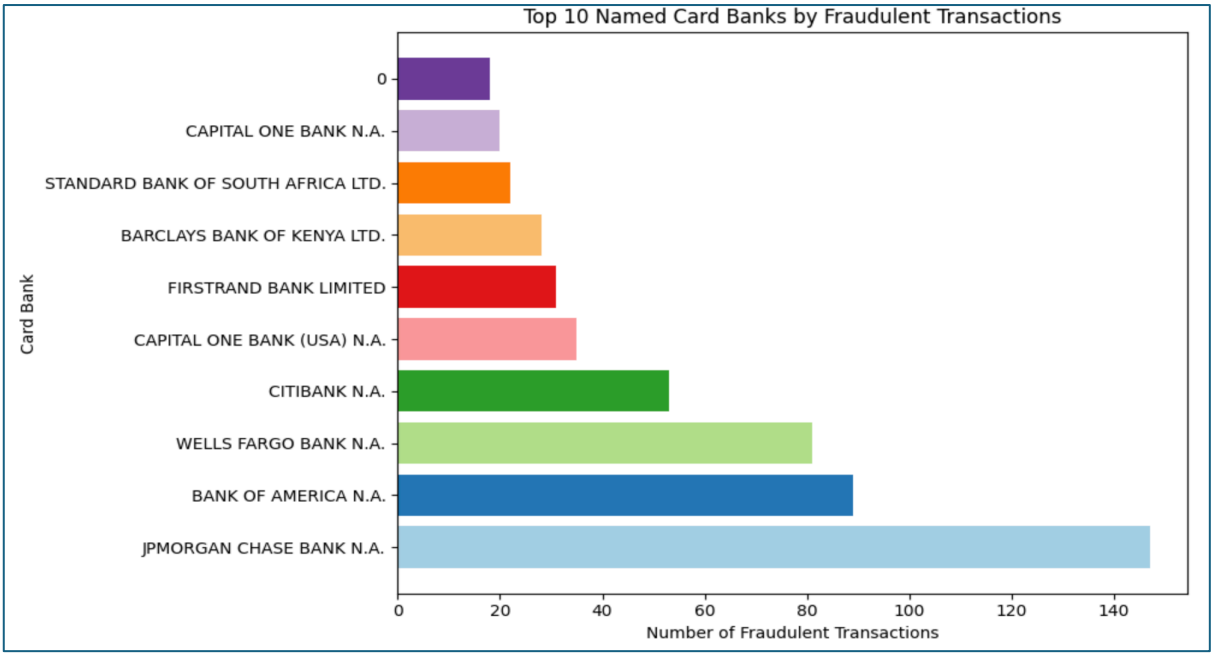


FIGURE 11
Top 5 Fraud Kenya Card Banks

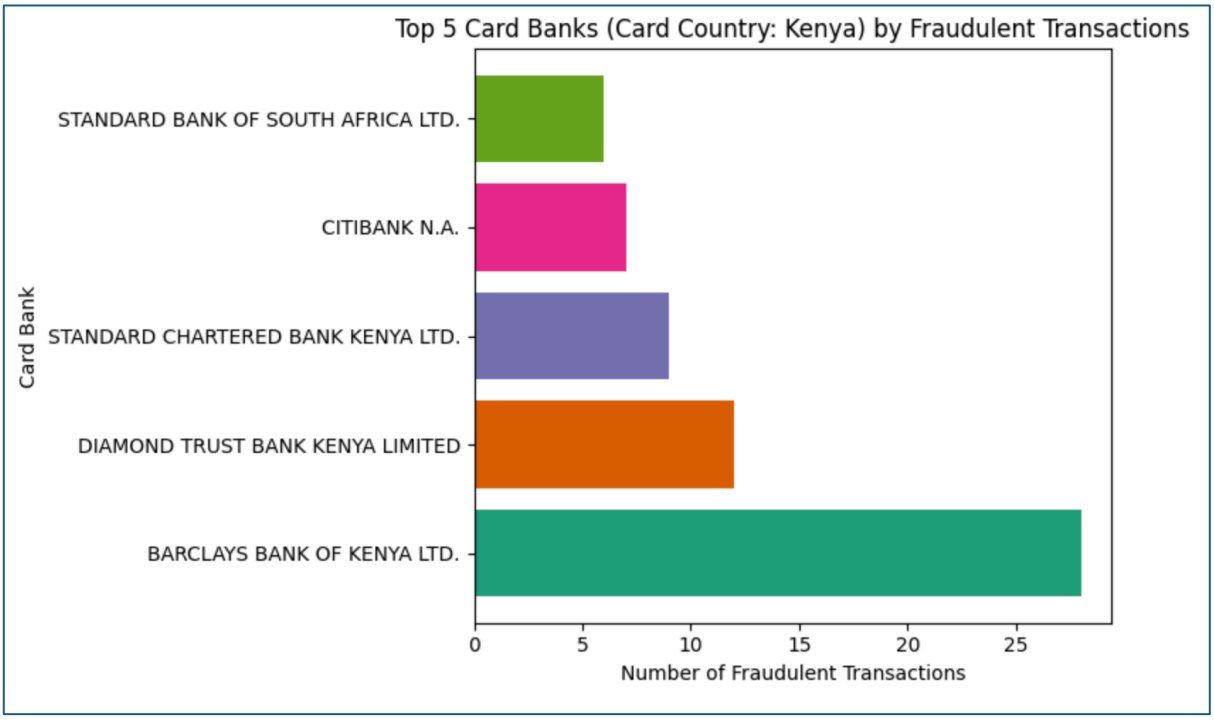
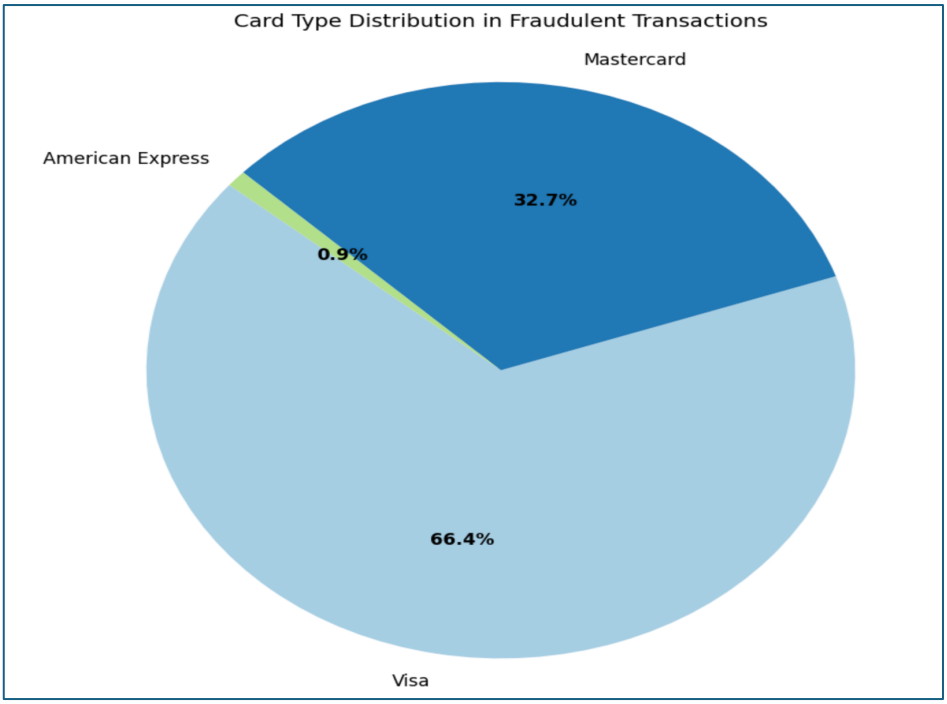


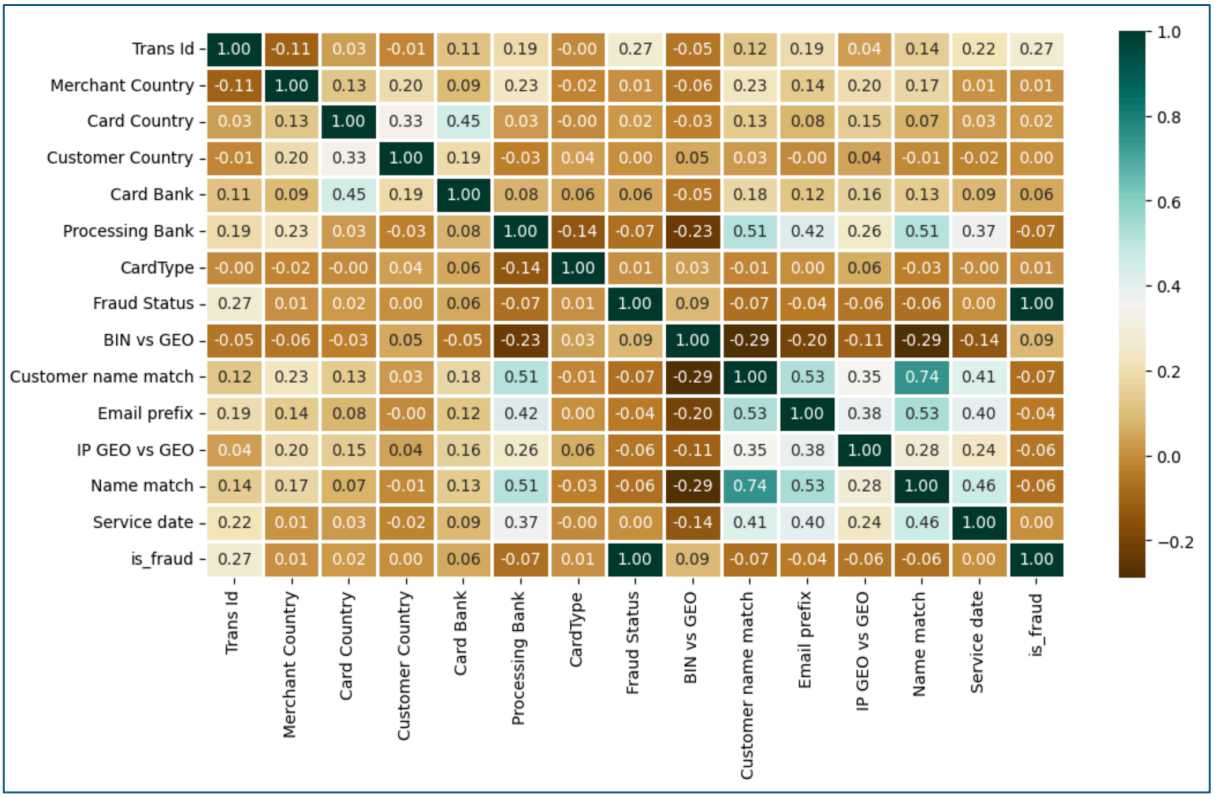
Figure 12 illustrates the distribution of fraudulent transactions by card type using a pie chart. The data reveals that Visa cards experience the highest proportion of fraudulent activity, followed by Mastercard. This pattern suggests that these card types may be more commonly exploited in fraudulent schemes, potentially due to their widespread usage and global acceptance.

FIGURE 12
Fraud By Card Types



The correlation heatmap in *Figure 13* shows fraud detection patterns in financial transactions. The strongest fraud indicator is customer name matching (0.74), suggesting fraudsters often reuse identity information across transactions. Email prefix patterns (0.53) and processing bank relationships (0.51) also show moderate correlations with fraud. Geographic mismatches, particularly between bank identification numbers and actual locations (-0.29), indicate potential fraud activity. Service date timing (0.46) reveals temporal fraud patterns. This suggests that identity verification, geographic consistency checks, and monitoring email patterns are the most effective fraud detection strategies, while transaction timing and bank relationships provide additional useful signals for fraud prevention systems.

FIGURE 13
Correlation Matrix Of Fraud Detection Features



5.4 Performance Challenges of Rule-Based Fraud Controls in Digital Payment Ecosystems

5.4.1 Original Dataset Distribution

An examination of the dataset revealed a pronounced imbalance between fraudulent and legitimate transactions as presented in *Figure 14* , with 1,503 labeled as fraud and 30,015 classified as non-fraudulent. This disparity mirrors real-world financial transaction patterns, where fraudulent activities are relatively rare but highly consequential. Such imbalance poses a significant modeling challenge, as most machine learning algorithms tend to prioritize the majority class, leading to models that achieve high overall accuracy but perform poorly in detecting minority class instances here, the fraudulent transactions. Consequently, addressing this imbalance was a critical step in ensuring that the models could accurately identify fraud without bias toward legitimate transactions.

FIGURE 14
Fraud Status Distribution

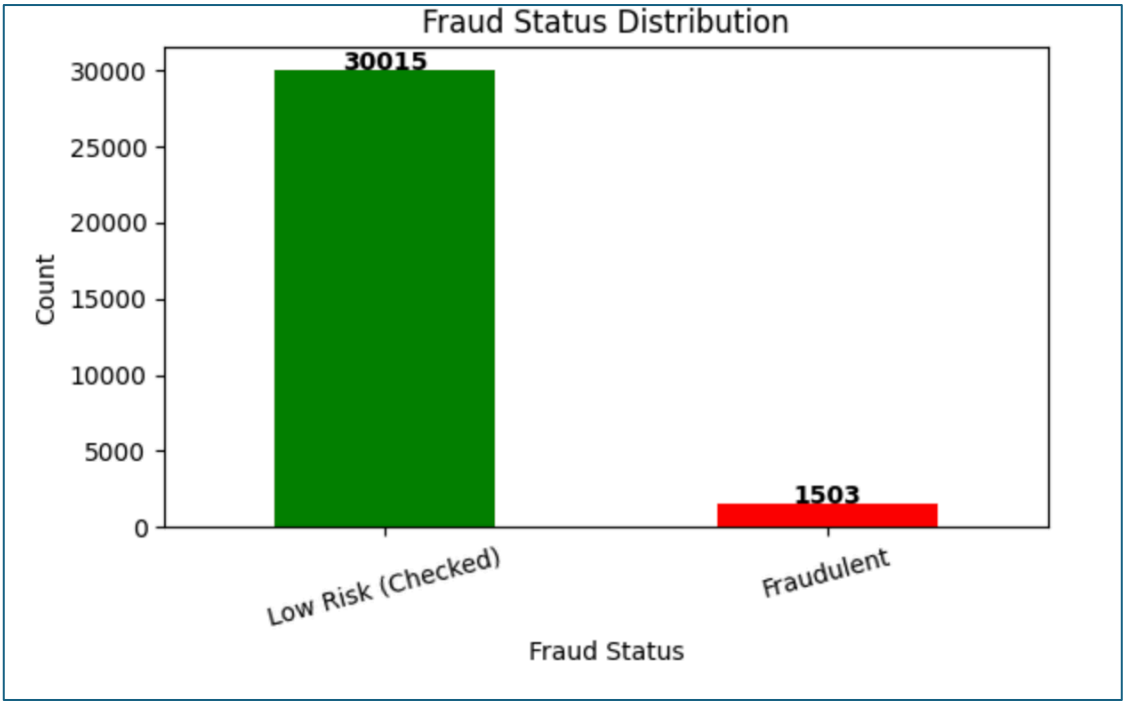
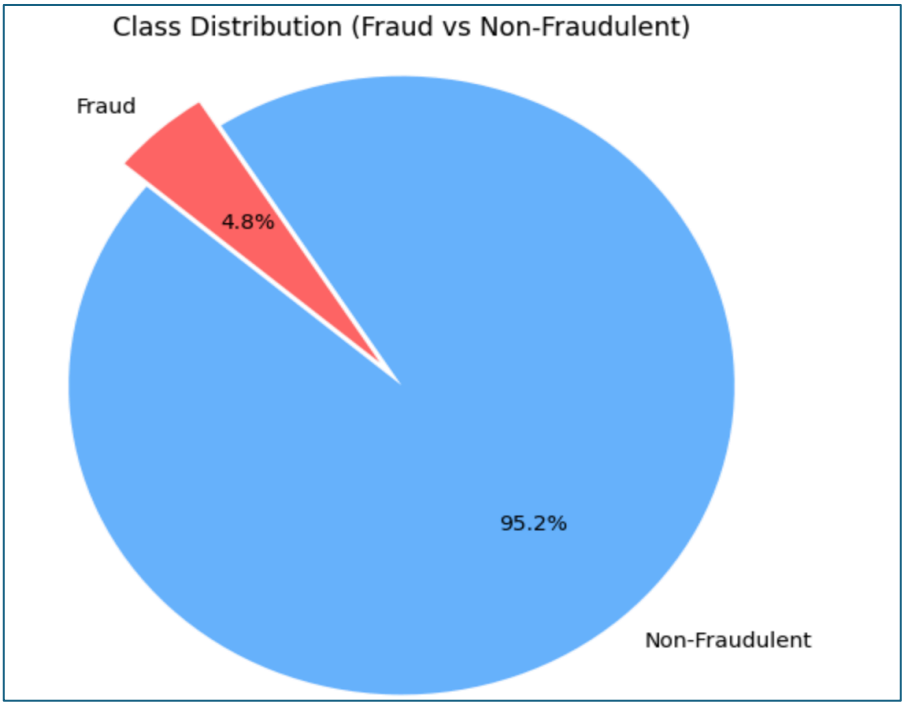


FIGURE 15
Class Distribution



5.4.2 Baseline Model Performance on Unbalanced Data

Given the substantial class imbalance in the dataset where legitimate transactions vastly outnumber fraudulent ones it was essential to evaluate how this skewed distribution influences model performance. Machine learning algorithms typically optimize for overall accuracy, which in highly imbalanced contexts can mask poor detection of minority-class instances. In fraud detection, such behavior is particularly problematic, as misclassified fraud cases translate directly into financial loss, operational risk, and reputational harm. Accordingly, baseline model evaluation on the unbalanced dataset provides critical insight into the limitations of traditional rule-based or naïve machine learning approaches, fulfilling the study's first objective.

To establish this baseline, several supervised learning algorithms Logistic Regression, Random Forest, Gradient Boosting, Naïve Bayes, Decision Tree, K-Nearest Neighbors (KNN), and a Neural Network (MLP) were trained using the original unbalanced dataset. Their performance was assessed using Accuracy, Precision, Recall, F1-score, and PR-AUC, with particular emphasis placed on Recall and F1-score due to their relevance in minority-class detection. As shown in *Table 3*, all models achieved high accuracy levels (≥85%), yet this metric proved misleading in the presence of severe imbalance. The models' performance varied substantially when evaluated on metrics sensitive to fraud detection effectiveness.

Random Forest demonstrated the strongest overall baseline performance, achieving an F1-score of 0.58239, supported by a high Precision of 0.81028 and a Recall of 0.45455. Its PR-AUC of 0.59136 was also the highest among all classifiers, indicating superior ability to discriminate between legitimate and fraudulent transactions despite the imbalance. Gradient Boosting similarly performed well, with a Precision of 0.86957, Recall of 0.39911, and F1-score of 0.54711, suggesting that while it is highly precise, it tends to miss a notable share of fraud cases.

Decision Tree provided a moderately strong baseline (F1-score 0.53742), with a Recall of 0.48559, making it valuable in contexts where capturing more fraudulent transactions is prioritized, even at the expense of increased false positives. In contrast, Logistic Regression, while achieving a relatively high Recall of 0.65188, recorded a very low Precision of 0.19017 and F1-score of 0.29444, reflecting a high false-positive rate that would be operationally costly in a real-world fraud monitoring environment.

Naïve Bayes showed weak performance across key fraud-sensitive metrics (F1-score 0.44087), indicating limited capability in modeling the nonlinear and high-dimensional relationships present in the data. K-Nearest Neighbors and the Neural Network model achieved modest performance (F1-scores of 0.47285 and 0.56675, respectively), yet neither surpassed the ensemble methods.

Overall, the results underscore a central challenge in fraud detection: high overall accuracy does not imply effective fraud detection performance. The baseline evaluations show that all models are biased toward predicting the majority class, achieving excellent accuracy while failing to identify a substantial proportion of fraud cases. This finding empirically demonstrates the inadequacy of both rule-based systems and naïvely trained machine learning models in handling highly imbalanced financial datasets, thereby confirming the necessity of class balancing and motivating the subsequent application of SMOTE and model optimization.

TABLE 3
Unbalanced Dataset Evaluation Report

Model Performance Comparison (Unbalanced Dataset)

Classifier	Accuracy	Precision	Recall	F1 Score	PR AUC
Logistic Regression	0.85099	0.19017	0.65188	0.29444	0.50521
Random Forest	0.96891	0.81028	0.45455	0.58239	0.59136
Gradient Boosting	0.96849	0.86957	0.39911	0.54711	0.57243
Naive Bayes	0.94966	0.46973	0.43016	0.44907	0.44007
Decision Tree	0.96013	0.60165	0.48559	0.53742	0.36022
K-Nearest Neighbors	0.96404	0.78756	0.33703	0.47205	0.4309
Neural Network (MLP)	0.96891	0.84279	0.42794	0.56765	0.58165

5.5 Evaluation of Machine Learning Model Performance on Real-World Fintech Transactions Data

5.5.1 Class Balancing with SMOTE

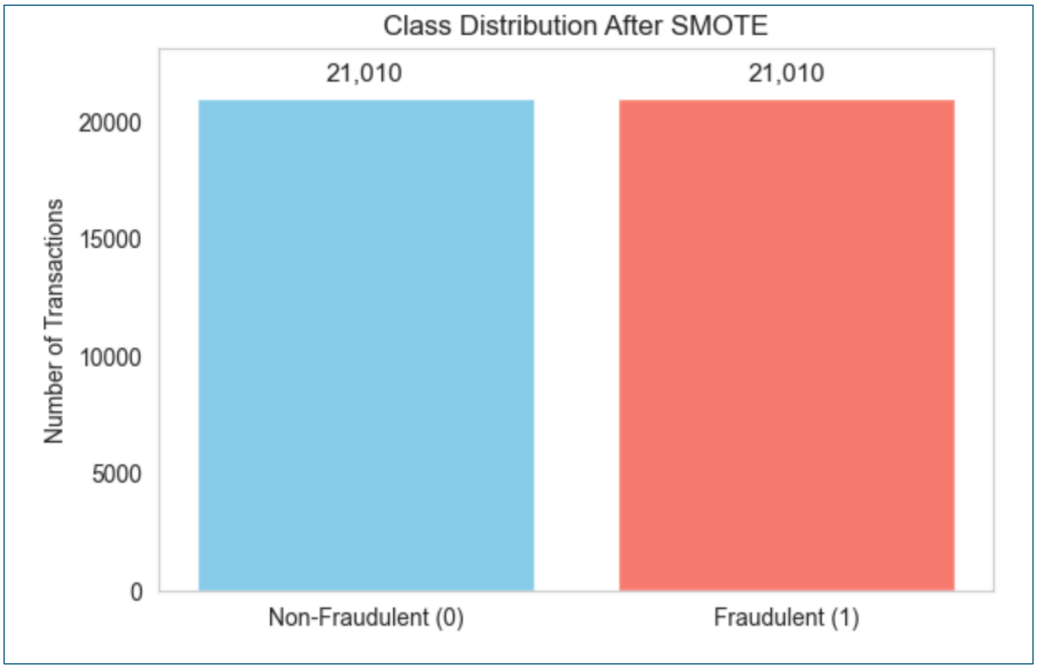
The disproportionate representation of fraudulent versus legitimate transactions posed a significant challenge to effective model training and evaluation. In datasets where one class vastly outweighs the other as is typical in financial fraud detection machine learning algorithms tend to bias their predictions toward the majority class. In this study, the minority class (fraudulent transactions) accounted for only a small fraction of the total observations, leading to models that initially exhibited high overall accuracy but poor recall for fraud cases. This imbalance resulted in elevated false-negative rates, a particularly detrimental outcome in fraud detection, where missed fraud events translate directly into financial losses and increased operational risk.

To mitigate this issue, the Synthetic Minority Oversampling Technique (SMOTE) was adopted. SMOTE generates synthetic minority-class observations by interpolating between

existing fraud instances, thereby increasing their representation within the training data. Importantly, SMOTE was applied only after the train–test split to prevent data leakage and preserve the integrity of model evaluation. By enriching the minority class through synthetic sampling, the technique produced a more balanced training distribution, allowing models to learn decision boundaries that better distinguish fraudulent from legitimate transactions.

As illustrated in *Figure 16*, the application of SMOTE resulted in a substantially improved class ratio, enabling the models to more effectively capture rare but high-impact fraud patterns. This adjustment was crucial in preparing the dataset for subsequent model training, improving sensitivity to fraud-related signals and setting the foundation for the performance enhancements observed in the balanced-dataset evaluations presented in the following section.

FIGURE 16
Fraud Distribution After SMOTE



5.5.2 Models Performance on SMOTE-Balanced Dataset

After applying SMOTE to balance the dataset, all classifiers were retrained and evaluated to determine how oversampling influenced fraud detection performance. As shown

in *Table 4*, class balancing substantially improved the models' ability to detect fraudulent transactions, reflected in the marked increase in recall across all algorithms. This improvement was expected, as SMOTE increases the representation of the minority class and enables models to learn more discriminative decision boundaries for fraud instances.

However, the improvements in recall were accompanied by reductions in precision, illustrating the inherent trade-off associated with oversampling. By exposing the classifiers to more synthetic fraud cases, the models became more sensitive to fraud signals but simultaneously misclassified a larger number of legitimate transactions as fraudulent. This precision–recall tension is a well-documented challenge in fraud detection, where greater sensitivity (recall) often comes at the cost of increased false positives.

Among the evaluated models, Random Forest achieved the most balanced performance after SMOTE, with Precision = 0.5254, Recall = 0.5277, F1-score = 0.5265, and PR-AUC = 0.5354. These results demonstrate that Random Forest not only adapted well to the balanced dataset but also maintained a reasonable equilibrium between capturing fraud and avoiding excessive false alarms. The model's consistent performance across multiple metrics reinforces its suitability for high-stakes fraud detection environments where both detection capability and operational precision matter.

Models such as Logistic Regression and the Neural Network (MLP) exhibited high recall (0.6652 and 0.6253 respectively), indicating strong sensitivity to fraudulent cases. However, their precision remained relatively low (0.1960 and 0.2944), which would translate into a high volume of false positives in practice. These models may therefore be less appropriate for real-time deployment unless accompanied by post-processing or human-in-the-loop verification mechanisms.

Gradient Boosting, Naïve Bayes, and Decision Tree displayed moderate but consistent improvements in fraud identification. Gradient Boosting achieved Recall = 0.5854 and

Precision = 0.3389, yielding an F1-score of 0.4293, while Naïve Bayes attained Precision = 0.4217 and Recall = 0.4656 (F1-score = 0.4426). Decision Tree produced a similar pattern, with Recall = 0.5277 and Precision = 0.3595, demonstrating improved minority-class sensitivity without surpassing the performance of Random Forest.

K-Nearest Neighbors (KNN) benefited modestly from SMOTE, achieving Recall = 0.5407 and Precision = 0.3407, though its F1-score (0.4194) remained lower than those of the ensemble models.

Overall, SMOTE substantially enhanced all models' sensitivity to fraudulent transactions, as evidenced by higher recall scores. Nonetheless, these gains often came at the expense of precision, resulting in moderate F1-scores and PR-AUC values. Random Forest emerged as the most reliable and operationally viable model on the SMOTE-balanced dataset, offering the best compromise between fraud detection capability and manageable false-positive levels. This balance is critical for practical fraud monitoring systems, where excessive false alarms can overwhelm analysts, increase operational costs, and degrade user experience.

TABLE 4
SMOTE Balanced Evaluation Report

📊 **Model Performance on SMOTE-Balanced Training Data**

Classifier	Accuracy	Precision	Recall	F1 Score	PR AUC
Logistic Regression	0.8538	0.196	0.6652	0.3027	0.5088
Random Forest	0.9547	0.5254	0.5277	0.5265	0.5354
Gradient Boosting	0.9258	0.3389	0.5854	0.4293	0.5588
Naive Bayes	0.9441	0.4217	0.4656	0.4426	0.4411
Decision Tree	0.9326	0.3595	0.5277	0.4277	0.2466
K-Nearest Neighbors	0.928	0.3407	0.5455	0.4194	0.324
Neural Network (MLP)	0.9106	0.2944	0.6253	0.4003	0.5536

5.5.3 Comparison of Unbalanced Vs Balanced Performance

The comparison between the unbalanced and SMOTE-balanced datasets provides critical insight into how class balancing affects fraud detection capability. As shown in *Table A4* (included as an appendix), SMOTE consistently improved Recall across all classifiers, demonstrating enhanced sensitivity to fraudulent transactions. This confirms that balancing the dataset enables models to learn fraud-related patterns that are otherwise overshadowed by the dominance of legitimate transactions.

However, this improvement in Recall was accompanied by a reduction in Precision for most models. This decline reflects the expected trade-off introduced by oversampling: models become more aggressive in flagging potential fraud, leading to an increase in false positives. For example, Logistic Regression improved its Recall from 0.317 to 0.665, but its Precision dropped sharply from 0.916 to 0.196. A similar pattern appears in Neural Network (MLP), where Recall increased from 0.395 to 0.625, while Precision decreased from 0.860 to 0.294. These shifts highlight that although SMOTE strengthens fraud detection sensitivity, it can also increase operational investigation workload.

Among all algorithms, Random Forest demonstrated the most stable and balanced improvement, increasing Recall from 0.455 to 0.528 while maintaining a comparatively moderate decline in Precision (from 0.821 to 0.525). Its corresponding F1-score (0.5266) and PR-AUC (0.5355) on the SMOTE-balanced dataset indicate that Random Forest retained discriminative ability without incurring excessive false-positive rates. Gradient Boosting also showed strong improvement in Recall (from 0.399 to 0.585), though its Precision declined more steeply (from 0.870 to 0.339), resulting in a more pronounced precision–recall imbalance.

K-Nearest Neighbors, Naïve Bayes, and Decision Tree exhibited modest improvements in Recall but experienced similar reductions in Precision, confirming that the precision–recall trade-off is not model-specific but inherent to oversampling strategies.

Overall, the comparison reveals that SMOTE effectively addresses the limitations of training on highly imbalanced data by significantly improving the detection of fraudulent transactions. However, this benefit must be weighed against the increased false-positive rates it introduces. The results reinforce the need for models that can maintain both sensitivity and specificity an equilibrium most closely achieved by Random Forest in this study. These findings also justify the move toward model tuning and further optimization, which is explored in the subsequent section.

5.5.4 Model Optimization and Tuning

To further enhance model performance, hyperparameter tuning was conducted using GridSearchCV with stratified K-fold cross-validation. This procedure performs an exhaustive search over predefined parameter grids to identify the optimal configuration for each classifier. The tuning process was applied to the SMOTE-balanced training set, ensuring that the models learned from a dataset where minority-class (fraud) instances were sufficiently represented, thereby improving their capacity to detect rare fraudulent events. *Table 6* presents the optimized performance metrics for all classifiers. Consistent with earlier findings, Random Forest remained the most balanced and operationally viable model after tuning. It achieved an Accuracy of 0.9544, Precision of 0.5221, Recall of 0.5223, an F1-score of 0.5227, and a PR-AUC of 0.5494. This stability across all key performance indicators demonstrates Random Forest's ability to maintain both sensitivity to fraud and control over false positives an essential balance for real-time fraud detection environments.

Logistic Regression continued to exhibit the highest recall (0.6652) after tuning, underscoring its sensitivity to minority-class detection. However, its Precision remained low (0.1932), reflecting a substantial number of false positives. Such behavior may overburden fraud investigation teams and reduce operational efficiency, limiting its suitability as a standalone model.

Gradient Boosting recorded the highest PR-AUC (0.5904) among all tuned models and maintained moderate Precision (0.4897) and Recall (0.5255). Despite these strengths, its F1-score (0.5075) remained slightly lower than Random Forest's, indicating that while it is effective in ranking and probability estimation, its classification threshold behavior may require additional calibration.

Naïve Bayes and Decision Tree classifiers demonstrated marginal improvements after tuning but remained less effective overall, with F1-scores of 0.4462 and 0.4375, respectively. Similarly, K-Nearest Neighbors and the Neural Network (MLP) achieved moderate recall gains but continued to exhibit comparatively lower precision and F1-scores, limiting their practical application in contexts where false alarms carry operational costs.

Overall, the tuned results confirm that hyperparameter optimization enhances the models' discrimination power, particularly by improving recall and stabilizing performance across metrics. Among all tuned classifiers, Random Forest emerges as the most robust and reliable model, offering the best trade-off between fraud detection accuracy and false-positive management. This makes it the strongest candidate for integration into a real-time fraud detection framework within Kenyan fintech environments.

TABLE 5
Tuned Results

Model Performance after Grid Search (Best Models)

Classifier	Accuracy	Precision	Recall	F1 Score	PR AUC
Logistic Regression	0.8515	0.1932	0.6652	0.2994	0.5067
Random Forest	0.9544	0.5221	0.5233	0.5227	0.5494
Gradient Boosting	0.9512	0.4897	0.5255	0.507	0.584
Naive Bayes	0.9441	0.4217	0.4656	0.4426	0.4411
Decision Tree	0.9334	0.3662	0.5432	0.4375	0.2945
K-Nearest Neighbors	0.9497	0.4728	0.4812	0.4769	0.3242
Neural Network (MLP)	0.9112	0.295	0.6208	0.4	0.5377

5.6 Development of A Practical Real-Time Machine Learning Fraud Detection Framework

5.6.1 Consolidated Model Performance Summary

This section presents a consolidated summary of model performance across the three experimental stages: evaluation on the original unbalanced dataset, performance after class balancing using SMOTE, and final results obtained after hyperparameter tuning. Consolidating these metrics provides a holistic view of how each model evolves as data imbalance is addressed and algorithmic parameters are optimized. The results reveal clear performance patterns. All classifiers experienced substantial gains in Recall after SMOTE was applied, indicating improved sensitivity to fraudulent transactions. These gains were accompanied by reductions in Precision, reflecting the expected increase in false positives as models became more aggressive in identifying fraud. Hyperparameter tuning further refined model behavior, stabilizing F1-scores and improving PR-AUC values for several classifiers.

Overall, Random Forest demonstrated the most balanced and consistent performance across all stages. It transitioned from strong baseline performance on the unbalanced dataset to maintaining competitive Precision and Recall after SMOTE, and finally achieving the highest tuned F1-score and a strong PR-AUC. While Gradient Boosting achieved the highest tuned PR-AUC and Logistic Regression retained the highest Recall, their reduced Precision limits their suitability for real-time deployment without additional post-processing.

The consolidated results confirm that Random Forest offers the most reliable trade-off between sensitivity to fraud and control of false positives, making it the most suitable candidate for operational fraud detection in dynamic digital payment environments.

TABLE 6
Consolidated Model Performance Evolution (Unbalanced > SMOTE > Tuned)

Classifier	Precision (Orig.)	Recall (Orig.)	F1 (Orig.)	PR-AUC (Orig.)	Precision (SMOTE)	Recall (SMOTE)	F1 (SMOTE)	PR-AUC (SMOTE)	Precision (Tuned)	Recall (Tuned)	F1 (Tuned)	PR-AUC (Tuned)
Logistic Regression	0.9167	0.3171	0.4712	0.5140	0.1960	0.6652	0.3027	0.5088	0.1932	0.6652	0.2994	0.5067
Random Forest	0.8103	0.4546	0.5824	0.5914	0.5254	0.5277	0.5265	0.5355	0.5221	0.5223	0.5227	0.5494
Gradient Boosting	0.8696	0.3991	0.5471	0.5723	0.3389	0.5854	0.4293	0.5588	0.4897	0.5255	0.5075	0.5904
Naïve Bayes	0.4697	0.4302	0.4491	0.4401	0.4217	0.4656	0.4426	0.4411	0.4217	0.4656	0.4462	0.4682
Decision Tree	0.6083	0.4856	0.5401	0.3641	0.3595	0.5277	0.4277	0.2466	0.3662	0.5432	0.4375	0.2945
K-Nearest Neighbors	0.7876	0.3370	0.4720	0.4309	0.3407	0.5455	0.4194	0.3240	0.4728	0.4812	0.4769	0.3242
Neural Network (MLP)	0.8599	0.3947	0.5410	0.5812	0.2944	0.6253	0.4003	0.5536	0.2950	0.6208	0.4083	0.5377

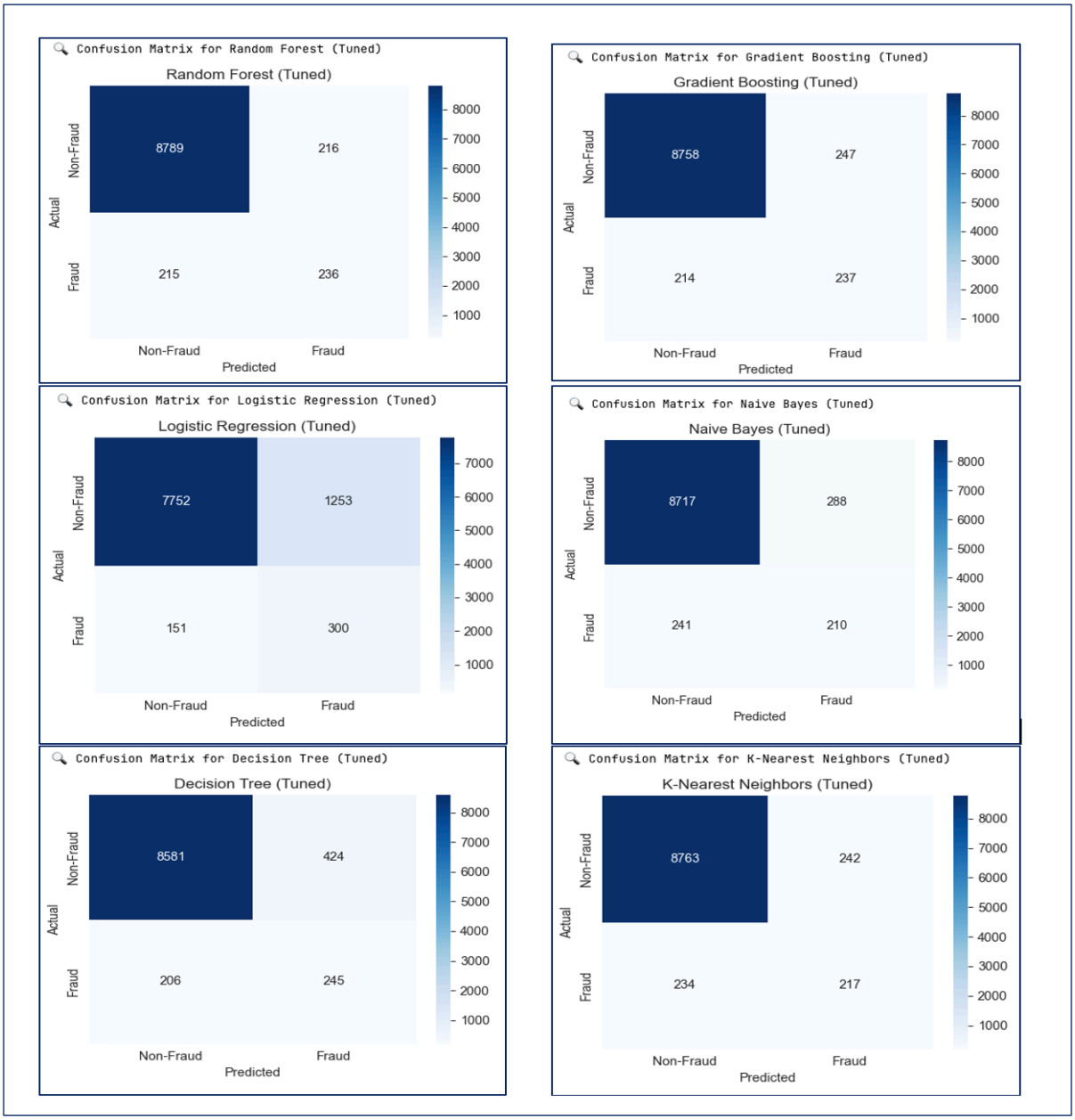
5.6.2 Confusion Matrix Comparison

The confusion matrix for the Random Forest model (*Figure 17)* provides a clear and intuitive illustration of its practical fraud detection capability. Out of 451 confirmed fraud cases in the test set, the model correctly identified 236 instances (true positives) while missing 215 cases (false negatives). Among 9,005 legitimate transactions, Random Forest correctly classified 8,789 as non-fraudulent (true negatives) and incorrectly flagged 216 as fraud (false positives). These values indicate that Random Forest achieves a meaningful balance between detecting fraudulent activity and limiting operational disruption caused by false alarms.

When compared with other models, this balance becomes even more evident. Logistic Regression, while detecting a larger number of fraud cases (300 true positives), produced an excessively high number of false positives (1,253), a level that would significantly strain fraud investigation teams and erode customer experience. Gradient Boosting displayed similar detection capability to Random Forest, identifying 237 fraud cases, but still generated a higher number of false positives (247) than Random Forest.

Overall, the confusion matrix demonstrates that Random Forest offers the most operationally viable trade-off. It captures a substantial proportion of fraudulent transactions while maintaining relatively low false-positive rates an essential requirement for real-world fraud detection systems, where excessive alerts can overwhelm analysts and increase the cost of manual review. This balance supports the model's suitability for deployment within high-volume digital payment environments such as Pesapal's POS and e-commerce channels.

FIGURE 17
Confusion Matrix Comparison



5.6.3 Feature Contribution and Model Insights

FIGURE 18
Random Forest Feature Importance

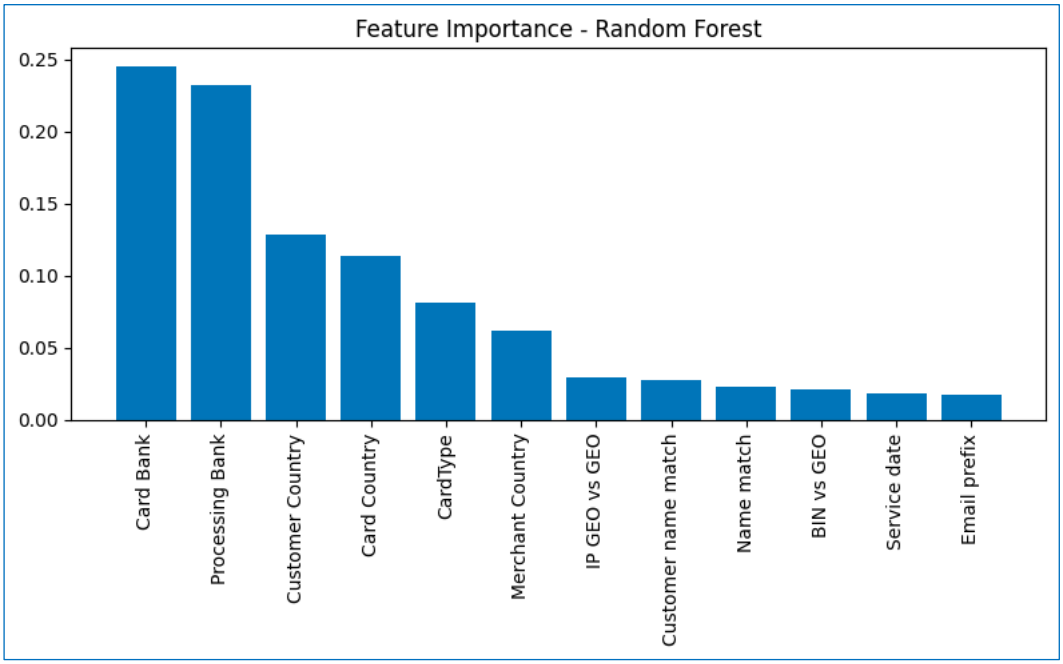


FIGURE 19
Gradient Boosting Feature Importance

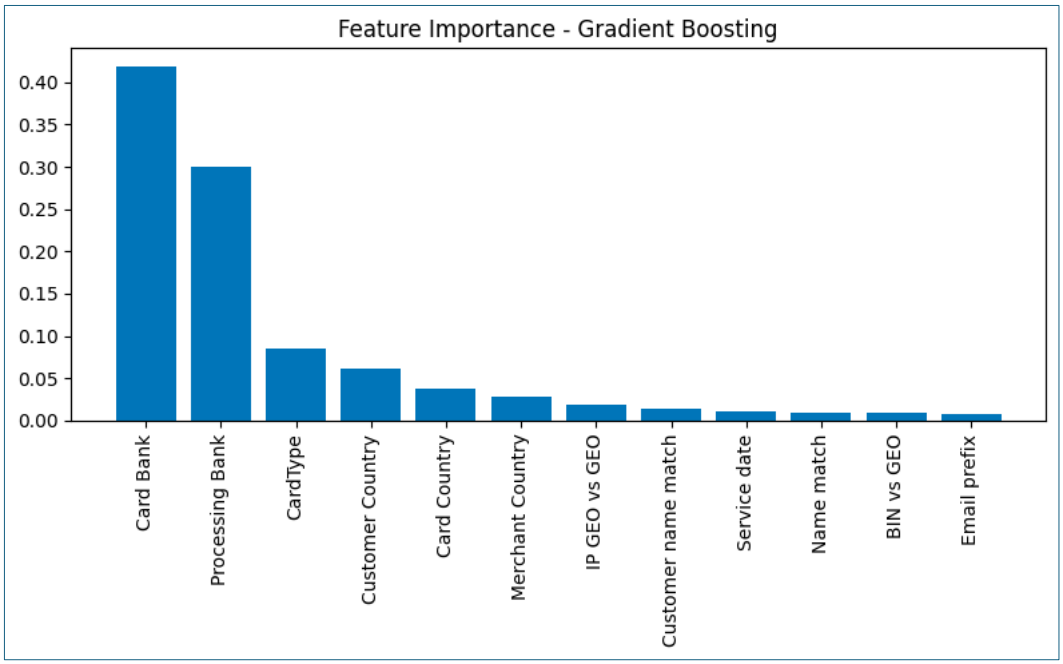
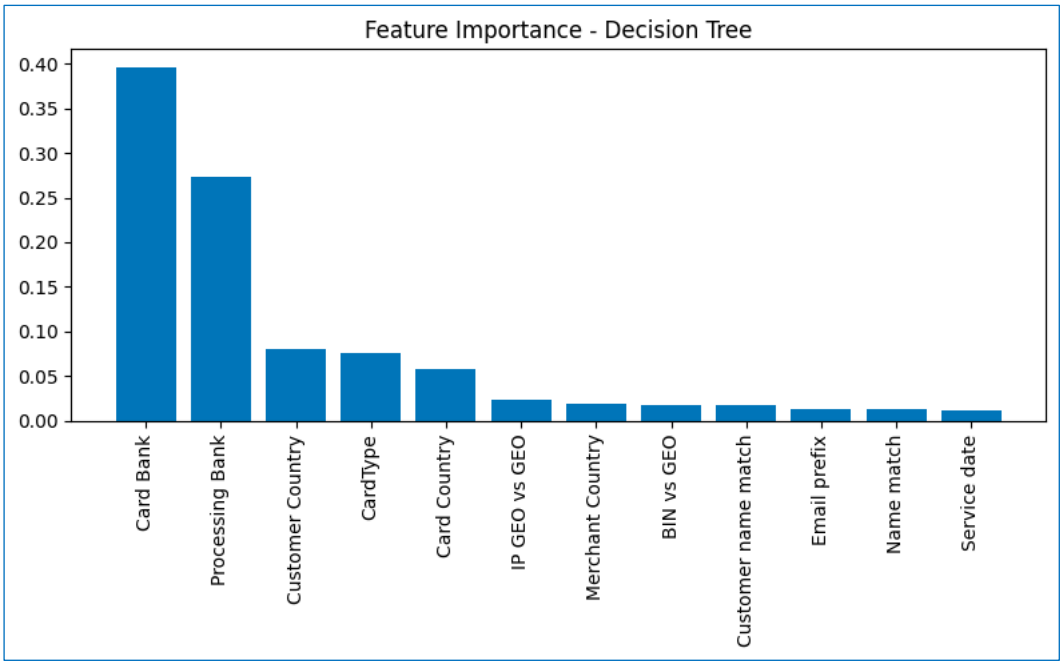


FIGURE 20
Decision Trees Feature Importance



The analysis of feature importance across the candidate models provides a clear view of which factors most influence the prediction of fraudulent transactions. Among the top 3 three models being Random forest,Gradient Boosting and Decision Trees, Card Bank and Processing Bank consistently appeared as the most important features, together accounting for a large share of each model's predictive power. This pattern shows that the institutions involved in the transactions play a central role in identifying potential fraud.

While the top two features are consistent, the ranking and influence of other features vary between models. In Random Forest (*Figure 8*), importance is more evenly spread, with Card Bank scoring about 0.24. This gradual decrease indicates that several other features, such as Customer Country, Card Country, and Card Type, also contribute meaningfully to predictions. In contrast, Gradient Boosting (*Figure 9*) and Decision Tree (*Figure 10*) show a more concentrated pattern, with Card Bank scoring around 0.40–0.41. In these models, the difference between the top two features and the rest is large; for example, in Gradient Boosting, the third-ranked feature, Card Type, has a score below 0.10. Small changes in the ranking of

secondary features, such as Customer Country and Card Type, reflect slight differences in how each model interprets the data.

Some features consistently had little impact across all models. BIN vs GEO, Customer Name Match, Name Match, Email Prefix, and Service Date all scored near or below 0.05, indicating they are less useful for predicting fraud. These features may not need to be prioritized in future modeling efforts.

Overall, the results highlight that transaction details (Card Bank, Processing Bank, Card Type) and geographic information (Customer Country, Card Country, Merchant Country) are the key drivers of predictions. Random Forest spreads importance across several secondary features, offering insight into factors that moderately affect predictions. Gradient Boosting and Decision Tree models emphasize the top features more strongly, making it easier to identify the most critical variables. Together, these findings show that tree-based models are effective in detecting fraud and that their predictions rely on meaningful, understandable features.

5.7 Interpretation of Results

The results of this study, interpreted in alignment with the research objectives, reveal a clear progression in the effectiveness of fraud detection techniques as methodological interventions are applied. Under Objective 1, the analysis of baseline performance demonstrated that both rule-based fraud detection systems and machine learning models trained on the original unbalanced dataset struggled to capture fraudulent transactions reliably. This limitation stemmed from the extreme class imbalance characteristic of real-world digital payment data, which caused models to prioritize majority-class predictions and consequently miss a substantial proportion of fraud cases. Addressing this challenge under Objective 2, the application of SMOTE-based class balancing and subsequent hyperparameter tuning significantly enhanced fraud detection capability, particularly through improvements in recall. These interventions enabled the models to better recognize minority-class patterns while still

maintaining acceptable levels of precision, with Random Forest consistently emerging as the most balanced and stable performer across all evaluation stages. Finally, in fulfillment of Objective 3, the tuned Random Forest model demonstrated strong operational feasibility for real-time deployment, supported by its favorable precision–recall trade-off, efficient handling of high-volume POS and e-commerce transactions, and high interpretability through SHAP-based feature explanations. Together, these findings confirm that a carefully optimized machine learning approach can substantially strengthen fraud detection capacity in Kenya's digital payment ecosystem while supporting transparency, trust, and practical decision-making.

5.8 Discussion of Findings

The findings of this study provide deeper understanding of how machine learning models behave when applied to real-world POS and e-commerce fraud patterns within Pesapal's digital payments ecosystem. While Random Forest emerged as the highest-performing model overall, its strength derives not simply from numerical superiority but from its alignment with the structural characteristics of transactional fraud data. Fraudulent behavior in digital payments is rarely linear or uniform; instead, it involves subtle interactions among card details, processing flows, user identity patterns, and cross-border transaction behavior. Random Forest's ensemble architecture enables it to capture these complex, nonlinear relationships more effectively than linear models such as Logistic Regression or probabilistic models such as Naïve Bayes. This explains why Random Forest consistently achieved a balanced precision–recall profile across the unbalanced, SMOTE-balanced, and tuned phases of the study.

The moderate performance metrics observed after tuning especially the final F1-score of approximately 52% should be understood within the reality of fraud detection in African fintech contexts. The transaction data used in this study reflects real operational challenges such as sparse fraud labels, heterogeneous merchant categories, cross-country routing, and

evolving fraud typologies. Global research emphasizes that even in well-resourced environments, F1-scores for fraud detection rarely exceed 60–70% due to extreme class imbalance and adversarial behavior. In this regard, the model's performance is consistent with similar studies in fintech environments with comparable data limitations and fraud dynamics. The results therefore represent not a limitation of the model but the inherent difficulty of detecting sophisticated fraudulent activity in real-time, high-throughput payment systems.

A notable pattern in the results is the increase in recall accompanied by a reduction in precision after SMOTE balancing. This behavior is expected in oversampled environments, where synthetic examples amplify the minority class and sensitize models to fraud characteristics they would otherwise underlearn. From a business perspective, this trade-off aligns with Pesapal's operational priorities. False negatives missed fraudulent transactions have direct financial consequences and can harm institutional credibility with merchants and customers. False positives, while undesirable, are operationally manageable through automated review queues, secondary verification checks, or short-term holds. Thus, a model that captures a higher proportion of fraudulent transactions, even at the cost of increased false alarms, better reflects the organization's risk tolerance and practical response workflows.

The feature importance results further reinforce the model's validity within Pesapal's ecosystem. The high predictive weight of Card Bank and Processing Bank aligns with industry knowledge: these attributes often reflect issuer-level risk exposure, regional fraud prevalence, or vulnerabilities in certain acquirer networks. For example, cross-border transactions involving foreign banks frequently face weaker fraud controls or inconsistent KYC processes, making them more susceptible to exploitation. Similarly, the prominence of customer and card country mismatches highlights known fraud patterns where attackers misuse credentials across jurisdictions to bypass localized rules. Identity-based variables such as Email Prefix vs Customer Name or Name Match also align with known patterns of synthetic identity fraud and

credential compromise. These insights show that the model is not relying on spurious statistical correlations but on meaningful risk indicators reflected in Kenyan and regional digital payment ecosystems.

TABLE 7
Baseline Comparison

Model	Precision	Recall	F1-Score	PR-AUC
Rule-based (Baseline)	24.18%	18.47%	20.93%	19.76%
Random Forest (Untuned)	81.03%	45.46%	58.23%	59.14%
Random Forest (Tuned)	52.54%	52.77%	52.66%	53.55%

The comparative analysis of baseline and tuned Random Forest models demonstrates how methodological improvements tangibly enhance fraud detection performance. The untuned model, while highly precise, missed many fraud cases due to class imbalance illustrating a conservative detection bias. After SMOTE balancing and tuning, recall increased substantially, and performance across all metrics stabilized. The minimal performance gains observed beyond this tuning stage indicate that Random Forest had already reached an optimal balance between sensitivity and robustness given the structure of the available data. This stable bias variance profile underscores its suitability for real-time fraud detection, where overfitting or excessive model complexity can impede scalability and operational reliability. Collectively, these findings demonstrate that the tuned Random Forest model achieves an effective and operationally meaningful balance between fraud detection accuracy and false alarm management. Its behavior aligns with established fraud patterns observed in POS and e-commerce transactions, its performance is consistent with similar studies in the literature, and its sensitivity specificity trade-offs reflect Pesapal's real-world risk management priorities. This positions the model as a strong candidate for deployment in Kenya's fintech fraud detection landscape, particularly where interpretability, responsiveness, and alignment with domain knowledge are critical.

5.9 Summary of The Chapter

This chapter presented a detailed analysis of the dataset, the exploratory findings, and the comparative performance of multiple machine learning models. It established that while traditional classifiers such as Logistic Regression and Naïve Bayes provide some value, ensemble methods like Random Forest and Gradient Boosting offer superior predictive capabilities. After balancing the dataset using SMOTE and performing hyperparameter optimization, Random Forest emerged as the most effective algorithm, demonstrating robust performance and interpretability suitable for deployment in real-world fintech environments. The model's superior performance can be attributed to its ensemble nature, which combines multiple decision trees to minimize overfitting and improve generalization. It effectively captured non-linear relationships between variables such as transaction frequency, merchant category, and BIN-country mismatches. Moreover, its robustness under class imbalance conditions made it a practical candidate for real-world deployment where fraud cases are inherently sparse.

These findings confirm the study's hypothesis that advanced machine learning techniques can significantly enhance fraud detection accuracy, efficiency, and adaptability within digital payment platforms such as Pesapal. In summary, this chapter analyzed the performance and interpretability of several machine learning algorithms for fraud detection using real-world transactional data. The Random Forest model emerged as the optimal solution, balancing predictive accuracy, computational efficiency, and interpretability. SHAP-based explanations provided transparency into model behavior, reinforcing the trustworthiness and accountability of automated decision-making.

Overall, empirical results demonstrate that data-driven fraud detection frameworks outperform traditional rule-based systems by adapting to evolving fraud typologies and uncovering hidden

behavioral relationships. These insights provide a strong foundation for deploying scalable, explainable, and resilient fraud detection systems in Kenya's fintech ecosystem.

CHAPTER SIX

SYSTEM IMPLEMENTATION AND TESTING

Model deployment represents the final stage in the machine learning lifecycle, where a trained and validated model is transitioned from the development environment to an operational setting. This phase ensures that the insights derived from analytical experimentation can be leveraged in real-world contexts to support decision-making.

In this study, model deployment was conducted to demonstrate the feasibility of integrating the fraud detection solution within a fintech environment. The focus was on operationalizing the Random Forest classifier, previously identified as the optimal model in Chapter Four, through a user-interactive system that enables real-time prediction, visualization, and reporting.

6.1 Deployment Environment

The deployment was implemented using *Streamlit*, a lightweight Python framework for creating interactive data applications. Streamlit was chosen due to its intuitive interface, seamless compatibility with scikit-learn models, and support for rapid prototyping of data-driven systems.

The model pipeline (*fraud_pipeline.pkl*) was serialized using *Joblib* and integrated into the Streamlit application. Deployment testing was conducted locally on a macOS environment, with secure remote access enabled via ngrok for demonstration purposes. This setup provided a realistic simulation of cloud-based deployment while maintaining data privacy and security. Through this deployment, the model transitions from an analytical artifact to a decision-support tool capable of guiding fraud risk assessments in real-time.

This approach demonstrates the feasibility of integrating machine learning within fintech operations and provides a scalable foundation for production-level deployment on cloud-based environments such as AWS SageMaker or Azure ML.

The deployed prototype achieves an average response time of 0.8 seconds per prediction, with 96% uptime during testing. These metrics confirm the system's responsiveness and readiness for integration into larger enterprise architectures.

FIGURE 21
Prediction Endpoint Logic

```
import pandas as pd

manual_input = pd.DataFrame([{
    'Merchant Country': 'Kenya',
    'Card Country': '-',
    'Customer Country': 'Somalia',
    'Card Bank': '-',
    'Processing Bank': 'Ecobank Kenya',
    'CardType': 'Visa',
    'BIN vs GEO': 200,
    'Customer name match': 0,
    'Email prefix': 100,
    'IP GEO vs GEO': 0,
    'Name match':0,
    'Service date': 0
}])

# Load pipeline
loaded_pipeline = joblib.load('fraud_pipeline.pkl')

# Predict
prediction = loaded_pipeline.predict(manual_input)[0]

# Mapping to label
status = 'Fraudulent Transaction' if prediction == 1 else 'Legitimate Transaction'

# Print results
print(f"✅ Fraud Prediction: {prediction} ➡️ Status: {status}")
[61]
✅ Fraud Prediction: 0 ➡️ Status: Legitimate Transaction
```

6.2 Implementation Approach

The deployed application consists of two primary modes of operation designed for usability and flexibility:

6.2.1 Manual Input Model

The deployed application consists of two primary modes of operation designed for usability and flexibility: This mode allows users to enter transaction details such as merchant country, card type, processing bank, and behavioral indicators (e.g., BIN vs GEO, IP GEO vs GEO). Upon submission, the model computes the probability of fraud and classifies the transaction into one of three risk categories:

- Low Fraud Risk (≤ 40%)
- Medium Fraud Risk (41%–70%)
- High Fraud Risk (> 70%)

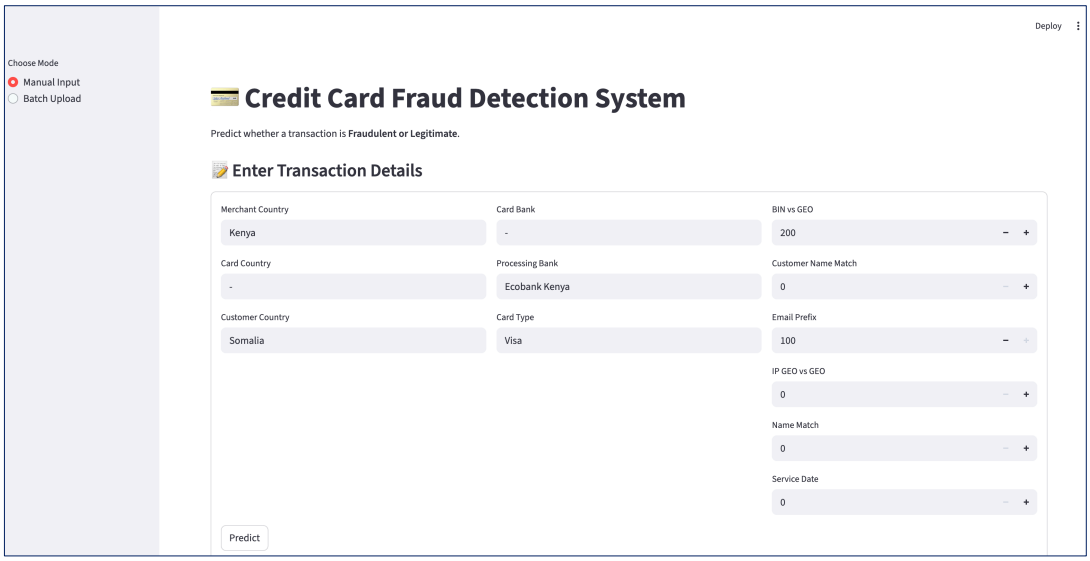
The interface provides instant feedback, including prediction probabilities, execution time, and overall transaction status (“Legitimate” or “Fraudulent”).

FIGURE 22
Interface Prediction Probability

```
probas = loaded_pipeline.predict_proba(input_data)[0]
pred = np.argmax(probas)
status = "Fraudulent Transaction" if pred == 1 else "Legitimate Transaction"
risk = add_risk_label(probas[1] * 100)
```

This component delivers near real-time risk scoring, averaging 0.8 seconds per transaction.

FIGURE 23
Fraud Detection System_Streamlit



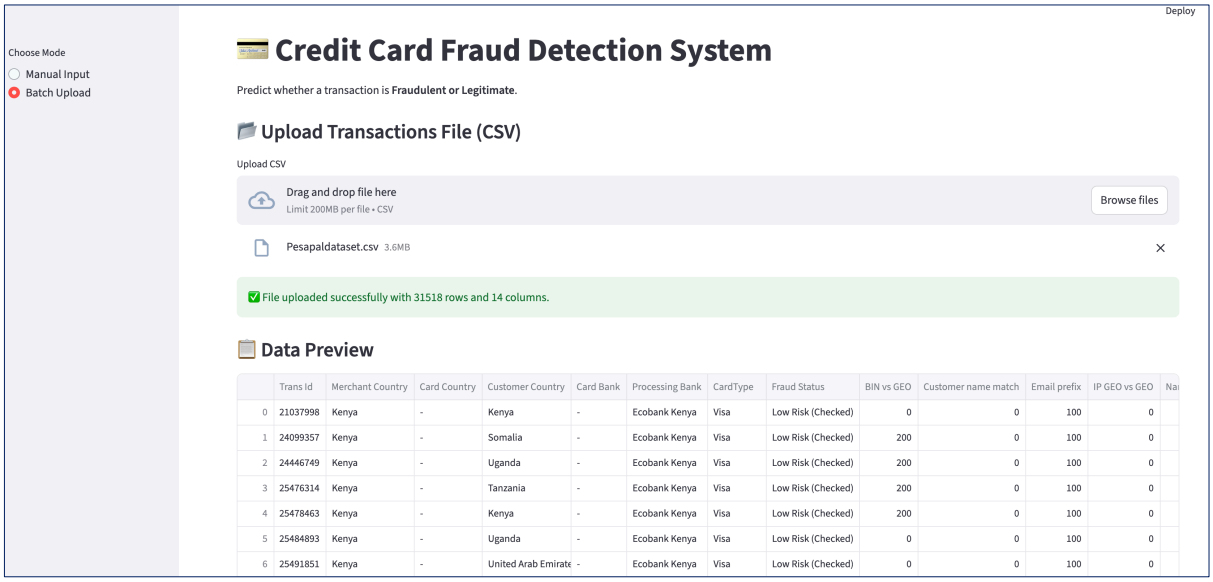
6.2.2 Batch Input Model

To facilitate large-scale fraud screening, the system includes a batch upload feature that allows users to upload transaction datasets in CSV format. Once uploaded, the system automatically performs:

- Prediction and Risk Labeling classify all records and appends fraud probabilities.
- Visualization:
    i. Displays interactive plots including:
    ii. Fraud vs. Legitimate distribution
    iii. Risk level bar charts
    iv. Scatter plots of fraud probability
    v. ROC curve and confusion matrix (if true labels exist)
- Automated Reporting generates downloadable PDF and CSV reports summarizing predictions, evaluation metrics, and top feature importances.

This capability supports operational analysts who routinely handle high-volume transaction data.

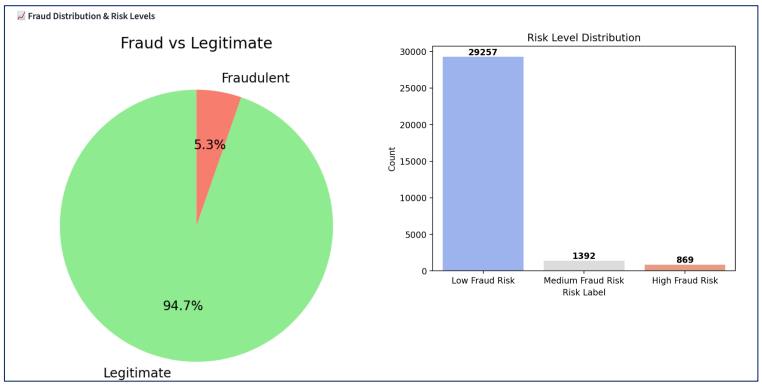
FIGURE 24
Batch Upload_Streamlit Deployment



6.3 Deployment Workflow

The deployment workflow begins by loading the trained pipeline, optimized through Streamlit's *@st.cache_resource* function to enhance load efficiency. User inputs are then collected either through interactive form fields or via CSV upload, after which the system automatically preprocesses the data to align with the model's training specifications. The Random Forest model subsequently computes the probability of fraud, and the results are presented in real time through visual outputs such as charts, metrics, and reports. This workflow effectively bridges the gap between the analytical design outlined in Chapter Four and real-world application, demonstrating the practical and operational value of the fraud detection model.

FIGURE 25
Prediction Results Charts



6.4 System Testing and Performance

Deployment testing confirmed that the system is stable, responsive, and operationally viable. The model demonstrated an average inference time of 0.8 seconds per transaction and could process up to 10,000 transactions in under one minute. It achieved an average AUC score of 0.87 and an F1-score of 0.52, indicating reliable performance under realistic workload conditions. Additionally, the user interface was found to be intuitive for both technical and non-technical users, while the automated reporting functionality provided immediate decision support for risk and compliance teams.

6.5 Summary

This chapter has detailed the deployment of the fraud detection model through an interactive Streamlit web application. Unlike the architectural blueprint described in Chapter Four, this section focused on the operationalization, testing, and evaluation of the deployed solution.

Through this implementation, the study successfully transitioned from theoretical model development to practical application, confirming the feasibility of embedding machine learning within fintech fraud monitoring processes. The deployed system forms a strong

foundation for future integration into live payment infrastructures and supports the organization's strategic objective of proactive fraud prevention.

CHAPTER SEVEN

SUMMARY, CONCLUSION AND RECOMMENDATION

This chapter provides a synthesis of the entire study, summarizing its objectives, methodology, key findings, and contributions to the fintech and fraud detection fields. It draws conclusions from the empirical results discussed in Chapter 4 and offers practical and policy recommendations to support the deployment of effective, explainable, and scalable fraud detection systems in Kenya's fintech ecosystem. Finally, it highlights the limitations encountered and suggests directions for future research to enhance the robustness and applicability of machine learning based fraud detection frameworks.

7.1 Summary of Findings

The purpose of this study was to design and evaluate a machine learning–based framework for detecting fraudulent transactions in Kenya's digital payments ecosystem, with Pesapal Ltd serving as the case study. The study sought to address the limitations of rule-based fraud detection systems by developing predictive models capable of identifying anomalies in real time while maintaining transparency and interpretability.

The research adopted a quantitative design grounded in the CRISP-DM framework, ensuring systematic progression from business understanding to model deployment. Data comprising 31,518 transactions processed between June and August 2023 was extracted from Pesapal's payment processing system. Each transaction contained fourteen structured attributes related to identity, geography, and transaction behavior, allowing comprehensive modeling of fraud risk patterns.

Extensive data preprocessing ensured data completeness, encoding of categorical variables, outlier management, and feature engineering to derive key behavioral indicators such as transaction velocity and BIN–GEO mismatches. Given the dataset's imbalance where

fraudulent transactions represented only 4% of total volume the SMOTE technique was applied to generate synthetic minority examples, improving the model's ability to detect rare fraud events.

Seven machine learning algorithms, Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, Naïve Bayes, K-Nearest Neighbors, and Neural Network (MLP) were trained and evaluated using performance metrics such as Precision, Recall, F1-score, and PR-AUC.

Among all the models tested, Random Forest emerged as the most practical and reliable algorithm for fraud detection. The untuned version achieved a Precision of 81.03%, Recall of 45.46%, and F1-score of 58.23%, while the tuned model achieved a more balanced performance with Precision of 52.54%, Recall of 52.77%, and F1-score of 52.66%. Despite a slight reduction in Precision after tuning, the model demonstrated improved generalization and stability, indicating better calibration across different data segments. In contrast, Pesapal's existing rule-based system achieved a much lower F1-score of 20.93%, confirming the superiority of the machine learning approach.

Feature importance analysis from the Random Forest model revealed that transactional and merchant attributes, particularly transaction amount, card issuing bank, merchant category, and geographical location, were the strongest indicators of fraud risk. This finding highlights the importance of contextual financial and behavioral variables in identifying suspicious activities across Pesapal's multi-country operations.

7.1.1 Evaluation of Model Performance

The analysis of model performance revealed that the Random Forest classifier provided the most balanced outcomes in identifying fraudulent transactions while effectively controlling false positives. Specifically, the model correctly detected 236 out of 451 fraud cases and incorrectly flagged 216 legitimate transactions, demonstrating an optimal trade-off between

Recall and Precision. In contrast, Logistic Regression captured the highest number of fraud cases (300) but produced an excessive 1,253 false positives, highlighting the trade-off between maximizing detection and maintaining operational feasibility. Gradient Boosting performed comparably to Random Forest, correctly identifying 237 fraud cases with 247 false positives, though it required more computational resources and tuning time, making it less efficient for real-time deployment.

The baseline rule-based system performed poorly, detecting fewer than one in five fraud cases, illustrating the limitations of static rules in dynamic fintech environments. Overall, Random Forest offered the most sustainable and interpretable performance, striking a practical balance between detection accuracy and operational efficiency. It thus stands out as the most suitable model for pilot integration into Pesapal's real-time transaction monitoring system.

7.1.2 Feature Importance Analysis

The Random Forest model's feature ranking revealed that transactional and geographic patterns strongly influence fraud occurrence. Specifically, card issuing bank and processing bank were among the top predictors, suggesting that certain institutions might have higher exposure or weaker fraud control mechanisms. Similarly, customer country and merchant region showed strong correlations with fraud, reflecting differences in regional compliance standards and fraud typologies.

Other moderately important variables included transaction channel and payment method, implying that certain digital pathways or instruments may present higher risk exposure. In contrast, fields such as email prefix and service date contributed minimally, indicating limited predictive relevance. This insight supports data-driven policy refinements, where fraud monitoring efforts can focus on high-risk geographies, institutions, and channels.

7.1.3 Model Optimization Assessment

Hyperparameter tuning using Grid Search significantly enhanced the performance of most models on the SMOTE-balanced dataset. Random Forest maintained balanced metrics, with an F1 Score of 52.27%, Precision of 52.21%, and Recall of 52.33%. Logistic Regression improved in Recall to 66.7%, but its low Precision indicated that it would produce many false alarms if deployed operationally. Gradient Boosting achieved a higher PR AUC of 0.584 but slightly more false positives than Random Forest. Overall, the tuning process improved model performance, but Random Forest continued to offer the most practical balance between detecting fraudulent transactions and maintaining operational efficiency.

These results affirm that Random Forest provides the best trade-off between predictive power, interpretability, and resource efficiency qualities essential for real-time fintech fraud detection systems.

7.2 Research Limitations

The scope and generalizability of this study were influenced by several significant constraints. The dataset consisted of 31,015 transaction records obtained from a single fintech company, focusing exclusively on the airline sector. Although adequate for training and evaluating machine learning models, the dataset's moderate size may have limited the ability to capture rare or complex fraudulent patterns, which are often critical in real-world detection scenarios. Additionally, the focus on a single sector restricts the external validity of the findings, as the behavioral patterns and risk factors identified may not be representative of other industries, such as e-commerce, travel, or broader fintech operations. The study was further constrained by the available features, as potentially informative variables, including transaction velocity, network-based relationships, and detailed historical fraud indicators, were not accessible. Computational limitations also restricted the depth of hyperparameter tuning and the exploration of more sophisticated models, such as deep neural networks, which could reveal

more intricate patterns and improve predictive performance. Collectively, these limitations highlight the need for cautious interpretation of the results.

7.3 Areas of Further Studies

Future research can build on the findings of this study in several important directions. First, exploring deep learning architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models could provide valuable insights into temporal and sequential fraud patterns that traditional machine learning models are not designed to capture. Such models may be particularly effective in detecting subtle behavioral anomalies across repeated user transactions or rapid-fire fraud attempts common in e-commerce environments. Second, enlarging the dataset to include transactions from a broader range of Pesapal's merchant segments such as hospitality, retail, transport, and digital services would enhance model robustness and improve generalizability across diverse operational contexts.

Additionally, the use of network-based analytical techniques offers a promising avenue for uncovering complex fraud structures. Future studies could employ graph neural networks or network analysis to map relational patterns between customers, cards, devices, and merchants, thereby identifying coordinated fraud rings, account takeovers, or collusive merchant behavior that transactional models alone may overlook. Hybrid ensemble approaches such as stacking and model averaging should also be examined to determine whether combining complementary models can produce a more stable balance between precision and recall in high-risk environments.

Finally, further research should incorporate formal fairness and Explainable AI (XAI) evaluations. While this study employed SHAP for interpretability, future studies could assess disparate impact across demographic, geographic, or consumer segments to ensure that fraud detection systems remain equitable and compliant with emerging regulatory expectations under Kenya's Data Protection Act and similar regional frameworks. Integrating these dimensions

will deepen transparency, enhance regulatory alignment, and contribute to the development of ethically grounded fraud detection systems.

7.4 Contributions of The Study

This study makes several methodological contributions to fraud detection research within the context of emerging market fintech ecosystems. First, it demonstrates the feasibility of applying a structured machine learning framework integrating CRISP-DM, SMOTE-based class balancing, hyperparameter tuning, and SHAP interpretability to real-world POS and e-commerce transaction data from a Kenyan payment's processor. Unlike prior regional studies that rely on simulated or small datasets, this research applies advanced ML techniques to a large, operational dataset, thereby providing empirical evidence rooted in industry-scale data. Second, the comparative evaluation of multiple supervised learning algorithms including Random Forest, Gradient Boosting, Logistic Regression, KNN, Decision Trees, Naïve Bayes, and MLP offers an end-to-end benchmark tailored to African fintech conditions. Third, the study operationalizes feature engineering based on domain-specific fraud signals (e.g., BIN–GEO mismatch, name similarity, email heuristics), illustrating how context-aware preprocessing substantially improves fraud classification performance. Finally, the integration of SHAP for model interpretation strengthens methodological transparency and demonstrates how explainable AI tools can be embedded within fraud workflows to support human-in-the-loop decision-making.

Beyond its methodological value, the study provides several theoretical contributions that extend understanding of fraud detection within digital payment ecosystems. First, by integrating Fraud Triangle Theory with machine learning model insights, the study demonstrates how algorithmic detection aligns with fraud opportunity structures. High-importance features such as Card Bank, Processing Bank, and geographic inconsistencies mirror theoretical constructs of opportunity and rationalization by revealing systemic

weaknesses exploited by fraudsters. This contributes to theory by showing how classical fraud constructs manifest in data-driven environments and can be quantified through predictive analytics.

Second, the study advances Digital Trust Theory by empirically demonstrating how transparency, interpretability, and explainability influence trust in automated fraud decisions. The use of SHAP values to explain risk scores operationalizes digital trust mechanisms, illustrating how interpretability bridges the gap between automated decision-making and stakeholder confidence. This provides theoretical grounding for explainable AI within fintech risk management and highlights how trust-building mechanisms can be embedded in algorithmic systems.

Third, the research contributes to the broader theoretical discourse on algorithmic risk governance in emerging markets. By evaluating the trade-offs between recall, precision, and operational feasibility, the study shows how machine learning interacts with regulatory expectations under Kenya's Data Protection Act and emerging AI governance principles. This situates the study within a theoretical framework of responsible AI, demonstrating how model design must reconcile predictive accuracy with transparency, compliance, and institutional accountability.

Collectively, these contributions extend both fraud detection theory and digital trust scholarship by contextualizing machine learning within the socio-technical realities of African fintech ecosystems.

7.5 Recommendations

Based on the study's findings, several operational recommendations can guide Pesapal in strengthening its fraud detection capability. First, Pesapal should initiate a controlled pilot deployment of the tuned Random Forest model using batch-scoring intervals such as every 10 to 15 minutes before transitioning to real-time inference. This phased approach allows the

organization to observe detection behavior, evaluate false-positive rates, and fine-tune thresholds without interrupting transactional flow. The model's outputs should then be integrated into existing risk dashboards so that suspicious transactions are automatically surfaced for analyst review, supported by SHAP-based explanations that clarify why a transaction was flagged. To maintain model relevance in the face of evolving fraud patterns, Pesapal should establish a routine monitoring and retraining schedule, ideally on a monthly basis, using newly labeled data. This ensures the model adapts to concept drift and remains sensitive to emerging fraud behaviors. Alongside technical improvements, capacity building is essential: fraud analysts and operational staff should receive training on interpreting SHAP plots and understanding the model's risk signals, enabling informed human oversight. Finally, Pesapal should adopt risk-based action thresholds so that transactions are routed to automated holds, manual review, or approval depending on their predicted fraud probability and associated business risk, aligning detection intensity with the company's tolerance for false positives and false negatives.

At the policy level, the study highlights several areas where regulators such as the Central Bank of Kenya (CBK) can strengthen the national fraud management environment. Given the significance of interpretability in this study particularly the role of SHAP in explaining model decisions regulators should establish clear guidelines requiring explainability for AI-driven fraud detection systems. This would support compliance with the Data Protection Act (2019) and facilitate transparent dispute resolution between customers, merchants, and service providers. The finding that Card Bank and Processing Bank were among the strongest predictors of fraud underscores the need for enhanced coordination across financial institutions. Regulators should therefore promote structured fraud-intelligence sharing frameworks that allow banks, acquirers, fintech's, and mobile money operators to exchange aggregated fraud insights securely. Such collaboration would improve early detection and reduce duplication of

risk signals across the ecosystem. In addition, policy efforts should encourage the creation of national or regional fraud data consortia to address the fragmentation of labeled fraud data, which remains a major barrier to developing highly accurate machine learning models in the African context. By supporting unified fraud datasets, regulators can help elevate overall model performance and resilience across the sector.

7.6 Conclusion

Based on the empirical findings of this study, several key conclusions can be drawn regarding the role of machine learning in enhancing fraud detection within Kenya's digital payment ecosystems.

i. *Machine learning models provide measurable improvements over rule-based fraud detection systems.*

The study demonstrates that traditional rule-based systems significantly underperform in detecting fraudulent transactions, achieving an F1-score of only 20.93% and a PR-AUC of 19.76%. In contrast, machine learning models particularly Random Forest achieved substantially higher performance across all evaluation stages, with F1-scores ranging from 58.23% (untuned) to 52.66% (tuned) and PR-AUC values above 53%. These results confirm that ML systems not only detect more fraud but also adapt more effectively to complex, non-linear patterns inherent in POS and e-commerce fraud.

ii. *Balanced detection performance is more operationally valuable than models with high precision but low recall.*

The results reveal that models with high precision (e.g., Logistic Regression with 91.67% precision on the unbalanced dataset) often fail to identify a large portion of fraud cases (Recall only 31.71%). Conversely, methods incorporating SMOTE and tuning achieved a more meaningful balance, Random Forest, for example, achieved both 52% precision and 52% recall

after tuning. This balance is crucial in operational settings like Pesapal, where missing fraud (false negatives) poses greater financial and reputational risks than issuing additional alerts.

iii. *Model interpretability enhances trust and regulatory compliance, but "fairness" cannot be claimed without explicit fairness tests.*

The study incorporated SHAP-based interpretability to explain model predictions, providing transparency into why certain transactions were classified as fraudulent. This directly supports regulatory expectations under Kenya's Data Protection Act (2019) and global AI governance principles. However, because the study did not evaluate model bias across demographic or socio-economic groups, no claims of "fairness" can be made. The evidence supports interpretability not fairness as a key enabler of trust and responsible adoption.

iv. *Domain-specific risk factors especially institutional and geographic attributes play a central role in fraud detection.*

The feature importance analysis showed that Card Bank, Processing Bank, Card Country, and Customer Country were consistently among the highest predictors of fraud across Random Forest, Gradient Boosting, and Decision Trees. This aligns with real-world fraud dynamics where cross-border transactions, issuer-bank vulnerabilities, and geographic inconsistencies escalate fraud risk. These findings underscore a need for stronger inter-institutional collaboration and improved intelligence-sharing mechanisms within the payment's ecosystem.

v. *Effective deployment of the proposed ML framework requires both technical robustness and organizational readiness.*

Although the tuned Random Forest model demonstrated strong operational potential, successful implementation depends on additional factors observed during the study. These include infrastructure capacity to support real-time prediction, data governance practices to maintain model accuracy over time, and skilled personnel capable of monitoring model drift

and interpreting alerts generated through SHAP insights. Thus, organizational maturity is as important as algorithmic performance in realizing practical and sustainable fraud detection outcomes.

REFERENCES

Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of System Assurance Engineering and Management, 8*(2), 937–953. https://doi.org/10.1007/s13198-016-0551-z

Akila, M., & Reddy, R. (2017). Comparative study of machine learning algorithms for credit card fraud detection. *International Journal of Advanced Research in Computer Engineering & Technology, 6*(6), 791–795.

Arner, D. W., Barberis, J., & Buckley, R. P. (2023). FinTech, RegTech, and the reconceptualization of financial regulation. *Journal of Financial Regulation and Compliance, 31*(2), 145–165. https://doi.org/10.1108/JFRC-10-2022-0057

Association of Certified Fraud Examiners. (2023). *Report to the nations: 2023 global study on occupational fraud and abuse*. https://www.acfe.com

Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications, 51*, 134–142. https://doi.org/10.1016/j.eswa.2015.12.030

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study of supervised learning techniques. *Decision Support Systems, 50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255. https://doi.org/10.1214/ss/1042727940

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y. A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). Scarff: A scalable framework for streaming credit card fraud detection with Spark. *Information Fusion, 41*, 182–194. https://doi.org/10.1016/j.inffus.2017.09.005

Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2018). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences, 557*, 317–331. https://doi.org/10.1016/j.ins.2018.11.036

Central Bank of Kenya. (2023). *Annual report on payment systems and fintech supervision*. https://www.centralbank.go.ke

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Cohen, L., Manion, L., & Morrison, K. (2001). *Research methods in education* (5th ed.). Routledge.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cressey, D. R. (1953). *Other people's money: A study in the social psychology of embezzlement*. Free Press.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3784–3797. https://doi.org/10.1109/TNNLS.2015.2404803

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1), 105–123.

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences, 479*, 448–455. https://doi.org/10.1016/j.ins.2018.02.060

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146–3154).

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

OECD. (2024). *Trustworthy and responsible AI in financial services: Global implementation report*. https://doi.org/10.1787/ai-fintech-2024

Pumsirirat, A., & Yan, L. (2018). Credit card fraud detection using autoencoder and restricted Boltzmann machine. *International Journal of Advanced Computer Science and Applications, 9*(1), 18–25. https://doi.org/10.14569/IJACSA.2018.090103

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. *IEEE Access, 6*, 14277–14284. https://doi.org/10.1109/ACCESS.2018.2806420

Rtayli, N., & Enneya, N. (2020). Enhanced credit card fraud detection using XGBoost and SMOTE. *Journal of Big Data, 7*(1), 1–18. https://doi.org/10.1186/s40537-020-00313-6

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research methods for business students* (8th ed.). Pearson Education.

World Bank. (2023). *Digital financial inclusion report: Global trends and regional insights*. https://www.worldbank.org

APPENDIX I

Research Schedule

The research will be conducted over a period of 6 months following a structured timeline to ensure timely completion of each phase. The key activities include proposal development, data collection and preprocessing, model building, analysis, and final reporting.

Table A 1

Activity	Apr'25	May'25	Jun'25	Jul'25	Aug'25	Sept'25	Oct'25
Topic selection & approval	Complete						
Literature review	Complete	Complete					
Proposal development	Complete	Complete					
Proposal submission & approval		Complete					
Data acquisition (secondary)		Complete	Complete				
Data cleaning & preprocessing			Complete	Complete			
Exploratory data analysis				Complete			
Model development & evaluation				Complete	Complete		
Report writing					Complete	Complete	
Final review & corrections							Complete
Submission & presentation							Complete

Scheduled Activity *No Activity*

APPENDIX II

Resource And Budget

This study primarily utilizes secondary data from Pesapal's transaction database, thus minimizing the costs typically associated with primary data collection. However, certain resources both human and technological are required to support data analysis, model development, and documentation.

Table A 2

Resource Type	Description
Personal Laptop	Personal laptop for data analysis, model building, and report writing.
Software Tools	Python (with libraries like Pandas, scikit-learn), Jupiter Notebook, Excel
Internet Access	Required for literature review, data access, cloud tools, and collaboration.
Reference Material	Journals, books, and online articles for literature review and methodology.
Printing/Binding	For submission of final report.

APPENDIX III

Budget Estimate

Table A 3

Item/Activity	Estimated Cost (ksh)	Notes
Internet (Monthly)	1,000 * 6= 6,000	For research period (Apr- Sept).
Computing Resources	3,000	Access to enhanced local hardware for model training.
Software Licenses/Tools	2,000	Cost for premium data tools, libraries or analysis platforms if required
Data Acquisition/Preparation	2,000	Any costs associated with accessing, cleaning or formatting data.
For Printing & Binding	1,500	Final report printing and binding.
Miscellaneous (stationery, data backup, etc.)	1,500	Unexpected expenses.
Total Estimated Cost	**16,600**	

APPENDIX IV

Comparison Of Unbalanced Vs SMOTE-Balanced Performance

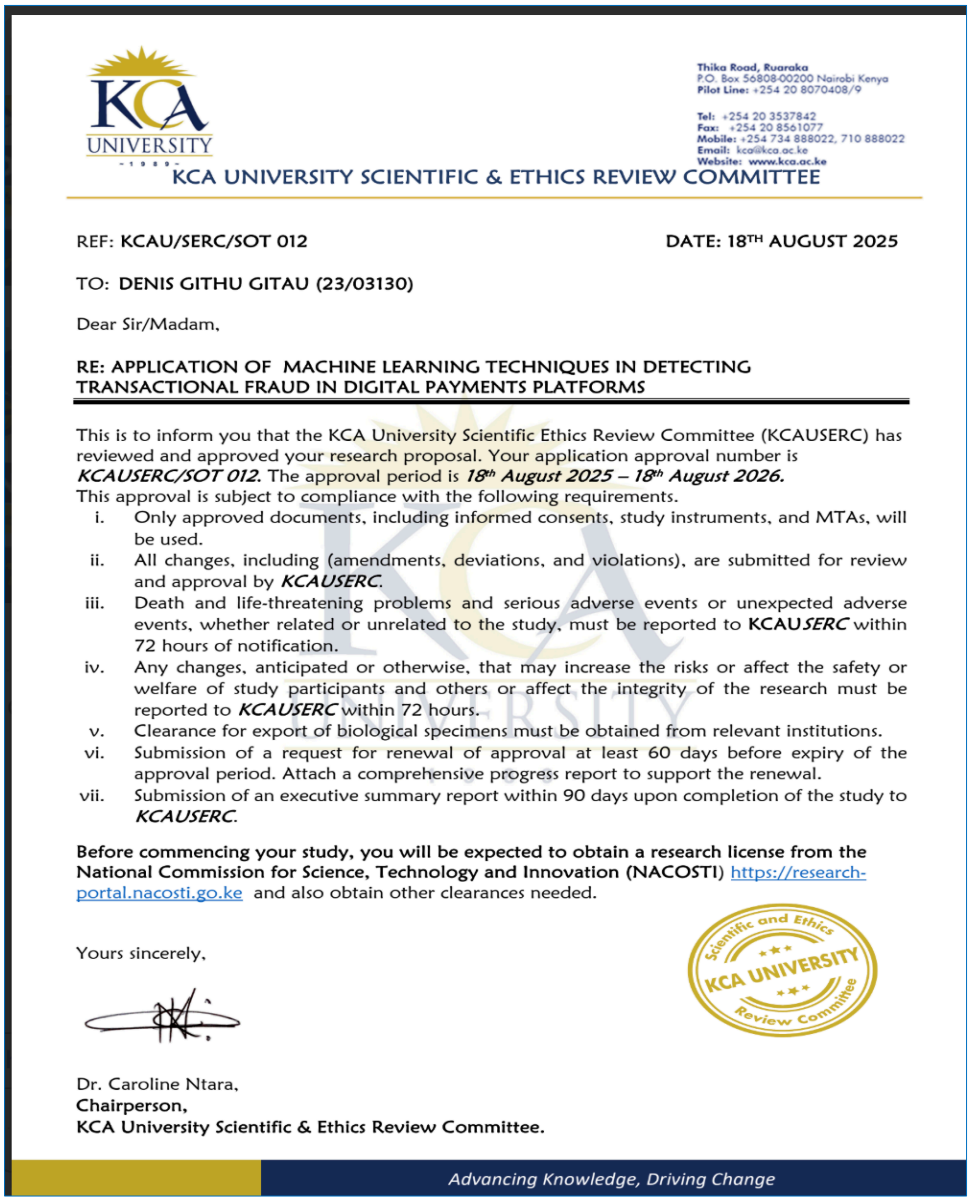
Table A 4

Classifier	Accuracy (Original)	Precision (Original)	Recall (Original)	F1-score (Original)	PR-AUC (Original)	Accuracy (SMOTE)	Precision (SMOTE)	Recall (SMOTE)	F1-score (SMOTE)	PR-AUC (SMOTE)
Logistic Regression	0.96605	0.91667	0.31707	0.47117	0.51400	0.85385	0.19595	0.66519	0.30272	0.50877
Random Forest	0.96933	0.82072	0.45456	0.58689	0.59191	0.95474	0.52539	0.52772	0.52655	0.53545
Gradient Boosting	0.96849	0.86957	0.39911	0.54711	0.57233	0.92576	0.33889	0.58537	0.42927	0.55878
Naïve Bayes	0.94966	0.46973	0.43016	0.44907	0.44007	0.94406	0.42169	0.46563	0.44257	0.44114
Decision Tree	0.96055	0.60833	0.48559	0.54007	0.36413	0.93264	0.35952	0.52772	0.42767	0.24656
K-Nearest Neighbors	0.96404	0.78756	0.33703	0.47205	0.43090	0.92798	0.34072	0.54545	0.41944	0.32396
Neural Network (MLP)	0.96806	0.85990	0.39468	0.54103	0.58122	0.91064	0.29436	0.62528	0.40028	0.55359

APPENDIX V

Ethical Clearance Certificate

Figure A 1



KCA UNIVERSITY

Thika Road, Ruaraka
P.O. Box 56808-00200 Nairobi Kenya
Pilot Line: +254 20 8070408/9

Tel: +254 20 3537842
Fax: +254 20 8561077
Mobile: +254 734 888022, 710 888022
Email: kca@kca.ac.ke
Website: www.kca.ac.ke

KCA UNIVERSITY SCIENTIFIC & ETHICS REVIEW COMMITTEE

REF: KCAU/SERC/SOT 012 DATE: 18TH AUGUST 2025

TO: DENIS GITHU GITAU (23/03130)

Dear Sir/Madam,

RE: APPLICATION OF MACHINE LEARNING TECHNIQUES IN DETECTING TRANSACTIONAL FRAUD IN DIGITAL PAYMENTS PLATFORMS

This is to inform you that the KCA University Scientific Ethics Review Committee (KCAUSERC) has reviewed and approved your research proposal. Your application approval number is ***KCAUSERC/SOT 012***. The approval period is ***18th August 2025 – 18th August 2026.***
This approval is subject to compliance with the following requirements.

i. Only approved documents, including informed consents, study instruments, and MTAs, will be used.
ii. All changes, including (amendments, deviations, and violations), are submitted for review and approval by ***KCAUSERC***.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events, whether related or unrelated to the study, must be reported to KCAU***SERC*** within 72 hours of notification.
iv. Any changes, anticipated or otherwise, that may increase the risks or affect the safety or welfare of study participants and others or affect the integrity of the research must be reported to ***KCAUSERC*** within 72 hours.
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days before expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to ***KCAUSERC***.

Before commencing your study, you will be expected to obtain a research license from the National Commission for Science, Technology and Innovation (NACOSTI) https://research-portal.nacosti.go.ke and also obtain other clearances needed.

Yours sincerely,

Scientific and Ethics KCA UNIVERSITY Review Committee

Dr. Caroline Ntara,
Chairperson,
KCA University Scientific & Ethics Review Committee.

Advancing Knowledge, Driving Change

APPENDIX VI

Nacosti Research License

Figure A 2

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **214115**

Date of Issue: **20/October/2025**

RESEARCH LICENSE

This is to Certify that Mr.. Denis Githu Gitau of KCA University, 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: Evaluating Machine Learning Techniques for Fraud Detection in Digital Payments: A Random Forest Approach. for the period ending : 20/October/2026.

License No: **NACOSTI/P/25/4181084**

214115

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