Suspicious Transaction Prediction in Kenyan Digital Payments: A Machine Learning Comparative Study with Imbalanced Data

Imam, Jihan Abdulrazak Swaleh

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistical Science of Strathmore University

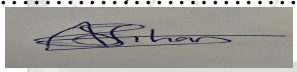
Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June 2025

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Name: **Imam , Jihan Abdulrazak Swaleh**

Signature: 

Date: May 23, 2025

Approval

The dissertation of **Imam Jihan Abdulrazak Swaleh** was reviewed and approved by the following:

Prof. Bernard Omolo,
Professor , Institute of Mathematical Sciences,
Strathmore University.

Dr. Godfrey Madigu,
Dean, Institute of Mathematical Sciences,
Strathmore University.

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University.

Abstract

The surge in digital payments in Kenya has heightened financial crime risks, including money laundering and terrorist financing. Despite regulatory mandates, Suspicious Transaction Reports (STRs) from Payment Service Providers (PSPs) remain below expectations. Traditional rule-based systems often fail to detect such activities, driving interest in machine learning (ML) methods like Random Forest, k-Nearest Neighbours, and Support Vector Machines. However, comparative research on these models, especially in handling severe class imbalance in Kenyan financial datasets, remains limited.

This study therefore evaluated the four ML algorithms (Random Forest, Support Vector Machine, k-Nearest Neighbours and Logistic Regression) for detecting suspicious transactions. To address class imbalance, the SMOTE-ENN re-sampling technique was applied. Factor Analysis for Mixed Data (FAMD) was used for dimensionality reduction, and model performance was assessed using F1-score and Matthews Correlation Coefficient (MCC).

Random Forest outperformed other models post-re-sampling (MCC 99.93%, F1-score 99.94%). Logistic Regression showed the greatest sensitivity to class imbalance, with MCC improving from 62.87% to 97.47%. kNN and SVM also recorded significant gains. Key predictors included Business Age, Score Rank, and Product Type.

The findings underscored the importance of using MCC and F1-score over accuracy when evaluating models on imbalanced datasets. They also supported the adoption of hybrid re-sampling techniques , specifically SMOTE-ENN , to enhance model performance, and highlight Random Forest as a particularly effective algorithm for fraud detection. Future research should explore advanced models such as XGBoost and leverage more diverse datasets to better capture evolving fraud patterns.

Keywords: suspicious transaction reporting; digital payments; machine learning; class imbalance; SMOTE-ENN; fraud detection; random forest; F 1-score; MCC.

Table of contents

List of Figures

List of Tables

List of Abbreviations

AML	Anti Money Laundering	AUC	Area Under Curve
CBK	Central Bank of Kenya	CFT	Counter Financing of Terrorism
DNFBP	Designated Non-Financial Business or Profession	ESAAMLG	Eastern and South Africa Anti Money Laundering Group
FinCEN	Financial Crimes Enforcement Network	FN	False Negative
FP	False Positive	FRC	Financial Reporting Centre
FSC	Financial Supervisory Commission	KNN	K-Nearest Neighbours
KYC	Know Your Customer	MCC	Matthews Correlation Coefficient
NN	Neural Network	POCAMLA	Proceeds of Crime and Anti Money Laundering Act
PSP	Payment Service Provider	ROC	Receiver Operating Characteristics
SAR	Suspicious Activity Report	SMOTE	Synthetic Minority Oversampling Technique
SMOTE-ENN	Synthetic Minority Oversampling Technique - Edited Nearest Neighbors	STR	Suspicious Transaction Reporting
SVM	Support Vector Machine	TN	True Negative
TP	True Positive	VASP	Virtual Asset Service Provider
FAMD	Factor Analysis for Mixed Data		

Acknowledgement

I would like to express my sincere gratitude to the Almighty God for granting me the strength, wisdom, and opportunity to undertake and complete this study.

I am deeply grateful to my family for their unwavering support throughout this journey. In particular, I extend my heartfelt appreciation to my parents for their constant encouragement, prayers, and belief in me, and to my siblings for their understanding, patience, and motivation during this time.

I am also thankful to my supervisor, Prof. Bernard Omolo, for his invaluable guidance, encouragement, and insightful feedback throughout the course of this research.

I am grateful to my classmates and friends for their support, collaboration, and encouragement, which made this journey more enriching.

In addition, I would like to acknowledge my work colleagues for their encouragement and for providing me with the time and flexibility to attend my classes.

Finally, I am thankful to Strathmore University for providing the academic environment, resources, and support necessary for the successful completion of this study.

Dedication

This dissertation is dedicated to God Almighty for giving me wisdom and good health.

Definition of Terms

Class Imbalance	A common issue in machine learning classification tasks where one class significantly outnumbers the other(s), leading to biased model performance (He and Garcia, 2009).
Digital Payments	Electronic financial transactions conducted without the use of physical cash, often via mobile money, internet banking, or digital wallets (Calderon et al., 2025).
Financial Crime	Illegal acts involving the misuse of the financial system, such as fraud, money laundering, terrorist financing, bribery, and corruption (Europol, 2020).
Money Laundering	The process of concealing the origins of illegally obtained funds, typically by passing them through a complex sequence of banking transfers or commercial transactions (Financial Action Task Force, 2023).
Payment Service Providers (PSPs)	Entities that operate payment systems enabling the clearing of payment instructions, netting or settlement of obligations arising from payment instructions, issuing payment instruments, or conducting payment services between payers and beneficiaries, supporting the circulation of money (Central Bank of Kenya, 2023).
Suspicious Transaction Report(STR)	A report that is submitted to the Financial Reporting Centre when a reporting institution becomes aware of transactions or activities that may indicate money laundering, terrorism financing, or other criminal conduct (Government of Kenya, 2023).
Terrorist Financing	Terrorist financing involves the solicitation, collection or provision of funds with the intention that they may be used to support terrorist acts or organizations. (Central Bank of Kenya, 2023).

Chapter 1

Introduction

1.1 Background to the Study

Suspicious Transaction Reporting (STR) is a process mandating financial institutions to report transactions that seem unusual or potentially linked to money laundering, terrorist financing, or other illicit activities. With the rise of digital payments, fraudulent transactions have become more prevalent hence the need for timely reporting. Submitting Suspicious Transaction Reports (STRs) to regulatory bodies is one of the measures implemented to combat the utilization of payment channels to carry out criminal activities.This in turn disincentivizes individuals from conducting money laundering or terrorist financing activity. Reporting such transactions also aids law enforcement agencies in apprehending wrongdoers. The increase in financial crimes has led to the establishment of international supervisory bodies like the Financial Action Task Force (FATF). FATF provides guidelines to combat money laundering and terrorist financing. In Kenya, regulatory bodies affiliated with payment services include the Central Bank of Kenya and the Financial Reporting Centre. Non-compliance with reporting requirements can result in penalties, fines, or sanctions. The Proceeds of Crime and Anti-Money Laundering Act, 2009 (POCAMLA), effective since June 28, 2010,and was revised in 2023 mandates Reporting Institutions to monitor and report suspicious transactions to the Financial Reporting Centre (FRC). FRC has automated reporting mechanisms via the GOAML to streamline the process and ensure swift and instant action. Despite these efforts,the conundrum still remains that these reports lack in quality and are rarely reported. Globally, Suspicious Transaction Report (STR) statistics indicate a significant rise. For instance, in the USA, SARs filed have increased by 50% from 2014 to

2020, while in Australia, the increase is more than 50% . In Kenya, local statistics reveal fewer than 10,000 STRs filed, primarily from banks.

It is important to note that banks are just one category of reporting institutions; others include forex bureaus, money remittance providers and the emerging category of Payment Service Providers (PSPs). PSPs offer payment services that facilitate online payments like mobile money, online bank transfers and card payments. The sectoral risk assessment report which was released by Central Bank of Kenya (2023) identified PSPs as high-risk entities for money laundering and terrorist financing.Therefore, the case of PSPs deserves special attention. The Central Bank of Kenya acknowledged the fundamental risks associated with this rapidly evolving sector. As PSPs introduce innovative payment solutions, new vulnerabilities for money laundering and terrorist financing emerge. Consequently, PSPs have been classified as "high risk" for such activities.

Banks, despite being considered to have more compliance resources, are still facing hefty fines for failing to meet AML/CFT requirements. If even well-established financial institutions with robust compliance structures are struggling, the stakes are even higher for PSPs, which are inherently more vulnerable to risk especially when it comes to Terrorist Financing. Payment Service Providers (PSPs) cannot afford to exercise complacency. It is imperative that they possess a clear understanding of their Suspicious Transaction Reporting (STR) obligations and are equipped with the necessary tools and systems to effectively identify and report suspicious activities. Strengthening risk management practices is not a mere compliance formality; it is a critical safeguard against significant financial, legal, and reputational repercussions.

1.1.1 Regulatory Fines and Penalties

Sanction Scanner highlighted that weaknesses in some bank's Anti-Money Laundering (AML) framework have led to severe penalties and reputational damage for businesses worldwide. An example is the Financial Crimes Enforcement Network (FinCEN) imposing a hefty penalty of $12,500,000 on Capital One. This penalty was attributed to their failure to

report thousands of suspicious activities and transactions from 2008 to 2014. Millions of dollars in suspicious transactions within a cash checking unit went unreported (san, 2021). Another instance involves ABN Amro, which agreed to a settlement of $574 million with Dutch authorities after criminals exploited its accounts for illegal gains. In response, ABN Amro overhauled its AML framework, significantly increasing personnel to enhance the detection and investigation of suspicious transactions. Similarly, Citibank Taiwan faced penalties from the Taiwanese Financial Supervisory Commission (FSC) for neglecting to assess money laundering risks and failing to detect and report suspicious conduct. These instances underscore the critical importance of robust AML frameworks in safeguarding financial institutions from regulatory scrutiny, financial penalties and reputational harm. Money Value Transfer Services such as Western Union have also received hefty penalties for failing to report suspicious transactions back in 2017. Western Union was fined $586 Million. The company was accused of allowing its services to be used for fraud and money laundering, and of failing to adequately report suspicious transactions to regulators. Focusing on Kenya, in 2018, five major banks—Standard Chartered Bank Kenya Ltd., KCB Group PLC, Equity Bank (Kenya) Ltd., Diamond Trust Bank Kenya Ltd., and Co-operative Bank of Kenya Ltd.—faced fines totalling nearly $4 million for failing to report suspicious transactions. These charges were related to the misappropriation of approximately $100 million from Kenya's National Youth Service (NYS). The prosecution of these banks was deferred, providing them with a deadline to enhance their compliance practices (Intelligence, 2020).

1.1.2 Kenya's Grey-listing and the Challenge of Suspicious Transaction Detection in Digital Payments

In 2024 Kenya was designated as a grey-listed area due to non-compliance and a suboptimal level of effectiveness in meeting certain technical compliance and immediate outcomes, as assessed by international bodies like the Financial Action Task Force and its subsidiaries, including the Eastern and Southern Africa Anti-Money Laundering Group (ESAAMLG). In a recent inspection exercise conducted in 2022, Kenya was flagged as non-compliant with

Recommendation 20, which specifically addresses the reporting of suspicious transactions, and demonstrated a low level of effectiveness in Immediate Outcome 4.

This pertains to ensuring that financial institutions, Designated Non-Financial Businesses and Professions (DNFBPs), and Virtual Asset Service Providers (VASPs) adequately implement Anti-Money Laundering/Combating the Financing of Terrorism (AML/CFT) preventive measures in line with their risks and diligently report suspicious transactions (ESAAMLG, 2022). The grey-listing of Kenya poses a significant risk, negatively impacting foreign investments and potentially hampering the nation's economic growth.

In the Second Enhanced Follow-Up Report and 1st Technical Compliance Re-Rating (April 2024), published in August, Kenya's rating for Recommendation 20 (Reporting of Suspicious Transactions) was upgraded from Non-Compliant to C (Compliant) (Eastern and Southern Africa Anti-Money Laundering Group (ESAAMLG), 2024). This improvement reflects changes made under the Proceeds of Crime and Anti-Money Laundering Regulations (POCAMLR), 2023, and amendments to the Prevention of Crime and Anti-Money Laundering Act (POCAMLA). Despite this rerating, Kenya remains on the grey list, indicating that further work is needed to fully address outstanding deficiencies in its anti-money laundering and counter-terrorism financing framework.

The rise of digital payments has revolutionized the financial landscape, bringing convenience and speed but also opening doors to illicit activities like money laundering and terrorist financing. Identifying suspicious transactions is crucial to safeguarding financial systems and curbing these harmful activities. While data analysis has emerged as a powerful tool for detection, a significant gap exists in our understanding of how to apply it effectively in the context of digital payments, particularly in Kenya.

Several factors contribute to this knowledge gap. First, the sensitive nature of the data involved in digital payments makes accessibility and ethical use a challenge. Second, the data itself often exhibits imbalance, meaning patterns tend to skew heavily towards normal transactions, making it difficult to identify rare anomalies indicative of suspicious activity. This challenge becomes even more pronounced in the Kenyan context, where the digital payments ecosystem is rapidly evolving and unique to the local socio-economic landscape.

1.1.3 The Evolving Landscape of Suspicious Transaction Detection: From Rule-Based Systems to Machine Learning

For years, financial institutions have relied on rule-based systems to detect suspicious transactions. These systems work by flagging transactions that break predefined rules, providing an initial defence against fraud. However, as fraudsters become more sophisticated, the limitations of these rule-based approaches become evident. Traditional rule-based systems operate on a single, trigger-based approach. When a transaction meets a specific criterion outlined within a rule, it's marked as suspicious. While this method is efficient to some extent, it falls short in capturing the intricate relationships between different data points within a transaction. These relationships, including factors like location, transaction amount, beneficiary information, and historical behaviour, are crucial for comprehensive fraud detection. Moreover, rule-based systems require constant manual oversight to uncover hidden connections within the data, a process that's prone to human error and resource-intensive. Rule based system require manual adjustments of scenarios by fraud experts and fails to capture the transactional correlations that would point to fraud (Shihembetsa et al., 2021). As noted by Šikman and Grujić (2021), machine learning algorithms leverage the knowledge learned from their training data to adapt to new scenarios without the need for explicit programming for every conceivable fraudulent pattern. This adaptability is in stark contrast to rule-based systems, which rely on human intervention to continually update and refine their rule sets in response to evolving fraudulent tactics.

Alotibi et al. (2022) examined the use of traditional systems in financial institutions, particularly within cryptocurrency exchanges, to identify illicit transactions. Their research revealed significant shortcomings in these conventional methods, characterized by high rates of missed detections and false positives, indicating ineffectiveness and susceptibility to biases. This underscores the imperative to enhance traditional systems for better detection of suspicious activities. In their study, Alotibi et al. (2022) also shed light on the adoption of machine learning approaches, which gained momentum around 2004, offering promising outcomes in combatting money laundering. These findings collectively emphasize the superiority of machine learning techniques over traditional methods in effectively identifying and address-

ing suspicious transactions. The effectiveness of machine learning in suspicious transaction detection isn't just theoretical. Studies conducted by KPMG International have shown that replacing rule-based systems with machine learning models can significantly improve the identification of suspicious activity, with detection rates increasing by up to 40% (Kulkarni, 2023). These real-world findings underscore the transformative potential of machine learning in safeguarding financial institutions from fraudulent activities.

Addressing this gap is critical. Without effective detection methods, suspicious transactions can slip through the cracks, fueling financial crime and undermining financial stability.

1.2 Research Objectives

1.2.1 General Objective

The main objective of this study was to identify a machine learning model that is robust to severe class imbalance, in order to enhance suspicious transaction detection in the Kenyan digital payments landscape by evaluating the impact of data imbalance on various algorithms.

1.2.2 Specific Objectives

1. Evaluate the predictive performance of machine learning models, including Logistic Regression (LR), Random Forest (RF) Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), in forecasting suspicious transaction reports within the Kenyan digital payments landscape by training the models using a designated training set and assessing their performance on a test set.

2. Evaluate the performance of the models identified above, before and after the severe data imbalance is addressed through a hybrid sampling method (SMOTE-ENN) by utilizing performance evaluation metrics such as the Matthews Correlation Coefficient (MCC), Accuracy and F1 score.

1.3 Problem Statement

In Kenya's fast-growing digital payments landscape, financial crime remains a pressing challenge, particularly with the recent international grey-listing due to low reporting of Suspicious Transaction Reports (STRs). While Payment Service Providers (PSPs) are key drivers of digital payment adoption, they also pose a high-risk for financial crime, yet consistently report fewer STRs. Despite existing regulations, Kenya's current approach to suspicious transaction detection rely on outdated, rule-based approaches for suspicious transaction detection, which are often inadequate in addressing the complexities of today's digital transactions.

Although advanced machine learning techniques have shown promise in improving suspicious transaction detection globally, their application in the Kenyan context remains to be comprehensively explored. Due to differing financial crime patterns and data characteristics, findings from other regions cannot be directly generalized to Kenya . Differing financial crime patterns and data characteristics necessitate the validation and adaptation of these techniques to Kenya's unique environment. Furthermore, current studies often neglect the issue of severe data imbalance, specifically the adoption of hybrid resampling methods, further limiting the effectiveness of detection models.

Therefore this study sought to bridge this gap by evaluating high-performing machine learning algorithms, such as SVM and Random Forest, and comparing them with established models like KNN in Kenya. To mitigate data imbalance, the study applied hybrid balancing technique, SMOTE-ENN , to enhance the models' ability to detect suspicious activity effectively.

By equipping stakeholders with tailored insights and tools, this research aimed to strengthen Kenya's AML (Anti-Money Laundering) efforts, reduce financial losses, improve compliance, and contribute to a more secure and transparent financial ecosystem by identifying a robust algorithm to enhance the detection of suspicious transaction reporting.

1.4 Research Questions

The study aimed to address the following research questions;

i. How effectively do machine learning models (LR, RF, SVM, KNN) predict suspicious transactions in Kenya's digital payments?

ii. Which model (LR, RF, SVM, KNN) demonstrates the highest predictive performance for detecting suspicious transactions?

iii. How does class imbalance impact the performance of these models?

iv. Does applying the hybrid SMOTE-ENN sampling technique improve model detection of suspicious transactions?

v. Which evaluation metrics best assess model performance on imbalanced datasets?

1.5 Significance of the Study

This study addressed a critical challenge in Kenya's evolving financial landscape: inadequate reporting of suspicious transactions (STRs). Despite regulations and penalties, low STRs, particularly from Payment Service Providers (PSPs), hinder efforts to combat money laundering, terrorist financing, and other illicit activities. This not only jeopardizes financial security but has also led to the international grey-listing for Kenya in 2024.

1. Addressing a national threat: By improving STR effectiveness, we can help safeguard Kenya's financial system and combat financial crime, protecting investments and fostering economic growth.

2. Empowering new players: PSPs are crucial to digital payments, but their high-risk profile requires robust STR practices. This study provides tools and knowledge to address their specific challenges.

3. Bridging the research gap: Existing studies are limited in the focus on the Kenyan context and data imbalance issues. This research explores top-performing algorithms, compares them with established models, and addresses data imbalance for improved applicability.

4. Developing practical solutions: This research aims to equip regulators, financial institutions, and PSPs with practical tools and insights to enhance STR reporting effectiveness.

Ultimately, this study's goal is to contribute to a more secure and transparent financial ecosystem for all stakeholders in Kenya, ensuring financial stability and promoting responsible digital innovation.

Chapter 2

Literature Review

2.1 Introduction

This section looks at the studies done thus far in the detection of suspicious transactions that includes money laundering and terrorist financing and how various machine learning models have been adopted both locally and internationally to improve the detection of suspicious activity reporting.

This area therefore provides an analysis of the studies done and the research gaps identified.

2.2 Machine Learning Models and Data Imbalance

2.2.1 Machine Learning Models

Machine learning (ML) has increasingly been employed to enhance suspicious transaction detection due to its adaptability and ability to learn from large, complex datasets. Several models, including Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), have been widely tested for their predictive performance.

In 2020, researchers Jullum et al. (2020) explored money laundering detection using machine learning, focusing on a real dataset from Norway. Their study formulated and validated an adaptable machine learning model, utilizing the power of XGBoost for Anti-Money Laundering (AML). Stressing the importance of incorporating all data, including non-reported alerts, the model demonstrated practicality in the intricate nature of Norwegian transactions.

Collaborating with AML experts, they enriched insights by summarizing two months of transaction history. The paper illuminates the significance of adopting machine learning for detecting money laundering but overlooks data imbalance considerations.

In 2023, Anggraeni and Harahap (2023) conducted a study in Indonesia by utilizing machine learning to detect suspicious transactions , in XYZ Bank. Analyzing three models—Decision Tree, Random Forest, and Gradient Boosting , the study recommended machine learning over rule-based approaches, highlighting its adaptability and ability to learn without manual intervention. Results favoured the Random Forest model for its superior accuracy, sensitivity, and recall, contributing to XYZ Bank's Suspicious Financial Transaction (SFT) prediction, reducing alerts, and enhancing compliance efficiency. While the algorithm minimizes false positives, the study suggested expanding datasets and exploring alternative approaches like unsupervised learning in future research. Notably, the impact of class imbalance on model performance and contextualizing findings in the Kenyan financial landscape remain unexplored in the study. Similarly in 2015, Liu et al. (2015) analysed the performance of Random Forest in detecting Financial Fraud using Chinese listed company data. The study employed Random Forest to detect financial fraud, utilizing data from Chinese listed companies. The model was evaluated against other statistical methods, including Logistic Regression, K-Nearest Neighbors, Decision Trees, and Support Vector Machines. Results indicated that Random Forest outperformed these models in terms of accuracy, particularly in reducing false negatives. There is need to put the comparative insights into the Kenyan Context and also address class imbalance using hybrid resampling techniques.

Lim et al. (2021) investigated how machine learning can outsmart credit card fraudsters. Their study compared popular models like SVM, KNN, ANN, and decision trees, finding them all to be more effective than traditional rule-based systems. They highlighted the importance of adaptability, emphasizing that the best machine learning models are constantly tweaked and improved to identify new fraud tactics. The research also points to the potential of combining different machine learning algorithms and tackling imbalanced data sets (where there's a lot more normal data than fraudulent transactions) for even better fraud detection in the future.

In 2018, Gyamfi and Abdulai (2018) conducted a study titled Bank fraud detection using SVM in Ghana utilizing credit card data. The study compared SVM and Back Propagation Network (BPN) performance and identified SVM-S as the superior model. The study failed to comprehensively address data imbalance and how it was resolved.There is also need to put the findings into the Kenyan context. Further,the study only focused on credit card transactions . Elsewhere in 2019, Minastireanu and Mesnita (2019) provided an analysis of the most used machine learning algorithms in online fraud detection especially in credit card data , financial fraud and e-commerce fraud. The study reviewed over 40 papers and concluded that the highest accuracy was achieved by supervised machine learning models which are SVM, ANN and Decision Tree. Despite the study reviewing several machine learning algorithms used in different papers and noting the existence of data imbalance, the paper failed to address how the reviewed papers addressed the data imbalance. There is also need to provide a practical application of the findings especially in the Kenyan context.

In his study, "Fraud Detection using Machine Learning:A Comparative Analysis of Neural Networks and Support Vector Machines," Gitonga (2018) undertook an examination of fraud detection and prevention tools, employing advanced machine learning methodologies such as Support Vector Machines (SVM) and Neural Networks (NN). The research focused primarily on credit card transaction data. Notably, SVM exhibited superior performance, while NN demonstrated improved computational efficiency. However, it is noted that the paper acknowledged the existence of class imbalance without a comprehensive exploration of its impact on model efficacy. Moreover, the analysis was confined solely to credit card transactions, thereby potentially overlooking valuable insights from other transaction types. Additionally, the study exclusively utilized SVM and NN, omitting the investigation of other potentially relevant models such as K-Nearest Neighbors (KNN) and Random Forest. Conducted in 2017, prior to the onset of the COVID-19 pandemic, the paper underscored the urgency for updates with recent data to ensure the continued relevance and applicability of its findings in contemporary fraud detection frameworks.

Eshiwani (2020) embarked on an intriguing exploration, diving into the realm of financial crimes through the intricate analysis of mobile money transactions using pattern recognition

techniques. Their primary aim was to craft a tool capable of pinpointing financial irregularities by tapping into the patterns within these transactions. They opted for the K-Nearest Neighbors (KNN) algorithm, achieving an impressive accuracy rate of 97%. Now, in the wake of the post-COVID era, there is a growing urgency to revisit such studies, especially to explore how various transaction channels like cards, banks, and mobile money stack up against each other. Notably, Eshiwani's paper tackled data imbalances head-on by employing undersampling techniques. It is also crucial to consider how hybrid resampling methods might influence the performance of KNN and Logistic regression in the Kenyan context (Eshiwani, 2020).

Kumar et al. (2020) analysed Anti Money Laundering detection using Naïve Bayes Classifier. The paper sought to develop a big data analytics model, centered on the Naïve Bayes classifier, for classifying banking transactions as legal or illegal, with a focus on detecting money laundering. While the methodology included data collection, cleaning, and analysis, as well as model implementation and validation, it overlooked addressing potential issues related to data imbalance and lacked comparison with alternative machine learning models. The absence of model comparison could limit insights into the effectiveness of the Naïve Bayes approach compared to other algorithms commonly used for similar classification tasks such as SVM, Random Forest and KNN, potentially affecting the robustness and generalizability of the findings. Addressing these gaps would strengthen the study's methodology and contribute to a more comprehensive understanding of the optimal approach for detecting money laundering in banking transactions. Liu et al. (2019) presented a novel approach to detecting suspicious bank card transactions by combining k-means clustering and random forest algorithms to address data imbalance challenges. Methodologically, it employed feature-weighted k-means clustering using Fisher's linear discrimination rate to effectively identify relevant features and random forest algorithm to enhance classification accuracy while avoiding overfitting. Evaluation using AUC and Recall metrics demonstrated significant improvements over traditional methods, with a 5% increase in AUC and 1% increase in F1-measure. However, the paper could have further explored the impact of hybrid methods such as SMOTE-ENN on random forest performance, providing additional insights into optimizing detection models for highly imbalanced datasets. There is need to put the study in the Kenyan context as well.

Shihembetsa et al. (2021) investigated the application of artificial intelligence algorithms for fraud detection in the Kenyan mobile banking sector. The research compared the performance of various machine learning algorithms, including Logistic Regression, Naive Bayes, and K-Nearest Neighbors (KNN). The study concluded that KNN exhibited superior performance compared to the other models . However, limitations exist within the research. Notably, the paper did not comprehensively address the issue of data imbalance, which can significantly impact the effectiveness of machine learning models. Additionally, the study's focus solely on mobile banking transactions within the banking industry restricts its generalizability to other financial sectors or transaction types (Shihembetsa et al., 2021).

2.2.2 Data Imbalance

Kumar et al. (2021) conducted a study on the classification of imbalanced data by reviewing the methods and applications. The study noted that imbalanced data have significant impact on predictive performance of machine learning models (Kumar et al., 2021). Essentially the rare event which is the minority class is not sufficient to correctly classify any incidents of the class of interest hence the need for data balancing techniques. Different data balancing techniques were assessed including undersampling, oversampling and algorithm based methods. Elsewhere Chawla et al. (2002) analysed the impact of data balancing techniques across different data sets in order to identify the most optimal technique. The study indicated that SMOTE approach can improve the accuracy of classification for a minority class. This is attributed to the fact that SMOTE provides a more related minority class samples to learn from hence allowing a learner to craft broader decision regions increasing coverage of the minority class (Chawla et al., 2002). It is noteworthy to point out that Chawla et al. (2002) also recommended the use of SMOTE and undersampling methods for better predictive performance.

Batista et al. (2004) evaluated ten different undersampling and oversampling methods to address class imbalances in training data. Using thirteen datasets from the UCI Machine Learning Repository, each with varying levels of imbalance, the research found that hybrid

data balancing techniques—particularly SMOTE-ENN and SMOTE-Tomek Link—yielded the best results for classification and prediction, especially for datasets with a very small number of rare events. Incorporating these data balancing techniques into the Kenyan context could significantly enhance the detection of suspicious activities by leveraging machine learning methods.

SMOTE-ENN has been extensively used in the medical field to improve the performance of data classification. Lamari et al. (2021) incorporated SMOTE-ENN across multiple medical datasets i.e appendicitis, diabetes and parkinsons and noted that classification was improved. Similarly, Hairani Hairani and Dadang Priyanto (2023) when working with the Pima Indian Diabetes dataset, the SMOTE-ENN technique was used to handle the problem of unbalanced data. The combination of SMOTE for oversampling and an ENN for filtering noisy cases was effective in enhancing the results achieved by the model. With Random Forest, the efficacy of SMOTE-ENN was found to be as high as 95% in terms of accuracy.This illustrated SMOTE-ENN as a useful approach for enhancing diabetes prediction in imbalanced datasets.There is need to identify how a similar hybrid technique will influence the performance of machine learning algorithms particularly in the detection of Suspicious Transactions.

Noviandy et al. (2023) indicated that SMOTE-ENN algorithm is effective in addressing imbalanced datasets, especially for fraud detection. By combining SMOTE, which oversamples the minority class, and ENN, which removes noisy majority class instances, SMOTE-ENN improved model accuracy. In their study, applying SMOTE-ENN to a credit card fraud dataset boosted the Adaptive Integrated Fraud Detection (AIFD) model's accuracy from 92.13% to 97.33%. This demonstrates its ability to enhance both training and testing, leading to more reliable fraud detection.

An in-depth analysis of the effects of class imbalance has been conducted extensively. However, a gap still remains in addressing how class imbalance impacts the detection of suspicious transactions within the Kenyan context.

2.3 Summary of Literature Review

The literature consistently supports the use of machine learning for detecting suspicious transactions, with models like Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression (LR) showing promise across various contexts. Random Forest has demonstrated strong predictive accuracy and robustness in financial fraud detection internationally (Anggraeni and Harahap, 2023; Liu et al., 2015), while studies in Kenya have highlighted SVM and KNN as particularly effective in identifying suspicious activity in mobile money and credit card transactions (Eshiwani, 2020; Gitonga, 2018; Shihembetsa et al., 2021). Logistic Regression remains widely used for its simplicity and interpretability, often serving as a baseline model (Shihembetsa et al., 2021).

A recurring challenge across these studies is the issue of class imbalance, where suspicious transactions are significantly underrepresented compared to legitimate ones. While many studies acknowledge this limitation, few take concrete steps to improve minority class representation, which is critical for optimizing model performance. Although some have applied basic resampling methods, more advanced hybrid techniques like SMOTE-ENN,which have proven effective in domains such as healthcare (Batista et al., 2004; Lamari et al., 2021; Noviandy et al., 2023), remain largely unexplored in the financial crime context, particularly in Kenya.

This study fills that gap by comparing the performance of RF, SVM, KNN, and LR on imbalanced data, both before and after applying SMOTE-ENN. Unlike prior studies which rarely apply hybrid resampling techniques in African financial datasets, this study integrates SMOTE-ENN with four ML models and empirically validates their performance using real transaction data from Kenya — a context underrepresented in current literature.The objective is to provide practical insights into the effectiveness of these models in detecting suspicious transactions within Kenya's evolving digital payments landscape.

Chapter 3

Methodology

3.1 Introduction

This chapter describes the data and the machine learning algorithms employed for analysis. It also explores approaches to handling the data imbalance and performance evaluation metric of the model.

3.2 Research Design

The analysis utilized quantitative methods and adopted supervised machine learning models to investigate the impact of data imbalance in the predictive performance of the machine learning models.The study used R statistical software to conduct the analysis.

3.3 Data Collection

The study utilized secondary data collected at an anonymous licensed payment services company with a specific focus on online transactions. The sample covered a time period within the timespan of 12 months ranging from October 2023 to September 2024. The data was anonymized and no personal information of the transactions was used in the study. This is to ensure the ethical standards of the study are not undermined.

3.4 Data Description

The dependent variables is whether the transaction was filed as a suspicious transaction or not.

Independent variables comprise of transaction details which are in both categorical and numerical forms.

Table 3.1: Data Description

No.	Name	Categories
1	Product Type	a. Card, b. Bank, c. Mobile
2	Amount	Numerical (N/A)
3	Time	a. Morning (7:00am-11:59am), b. Afternoon (12:00pm-3:59pm), c. Evening (4:00pm-6:59pm), d. Night (7:00pm-6:59am)
4	Industry	a. Manufacturing, b. Hotel, Food, and Accommodation, c. Marketing and Advertisement, d. Architecture, Construction, and Civil, e. Travel and Tourism, f. Healthcare, g. Transportation, h. Real Estate and Property Management, i. Renewable Energy, j. Financial Services, k. Wholesale or Retail, l. Insurance and Related Activity, m. Agribusiness, n. Telecommunications, o. Education, p. Arts, Entertainment, and Recreation, q. Information Technology, r. Cosmetics and Personal Care
5	Business Age	Numerical (N/A)
6	KYC Risk Scoring	a. Low, b. Medium, c. High
7	Transaction Score Rank	a. Low, b. Medium, c. High
8	Nature of the Transaction	a. In, b. Out
9	Currency	a. KES, b. USD

3.5 Data Analysis

The study provided a comparative analysis of various machine learning algorithms i.e Support Vector Machine, Logistic Regression, K-Nearest Neighbour and Random Forest.

To investigate the impact of data imbalance on the predictive performance of the mentioned models, the comparative analysis was divided into two analysis i.e model performance before and after data imbalance is addressed . The data resampling techniques that was adopted was a hybrid undersampling and oversampling method called SMOTE- ENN.

3.6 Data Resampling Techniques

Addressing class imbalance in machine learning datasets is crucial for improving the detection of suspicious activities, often requiring the use of oversampling techniques to compensate for the underrepresentation of minority classes . Random Oversampling involves making copies of certain instances randomly, while SMOTE (Synthetic Minority Oversampling Technique) involves creating new instances by using mean values between minority class examples, which effectively addresses imbalance (Chawla et al., 2002).

Oversampling is one of the procedures used when dealing with imbalanced datasets, although it can lead to overfitting a model by replicating data points from the minority class, which limits its generalization. On the other hand undersampling, a method for cutting down the size of the majority class, can remove the number of observations, thus, reducing the sample size (Figure 3.1). A number of the common undersampling techniques consist of methods such as Tomek Link and Edited Nearest Neighbour (ENN) .

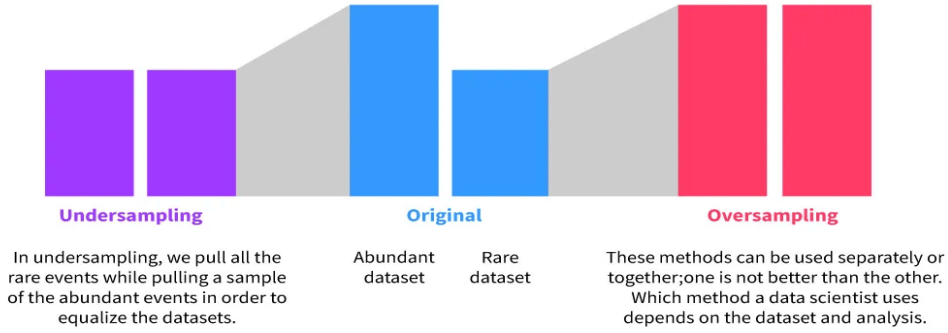


Figure 3.1: Data Balancing Techniques: Undersampling and Oversampling

Both oversampling and undersampling as individual resampling techniques tend to have their own limitations. However, hybrid resampling techniques, such as SMOTE-ENN, are more balanced than individual ones. These methods are very effective in the way that they bring together the positive qualities of both methods (Batista et al., 2004). The methods are particularly useful when dealing with data imbalance in machine learning. In this research, a hybrid methodology was used because there is very limited research in this specific context.

3.6.1 SMOTE-ENN

SMOTE-ENN was proposed by Batista et al. (2004) with an aim of correcting the problems associated with applying oversampling or undersampling only . This hybrid method involves SMOTE (Synthetic Minority Over-sampling Technique) for the purpose of creating synthetic examples in the form of the minority class and in this way balancing the dataset by creating new instances of the rare events. Following oversampling, the ENN (Edited Nearest Neighbour) is used to undersample in order to eliminate noisy, redundant or mislabeled instances on both the minority as well as the majority classes so as to improve data quality and distinction.

Through SMOTE-ENN, this study avoided the problems arising from singular balancing techniques such as risk of overfitting when doing oversampling, and losing other valuable

patterns when undersampling. First, this approach helps to minimize and balance the data and second, it cleans the data so that the machine learning models that are being built can be more robust when it comes to detecting the rare incident. Besides, this technique will be useful in enhancing generalization and performance of the model since noise reduction will improve the quality of the data.

3.7 Machine Learning Models

3.7.1 Motivation for Algorithm Selection

The selection of machine learning algorithms for this study was based on their proven effectiveness in fraud detection and their relevance to prior research in both global and Kenyan contexts.

Random Forest (RF) was selected for its robustness, high accuracy, and ability to reduce overfitting through ensemble learning. Several studies affirm its strong predictive performance in financial fraud detection. For example, Anggraeni and Harahap (2023) found RF to outperform other models in detecting suspicious transactions at an Indonesian bank, while Liu et al. (2015) noted its superior accuracy in identifying financial fraud in Chinese listed companies. In addition, Liu et al. (2019) successfully combined RF with clustering techniques to address data imbalance challenges, demonstrating its versatility.

Support Vector Machine (SVM) is particularly effective in handling imbalanced datasets and modeling complex, non-linear decision boundaries. Its effectiveness was validated in studies by Gitonga (2018) and Gyamfi and Abdulai (2018), both of which found SVM to outperform other models in detecting credit card fraud in Kenya and Ghana, respectively.

K-Nearest Neighbors (KNN) was chosen for its simplicity and ability to capture local data patterns. In the Kenyan context, Eshiwani (2020) achieved 97% accuracy using KNN to detect anomalies in mobile money transactions, demonstrating its practical utility. Similarly,

Shihembetsa et al. (2021) identified KNN as the top-performing model for mobile banking fraud detection in Kenya, despite limitations in addressing class imbalance.

Logistic Regression (LR) was included as a baseline model due to its interpretability and foundational role in classification tasks. It remains widely used in fraud detection studies for its transparency and ability to serve as a benchmark against which to compare more complex models (Shihembetsa et al., 2021).

Together, these models which span from tree-based, margin-based, instance-based, and probabilistic approaches enable a comprehensive comparative analysis. This diverse selection increases the robustness of the study's findings and aligns with global and local precedents in machine learning applications for financial anomaly detection.

3.7.2 Random Forest

Random Forest is a combination of several classifiers or predictors where all the classifiers or predictors are trained to predict the outcome of a given data set in order to create a highly accurate general classifier or predictor. It operates in a way that constructs a Forest of Trees. These trees are built separately in order to improve the final model.

Random Forest has been widely used in financial fraud detection tasks due to its high accuracy and ability to handle large datasets with complex interactions. Liu et al. (2015) developed a financial fraud detection model using Random Forest, demonstrating its effectiveness in detecting anomalies in financial data .

Figure 3.2: Machine learning models: Random Forest

Bagging Principle

Random Forest employs the bagging (bootstrap aggregating) technique and random variable selection, both introduced by Breiman (2001), making it a powerful and widely adopted ensemble learning method. It entails developing more than one tree established on bootstrap resamples of the data, which are samples acquired by drawing replaced sampling from the original data set.

Random Variable Selection

The nodes in both decision trees are split according to random samples of predictors. These subsets are of size m, where m is a fixed integer and y is an identifier for the particular subset. This random variable selection is useful in preventing the individual trees from becoming too interaction which adds diversity into the decision making process.

Majority Voting (Classification)

Using the Random Forest method, the new observations are assigned to the previously defined classes by a majority voting of the created trees. The predicted class to the new data point is the class with the highest votes in the class.

Averaging (Regression)

In regression tasks, the final predictions are constructed by averaging over the predicted response values with the use of all trees in the ensemble. This kind of ensemble averaging averages out the prediction variance.

Error Rate Assessment

The accuracy of the Random Forest model is usually measured when the model is making prediction on out of bag or OOB data. These are data points that are in the original dataset but are not included in the bootstrap sample of a particular tree. Every decision tree in the forest is calibrated with the OOB data to determine the performance of each tree.

The Random Forest models can be created using the function in R known as 'randomForest' belonging to the 'randomForest' package. It can be used for regression analysis, when response variable is continuous, and for classification analysis, when response variable is categorical.

3.7.3 Support Vector Machine

Support Vector Machine, commonly abbreviated as SVM, is a powerful machine learning algorithm used for both classification and regression tasks. It is particularly renowned for its ability to create a hyperplane that distinctly separates data points in high-dimensional space.

The foundational work on SVMs was introduced by Cortes and Vapnik (1995), where the concept of maximizing the margin between classes using a hyperplane was formalized.

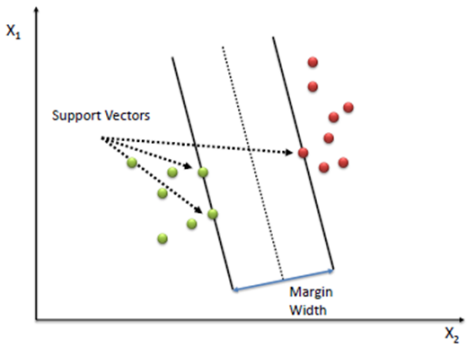


Figure 3.3: Machine learning models: Support Vector Machine

This hyperplane is chosen to maximize the margin, which is the distance between data points of different classes, ensuring robust classification. SVM's objective is to find this optimal hyperplane, and its core concepts can be described with the equation indicated below:

$$w \cdot x + b = 0, \tag{3.1}$$

where w represents the weights, x is the input vector, and b is the bias or intercept term.

This hyperplane serves as the decision boundary. Margin calculation is also a necessary metric. The margin is the distance between the hyperplane and the nearest data points. The margin is calculated as:

$$Margin = \frac{2}{||w||}, \tag{3.2}$$

where the denominator represents the Euclidean norm of the weight vector w.
SVM aims to maximize the margin while minimizing classification errors. This is formulated using the hinge loss:

$$Hinge\ Loss = \sum_{i=1}^{n} max(0, 1 - y_i(w.x_i + b)), \tag{3.3}$$

where n is the number of data points, y_i is the class label (either -1 or 1), x_i is the data point, and w.x+b is the decision boundary. To prevent overfitting and balance the margin and loss, SVM includes a regularization term. The objective function becomes;

$$ObjectiveFunction = \frac{1}{2}||w||^2 + C \cdot (\text{Hinge Loss}). \tag{3.4}$$

Here is the regularization parameter that controls the trade-off between maximizing the margin and minimizing the hinge loss. SVM's objective is to find the optimal w and b that minimize the objective function. This is typically solved using quadratic programming techniques.

In summary, SVM's strength lies in its ability to handle complex, high-dimensional data by creating an optimal hyperplane that maximizes the margin between classes. This hyperplane, is found by optimizing the hinge loss with a regularization term, providing a robust framework for classification tasks .

Kumar et al. (2022) also noted that one of SVM's key strengths is its ability to handle high-dimensional data and imbalanced datasets, which is crucial in fraud detection. By creating a decision boundary that maximizes the margin between classes, SVM helps effectively distinguish between fraudulent and legitimate transactions, even when the difference between them is subtle and the dataset is imbalanced

3.7.4 Logistic Regression

Logistic Regression is one of the most popular machine learning algorithms that falls under the supervised learning category and is used mainly for binary classification problems (Hosmer Jr et al., 2013). It outlines the likelihood of an event taking place with reference to the predictor variables.

Following are the description of the key concepts/equations of logistic regression. The logistic regression model uses the logistic function or the sigmoid function in order to

establish the probability of the event. The sigmoid function is defined as:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(w \cdot x + b)}}, \tag{3.5}$$

where Y is the event which we want to predict and P(Y) is the likelihood of occurrence of the event Y when given the input vector X , while w and b are the weight and the bias, respectively.

Log-Odds Transformations

The log-odds (logit) transformation is often used to express the logistic function in terms of linear equations:

$$\frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = w.x + b. \tag{3.6}$$

This transformation maps the data in a linear fashion where the log-odds of the event is equal to the linear function of the predictor variables.

Maximum Likelihood Estimation

Logistic regression uses maximum likelihood estimation to estimate the parameters of the model, which are w and b . The likelihood function is:

$$L(w, b) = \prod_{i=1}^{n} P(Y_i - 1|X_i)^{Y_i} . (1 - P(Y_i - 1|X_i))^{1-Y_i}. \tag{3.7}$$

The aim of any training algorithm is to identify the w and b that will maximize the likelihood function; in other words, find the values of w and b that will make the likelihood of the observed data as high as possible (Hosmer Jr et al., 2013).

Cost Function (Log-Loss)

To fit the logistic regression model, there is a cost function used which is also referred to as

log-loss or cross entropy loss. The cost function is given by;

$$Cost\ Function = \frac{1}{n}\sum_{i=1}^{n}[Y_i log(P(Y_i-1|X_i)) + (1-Y_i) log(1-P(Y_i-1|X_i))]. \tag{3.8}$$

This cost function quantifies the difference between the probability estimates produced by the model and the actual classification labels in the training set.

Logistic regression is widely applied in binary classification scenarios, such as fraud and suspicious transaction detection, due to its interpretability and probabilistic framework as noted in a study by Jurgovsky et al. (2018) . The ability to output probabilities makes it especially useful in financial crime detection, where understanding the risk level is as important as the classification itself.

3.7.5 K-Nearest Neighbour

The K-Nearest Neighbors (KNN) algorithm was one of the first algorithms for classification and regression, which was developed Fix and Hodges Jr. (1951). It works on the principle of labelling or valuing the new data point with respect to K nearest neighbors in the training dataset.

Mathematically, for a training dataset X with n data points, each represented by a d-dimensional feature vector X_i and corresponding labels or values Y:

$$\text{distance}(x, X_i) = \sqrt{\sum_{j=1}^{d}(x_j - X_{ij})^2}. \tag{3.9}$$

When a new data point x is added to the container, KNN computes distances to all the points in the set X and then finds the first K of them. In classification, the average of the nearest neighbours' class labels is calculated and assigned, while for regression, the mean or weighted mean of the nearest neighbours' values is computed.

Such aspects are the decision regarding which distance measure to use (Euclidean, Manhattan and all others) and the decision of which value of K to choose. K values are higher to reduce noise impacts, and values are in odd number to avoid tied ranks.

KNN is non-parametric, which means that it does not require the data distribution to follow any specific probability distribution; hence suitable for a wide range of datasets. But this can be very costly in terms of computational and also may perform poorly in large dimensional spaces as noted by Beyer et al. (1999) .

In the context of fraud detection, KNN has been applied successfully to detect anomalous financial patterns by identifying transactions that deviate from established behavioural norms. Studies such as Eshiwani (2020) have demonstrated the use of pattern recognition techniques, including KNN, for detecting financial crimes in mobile money platforms . Similarly, Shihembetsa et al. (2021) explored the role of artificial intelligence algorithms such as KNN in enhancing fraud detection in Kenya's banking industry . These studies highlight the effectiveness of KNN in identifying suspicious activities based on historical transaction patterns.

In R, KNN is implemented in the "class" package using the function `knn()`.

3.8 Performance Evaluation Metrics

The Performance Evaluation Metric was utilised in the assessment of the models before and after application of SMOTE-ENN. Before we take a look at the metrics used, it is necessary to explain the concept of a confusion matrix is.

A confusion matrix is a very useful matrix that is used in determining the performance of a classification model be it binary or multi-class. This matrix provides a breakdown of the model's predictions, highlighting the number of accurate classifications: Scholars also classify the models based on their performance metrics namely True Positives (TP) and True Negatives (TN). Also, it shows cases of model errors such as false positive (FP) which are false alarms and false negative (FN) which are cases that were missed.

The usage of these elements makes it possible to estimate other important characteristics of the classification quality, such as precision and recall.

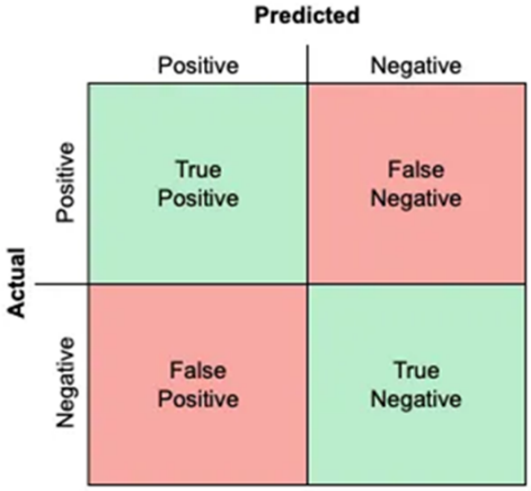


Figure 3.4: Confusion Matrix

3.8.1 Accuracy

Accuracy is a statistical measure used to evaluate the overall correctness of predictions or classifications made by a model accuracy. Anything greater than 70% is a great model performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.10}$$

However, accuracy alone can be misleading, especially in imbalanced datasets as noted by He and Garcia (2009) and Chicco and Jurman (2020).

3.8.2 F1-Score

F1-score evaluates the performance of a classification model, particularly in scenarios where you want to balance the trade-off between precision and recall. A higher F1- score suggests

better model performance (0.7 and above). Note that ;

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall}\ , \tag{3.11}$$

$$Precision = \frac{TP}{TP+FP}\ , \tag{3.12}$$

and

$$Recall = \frac{TP}{TP+FN}\ . \tag{3.13}$$

3.8.3 MCC

Mathew Correlation Coefficient (MCC) is a balanced measure that considers true positives, true negatives, false positives, and false negatives. MCC is ;

$$MCC = \frac{(TN \times TP) - (FN \times FP)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}\ . \tag{3.14}$$

A correlation of:

C = 1 indicates perfect agreement,

C = 0 is expected for a prediction no better than random, and

C = -1 indicates total disagreement between prediction and observation

The MCC gives a better perspective of the classification performance compared to accuracy, especially in cases of imbalanced data sets since it takes into consideration both the false positives as well as false negatives. Its strength lies in providing a single-valued metric that better reflects classification quality in cases like suspicious activity detection, where class imbalance is significant as noted by Chicco and Jurman (2020) and Jullum et al. (2020).

3.8.4 Utilization and Dissemination of the Findings of the Study

The findings of this study will play a crucial role in enhancing the detection of suspicious transaction reports (STRs) in Kenya. By identifying key influencing factors within the digital payments ecosystem, the study will provide valuable insights for relevant stakeholders, including financial institutions, regulators and policy makers. Additionally, the research will contribute to advancing the understanding of data imbalance challenges and promoting the adoption of machine learning techniques to improve STR detection.

To ensure broad accessibility and impact, the study's findings will be disseminated through online platforms, industry forums, and targeted sensitization efforts with key stakeholders. These efforts will help drive awareness, encourage adoption, and foster collaboration in strengthening financial security measures in Kenya. Moreover, the final report will be available for reference by the relevant stakeholders from the industry and government, ensuring that the insights generated can be effectively utilized to inform decision-making and policy development.

3.8.5 Ethical Considerations

This study adhered to strict ethical standards to ensure data privacy, confidentiality, and regulatory compliance. All transaction data was anonymized to protect sensitive details, such as merchant names and transaction IDs, and to prevent unauthorized access. Ethical approval was obtained from relevant authorities, with informed consent sought where necessary.

The study also upheld fairness and transparency in model development. Findings will be responsibly reported to prevent misuse and support ethical financial crime detection practices.

Chapter 4

Results and Interpretation

4.1 Introduction

This chapter explores the dataset by analyzing descriptive statistics, such as proportions for categorical variables and mean values for numerical data. It also covers the data preprocessing steps, including variable selection and encoding of categorical features, before applying machine learning models—SVM, Random Forest, Logistic Regression, and KNN—on the imbalanced dataset. To address the imbalance, the dataset was resampled using SMOTE-ENN, a hybrid technique that combines oversampling and undersampling. The same models were then trained on the resampled dataset that has increased representation of the minority class. Finally, the findings were interpreted, key variables identified, and the implications discussed.

4.2 Descriptive Statistics

The dataset consisted of 67,491 transactions with 10 variables, comprising both categorical and numerical features. The categorical variables included STR.Filed, Product.type, Currency, Nature.of.Transaction, Industry, KYC.Risk, Score.Rank, and Time.of.Day, while the numerical variables were Amount and Business.age. Notably, no missing values were detected in the dataset.

However, the dataset exhibited a significant class imbalance, particularly in the STR.Filed variable, where 99.75% of the observations were labelled as "NO" and only 0.25% as "YES."

This severe imbalance posed challenges for model performance and required appropriate handling. The tables below provide a detailed summary of the key statistics for each variable.

Table 4.1 reveals an extreme low incidence of STR filings, with 67,322 transactions representing (99.75%) not flagged for suspicious activity, and 169 transactions (0.25%) resulting in STR filings. This indicated a low prevalence of transactions deemed potentially illicit within the dataset.

Table 4.1: Summary of STR.Filed

Category	**Count**	**Percentage**
NO	67,322	99.75%
YES	169	0.25%

Table 4.2 demonstrates significant concentration in mobile money transactions, accounting for 50,859 transactions (75.36%), followed by bank transfers with 15,609 transactions (23.13%). Card payments represented a minor segment, with 1,023 transactions (1.52%). This distribution underscored the dominance of mobile money as a transactional modality within the system.

Table 4.2: Summary of Product Type

Category	**Count**	**Percentage**
BANK_TRANSFER	15,609	23.13%
CARD_PAYMENT	1,023	1.52%
MOBILE_MONEY	50,859	75.36%

The currency composition of transactions in Table 4.3 was predominantly Kenyan Shillings (KES), representing 66,711 transactions (98.84%). US Dollar (USD) transactions constituted a negligible proportion, with 780 transactions (1.16%), reflecting a strong reliance on the local currency.

Table 4.3: Summary of Currency

Category	Count	Percentage
KES	66,711	98.84%
USD	780	1.16%

The nature of transactions in Table 4.4 was characterized by a higher volume of outflows compared to inflows, with outflows accounting for 46,159 transactions (68.39%) and inflows representing 21,332 transactions (31.61%), indicating a net flow of funds out of the system.

Table 4.4: Summary of Nature of Transaction

Category	Count	Percentage
IN	21,332	31.61%
OUT	46,159	68.39%

The industry sector analysis in Table 4.5 revealed a heterogeneous distribution of transactional activity, with the transportation sector exhibiting the highest transaction volume (19,388 transactions, 28.73%), followed by marketing and advertisement (11,538 transactions, 17.10%) and telecommunications (9,497 transactions, 14.07%). Conversely, sectors such as cosmetics and personal care, wholesale or retail, and insurance and related activity demonstrated minimal transactional activity.

Table 4.5: Summary of Industry

Category	Count	Percentage
Agribusiness	3,986	5.91%
Architecture, Construction	4,216	6.25%
Arts, Entertainment	168	0.25%
Cosmetics	33	0.05%
Education	361	0.53%
Financial Services	954	1.41%
Healthcare	3,515	5.21%
Hotel, Food	2,599	3.85%
IT	811	1.20%
Insurance	73	0.11%
Manufacturing	3,849	5.70%
Marketing	11,538	17.10%
Real Estate	5,153	7.64%
Renewable Energy	364	0.54%
Telecommunications	9,497	14.07%
Transportation	19,388	28.73%
Travel	833	1.23%
Wholesale/Retail	153	0.23%

The KYC risk assessment (Table 4.6) indicated a predominantly low-risk profile, with low-risk profiles constituting 46,901 transactions (69.49%), followed by medium-risk profile with 19,937 transactions (29.54%), and high-risk profile representing a small fraction, with 653 transactions (0.97%). This distribution suggested low incidence of high-risk clientele.

Table 4.6: Summary of KYC Risk

Category	Count	Percentage
High	653	0.97%
Low	46,901	69.49%
Medium	19,937	29.54%

Table 4.7 provides a breakdown of the Score Rank distribution within the dataset. The majority of the transactions fall under the Low Risk category (57.63%), followed by Medium Risk (42.19%). A very small proportion of transactions are classified as High Risk (0.18%). This distribution indicated that the dataset is predominantly composed of low and medium-risk transactions, with high-risk cases being rare.

Table 4.7: Summary of Score Rank

Category	Count	Percentage
High Risk	120	0.18%
Low Risk	38,895	57.63%
Medium Risk	28,476	42.19%

Table 4.8 presents the distribution of transactions across different times of the day. The Afternoon period had the highest proportion of transactions (32.27%), followed by Morning (23.52% and Evening (23.43%). The Night period had the lowest transaction count (20.78%). This suggested that transaction activity was highest during the daytime and decreased towards the night.

Table 4.8: Transaction Distribution by Time of Day

Category	Count	Percentage
Afternoon	21,777	32.27%
Evening	15,815	23.43%
Morning	15,871	23.52%
Night	14,028	20.78%

Table 4.9 provides summary statistics for the numerical variables Amount and Business Age in the dataset. The transaction amounts ranged from 1 to 12.9 billion, with a median of 4,000. However, the mean transaction amount was significantly higher at 213,300, indicating a right-skewed distribution, likely influenced by a few very large transactions.

For Business Age, the dataset included businesses ranging from 0.08 to 74 years old. The median age was 11 years, while the mean was 10.93 years, suggesting a relatively even distribution without significant skewness.

Table 4.9: Descriptive Statistics for Numerical Variables

Statistic	Amount	Business Age
Min	1	0.08
1st Quartile	1,500	2.00
Median	4,000	11.00
Mean	213,300	10.93
3rd Quartile	13,270	12.00
Max	12,900,000,000	74.00

Table 4.10 presents the Cramér's V correlation values between categorical variables in the dataset, indicating the strength of associations between them.

Table 4.10: Correlation Matrix of Categorical Variables

Variable	STR.Filed	Product.type	Currency	Nature.of.Transaction	Industry	KYC.Risk	Score.Rank	Time.of.Day
STR.Filed	1.000	0.399	0.458	0.072	0.910	0.501	0.317	0.049
Product.type	0.399	1.000	0.872	0.381	0.677	0.313	0.332	0.074
Currency	0.458	0.872	1.000	0.159	0.760	0.273	0.405	0.047
Nature.of.Transaction	0.072	0.381	0.159	1.000	0.771	0.709	0.375	0.172
Industry	0.910	0.677	0.760	0.771	1.000	0.792	0.410	0.203
KYC.Risk	0.501	0.313	0.273	0.709	0.792	1.000	0.241	0.097
Score.Rank	0.317	0.332	0.405	0.375	0.410	0.241	1.000	0.095
Time.of.Day	0.049	0.074	0.047	0.172	0.203	0.097	0.095	1.000

A strong correlation was observed between Industry and STR.Filed (0.910), suggesting that certain industries were more likely to be involved in suspicious transaction reporting. Similarly, Product Type and Currency (0.872) showed a high association, indicating that certain

payment methods were linked to specific currencies. The relationship between Industry and Nature of Transaction (0.771) suggested that businesses within particular industries have preferred transaction flows, either inbound or outbound. Additionally, Industry and KYC Risk (0.792) highlighted that some industries may carry a higher level of risk based on Know Your Customer (KYC) assessments.

Moderate correlations were seen between KYC Risk and Nature of Transaction (0.709), implying that certain transaction types may be more prone to higher risk. Similarly, Score Rank and Currency (0.405) indicateed that the choice of currency could impact the ranking or scoring of entities within the dataset. On the other hand, weaker correlations were observed in factors such as STR.Filed and Time of Day (0.049), suggesting that the time at which a transaction occurs had minimal influence on whether it was flagged as suspicious. Likewise, Product Type and Time of Day (0.074) showed a low dependency, implying that payment method selection was not strongly influenced by the timing of transactions. Overall, these insights provided valuable information for risk assessment, transaction monitoring, and potential fraud detection strategies.

Figure 4.1 below presents the correlation matrix for continuous variables in the dataset, specifically Amount and Business Age. The correlation between Amount and Business Age (-0.003) was extremely weak and close to zero, indicating that there was no significant relationship between the transaction amount and the age of the business. This suggested that older or newer businesses do not necessarily conduct transactions of higher or lower amounts in any predictable pattern.

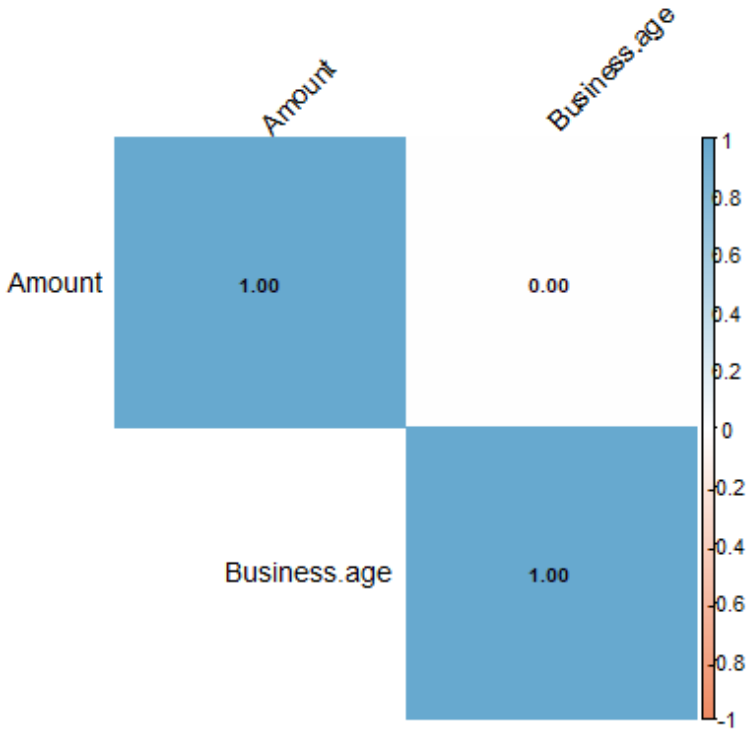


Figure 4.1: Correlation between Business Age and Amount

Boxplots were used to explore the distribution of the numerical variables Business Age and Amount, and to identify any potential outliers (Figure 4.2). The plots highlighted the presence of extreme values, particularly in Amount, which is expected in real-world financial data. These outliers were retained, as they likely represent legitimate transactions and may carry useful signals for model training. Their presence also reinforced the need for appropriate resampling techniques to ensure underrepresented classes are not overshadowed by dominant patterns during classification.

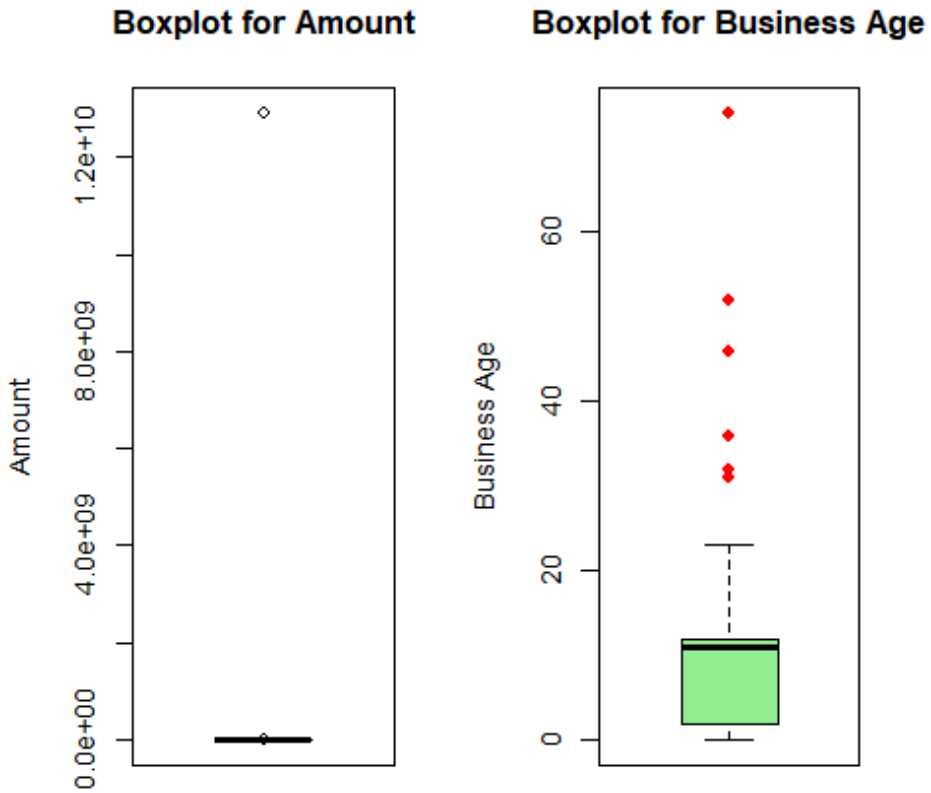


Figure 4.2: Box plot Business Age and Amount

Before applying machine learning models, highly correlated variables—Industry, Currency, and KYC Risk—were removed from the dataset to prevent inflated metrics and potential overfitting as indicated in Table 4.10. This preprocessing step ensures that the machine learning models are not biased by redundant information, leading to more reliable and generalizable predictions.

4.3 Modelling Setup

Before training the models, the dataset underwent several preprocessing steps to ensure optimal performance and reduce the risk of overfitting. First, the 'Industry,' 'Currency,' and

'KYC.Risk' variables were removed, due to the correlation. The target variable, 'STR.Filed,' was converted into a factor to align with classification modeling requirements. Other categorical variables, including 'Product.type,' 'Nature.of.Transaction,' 'Score.Rank,' and 'Time.of.Day,' were also transformed into factors to ensure proper handling during analysis.

For numerical preprocessing, the 'Amount' variable was cleaned by removing commas and converting it into a numeric format. To make the dataset compatible with machine learning algorithms, one-hot encoding was applied to categorical variables, while numerical features were standardized through centering and scaling. This helped normalize the data and improve model performance.

Once preprocessing was complete, the dataset was split into training and testing sets using an 80-20 ratio, ensuring that the distribution of the target variable remained balanced. To further mitigate overfitting and improve generalization, 10-fold cross-validation was applied during model tuning. This approach ensured robust performance evaluation

4.4 Machine Learning before Data Resampling

4.4.1 K-Nearest Neighbors (KNN) Model

The kNN model was trained with hyperparameter tuning to identify the optimal value of k. The best k value was found to be 5. Upon evaluation, the kNN model achieved an accuracy of 99.96%, with a balanced accuracy of 95.44%. The model demonstrated a sensitivity of 90.91% and a specificity of 99.98%, indicating stronger performance in detecting negative cases. The F1-score was recorded at 90.91%, and the Matthews Correlation Coefficient (MCC) was 90.89%, further affirming the model's ability to distinguish between the two classes effectively.

Table 4.11: Confusion Matrix for kNN Model

	Actual NO	**Actual YES**
Predicted NO	13461	3
Predicted YES	3	30

4.4.2 Support Vector Machine (SVM) Model

The SVM model with a radial basis function kernel was trained and evaluated on the test dataset. The model achieved an accuracy of 99.89% and a balanced accuracy of 80.30%. However, the sensitivity was relatively lower at 60.61%, while specificity was significantly high at 99.99%. The F1-score of the SVM model was recorded at 72.73%, and the MCC stood at 74.18%. Although the model exhibited high precision, its recall performance suggests that it may struggle to detect positive cases effectively.

Table 4.12: Confusion Matrix for SVM Model

	Actual NO	**Actual YES**
Predicted NO	13462	13
Predicted YES	2	20

4.4.3 Random Forest Model

The Random Forest model exhibited exceptional performance, achieving an accuracy of 99.99% with a balanced accuracy of 96.97%. The sensitivity of the model was 93.94%, while specificity reached 100%. The F1-score of the model was 96.88%, and the MCC was recorded at 96.92%, making it the best-performing model in this study.

Table 4.13: Confusion Matrix for Random Forest Model

	Actual NO	**Actual YES**
Predicted NO	13464	2
Predicted YES	0	31

The feature importance analysis indicated that 'Business.age' and 'Amount' were the most significant predictors, followed by 'Product.type' and 'Score.Rank.'

The variable importance analysis (Table 4.14) from the Random Forest model highlighted Business Age as the most influential predictor, with a score of 100.00, indicating its strong impact on the classification of the target variable. Transaction Amount (27.30) and Product Type - Card Payment (22.28) also played significant roles in determining whether an STR is filed.

Table 4.14: Variable Importance in the Random Forest Model

Variable	**Importance Score**
Business Age	100.00
Amount	27.30
Product Type - Card Payment	22.28
Score Rank - High Risk	5.50
Time of Day - Night	2.89
Score Rank - Medium Risk	1.37
Product Type - Bank Transfer	0.69
Time of Day - Evening	0.62
Time of Day - Morning	0.22
Product Type - Mobile Money	0.20
Time of Day - Afternoon	0.17
Score Rank - Low Risk	0.12
Nature of Transaction - Out	0.01
Nature of Transaction - In	0.00

Score Rank - High Risk (5.50) and Time of Day - Night (2.89) showed moderate influence, suggesting that high-risk scores and nighttime transactions contributed to classification decisions. Lower importance scores were observed for Score Rank - Medium Risk (1.37) and Product Type - Bank Transfer (0.69), indicating they had smaller but still measurable effect.

Less influential predictors included Time of Day - Evening (0.62), Time of Day - Morning (0.22), and Product Type - Mobile Money (0.20). The lowest-ranked features, Nature of Transaction - Out (0.01) and Nature of Transaction - In (0.00), contributed minimally to the model's decision-making process.

Overall, the results emphasized the critical role of business age, transaction amount, and product type in predicting STR filings, while time-based variables and transaction types exhibited lower significance.

4.4.4 Logistic and Lasso Regression Model

The Logistic Regression model achieved an accuracy of 99.85%, with a balanced accuracy of 71.21%. While its specificity was high (99.99%), the sensitivity remained low at 42.42%, indicating challenges in identifying positive cases. The model's F1-score was 58.33%, and the MCC (Matthews Correlation Coefficient) was 62.87% , reflecting moderate predictive power.

Table 4.15: Confusion Matrix for Logistic Regression Model

	Actual NO	**Actual YES**
Predicted NO	13463	19
Predicted YES	1	14

Analysis of feature importance revealed that variables such as business age, high-risk score rank, and time of day had a strong influence on model predictions. To enhance feature selection and reduce redundancy, Lasso Regression was applied. This technique uses L1 regularization to shrink less relevant coefficients to zero, effectively filtering out weaker predictors.

Although Lasso Regression had a slightly lower overall accuracy (99.82%) compared to standard logistic regression, it achieved a balanced accuracy of 66.66%. While sensitivity dropped to 33.33%, precision remained high at 84.61%, and the Matthews Correlation Coefficient reached 53.05%, indicating moderate predictive strength.

One of the key advantages of Lasso is its ability to retain only the most impactful variables. In this study, it highlighted the importance of transaction amount, card-based product types, and business age in predicting suspicious transaction reports—factors that were not consistently emphasized by traditional logistic regression. These results demonstrate the value of regularized models in identifying meaningful patterns within imbalanced financial datasets.

Table 4.16 summarizes the most significant variables identified by each model, showcasing Lasso's effectiveness in refining feature selection for fraud detection.

Table 4.17 below indicates a summary of the performance metrics for all the machine learning models.

Table 4.16: Significant Variables Identified in Each Model

Variable	**Logistic Regression**	**Lasso Regression**
Product Type - Card Payment	Not Significant	Significant
Amount	Not Significant	Significant
Business Age	Significant	Significant
Score Rank - High Risk	Significant	Significant
Time of Day - Night	Not Available	Significant
Time of Day - Evening	Significant	Not Significant
Time of Day - Morning	Significant	Not Significant

Table 4.17: Performance Comparison of Different Models

Model	**Accuracy**	**Balanced Accuracy**	**Sensitivity**	**Specificity**	**MCC**	**F1-Score**
kNN (k=5)	0.9996	0.9544	0.9091	0.9998	0.9089	0.9091
SVM (Radial)	0.9989	0.8029	0.6061	0.9999	0.7418	0.7273
Random Forest	0.9999	0.9697	0.9394	1.0000	0.9692	0.9688
Logistic Regression	0.9985	0.7121	0.4242	0.9999	0.6287	0.5833
Lasso Regression	0.9982	0.6666	0.3333	0.9999	0.5305	0.4783

The performance comparison showed that Random Forest was the best model, achieving the highest accuracy (99.99%), balanced accuracy (96.97%), and F1-score (96.88%), making it the most reliable choice for classification. kNN (k=5) also performed well, with strong accuracy (99.96%) and balanced accuracy (95.44%). SVM (Radial) showed moderate performance but struggled with sensitivity (60.61%). Logistic and Lasso Regression performed the worst, with low balanced accuracy and sensitivity, making them less effective for imbalanced data. Lasso Regression was the weakest model, highlighting the need for techniques that handle class imbalances effectively.

4.5 Model setup with SMOTE-ENN

To enhance model performance and mitigate overfitting, the dataset underwent preprocessing and resampling, with a focus on increasing minority class representation. Initially, irrelevant columns ('Industry', 'Currency', and 'KYC.Risk') were removed.

Categorical variables, including the target variable STR.Filed, were converted to factors to ensure proper handling during analysis. The Amount variable was cleaned by removing commas and converting it to numeric format.

To effectively manage mixed categorical and numerical features, Factor Analysis for Mixed Data (FAMD) was applied. FAMD is a dimension reduction technique that is designed to handle datasets containing both numerical and categorical variables. It generates principal components that preserve key information without excessive dimensionality, allowing for a more nuanced analysis of mixed data types.

Following dimensionality reduction, Synthetic Minority Over-sampling Technique (SMOTE) was applied to synthetically augment the minority class using K=7 nearest neighbors and a duplication size of 50. These values were carefully selected to balance improved minority class representation without compromising model generalizability. Excessive duplication size led to overly optimistic model metrics (MCC and F1-score of 1) with the Random Forest, indicating potential overfitting due to excessive oversampling. Subsequently, Edited Nearest Neighbors (ENN) with K=3 nearest neighbors was used to remove noisy or ambiguous samples. Instances were excluded if fewer than half of their neighbors shared the same label, which helped clean the dataset by reducing overlap and noise that could otherwise degrade model performance.

This combined SMOTE-ENN approach optimized the trade-off between minority class representation and data quality, resulting in a dataset with 67,318 (88.64%) majority class (NO) and 8,610 (11.36%) minority class (YES) instances. The resampled data was split into training (80%) and test (20%) sets, maintaining class proportions. Finally, a 10-fold cross-validation framework was implemented during model training to further mitigate overfitting and ensure robust evaluation of model generalization.

4.6 Machine Learning after SMOTE-ENN

4.6.1 Random Forest

The Random Forest model demonstrated near-perfect classification performance with an accuracy of 99.99% and an MCC of 99.93%. Sensitivity (Recall) is 99.88%, indicating that the model correctly identified almost all positive cases. Specificity is 1.0000, meaning no false positives were recorded. The high F1-score (99.94%) confirmed the model's robustness in handling both classes effectively.

Table 4.18: Confusion Matrix for Random Forest

	Actual NO	**Actual YES**
Predicted NO	13463	2
Predicted YES	0	1719

4.6.2 Support Vector Machine (SVM)

The SVM model achieved an accuracy of 99.58% with an MCC of 97.93%. Sensitivity is 99.77%, showing strong performance in detecting positive cases. Specificity is 99.55%, indicating a small number of false positives. The F1-score was 98.17%, confirming reliable classification despite minor misclassification.

Table 4.19: Confusion Matrix for SVM

	Actual NO	**Actual YES**
Predicted NO	13403	4
Predicted YES	60	1717

4.6.3 Logistic Regression

Logistic Regression performed well with an accuracy of 99.49% and an MCC of 97.47%. Sensitivity (99.13%) suggests it correctly identified most positive cases, while specificity

(99.53%) was slightly lower than SVM and Random Forest. The F1-score of 97.77% showed strong overall performance, though slightly less effective than Random Forest.

Table 4.20: Confusion Matrix for Logistic Regression

	Actual NO	**Actual YES**
Predicted NO	13400	15
Predicted YES	63	1706

4.6.4 K-Nearest Neighbors (KNN)

KNN provided an excellent classification performance with an accuracy of 99.94% and an MCC of 99.70%. Sensitivity (99.54%) ensured minimal false negatives, and specificity (99.99%) kept false positives low. The F1-score of 99.74% indicated robust classification abilities comparable to Random Forest.

Table 4.21: Confusion Matrix for KNN

	Actual NO	**Actual YES**
Predicted NO	13462	8
Predicted YES	1	1714

As presented in Table 4.22, Random Forest proved to be the most robust model, consistently achieving high performance across all evaluation metrics both before and after applying SMOTE-ENN. Its superior sensitivity, F1 score, and Matthews Correlation Coefficient (MCC) affirm its reliability in detecting suspicious transactions.

Logistic Regression exhibited substantial sensitivity to class imbalance, recording a sensitivity of only 42.42% and an MCC of 62.87% prior to resampling. However, its performance improved significantly following the application of SMOTE-ENN, underscoring the effectiveness of resampling techniques for such classifiers. Lasso Regression also benefited considerably from SMOTE-ENN, showing notable improvements across all metrics. While its performance was initially lower than that of Logistic Regression, the post-resampling results for both models became nearly identical, indicating that SMOTE-ENN effectively mitigated performance disparities between the two.

K-Nearest Neighbors (kNN) and Support Vector Machine (SVM) demonstrated relatively stable performance, with kNN performing particularly well even before resampling.

Table 4.22: Model Performance Before and After SMOTE-ENN

Model	Dataset	Acc.	Bal. Acc.	Sens.	MCC	F1
kNN (k=5)	Before	0.9996	0.9544	0.9091	0.9089	0.9091
	After	0.9994	0.9976	0.9954	0.9970	0.9974
SVM	Before	0.9989	0.8029	0.6061	0.7418	0.7273
	After	0.9957	0.9963	0.9971	0.9791	0.9814
Random Forest	Before	0.9999	0.9697	0.9394	0.9692	0.9688
	After	0.9999	0.9994	0.9988	0.9993	0.9994
Logistic Reg.	Before	0.9985	0.7121	0.4242	0.6287	0.5833
	After	0.9949	0.9933	0.9913	0.9747	0.9777
Lasso Reg.	Before	0.9982	0.6666	0.3333	0.5305	0.4783
	After	0.9946	0.9921	0.9890	0.9735	0.9765

4.6.5 Variable Importance After SMOTE ENN

The most important variable for the Random Forest model is **Dim.2**, followed by **Dim.9** and **Dim.10**. These dimensions contributed significantly to model predictions (Table 4.23).

Table 4.23: Variable Importance from Random Forest

Variable	Importance Score
Dim.2	100.000
Dim.9	67.869
Dim.10	55.325
Dim.8	51.167
Dim.1	35.555
Dim.5	21.373
Dim.3	16.910
Dim.6	14.583
Dim.7	7.624
Dim.4	0.000

Table 4.24 highlights the contribution of each original variable to the extracted dimensions. Notably, **Score.Rank** contributed significantly to **Dim.8** (60.37%), while **Product.type** dominated **Dim.3** (55.21%).

Based on both Random Forest and Lasso Regression, Dim.2, Dim.9, and Dim.10 emerged as the most important dimensions in predicting the outcome. Additionally, the FAMD Variable Contribution analysis highlighted the key variables associated with these dimensions, such as Product.type, Score.Rank, and Business.age.

These insights confirmed the significance of these features in the dataset and reinforced their importance in model interpretation and decision-making.

Table 4.24: FAMD Variable Contributions

Variable	**Dim.1**	**Dim.2**	**Dim.3**	**Dim.4**	**Dim.5**	**Dim.6**	**Dim.7**	**Dim.8**	**Dim.9**	**Dim.10**
Amount	0.065	1.883	0.612	52.948	28.284	9.741	1.776	4.687	0.003	0.001
Business.age	33.900	0.795	1.755	1.692	0.107	0.225	0.010	17.406	35.292	8.818
Product.type	9.048	42.741	55.208	1.948	1.110	5.711	0.912	11.027	39.469	32.826
Nature.of.Transaction	23.824	24.608	2.431	1.384	0.201	1.120	2.065	5.231	0.260	38.877
Score.Rank	26.101	26.657	27.009	1.329	0.261	0.202	16.894	60.372	21.895	19.280
Time.of.Day	7.062	3.316	12.986	40.699	70.038	83.002	78.344	1.278	3.080	0.196

Table 4.25 provided a comparative summary of the most influential features identified by each model before and after the application of SMOTE-ENN. For Random Forest, the most predictive attributes prior to resampling were Business Age, Transaction Amount, and Product Type – Card. After resampling, the model relied more heavily on the FAMD-generated dimensions, particularly Dim.2, Dim.9, and Dim.10, which reflected important combinations of original variables.

Lasso Regression also demonstrated a shift in feature reliance after SMOTE-ENN. While it originally emphasized individual features such as Amount, Product Type, and Business Age, its post-resampling predictions were primarily driven by the same FAMD dimensions highlighted in the Random Forest model. This indicates a meaningful benefit from resampling and dimensionality reduction .

Interestingly, Logistic Regression exhibited several significant predictors prior to SMOTE-ENN—including Business Age and Score Rank – High Risk, but failed to identify any strongly influential variables after resampling. This suggested that transforming the input space may have reduced the impact of certain linear features, highlighting the differing sensitivities of linear and regularized models to synthetic data generation and mixed-type dimension reduction techniques.

Table 4.25: Key Predictive Features Before and After SMOTE-ENN Across Models

Method	**Before SMOTE-ENN**	**After SMOTE-ENN**
Random Forest	Business Age (100.00) Amount (27.30) Prod. Type - Card (22.28)	Dim.2 (100.00) Dim.9 (67.87) Dim.10 (55.32)
Logistic Regression	Business Age Score Rank - High Risk Time of Day - Eve. Time of Day - Morn.	*None Significant*
Lasso Regression	Amount Prod. Type - Card Business Age Score Rank - High Risk Time of Day - Night	Dim.2 Dim.9 Dim.10
FAMD Contributions: Dim.2 → Prod. Type (42.7%), Score Rank (26.7%) Dim.9 → Business Age (35.3%), Prod. Type (39.5%) Dim.10 → Nature of Trans. (38.9%), Prod. Type (32.8%)		

4.6.6 Summary of Results

This chapter presented the process and outcomes of utilizing machine learning models to detect suspicious transactions. The analysis began with rigorous data preprocessing. Variables such as Industry, Currency, and KYC Risk were excluded due to potential multicollinearity. The target variable STR.Filed was converted into a factor, alongside categorical variables such as Product Type, Nature of Transaction, Score Rank, and Time of Day. The Amount variable was cleaned and standardized, and categorical variables were transformed through one-hot encoding. The dataset was then divided into training and testing subsets using an 80:20 split while maintaining the original class distribution. Model training was conducted using 10-fold cross-validation to enhance generalizability and reduce variance.

Five classification algorithms were evaluated on the original dataset before applying any resampling technique. These included k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest, Logistic Regression, and Lasso Regression. Among them, the Random Forest model achieved the best performance, recording an accuracy of 99.99%, balanced accuracy of 96.97%, sensitivity of 93.94%, specificity of 100%, and an F1-score

of 96.88%. It also attained the highest Matthews Correlation Coefficient (MCC) of 96.92%, indicating strong predictive reliability. Variable importance analysis revealed that Business Age, Amount, and Product Type – Card Payment were the most influential features in detecting suspicious activity.

The kNN model also performed relatively well, with an accuracy of 99.96% and balanced accuracy of 95.44%. In contrast, the SVM model, while attaining high overall accuracy (99.89%), demonstrated a lower sensitivity (60.61%), indicating challenges in detecting the minority class. Logistic Regression and Lasso Regression were the least effective, with sensitivities of 42.42% and 33.33%, respectively. Logistic Regression was particularly affected by the underrepresentation of the minority class, underscoring its susceptibility to data imbalance and its reliance on sufficient examples from both classes for effective learning. Although Lasso Regression underperformed in terms of recall, it was useful for feature selection, identifying Amount and Business Age as key predictors.

To improve model performance and address class imbalance, resampling was conducted using SMOTE-ENN. Prior to resampling, Factor Analysis for Mixed Data (FAMD) was used for dimensionality reduction while preserving the integrity of both numerical and categorical variables. SMOTE generated synthetic instances of the minority class using seven nearest neighbors and a duplication rate of 50, while ENN removed noisy and ambiguous observations based on three nearest neighbors. The resampled dataset was then split, and models were retrained using 10-fold cross-validation.

Following resampling, all models demonstrated improved sensitivity and overall performance, with the Random Forest model showing the most significant gains. It achieved nearly perfect results: an accuracy of 99.99%, sensitivity of 99.88%, specificity of 100%, MCC of 99.93%, and an F1-score of 99.94%. Variable importance analysis after resampling confirmed the continued relevance of Business Age and Amount, while also highlighting the increasing significance of Product Type – Card Payment, Nature of Transaction – Wallet Transfer, and Score Rank – Low Risk as key predictors. These findings illustrate the effectiveness of SMOTE-ENN in improving minority class representation and model interpretability.

Overall, Random Forest emerged as the most robust and reliable model for detecting suspicious transaction reports (STRs), demonstrating strong classification performance and adaptability in handling imbalanced and complex financial datasets.

Chapter 5

Discussions, Conclusions and Recommendations

5.1 Introduction

This chapter presents an in-depth discussion of the study's findings, linking them to the research objectives and existing literature. It highlights key insights derived from the analysis, including the impact of data imbalance on model performance and the effectiveness of machine learning techniques in detecting suspicious transactions. Additionally, the chapter outlines the study's limitations, acknowledging areas where constraints such as data availability or methodological choices may have influenced the results. Finally, recommendations are provided for future research and practical applications, offering insights on how financial institutions and regulators can enhance fraud detection strategies in Kenya's evolving digital payment landscape.

5.2 Discussion

This study aimed to enhance the performance of machine learning models for detecting suspicious transactions within Kenya's digital payment ecosystem by incorporating the SMOTE-ENN hybrid resampling technique. In doing so, it also sought to provide comparative insights into how different models respond to class imbalance and dimensionality reduction. The discussion that follows begins with a summary of the dataset's key characteristics and imbalances, then interprets model performance before and after resampling, and finally explores the practical implications of these findings for financial crime detection.

The dataset exhibited severe class imbalance, with 99.75% of transactions labelled as "NO" and only 0.25% as "YES," indicating a low incidence of suspicious activity. Mobile money transactions dominated (75.36%), and Kenyan Shillings (98.84%) was the primary currency, highlighting the reliance on local financial systems. Outflows (68.39%) exceeded inflows (31.61%), suggesting a net movement of funds out of the system. Risk assessment showed that low-risk transactions (69.49%) were predominant, while high-risk transactions accounted for less than 1%, reflecting a largely compliant customer base. Certain industries, such as transportation (28.73%), marketing (17.10%), and telecommunications (14.07%), recorded higher transaction volumes, whereas cosmetics, retail, and insurance had minimal activity. The presence of a strong correlation between Industry and STR.Filed (0.910) suggested that specific industries were more likely to be flagged for suspicious activity. Additionally, a high association between Product Type and Currency (0.872) indicated distinct financial behaviors. However, the minimal influence of Time of Day on STR.Filed (0.049) suggested that suspicious transactions occurred consistently throughout the day.

The highly skewed distribution of transaction amounts, with extreme values, indicated the presence of large-value transactions that might have required further scrutiny. Business age was evenly distributed, with a median of 11 years, suggesting a mix of both established and newer businesses. Outliers in transaction amounts and business age appeared to represent exceptional cases rather than anomalies. These findings pointed to the need for targeted risk assessment in specific industries, enhanced monitoring of large-value transactions, and specialized fraud detection measures for mobile money platforms. The dominance of low-risk transactions suggested the effectiveness of existing compliance measures. However, the extreme class imbalance posed challenges for predictive modeling, particularly in identifying the minority class.

The results highlighted the limitations of accuracy as an evaluation metric for imbalanced datasets. While all models achieved exceptionally high accuracy (above 99.8%), this metric proved to be misleading due to the overwhelming dominance of the majority class. This aligned with findings from He and Garcia (2009), who emphasized that accuracy can create a false sense of model effectiveness in such cases. Similarly, Japkowicz and Shah (2011) argued

that alternative metrics like MCC and F1-score are better suited for assessing performance in imbalanced classification tasks, as they account for both false positives and false negatives. More reliable metrics, such as the Matthews Correlation Coefficient (MCC) and F1-score, provided a clearer picture of each model's effectiveness in distinguishing between classes.

The Random Forest model outperformed all others, achieving an F1-score of 96.88% and an MCC of 96.92%, indicating strong predictive capability for both classes. The kNN model also performed well, with an F1-score of 90.91% and an MCC of 90.89%, showing a good balance between precision and recall. In contrast, the SVM model, despite its high accuracy of 99.89%, recorded a lower F1-score (72.73%) and MCC (74.18%), reflecting its weaker ability to detect the minority class. Similarly, Logistic Regression and Lasso Regression emerged as the least effective, with MCC values of 62.87% and 53.05%, and F1-scores of 58.33% and 47.83%, respectively. Their low sensitivity (42.42% and 33.33%) further confirmed their struggle in identifying positive cases. These discrepancies between accuracy and MCC/F1-score emphasized why accuracy alone was unreliable in imbalanced datasets—it overestimated model performance by favoring the majority class.

The results demonstrated that MCC and F1-score were more robust evaluation metrics in this context, as they accounted for both false positives and false negatives, providing a more balanced assessment. This aligned with previous research, such as Chicco and Jurman (2020) which highlighted that MCC is particularly effective in imbalanced datasets as it considers all four confusion matrix elements, unlike accuracy, which can be misleading in such cases. Similarly, Powers (2011) emphasized that F1-score, while widely used, is still limited compared to MCC when dealing with severe class imbalances .

The strong performance of Random Forest suggested that it was the best model for this dataset, but the low sensitivity of other models highlighted the need for imbalance-handling techniques to improve classification outcomes. Addressing data imbalance through methods like SMOTE-ENN or cost-sensitive learning proved to be essential in enhancing model reliability and ensuring fair detection of STR filings.

Before applying SMOTE-ENN, the dataset was highly imbalanced, with the minority class representing only 0.25% of all transactions.

After resampling, this proportion increased to approximately 11.36%, significantly improving the model's sensitivity to suspicious transactions. This resampling ratio was deliberately chosen , not to achieve a 50-50 class balance, but to optimize model performance while preserving the natural distribution of financial transaction data. A perfectly balanced dataset could have introduced overfitting or unrealistic synthetic patterns, reducing the model's generalizability. By maintaining a more representative 88.64 – 11.36 class ratio, the models were better equipped to detect rare but critical cases without compromising data integrity or realism.

The Random Forest model achieved an MCC of 99.93% and an F1-score of 99.94%, indicating that it effectively distinguished between fraudulent and non-fraudulent transactions. Similarly, kNN recorded an MCC of 99.70% and an F1-score of 99.74%, demonstrating its robustness in classification. SVM and Logistic Regression also showed notable performance improvements, with MCC values of 97.93% and 97.47%, and F1-scores of 98.17% and 97.77%, respectively.

These findings implied that Random Forest was highly effective in handling class imbalance while maintaining high predictive accuracy. The model's near-perfect classification performance suggested that ensemble methods were particularly useful in learning complex patterns within structured financial datasets.

This aligned with findings by Aburbeian and Ashqar (2023), who implemented the Random Forest algorithm on a credit card transaction dataset while addressing class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). Their enhanced Random Forest classifier achieved an accuracy and F1-score of approximately 98%, demonstrating its proficiency in distinguishing between fraudulent and non-fraudulent transactions .This study adopted SMOTE-ENN and achieved near perfect results (MCC 99.93%) implying that hybrid resampling techniques can potentially lead to better metrics than individual ones. Similarly, research by Ye et al. (2019) explored the detection of fraudulent financial statements using Random Forest combined with SMOTE. Their study showed that Random Forest outperformed other models, including Artificial Neural Networks and Support Vector Machines, in handling imbalanced datasets and accurately identifying fraudulent activities .

These studies reinforce the notion that ensemble methods like Random Forest are particularly effective in financial fraud detection, as they can learn complex patterns and mitigate the impact of class imbalance.

Additionally, the strong performance of kNN, SVM, and Logistic Regression highlighted the potential for alternative models in different scenarios. kNN's ability to classify cases with high precision suggested that distance-based learning methods could be valuable, particularly when computational efficiency was not a constraint. SVM's performance, with an MCC of 97.93%, indicated that boundary-based classification remained a viable approach, especially in cases where feature interactions were complex. Although Logistic Regression performed slightly lower than the other models, its MCC of 97.47% and F1-score of 97.77% implied that even simple linear models benefited significantly from data balancing techniques.

Another critical implication was the role of Factor Analysis for Mixed Data (FAMD) in handling categorical and numerical variables before resampling. The high importance of Dim.2, Dim.9, and Dim.10 suggested that dimensionality reduction techniques preserved essential information while improving model interpretability. This finding was particularly useful for organizations dealing with large, heterogeneous datasets, as it demonstrated that FAMD effectively transformed data without losing predictive power.

Finally, these results emphasized the need for businesses, regulators, and analysts to prioritize proper data preprocessing, particularly in fraud detection systems, anti-money laundering (AML) monitoring, and financial crime prevention. Implementing robust balancing techniques and selecting appropriate performance metrics, such as MCC and F1-score, could lead to more effective risk management strategies. The study reinforced that handling data imbalance was not just a technical improvement but a crucial step in ensuring fairness, accuracy and reliability in predictive analytics.

5.3 Conclusion

This study identified Random Forest as the most effective model for detecting suspicious transactions in Kenya's digital payments landscape, particularly after addressing severe class imbalance. Before resampling, accuracy was misleading, with models struggling to detect the minority class. After applying SMOTE-ENN, all models improved significantly, with Random Forest achieving the highest MCC (99.93%) and F1-score (99.94%). The findings underscored the importance of using MCC and F1-score for imbalanced datasets and demonstrated that the hybrid resampling technique, SMOTE-ENN, enhances fraud detection.

These insights can help financial institutions strengthen risk management and fraud prevention strategies.

5.4 Limitations of the Study

This study had several limitations. First, the dataset was limited to Kenyan digital payments and focused on transactions processed by a single payment service provider. As such, the findings may not be fully generalizable to other financial ecosystems with different transaction patterns or fraud typologies.

Second, fraud techniques evolve rapidly as perpetrators adapt to detection mechanisms. Consequently, models trained on historical data may become less effective over time. This necessitates regular model retraining and ongoing adaptation to new fraud patterns to sustain detection accuracy.

Lastly, while best practices such as Factor Analysis of Mixed Data (FAMD), data splitting, and K-fold cross-validation were applied to reduce overfitting, the near-perfect model performance suggests there may still be residual overfitting. Further validation using external datasets is recommended to confirm the models' generalizability and robustness.

5.5 Recommendations for Future Studies

Future research should focus on applying hyperparameter tuning techniques to mitigate overfitting and enhance model generalization. Exploring advanced ensemble learning methods such as XGBoost, LightGBM, or model stacking could yield performance improvements over the Random Forest baseline. Additionally, evaluating alternative hybrid resampling strategies and cost-sensitive learning approaches may further improve the detection of minority classes in imbalanced datasets.

Researchers are also encouraged to investigate how different resampling distributions affect model performance. This can be achieved through simulation studies that systematically vary class ratios and sampling techniques to assess their impact on predictive accuracy and stability. Such simulations would provide deeper insights into the robustness of resampling methods under varying data conditions.

Incorporating deep learning-based feature extraction or embedding techniques for mixed data types could enhance the model's ability to capture complex, non-linear relationships. Finally, validating models on diverse financial datasets beyond the Kenyan digital payments ecosystem is recommended for assessing their generalizability across different transaction environments and fraud typologies.

References

(2021). Anti-money laundering (aml) fines of 2021. https://sanctionscanner.com/blog/anti-money-laundering-aml-fines-of-2021-561. Accessed: 2024-06-25.

Aburbeian, M. and Ashqar, H. (2023). Random forest model for credit card fraud detection using smote. *arXiv preprint arXiv:2303.06514*.

Alotibi, J., Almutanni, B., Alsubait, T., Alhakami, H., and Baz, A. (2022). Money laundering detection using machine learning and deep learning. *International Journal of Advanced Computer Science and Applications*, 13(10):732–738.

Anggraeni, D. and Harahap, S. N. (2023). Utilization of machine learning to detect the possibility of suspicious financial transactions. *Fair Value: Jurnal Ilmiah Akuntansi dan Keuangan*, 5(8):3277–3283.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In *International Conference on Database Theory*, pages 217–235. Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Calderon, M. et al. (2025). Digital payments and their role in enhancing financial transactions efficiency. *International Journal of Economics and Financial Issues*, 15(1):183–190. Accessed: 2025-05-21.

Central Bank of Kenya (2023). Central Bank of Kenya Licensees Sectorial Risk Assessment. Cbk report, Central Bank of Kenya.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Eastern and Southern Africa Anti-Money Laundering Group (ESAAMLG) (2024). Second enhanced follow-up report and 1st technical compliance re-rating report: Kenya. Published in August 2024.

ESAAMLG (2022). Anti-money laundering and counter-terrorist financing measures - kenya, second round mutual evaluation report. Technical report, ESAAMLG, Dar es Salaam. Accessed: 2024-06-25.

Eshiwani, M. M. (2020). *Detecting financial crimes using pattern recognition techniques: case of mobile money transactions*. PhD thesis, Strathmore University.

Europol (2020). Financial and economic crime. https://www.europol.europa.eu/crime-areas-and-statistics/crime-areas/financial-crime. Accessed May 2025.

Financial Action Task Force (2023). Money laundering. https://www.fatf-gafi.org/en/topics/money-laundering.html. Accessed May 2025.

Fix, E. and Hodges Jr., J. L. (1951). Discriminatory analysis. nonparametric discrimination: Consistency properties. Technical Report Project 21-49-004, Report No. 4, USA Air Force School of Aviation Medicine, Randolph Field, Texas.

Gitonga, J. T. (2018). *Fraud detection using machine leaning: a comparative analysis of neural networks & support vector machines*. PhD thesis, Strathmore University.

Government of Kenya (2023). The Proceeds of Crime and Anti-Money Laundering Regulations, 2023. Accessed: 2025-05-21.

Gyamfi, N. K. and Abdulai, J.-D. (2018). Bank fraud detection using support vector machine. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 37–41. IEEE.

Hairani Hairani, H. and Dadang Priyanto, D. (2023). A new approach of hybrid sampling smote and enn to the accuracy of machine learning methods on unbalanced diabetes disease data. *A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data*, 14(8):585–890.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Intelligence, S. G. M. (2020). Kenya fines 5 lenders for violating anti-money laundering rules. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/kenya-fines-5-lenders-for-violating-anti-money-laundering-rules-57434329. Accessed: 2024-06-25.

Japkowicz, N. and Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.

Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., and Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, 23(1):173–186.

Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., and Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert systems with applications*, 100:234–245.

Kulkarni, S. (2023). Machine learning for anti-money laundering (ml for aml). https://kpmg.com/be/en/home/insights/2023/08/lh-machine-learning-for-anti-money-laundering.html. Accessed: 2024-06-20.

Kumar, A., Das, S., and Tyagi, V. (2020). Anti money laundering detection using naïve bayes classifier. In *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 568–572. IEEE.

Kumar, P., Bhatnagar, R., Gaur, K., and Bhatnagar, A. (2021). Classification of imbalanced data: review of methods and applications. In *IOP Conference Series: Materials Science and ngineering*, volume 1099, page 012077. IOP Publishing.

Kumar, S., Gunjan, V. K., Ansari, M. D., and Pathak, R. (2022). Credit card fraud detection using support vector machine. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*, pages 27–37. Springer.

Lamari, M., Azizi, N., Hammami, N. E., Boukhamla, A., Cheriguene, S., Dendani, N., and Benzebouchi, N. E. (2021). Smote–enn-based data sampling and improved dynamic ensemble selection for imbalanced medical data classification. In *Advances on Smart and Soft Computing: Proceedings of ICACIn 2020*, pages 37–49. Springer.

Lim, K. S., Lee, L. H., and Sim, Y.-W. (2021). A review of machine learning algorithms for fraud detection in credit card transaction. *International Journal of Computer Science & Network Security*, 21(9):31–40.

Liu, C., Chan, Y., Alam Kazmi, S. H., and Fu, H. (2015). Financial fraud detection model: Based on random forest. *International Journal of Economics and Finance*, 7(7):178–188.

Liu, Y., Tang, Z., and Zheng, W. (2019). Suspicious bank card transaction recognition based on k-means clustering and random forest algorithm. In *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pages 332–336. IEEE.

Minastireanu, E.-A. and Mesnita, G. (2019). An analysis of the most used machine learning algorithms for online fraud detection. *Informatica Economica*, 23(1):5–16.

Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., and Idroes, R. (2023). Credit card fraud detection for contemporary financial management using xgboost-driven machine learning and data augmentation techniques. *Indatu Journal of Management and Accounting*, 1(1):29–35.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Shihembetsa, E. et al. (2021). *Use of artificial intelligence algorithms to enhance fraud detection in the Banking Industry*. PhD thesis, University of Nairobi.

Šikman, M. M. and Grujić, M. (2021). Relationship of anti-money laundering index with gdp, financial market development, and human development index. *NBP. Nauka, bezbednost, policija*, 26(1):21–33.

Ye, M., Xiang, G., and Gan, X. (2019). Fraud detection in financial statements using machine learning methods. *IOP Conference Series: Materials Science and Engineering*, 612(5):052051.

Appendix A

Similarity Report

Suspicious Transaction Prediction in Kenyan Digital Payments A Machine Learning Comparative Study with Imbalanced Data.pdf

ORIGINALITY REPORT

17%	13%	14%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Indian Institute of Management, Bangalore Student Paper	2%
2	rstudio-pubs-static.s3.amazonaws.com Internet Source	1%
3	Submitted to Strathmore University Student Paper	1%
4	Submitted to University of New South Wales Student Paper	1%
5	Submitted to Queen's University of Belfast Student Paper	1%
6	eitca.org Internet Source	1%
7	Submitted to University of Utah Student Paper	1%
8	doctorpenguin.com Internet Source	<1%
9	www.spglobal.com Internet Source	<1%
10	www.coursehero.com Internet Source	<1%
11	Submitted to La Trobe University Student Paper	<1%

		<1 %
12	journalofcmsd.net Internet Source	<1 %
13	Submitted to University of Wales Institute, Cardiff Student Paper	<1 %
14	repository.universitasbumigora.ac.id Internet Source	<1 %
15	iccs.ac.in Internet Source	<1 %
16	Submitted to London Business School Student Paper	<1 %
17	Submitted to University of Northumbria at Newcastle Student Paper	<1 %
18	Submitted to University of Limerick Student Paper	<1 %
19	Submitted to Monash University Student Paper	<1 %
20	Deb, Dipok. "Application and Analysis of Machine Learning and Deep Learning Algorithms in Detection of DDoS Cyberattacks", The University of Texas Rio Grande Valley, 2024 Publication	<1 %
21	www.mdpi.com Internet Source	<1 %
22	Submitted to Australian National University Student Paper	<1 %

23	Wellington Pinheiro dos Santos, Juliana Carneiro Gomes, Valter Augusto de Freitas Barbosa. "Swarm Intelligence Trends and Applications", CRC Press, 2022 Publication	<1 %
24	Submitted to Loughborough University Student Paper	<1 %
25	ia802900.us.archive.org Internet Source	<1 %
26	Ines Belgacem. "Whistleblowing Disclosure as a Shield Against Earnings Management: Evidence from the Insurance Sector", Journal of Risk and Financial Management, 2025 Publication	<1 %
27	Submitted to University of Teesside Student Paper	<1 %
28	etheses.whiterose.ac.uk Internet Source	<1 %
29	Submitted to University of Sydney Student Paper	<1 %
30	Dinesh Goyal, Bhanu Pratap, Sandeep Gupta, Saurabh Raj, Rekha Rani Agrawal, Indra Kishor. "Recent Advances in Sciences, Engineering, Information Technology & Management - Proceedings of the 6th International Conference “Convergence2024” Recent Advances in Sciences, Engineering, Information Technology & Management, April 24–25, 2024, Jaipur, India", CRC Press, 2025 Publication	<1 %
31	Strickett, Mark. "Logistic Regression Methods Versus Machine Learning Techniques in	<1 %

	Status and Severity Prediction of South African Covid-19 Laboratory Test Data", University of the Witwatersrand, Johannesburg (South Africa), 2025 Publication	
32	Submitted to University of Adelaide Student Paper	<1 %
33	Submitted to University of Minnesota System Student Paper	<1 %
34	Submitted to Liverpool John Moores University Student Paper	<1 %
35	kitasato.repo.nii.ac.jp Internet Source	<1 %
36	Submitted to University of Hertfordshire Student Paper	<1 %
37	Submitted to Taylor's Education Group Student Paper	<1 %
38	Shahab Emaani, Abbas Saghaei. "Driver Anomaly Detection in Cargo Terminal", Heliyon, 2024 Publication	<1 %
39	Submitted to University of San Diego Student Paper	<1 %
40	Submitted to Swinburne University of Technology Student Paper	<1 %
41	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International	<1 %

	Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	
42	Muskan Garg, Sunghwan Sohn. "CareD: Caregiver's Experience with Cognitive Decline in Reddit Posts", 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), 2023 Publication	<1 %
43	Submitted to University of Durham Student Paper	<1 %
44	Submitted to University of Fort Hare Student Paper	<1 %
45	Submitted to University of Strathclyde Student Paper	<1 %
46	eprints.utm.my Internet Source	<1 %
47	Submitted to Birkbeck College Student Paper	<1 %
48	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	<1 %
49	Submitted to University of Canberra Student Paper	<1 %
50	saeb.feaa.uaic.ro Internet Source	<1 %
51	www.bio-conferences.org Internet Source	<1 %
	www.journal.ikopin.ac.id	

52	Internet Source	<1 %
53	Mabrouka Salmi, Dalia Atif, Diego Oliva, Ajith Abraham, Sebastian Ventura. "Handling imbalanced medical datasets: review of a decade of research", Artificial Intelligence Review, 2024 Publication	<1 %
54	Abdelaziz Testas. "Distributed Machine Learning with PySpark", Springer Science and Business Media LLC, 2023 Publication	<1 %
55	B. Sundaravadivazhagan, Sekar Mohan, Balakrishnaraja Rengaraju. "Recent Developments in Microbiology, Biotechnology and Pharmaceutical Sciences - International Conference on Recent Development in Microbiology, Biotechnology and Pharmaceutical Science", CRC Press, 2025 Publication	<1 %
56	P.V. Mohanan. "Artificial Intelligence and Biological Sciences", CRC Press, 2025 Publication	<1 %
57	academic-accelerator.com Internet Source	<1 %
58	arxiv.org Internet Source	<1 %
59	heca-analitika.com Internet Source	<1 %
60	Mitja Steinbacher, Matej Steinbacher, Matjaz Steinbacher. "Using CNN to Model Stock Prices", Computational Economics, 2025 Publication	<1 %

61	Palese, Michael E.. "Landslide Hazard Assessment Framework for Cut-Slopes Along Railroad Rights-of-Way Using Statistical Analysis of Images", University of Delaware, 2023 Publication	<1 %
62	Submitted to Binus University International Student Paper	<1 %
63	Sai Kiran Oruganti, Dimitrios A Karras, Srinesh Singh Thakur, Janapati Krishna Chaithanya, Sukanya Metta, Amit Lathigara. "Digital Transformation and Sustainability of Business", CRC Press, 2025 Publication	<1 %
64	www.econjournals.net.tr Internet Source	<1 %

Exclude quotes	Off	Exclude matches	< 25 words
Exclude bibliography	Off		

Appendix B

Ethical Clearance Confirmation

21st March 2025

Ms Swaleh Jihan,
jihan.abdulrazak@strathmore.edu

Dear Ms Swaleh,

RE: Suspicious Transaction Prediction in Kenyan Digital Payments: A Machine Learning Comparative Study with Imbalanced Data

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2694/25.** The approval period is from **21st March 2025 to 20th March 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
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 SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

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

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu