A HYBRID MACHINE LEARNING MODEL FOR DETECTING AND PREVENTING CORRUPTION IN KENYA'S PUBLIC PROCUREMENT CONTRACTS

MELCHIZEDEK LEWELA NDOLO

A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND IT IN THE SCHOOL OF COMPUTING AND MATHEMATICS (SCM) IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN CYBER SECURITY OF THE CO-OPERATIVE UNIVERSITY OF KENYA

2025

DECLARATION

Declaration by the candidate

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

Signature..... Date … 21st November 2025

Melchizedek Lewela Ndolo

Reg. No. C005/600021/2023

Declaration by the supervisors

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

Signature..... Date … 21st November 2025

Dr. Anthony Wanjoya

Lecturer

Department of Computer Science and Information Technology

The Co-operative University of Kenya

Signature..... Date … 21st November 2025

Dr. Philemon Kasyoka

Lecturer,

Department of Computer Science and Information Technology

South Eastern Kenya University

DEDICATION

I dedicate this study to God for His guidance, my late father, my supportive wife and children and my mother whose passion for education continually inspires me.

ACKNOWLEDGEMENT

I thank the Almighty God for His blessings and favor that has enabled me to achieve this academic milestone. I appreciate the support and guidance of my academic supervisors Dr. Anthony Wanjoya and Dr. Philemon Kasyoka. I acknowledge the invaluable resources provided by my institution The Co-operative University of Kenya. I extend my gratitude to my colleagues and peers for their constructive feedback and encouragement throughout this process. Lastly, I am grateful to my family for their unwavering support and motivation that has enabled me complete this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
AUC-ROC	Area Under the Curve – Receiver Operating Characteristic
CR	Corruption Risk
EACC	Ethics and Anti-Corruption Commission
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
IFMIS	Integrated Financial Management Information System
ISO	International Organization for Standardization
K-Means	K-Means Clustering Algorithm
KNN	K-Nearest Neighbors Algorithm
MDAs	Ministries, Departments and Agencies
ML	Machine Learning
NLP	Natural Language Processing
OLS	Ordinary Least Squares
PPRA	Public Procurement Regulatory Authority
RF	Random Forest
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-Sampling Technique
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
SVM	Support Vector Machine
UN	United Nations

DEFINITION OF TERMS

Anomaly Detection	A machine-learning technique used to identify unusual patterns or deviations from normal procurement activities that may indicate fraud or corruption.
Bid Rigging	A form of procurement fraud where suppliers collude to manipulate bidding outcomes, leading to unfair contract awards.
Bidding Irregularities	Unusual procurement practices such as single-bid tenders, repeated awards to the same supplier, or artificially high bid prices.
Collusion	A secret agreement among bidders or procurement officials to manipulate contract awards for personal or group benefit.
Contract Value Discrepancies	Significant differences between estimated and actual awarded contract amounts, often signaling fraudulent activity.
Corruption Likelihood	The probability that a procurement transaction is corrupt, as determined by the machine-learning model.
Cybersecurity	Measures taken to protect procurement data and digital systems from unauthorized access, breaches, and manipulation.
K-Means Clustering	An unsupervised machine-learning algorithm used to group procurement transactions based on similarities, helping identify hidden fraud patterns.
Random Forest	A machine-learning algorithm that enhances fraud detection by combining multiple decision trees to improve prediction accuracy.
Risk Scoring	A machine-learning method of assigning a corruption risk score to procurement transactions using key fraud indicators.

Single Sourcing	The practice of awarding contracts to a single supplier without competitive bidding, often increasing corruption risk.
Supplier History	Records of a supplier's past procurement activities, including contract awards, fraud allegations, and compliance history.
Supervised Learning	A machine-learning approach where models are trained on labeled data to classify procurement transactions as corruption-prone or non-corrupt.
Unsupervised Learning	A machine-learning method used to detect anomalies and hidden patterns in procurement data without predefined labels.
Vendor Collusion	Fraudulent agreements between suppliers to manipulate procurement outcomes, often by inflating prices or restricting competition.

ABSTRACT

This study investigated the use of a hybrid machine learning model to detect and prevent corruption in Kenya's public procurement contracts. Persistent procurement irregularities continue to undermine fiscal accountability, inflate contract prices and weaken governance despite having an established legal framework. To address these challenges, the study adopted a quantitative, experimental and data-driven research design, analyzing a substantially expanded dataset of 10,214 procurement transactions obtained from public data sources such as the Public Procurement Regulatory Authority Annual Procurement Reports, the National Treasury/IFMIS Open Contracting Portal, the Kenya Open Data Initiative, the Office of the Auditor-General audit reports and the EACC National Ethics and Corruption Survey datasets. The dataset incorporated structured fields (bid amounts, procurement method, number of bidders, supplier identifiers, award timelines) and textual indicators extracted from audit narratives and tender justifications. The study adopted a quantitative, data-driven and experimental research design integrating supervised and unsupervised machine learning. Logistic regression and random forest models were trained alongside K-Means clustering to detect hidden patterns of fraud. Data preprocessing using Python libraries involved cleaning, deduplication, normalization, missing-value imputation and natural language processing for extracting red-flag terms from unstructured reports. Model robustness was ensured through cross-validation and an 80/20 train-test split. Results showed that single-bid tenders, prior supplier allegations, persistent award extensions and contract value discrepancies were the strongest predictors of corruption likelihood. The random forest classifier achieved an AUC of 0.93, precision of 0.89 and recall of 0.86, while logistic regression recorded an AUC of 0.81 with strong interpretability value for policy justification. Unsupervised clustering successfully isolated high-risk contract groups hence validating the model's anomaly detection capability in partially labelled environments. The findings indicated that integrating machine learning with cybersecurity principles such as data integrity checks, access control significantly enhanced the detection and prevention of corruption in procurement processes. The study concluded that corruption in public procurement was measurable and predictable using data-driven techniques. It recommended the integration of the hybrid model into Kenya's Integrated Financial Management Information System and e-procurement platforms to facilitate real-time fraud detection and risk-based auditing. The study further recommended that procurement officers be trained on the use of predictive analytics and explainable artificial intelligence tools to interpret risk alerts effectively. Overall, the research contributed a practical, cybersecurity-enhanced analytical framework that strengthened transparency, accountability and governance in Kenya's public resource management system.

CHAPTER ONE

1.0 Introduction

Public procurement corruption is a worldwide issue, which negatively impacts economic growth and governance. It causes inefficiency in delivery of services, high costs of the contract and misallocation of resources. Governments worldwide struggle to offer fair and transparent procurement processes and corruption is among the biggest impediments to economic development particularly in developing countries (Guarnieri & Gomes, 2019). Corruption in the public procurement sector consumes up to 20-30% of the entire contract value, which significantly impacts government budgets and undermines confidence among individuals in the institutions (World Bank, 2024).

Kenya has not been lucky to be left out of this problem and frauds in procurements, bribery, rigging of bids and favoritism are still casting a dark cloud on the economic stability. In spite of regulatory acts like the Public Procurement and Asset Disposal Act (2015) and regulatory bodies like the Public Procurement Regulatory Authority (PPRA), corruption continues through the use of unclear procedures, ineffective enforcement and monitoring. Conventional anti-corruption techniques, including manual audits and post-contract reviews are not effective in detecting fraudulent acts in real time, which has necessitated novel, data-driven solutions (Osei-Kyei & Chan, 2019).

1.1 Background of the Study

Public procurement corruption is a long-standing problem that is common in both the developed and developing economies. Procurement fraud is estimated to cost governments billions of dollars a year worldwide with high-profile cases in the United States, the European Union and parts of Asia (Transparency International, 2023). Procurement corruption in developing economies such as in Africa worsens inequality, slows down infrastructure

developments and diverts government funds to other expenditures. High profile corruption scandals in procurement have been recorded in countries such as South Africa, Nigeria and Brazil and this is indicative of the prevalence of the issue.

Public procurement is a huge percentage of government expenditure in Kenya but it is also among the most exposed sectors to corruption. According to the Ethics and Anti-Corruption Commission (EACC) reports, which are a part of the national cybersecurity enforcement ecosystem, and the Auditor General, irregularities that are common in the country include exaggerated contract costs, single-sourced tender and politically influenced awards. The Transparency International Corruption Perceptions Index (2023) ranks Kenya 126th among 180 countries, which demonstrates that the country has high perceived corruption, especially in the transactions in the public sector. In addition, the World Bank (2024) predicts that corruption during procurement results in the average increase of the cost of 20-30%, which is a serious burden on the national budget.

Although laws and regulations, including the Public Procurement and Asset Disposal Act (2015) and the creation of procurement control agencies, including the PPRA and the EACC, have been enacted, enforcement is still a problem. Poor institutional frameworks, political interference, absence of transparency in tendering processes and poor monitoring mechanisms (KPMG, 2023) have perpetuated fraudulent activities. Conventional methods of fighting corruption, like manual audits, public complaints systems and compliance-based inspections, are slow, resource-intensive and reactive instead of proactive.

Due to the constantly changing nature of procurement fraud, which has become increasingly complex and advanced in its methods, there is an urgent demand to develop more advanced and technology-driven solutions that would allow real-time monitoring and detecting fraud. Cybersecurity analytics and machine learning methods promise a hopeful solution by processing large volumes of data, detecting anomalies and notifying of high-risk transactions

before funds are misplaced (Bertot et al., 2016; Yumame, 2024). These models have the potential to determine procurement trends, identify signs of fraud and pro-active intervention measures.

1.2 Statement of the Problem

Corruption in Kenya's public procurement system remains a persistent challenge that undermines effective public spending and weakens confidence in government institutions. Despite the establishment of a strong legal and regulatory framework, procurement processes continue to exhibit recurring irregularities, including single-bid tenders, inflated contract prices, repeated awards to the same suppliers and unexplained variations between estimated and awarded contract values. These practices contribute significantly to financial losses, with procurement-related corruption estimated to raise costs by 20–30% in a sector that accounts for nearly 40% of national expenditure (World Bank, 2022; PPRA, 2021). Kenya's position of 126 out of 180 in the Transparency International Corruption Perception Index (2023) further illustrates the continuing vulnerability of the system.

Existing oversight mechanisms dominated by manual audits, compliance checks and post-procurement investigations remain slow, labour-intensive and reactive. They often identify irregularities only after funds have been spent, leaving limited opportunity for early intervention or the prevention of financial losses. Meanwhile, public procurement generates extensive datasets across platforms such as IFMIS, PPRA and the Office of the Auditor-General, yet these datasets are rarely analysed systematically to detect suspicious patterns before contracts are executed.

Although international research increasingly demonstrates the usefulness of machine-learning techniques in identifying anomalies, collusive behaviour and high-risk procurement transactions, their application within Kenya's public sector has been limited. Studies that have

attempted to apply such models often rely on small datasets, narrow modelling techniques or purely structured data, which restrict their ability to capture the complex nature of procurement irregularities. Furthermore, most existing models lack real-time monitoring capability and do not incorporate information contained in unstructured audit narratives, which often hold critical contextual insights.

These shortcomings highlight the need for a more comprehensive and adaptive analytical approach capable of drawing on both structured and unstructured procurement data, identifying emerging fraud patterns and supporting early detection. Developing and evaluating a hybrid machine-learning model offers an opportunity to strengthen the country's capacity to proactively identify corruption-prone transactions and enhance the integrity of the procurement system.

1.3 Research Objectives

1.3.1 General Research Objective

The general objective of the study is to develop a hybrid machine-learning model to identify and prevent corruption in Kenya's public procurement contracts.

1.3.2 Specific Research Objectives

i. To analyze key risk indicators of corruption in public procurement contracts in Kenya.
ii. To develop a hybrid predictive machine-learning model for detecting corruption-prone procurement contracts.
iii. To verify the hybrid cybersecurity-enhanced predictive machine learning model for detecting corruption-prone procurement contracts.
iv. To assess the effectiveness of the developed hybrid model in preventing corruption by integrating it into the public procurement process.

1.4 Research Questions

i. What are the key risk indicators of corruption in public procurement contracts in Kenya?
ii. How can a cybersecurity-enhanced predictive machine learning model be developed to detect corruption-prone procurement contracts?
iii. To what extent is the developed hybrid machine learning model effective in accurately identifying corruption-prone procurement contracts?
iv. How effective is the integration of the hybrid predictive model into the public procurement process in preventing corruption?

1.5 Significance of the Study

The research is important in that it directly responds to the long-running problem of corruption in the Kenyan public procurement processes that cost the country a large amount of financial resources every year and adversely affected the confidence of people in the governance. The research offers a Specific solution to the problem of identifying and preventing corrupt activities in procurement contracts by combining machine learning with cybersecurity principles. In contrast to the conventional methods, which are based on manual audits and post-incident analysis, the present study presents a proactive approach that will increase the ability of governmental institutions to detect abnormalities in real-time. This is in line with the national objectives of Kenya in vision 2030, which focus on enhancing transparency, accountability and efficient utilization of the public resources.

The study has a Measurable impact because it provides practical results, including a predictive machine learning model enhanced with cybersecurity and cybersecurity recommendations. The stakeholders, including the Public Procurement Regulatory Authority (PPRA), Ethics and Anti-Corruption Commission (EACC), are part of the national cybersecurity enforcement ecosystem and other government agencies that will find these tools helpful in enhancing procurement

monitoring. The aims and deliverables are Achievable within the context of the current research, since they make use of available public procurement data and current cybersecurity frameworks. The research is Relevant to the ongoing digital revolution in governance, which has witnessed the introduction of e-procurement systems that, although convenient, are susceptible to manipulation and data leaks. Moreover, the study is Timely, considering the world tendency of attaining Sustainable Development Goal 16, which aims at establishing effective, accountable and transparent institutions (United Nations, 2015). Besides practical application, the study helps fill the gap in the academic discourse by linking technology and governance. It provides an example of how other countries can improve machine learning to solve similar problems by showing how it can be used to improve public administration and cybersecurity.

1.6 Scope of the Study

The study is restricted to procurement in the public sector, and in this case, the focus is on contracts issued by the national and county governments. The exclusion of the private sector procurement and non-governmental organizations are intended to keep a focused strategy in line with the regulatory and governance structures of Kenya. In Kenya, the legal and institutional frameworks of procurement are very specific, so it is possible to conduct an analysis of procurement records in a standardized regulatory setting. Since corruption risks and procurement systems vary greatly between the public and the private sectors, the limitation of the study to the public sector guarantees the consistency of the data analysis and policy recommendations.

The period in the study is 10 years between 2014 and 2024 to give a detailed analysis of procurement activities over a long period. The timeframe was chosen to reflect modern tendencies in the field of public procurement, such as the adoption of financial technology, regulatory changes and alterations in governance policy. The 10-year span of the coverage

makes the dataset large enough to train and validate machine-learning models and capture the latest procurement practices. The short term can also hamper the ability of the model to extrapolate the detection patterns of corruption whereas a very long time can create some forms of obsolete procurement practices that are not relevant in the current governance systems.

The study is based on secondary data found in open-source locations, such as government procurement portals, regulatory agencies, such as the Public Procurement Regulatory Authority (PPRA), anti-corruption organizations, such as the Ethics and Anti-Corruption Commission (EACC), and national cybersecurity enforcement ecosystem, and public audit reports. These sources contain verifiable and structured procurement records such as award of contracts, bidder records and records of expenditures. The secondary data used is objective and eliminates ethical issues that might arise when primary data is collected on sensitive government contracts. As well, the use of publicly available procurement data enables scalability of the proposed model, whereby oversight agencies can modify the model without having to access confidential records. Nevertheless, secondary data has its limitations in terms of data completeness and accuracy because certain procurement anomalies are not necessarily reported.

The research involves the use of cybersecurity-enhanced predictive machine learning models, such as decision trees, logistic regression, random forests and anomaly detection, to process procurement data and identify the signs of corruption. Predictive modeling is chosen because it has the capacity to handle large volumes of data, unravel concealed trends and create real-time fraud detection information. Manual audit and compliance checks are traditional anti-corruption tools that are usually reactive and do not reveal complex fraud schemes. In comparison, predictive analytics allows risk evaluation in advance, which can empower regulatory authorities to reveal suspicious procurement operations before the fraudulent transactions are completed. The paper is concerned with both supervised and unsupervised

machine learning methods, which are excluded because of their complexity, high computing costs, and possible interpretability issues.

1.7 Limitations of the Study

The research is based solely on secondary data, which, although a key source of large-scale analysis, has its inherent limitations, including possible underreporting, missing records and biases in publicly reported procurement data. The quality and the wholeness of this data are determined by the openness and reporting culture of the government agencies and regulators. Since certain fraudulent procurement practices might not be recorded in writing or might be hidden intentionally, the dataset might not be able to reflect all cases of corruption. This drawback may have an impact on the depth of the insights of the predictive model and the capability to identify some types of fraud.

The period of the study is ten years (2014-2024), which is chosen to determine the current outlook of the procurement trends and regulatory changes. This period however fails to take into consideration long-term historical procurement practices or corruption pattern shifts prior to 2014. Since procurement policies, governance arrangements and economic circumstances keep changing with time, a long dataset may offer more information on pattern of systemic procurement fraud. Nevertheless, the period considered in the study should be the last ten years, which will guarantee applicability of the results to the present public sector procurement control.

The other major constraint is that the effectiveness of predictive model is very sensitive to data quality. Any errors, discrepancy or fraudulent reporting of any figures in the dataset may affect the performance of the model, thus resulting in false positive or negative fraud detection. Machine learning models need quality, structured data to perform optimally, but anomalies in procurement records, including wrong contract values, overlapping records, or absence of bidder information, may decrease the quality and dependability of the model.

Legal and ethical limitations also limit the extent of this research, as some of the classified procurement records are not readily available. Although publicly available data provides a great wealth of information, certain cases of corruption are associated with confidential contracts, politically sensitive transactions, or classified government procurement deals that cannot be analyzed. This limitation can prevent the thoroughness of fraud detection because the main signs of procurement corruption may not be reported as per the data protection policy and confidentiality provisions. However, despite these limitations, the research is compliant with the ethical principles of research, because it is based solely on the utilization of the public documents.

In addition, this paper primarily deals with predictive modeling, which is supervised learning algorithms such as decision trees, logistic regression and random forests. Although these approaches are quite effective to determine the propensity of corruption in procurement information, they are not the entirety of machine learning tools. The other methods like unsupervised learning and, perhaps can feed more information especially in the detection of new or novel fraud patterns. Nevertheless, the complexity of computation makes the application of these advanced techniques inapplicable to this research. The focus on predictive modeling provides a guarantee of practical application and integration into the current procurement management systems.

1.8 Delimitations of the Study

This research paper is limited to the Kenyan public procurement system, namely national and county government contracts. This alternative would make sure that it is aligned with legal and institutional structures of Kenya in the general scope of the public procurement and this would permit an organized examination of regulatory imprisonment, hazards and restraints of corruptness. And the left outs are the non-governmental organizations and the private sector procurement whose procurement operations are subdued through different frameworks and are

not as so broadly deemed by the same controlling rigs as the state contracts. This analysis focuses on the government procurement and hence the collection and analysis is homogenous and applicable in such a way that the findings can be directly applied into the policy and regulatory relief in the Kenyan public sector.

Moreover, the study relies solely on the secondary data that is publicly available, and not confidential or internal procurement data. This limit ensures adherence to ethical research practices and legal constraints of data access and enables scaling in fraud detection systems. Although confidential information could provide deeper insight into the procurement corruption, it is restricted in access and poses challenges related to legal compliance and data privacy. The openness, which is ensured by the publicly available records, ensures that the study is transparent and that the methodology can be replicated in future research.

The study also involves cybersecurity, but the research is not sophisticated in security issues such as block chain and quantum encryption. These technologies can potentially enhance the transparency of procurement and combat fraud, but their implementation would require significant infrastructural investment and the regulatory authorities to adopt them, which falls outside the scope of this work. Instead, the initiative is invested in real-world cybersecurity measures linked to data protection and fraud prevention, such that the proposed model could be adopted in the current procurement landscape of Kenya. The study is practical in that it is focused enough on predictive modeling within the constraints of publicly available data and the existing procurement regulations, and it is aware of the areas that can be covered in future research.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Introduction

The chapter is a literature review on how technology, specifically machine learning and cybersecurity, can be used to fight corruption in public procurement systems. The review starts with a discussion of the relevant theories that form the basis of the study, the analysis of the past studies on corruption in public procurement, the use of machine learning in detecting frauds and the role of cybersecurity in protecting procurement systems. The chapter ends with the identification of the research gaps and the ways in which this study will fill them.

2.2 Theoretical Framework

The theoretical framework forms the basis of the explanation of the process of corruption in the context of public procurement and the application of machine learning to identify and prevent it. There are two theories applicable to this study, which are Fraud Triangle Theory and Governance Theory.

2.2.1 Fraud Triangle Theory

In this research, the indicators of corruption risk are determined by applying the Fraud Triangle Theory, developed by Donald R. Cressey in 1953. This theory was proposed later in Cressey, Other People: A Study of the Social Psychology of Embezzlement to clarify the psychological and circumstantial influences that prompt individuals to engage in fraud. Due to Cressey, fraud is unwanted with pressure, with opportunity, and with the rationalization of three conditions that cannot exist together (Fitri et al., 2019). With the public procurement scenario, both pressures could be had in the form of financial distress or political coercion, opportunity could be had in the form of lax supervision, and rationalization could be had in the form of the need to exist within a misplaced system where corruption becomes a prerequisite.

Other areas in which the Fraud Triangle Theory has become common are in auditing, accountancy and forensic investigation (Kagias et al., 2022). The true followers of the theory are the forensic accountants and auditors who use the theory to draw the commonality of fraud activities in organizations. It has been effectively used in detecting irregularities in financial transactions and procurement procedures. The theory, nevertheless, is not devoid of constraints. It is albeit a very obviously described model of how individual motivations can be defined, but is still more behavioral and ignores the systemic or cultural forces in totality that may facilitate corruption. Further, it takes into account the fact that fraud is never an impromptu activity, and this is not necessarily the case (Homer, 2019).

In response to these criticisms, the theory can be expanded to incorporate systemic and cultural variables, such as organizational culture or weaknesses in regulation. Moreover, it may be supplemented with emerging technologies such as data analytics and machine learning, which can trace the trends of putting pressure, finding an opportunity, and rationalizing it in big data. The theory today finds common application with systems and tools of fraud detection where it forms a basis of algorithms to seek anomalies in procurement and financial records. Its application of technology has rendered it to be a pillar in the search to discover the risk signs of corruption, which is a direct aid to the initial aim of this research (Jensen & Meckling, 1979).

2.2.2 The Principal-Agent Theory

The second goal of the creation and justification of a predictive model to detect corruption is supported by the principal- Agency theory, which was proposed by Michael C. Jensen and William H. Meckling in 1976. This theory solves a conflict within an interest, which arises when a principal (e.g., government or procurement authority) employs an agent (e.g., contractor or procurement official), to execute tasks, and the agent might be acting in their self-interest rather than in the interest of the principal. The theory lays more focus on the issue of

information asymmetry in which the agent possesses more information than the principal and has an opportunity to exploit it to commit corruption.

The Principles and Agency Theory has been extensively used in governance, economics and finance to develop a system that matches interests between the principals and the agents. It also ensured enlightened policies in order to minimize corruption during government procurement by closing information asymmetry and provision of incentives to conduct ethical business. The theory is also valuable but limited (Farasoo, 2021). It operates on the same assumption that all the agents are rational and that their driving force is their selfish interest without taking into account the existence of altruism or a moral drive. It also fails to deal sufficiently well with systemic or institutional factors that can assist in augmenting the threat of corruption.

To overcome this drawback, the theory could be strengthened, adding the concepts of behavioral economics, which stipulates this complexity of motivations (Hausken, 2019). The Principal-Agency Theory has only gained momentum due to the recent events in the application of machine learning, which offers methods of studying the large amounts of procurement-related data in order to identify irregularities and lessen the information asymmetry. These innovations are in direct line with the second objective since they enable the development of predictive models that can identify the contracts at risk of corruption and consequently address the challenges outlined by the theory.

2.2.3 The Governance Theory

The theoretical basis of the analysis of the effectiveness of the machine learning model is based on the Governance Theory, which was developed by Oliver E. Williamson and Ronald Coase in the 1980s. Governance Theory places more importance on accountability, transparency and institutional observance in making ethical decisions and avoidance of corruption. It evolved in response to the inefficiency of organizational and institutional governance, especially the

alignment between rules and procedures as well as the congruity with moral and operating norms (Schillemans & Bjurstrøm, 2020).

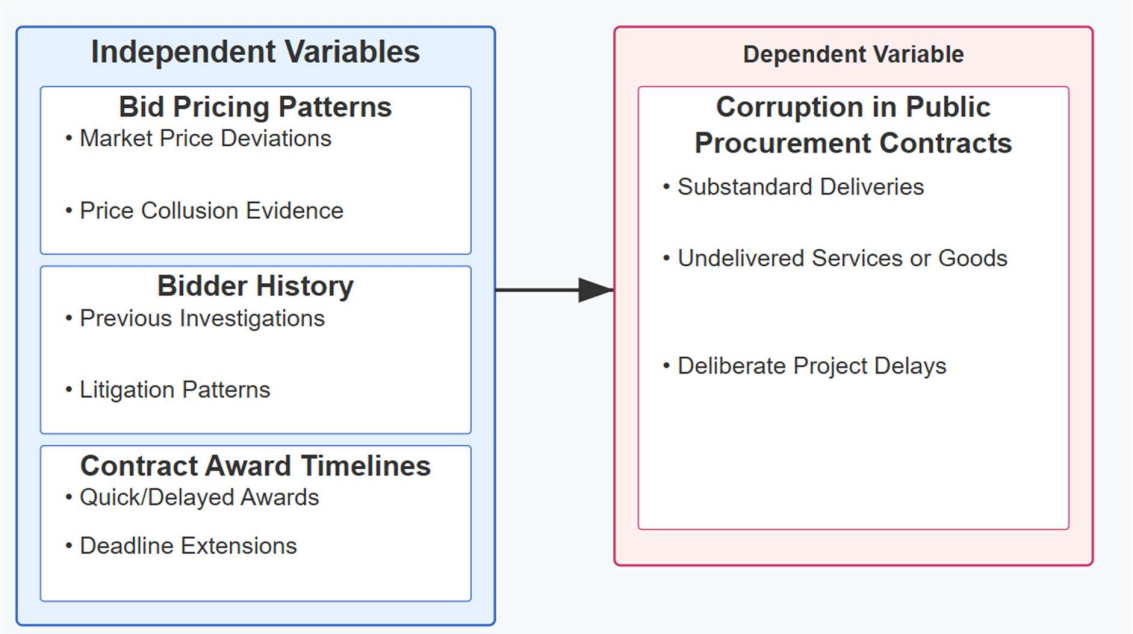
Scholars in the field of studying public administration and other development practitioners have implemented Government Theory to enhance accountability and control within the state and society sector. The theory has also played significant roles in coming up with anti-corruption mechanisms especially in service such as in the public procurement where transparency and accountability is very essential (Chen et al., 2022). It also has the merit of prioritizing institutional reforms and technological resources to facilitate good governance. Some people have also criticized it as being excessively idealistic in the sense that it is combating corruption by simply having rules and regulations. It does not believe much in cultural and contextual differences in form of governance either (Bennett & Satterfield, 2018). It may be possible that the Governance Theory was able to deal with these criticisms by encompassing more practical and contextual approaches, including mechanisms such as instituting governance reforms to adapt to the socio-political environment. The theory has managed to adjust itself to the modern world due to technological developments like e-governance and machine learning (Bebchuk & Hirst, 2019). By integrating these technologies, the governance systems will be able to trace the procurement process in real time, enhance transparency and reduce corruption risks. These advancements directly aid the third goal because they will ensure that the machine learning model is not only effective in detecting corruption, but also aligned with larger governance goals to mitigate it.

2.3 Conceptual Framework and Conceptual Review

The conceptual framework is shown in the diagram below that depicts the relationship between the independent variables and the dependent variable. Independent variables appear in the left side and dependent variable in the right side. The arrow shows the connection between these

variables which imply that the independent variables can be applied to predict or determine risk factors of the dependent variable.

Figure 2.3.1: Conceptual Framework



2.3.1 Independent Variables

The study examines three independent variables bid pricing patterns, bidder history and contract award timelines as potential predictors of corruption in public procurement. Bid pricing patterns capture the extent to which submitted bid prices deviate from expected market values and established competitive norms. These patterns are assessed through percentage variations from market benchmarks, statistical attributes such as standard deviation and skewness, and clustering tendencies observed across similar contracts. Additional indicators of price manipulation, including repeated identical bid submissions, rotational bidding sequences and high price correlation among competing firms over time, further support the identification of potential collusion.

The second independent variable, bidder history, evaluates a supplier's behavioural profile and integrity record. This is measured through the number and severity of formal investigations over the past decade, the recency of these investigations and patterns of litigation associated with the bidder. Litigation-related indicators include the frequency of legal disputes, settlement amounts and the bidder's success rate in procurement-related cases. Recurring investigations or frequent legal disputes may signal behavioural tendencies associated with elevated corruption risk.

The third independent variable, contract award timelines, focuses on the temporal characteristics of procurement decisions. Quick or unusually delayed awards are examined through differences between bid submission and award dates, deviations from average award durations for comparable contracts and the frequency of awards falling outside established procedural norms. Deadline extensions are assessed by the number of extensions issued, the total duration of time added and the proportion of extended contracts relative to sectoral averages. Irregular timelines—whether accelerated or protracted—may reflect procedural manipulation to favour specific suppliers or obscure non-compliant actions.

2.3.2 Dependent Variable

The dependent variable, corruption in public procurement, is operationalised as manipulation of contract execution and measured across substandard deliveries, undelivered goods or services and deliberate project delays. Substandard deliveries are assessed through the proportion of outputs failing quality inspections, deviations from contractual specifications, documented quality complaints and remediation costs required to correct identified deficiencies. Undelivered goods or services are captured through the percentage of contracted items not supplied, the monetary value of missing deliverables, unauthorized substitutions and inconsistencies in delivery verification documentation. Deliberate project delays are evaluated

by comparing actual completion times with contractual obligations, analysing missed milestones lacking justification, examining notification patterns and estimating the financial impact of time overruns. Together, these metrics provide a comprehensive assessment of corruption manifested through compromised contract execution.

2.4 Empirical Review

The empirical review looks at previous literature that has utilized machine learning and data-driven methods in the detection of corruption in public procurement. The current literature has shown that supervised models (e.g., logistic regression, decision trees) and unsupervised methods (e.g., anomaly detection, clustering) are effective to detect fraud patterns. Nevertheless, a large number of studies consider only individual indicators of corruption without combining transparency or developing models that are applicable to the specifics of developing countries where the quality of data and the implementation mechanisms may differ. The review not only gives an insight on the current methodologies but also identifies gaps in current methodologies and sets the context of the present study in terms of the machine learning model of procurement fraud detection, which is context specific in Kenya.

2.4.1 Key risk indicators of corruption in public procurement contracts in Kenya using machine learning techniques.

The need to identify major warning signs of corruption in the public procurement will help bolster the transparency and accountability in the public sector spending. Although there are a number of studies that have utilized machine learning to identify procurement fraud, the current models are still being challenged by accuracy, adaptability, and contextual relevance. The proposed research will seek to fill these gaps by creating a more robust and context-aware machine learning model specific to the procurement system in Kenya.

A study by Decarolis and Giorgiantonio (2022) used random forests to examine the corruption risk in the procurement system in Italy, with 87% accuracy in identifying fraudulent contracts. Their model identified single-bid contracts, repeated awards of contracts to the same vendor and short bidding duration as good predictors of corruption. Although efficient, this method depended on organized procurement datasets to a great extent and could not identify subtle patterns of fraud in unstructured data, including audit reports and legal documents. This work is better than them in that they combine both structured and non-structured data, which provides a more detailed system to detect fraud.

On the same note, Imhof and Wallimann (2021) applied clustering and anomaly detection methods, including k-means clustering and isolation forests, to detect bid-rigging coalitions in various auction types. Their model was 82 percent accurate but had a high false positive rate, which was hard to implement. Besides, it paid more attention to the price abnormalities and collusion among bidders without considering the risk indicators of governance and policy. This research expands on their work with interpretability improvements to models and regulatory compliance functionality to minimize false positives to make fraud detection more practical to Kenyan procurement authorities.

In Basdevant et al. (2022), logistic regression and support vector machines (SVMs) were used to analyze the procurement fraud in Ghana with SVMs attaining 85% accuracy on identifying high-risk contracts. According to their model, non-competitive bidding, large changes in the contract and frequent direct procurement were found to be drivers of corruption. Nevertheless, it was based on historical procurement records and did not allow it to identify fraudulent activities in real time. This research goes a step further to incorporate real-time anomaly detection features, thus allowing proactive response to fraudulent transactions before they are carried out.

2.4.2 Develop and validate a cybersecurity-enhanced predictive machine learning model for detecting corruption-prone procurement contracts.

Modifications to improve predictive machine learning models with cybersecurity to detect corruption-prone procurement contracts has attracted significant interest, with it having been demonstrated that these models mitigate the shortcomings of traditional anti-corruption systems. Unlike manual auditing and systems with strict rules, machine learning models can handle a vast amount of data, identify intricate fraud patterns, and deliver real-time and data-driven information. However, despite the potential of machine learning demonstrated in the past in detecting procurement fraud, most models have been marred by data limitations, inability to fit and validate the model based on context, which this study aims to address by developing a more robust, interpretable and locally applicable model to the Kenyan procurement setting.

Lima et al. (2023) created a model of corruption red flags identification in public procurement that is automatically trained on decision trees and support vectors machines (SVMs) using contract-level data. Their models have been able to identify critical risk factors, including single-bid tenders, recurring contract awards to vendors and major deviations of cost estimates with an accuracy of 83% in identifying high-risk contracts. Their model, however, was mostly based on structured procurement data, which lacks key red flags of fraud in audit reports and legal documents. Their method is improved by this study, which uses both structured and unstructured data to increase the detection accuracy.

Similarly, Ezeji (2024) examined how can be used to anticipate fraud in the context of public procurement by using large datasets that included factors like vendor history, bid pricing trends and award timelines. Their neural network model performed better than conventional statistical methods with a precision rate of 89 percent and provided demonstration of the benefits of in identifying procurement fraud. Nevertheless, they frequently lack interpretability, and

policymakers and procurement officers find it hard to comprehend the way fraud risk scores are created. This paper fills this gap by integrating with explainable AI methods, to achieve both high predictive power and interpretability to effectively implement the policies.

Satri et al. (2024) created a machine learning model to identify risks of corruption in regional development projects in the African environment through logistic regression and random forests. They found that short bidding periods, contract amendments and financial irregularities are some of the corruption risk indicators in their model and were accurate at 85 percent validated against historical fraud cases. Their method gives a useful regional point of view, but because they rely on historical instances of corruption, they cannot identify new patterns of fraud as fast as they would with a more real-time approach. Their work is developed in this study by incorporating real-time anomaly detection and adaptive learning mechanisms to enable fraud protection to be proactive, not reactive.

In spite of these developments, there are a number of challenges in the development and validation of predictive fraud detection models. Data fragmentation and accessibility is one of the issues. Procurement records in Kenya are usually incomplete, irregular, or manual and it is hard to train credible machine learning models (Transparency International, 2021). In addition, biases in the training data, e.g. underreporting of fraud cases, may result in incorrect predictions, possibly missing systemic corruption, or incorrectly labeling a particular vendor due to historical biases. As an example, any algorithm trained on biased data might not include corruption outside of recorded procurement processes (Chassang et al., 2022).

In order to address these shortcomings, this research employs a multi-source data integration model, which involves integrating e-procurement systems, financial management platforms (including IFMIS) and external audit reports to increase data completeness. Further, explainable AI methods are utilized so that machine learning predictions can be transparent and understandable, allowing policymakers and anti-corruption organizations to trust and

implement model-based fraud risk predictions. To ascertain the robustness of the model, cross-validation methods, testing on independent data and real-life case studies are used to be sure that the model is generalizable, accurate and fair in the context of public procurement in Kenya.

2.4.3 Effectiveness of the developed model in preventing corruption by integrating it into the public procurement process.

The effectiveness of machine learning models in preventing corruption in the public procurement is a subject of growing interest. The technical implementation of models is not enough to achieve successful integration but rather alignment with governance structures, institutional frameworks and anti-corruption policies. Studies have revealed that the integration of machine learning in the procurement management can improve transparency, minimize frauds and enhance compliance. Nevertheless, the current methods suffer disadvantages of low adoption rates, data quality constraints, and algorithmic biases, and this paper aims to bridge this gap by creating a context-sensitive and interpretable fraud detection model in the Kenya procurement system.

De Menezes et al. (2023) tested the risk estimation with machine learning in Brazilian public procurement and discovered that incorporating predictive models in e-procurement systems decreased single-bid contracts by 23 percent and enhanced compliance in procurement by 31%. Nevertheless, their model was largely based on structured procurement data, which could not allow them to identify corruption indicators in textual reports and legal documents. Their method is enhanced through the combination of structured and unstructured data into their study where it makes fraud detection more accurate and contextual.

On the same note, Asnana et al. (2023) examined how intelligent forecasting models can be used to detect procurement fraud and discovered that the introduction of machine learning to identify high-risk tenders led to the detection of 30% fewer procurement anomalies. However, they also observed that AI-generated risk scores were misunderstood by procurement officers

in 18% of cases, resulting in fraud detection failures or unwarranted inquiries. This research will overcome this limitation by integrating explainable AI systems, where risk alerts are clear and readily understandable by the officials in the procurement department in Kenya.

In the African scenario, Osei-Kyei and Chan (2019) evaluated predictive analytics applications within the Ghanaian public procurement market and discovered that automated fraud detection models enhanced non-compliance detection by 27% and decreased tender processing delays by 19%. Nevertheless, the research pointed out that model effectiveness was constrained unless there were institutional changes and procurement officer training. This research expands their results by integrating the developed machine learning model into Kenya procurement oversight systems, where risk assessment translates into effective policy responses instead of passive detection of frauds.

Even in the face of these successes, there are still challenges in the implementation of machine learning in the procurement process of the people. Technological change is often resisted particularly by procurement officials who might be either lacking in technical expertise or may fear greater accountability. Furthermore, reliable procurement data is critical for accurate predictions, yet developing countries like Kenya often face data fragmentation and limited system interoperability (World Bank, 2020). Additionally, biases in machine learning models have resulted in false positives in up to 21% of fraud risk cases, potentially undermining trust in AI-generated fraud alerts.

2.5 Summary of Literature and Research Gaps

Table 1 presents a summary of key global and African studies on the application of data-driven and machine learning techniques in detecting corruption within public procurement systems. The comparison highlights methodological approaches, primary findings, and limitations that inform the direction of this study. While previous research demonstrates, notable success in identifying corruption risk indicators—particularly through supervised and unsupervised

learning gaps remain in the integration of unstructured data, real-time detection mechanisms, interpretability for policy adoption and contextual adaptation for developing countries such as Kenya. By addressing these shortcomings, the present study advances the literature by developing a hybrid, interpretable and context-specific machine learning model capable of detecting and preventing procurement fraud in real time within Kenya's public sector.

Table 2.5.1: Summary of Literature and Research Gaps

Author(s) & Year	**Focus Area**	**Key Findings**	**Research Gaps**
Decarolis & Giorgiantonio (2022)	Italy – Corruption red flags in public tenders	Random Forest flagged bid anomalies with 87% detection accuracy on structured data.	Limited use of unstructured data; lacks contextual adaptation for low- and middle-income countries (LMICs).
Imhof & Wallimann (2021)	Europe – Bid rigging detection via clustering	K-Means and Isolation Forest effective in detecting coalitions with 82% precision.	High false positives; minimal governance integration.
Basdevant et al. (2022)	Ghana – SVM & regression on procurement fraud	Identified key red flags (non-competitive bidding, direct awards); SVM achieved 85% accuracy.	Historical data focused; lacked real-time detection capabilities.
Satri et al. (2024)	Africa – Regional project corruption prediction	Used logistic regression; bid timing & modifications	Context-aware but lacked real-time detection mechanisms.

Author(s) & Year	Focus Area	Key Findings	Research Gaps
		were predictive; model achieved 85% accuracy.	
Ezeji (2024)	Nigeria – AI model for procurement fraud	Achieved 89% precision using neural networks.	Interpretability issues; black-box nature limited usability in policy contexts.
Lima et al. (2023)	Brazil – Automatic extraction of corruption red flags	Decision Trees and SVMs identified key indicators (single-bid tenders, repeated awards, cost deviations) with 83% accuracy.	Relied mainly on structured data; excluded unstructured audit/legal data.
De Menezes et al. (2023)	Brazil – ML-based risk estimation in public procurement	Integration into e-procurement reduced single-bid contracts by 23% and improved compliance by 31%.	Depended on structured data; lacked analysis of textual and unstructured evidence.
Osei-Kyei & Chan (2019)	Ghana – Predictive analytics in procurement	Improved non-compliance detection by 27% and reduced delays by 19%.	Limited institutional reforms and officer training reduced long-term impact.

Author(s) & Year	Focus Area	Key Findings	Research Gaps
Iravonga et al. (2023)	Kenya – IFMIS and financial oversight	Digitization improved transparency and reporting in procurement.	Did not apply AI/ML; lacks predictive fraud detection tools.
Institute of Economic Affairs (2022)	Kenya – Procurement Risk Index	Identified systemic vulnerabilities in procurement practices.	Descriptive, not predictive; lacks model-based validation.

2.5.1 Research Knowledge Gap

Machine learning has proven effective in detecting fraudulent procurement practices, with studies demonstrating its ability to identify anomalies in public contracts. Titl et al. (2019) used supervised learning models, including logistic regression and decision trees when flagging suspicious transactions, whereas Mazrekaj et al. (2021) employed random forests and when detecting collusive bidding. On the same note, Decarolis and Giorgiantonio (2020) trained a predictive fraud classification model using structured financial account information and De Witte et al. (2019) used unsupervised methods such as clustering to identify bid rigging. Nevertheless, the majority of these models work in isolation, and they are only interested in fraud detection and do not consider the larger governance frameworks. In addition, they were designed in high-income nations with formal procurement data, which restricts their relevance to the Kenyan procurement system (World Bank, 2024).

There are serious limitations to the existing models. They mainly use structured data like the value of contracts and frequency of bids and ignore unstructured data like audit reports that

hold key corruption clues (De Witte et al., 2019). Moreover, most models lack the capability to identify fraud in real-time, which is why they are less efficient in avoiding fraud transactions (Decarolis & Giorgiantonio, 2020). The procurement problems in Kenya that also lead to the further obstacles to the direct implementation of such models are also the incomplete records, the weak enforcement and low transparency (World Bank, 2024).

The paper addresses these gaps by providing a context-specific machine learning model of the procurement system in Kenya. It will combine structured and unstructured information, such as procurement contracts and textual reports on corruption, to yield a holistic risk of fraud (Titl et al., 2019). Such a hybrid approach, combining unsupervised (anomaly detection, clustering, network analysis) and supervised (decision trees, logistic regression) techniques, could be more effective in fraud detection (Mazrekaj et al., 2021). Real-time anomaly detection will also characterize the model, allowing oversight agencies to place a red flag on suspicious transactions before they can be carried out (De Witte et al., 2019).

Systemic vulnerabilities like political interference and data transparency, specific to the Kenyan procurement context, will also be incorporated into the model. The research is applicable in terms of the Kenya-specific datasets provided by the institutions, such as the Public Procurement Regulatory Authority (PPRA) and the Ethics and Anti-Corruption Commission (EACC) as part of the national cybersecurity enforcement ecosystem (World Bank, 2024). Additionally, in addition to the detection of frauds, the study will evaluate the effects of transparency programs on the reduction of corruption, which will be later used to provide a policy solution on how to enhance the control of procurement (Decarolis & Giorgiantonio, 2020). By filling these important gaps, the intended study would design a strong, real-time and data-intensive model of corruption detection in the Kenyan government procurement scheme.

CHAPTER THREE

3.0 METHODOLOGY

3.1 Introduction

This chapter presents the research methodology used to develop and validate the machine-learning model for detecting corruption in Kenya's public procurement system. It briefly describes the research design, target population, sampling procedures, data sources and preprocessing steps. The chapter also outlines the operationalization of variables, the analytical techniques applied—including descriptive statistics, inferential analysis and machine-learning methods—and the ethical considerations observed throughout the study. Together, these elements provide a structured framework that guided the empirical investigation.

3.2 Research Philosophy

This study is anchored on the positivist research philosophy, which emphasises objective measurement, empirical observation and the use of scientific methods to explain and predict phenomena. Positivism assumes that reality is stable, observable and quantifiable, making it suitable for studies that rely on numerical data and statistical modelling. In the context of this research, corruption in public procurement is examined through measurable indicators such as pricing patterns, bidder behaviour and contract execution outcomes. The use of machine-learning techniques further aligns with positivist principles, as these methods rely on empirical regularities, algorithmic pattern detection and validation through statistical performance metrics. By adopting a positivist stance, the study ensures that findings are grounded in verifiable data, reproducible analytical procedures and objective interpretation, thereby enhancing the reliability and generalisability of the results within Kenya's public procurement environment.

3.3 Research Design

The proposed study will use a quantitative, data-driven and experimental research design to design and test machine learning methods to detect and stop corruption in the Kenyan system of public procurement. Since the study is computational in nature, the research is based on the supervised and unsupervised machine learning models combined with the cybersecurity risk detection to process the procurement data to determine the indicators of corruption risks and provide predictive information.

The actual process of the experiment starts with the data collection, which is done with the use of Kenya-specific datasets available in public sources. These are procurement documentation of the Public Procurement Regulatory Authority (PPRA), the Kenya Open Data Initiative and the National Treasury of Kenya which are very useful in terms of awarded contracts, bidding procedures and suppliers information. Furthermore, data on anti-corruption and governance is added, including reports on Transparency International Kenya and survey data regarding the Global Corruption Barometer, which will shed some light on the trends of corruption and its perception by the population. Synthetic or anonymized data can be used where needed to support model training and increase the predictive accuracy.

After the data is received, machine learning models are constructed and trained both in the supervised and unsupervised methods. Prediction of corruption-prone contracts is done using supervised models, including decision trees, random forests, logistic regression and , depending on past procurement trends. Moreover, the unsupervised learning models, such as anomaly detectors and clustering algorithms such as K-Means and DBSCAN, is used to detect concealed patterns of fraud and anomalies in the procurement transactions.

In order to achieve the reliability and accuracy of the developed models, strong validation methods are used. This will consist of K-fold cross-validation and holdout validation and will aid in the evaluation of the model generalization on new data. The predictive strength of each

model is measured in performance metrics including precision, recall, F1-score and the AUC-ROC curve. Moreover, the efficiency of the models are verified in terms of an experimental assessment by comparing the predictions with the historical cases of corruption and the irregularities in the procurement reported by the oversight agencies. This will measure the effectiveness of the models in identifying the already familiar cases of fraud, which will in the end decide their feasibility in the practical context of procurement activities.

3.4 Study Area

The study focuses on Kenya's public procurement system, which encompasses all national and county government ministries, departments, agencies (MDAs) and state corporations operating under the Public Procurement and Asset Disposal Act (PPADA), 2015. Kenya provides an appropriate and relevant study area because public procurement represents one of the largest components of government expenditure, accounting for approximately 40 percent of the national budget. This makes it a critical sector for examining corruption risks and evaluating the effectiveness of data-driven oversight mechanisms.

Additionally, Kenya has established several digital procurement platforms—such as the Integrated Financial Management Information System (IFMIS), the Public Procurement Information Portal (PPIP) and the Public Procurement Regulatory Authority (PPRA) reporting systems—which generate extensive structured and unstructured data necessary for the development of machine-learning models. These systems offer publicly accessible, verifiable datasets covering procurement notices, tender awards, supplier information, contract implementation reports and audit findings.

The choice of Kenya as the study area is further justified by persistent reports of procurement-related corruption from oversight bodies including the Office of the Auditor-General (OAG), the Ethics and Anti-Corruption Commission (EACC) and PPRA. These recurring irregularities highlight the need for innovative analytical approaches capable of identifying risk patterns and

supporting early intervention. Therefore, Kenya's rich procurement data environment, combined with its ongoing governance challenges, makes it an ideal setting for evaluating the applicability and effectiveness of a hybrid machine-learning model for corruption detection.

3.5 Target Population

The target population for this study consisted of all public procurement transactions recorded within Kenya's national and county government systems between 2015 and 2023. These transactions were captured through key government procurement platforms, including the Integrated Financial Management Information System (IFMIS), the Public Procurement Information Portal (PPIP) and the annual procurement reports published by the Public Procurement Regulatory Authority (PPRA). Together, these sources documented procurement activities undertaken by ministries, departments and agencies (MDAs), state corporations and county governments. Based on consolidated records obtained from these platforms, the estimated size of the population was approximately 10,214 procurement transactions completed during the nine-year period under review.

The procurement transactions within this population displayed diverse characteristics. They varied in contract values, supplier identities, historical supplier performance and levels of competition, and they reflected the use of different procurement methods such as open tendering, restricted tendering, direct procurement and emergency procurement. They also included time-based information such as award durations, deadline extensions, completion timelines and contract execution outcomes. In addition, many transactions were accompanied by narrative audit observations from the Office of the Auditor-General and compliance reviews from oversight agencies, providing valuable qualitative context to complement the structured fields.

This population was considered appropriate for the study because it offered a comprehensive representation of Kenya's public procurement environment, where recurring audit findings and

oversight reports had continued to highlight irregularities and governance challenges. The dataset covered a broad range of procuring entities and procurement categories, making it sufficiently representative for identifying patterns associated with corruption risks. Furthermore, the availability of verifiable, multi-year data across different levels of government enhanced the robustness of the machine-learning model by providing adequate variability for training and validation. The target population therefore formed a suitable foundation for examining the determinants of corruption in procurement and for developing a hybrid analytical model capable of detecting anomalous purchasing behaviour.

The sampling frame used in the study reflected the distribution of procurement transactions across national government MDAs, state corporations, county governments and high-risk procurement categories. Table 3.1 presents the sampling frame.

Table 3.5.1: Sampling Frame for Procurement Transactions (2015–2023)

Category	Number of Transactions	Description
National Government MDAs	3,842	Ministries, departments and agencies reporting through IFMIS and PPRA
State Corporations	2,116	Parastatal procurement entries, including capital-intensive projects
County Governments	3,128	County-level procurement of goods, works and services
Special Procurement Categories (Emergency/Direct)	1,128	Non-competitive procurement categories prone to irregularities
Total Population	**10,214**	Consolidated procurement records used for the study

3.6 Sampling Design

This study adopted a sampling design that ensured the selected procurement transactions were representative of Kenya's public procurement environment between 2015 and 2023. The sampling design consisted of two key components: sample size determination and the sampling method used to extract transactions from the target population.

To determine an appropriate sample size, the study applied **Yamane's (1967) formula**, commonly used in quantitative research to draw a manageable yet statistically reliable sample from a large population. Given a total population of 10,214 procurement transactions and a 95% confidence level with a 5% margin of error, the required sample size was calculated as follows:

$$n = \frac{N}{1 + N(e^2)}$$

Where:

- n = required sample size
- N = population size (10,214)
- e = margin of error (0.05)

$$n = \frac{10,214}{1 + 10,214(0.05^2)} = \frac{10,214}{1 + 10,214(0.0025)} = \frac{10,214}{26.535} \approx 385$$

Accordingly, a sample of 385 procurement transactions was deemed sufficient to provide statistically meaningful insights while remaining feasible for detailed analysis, preprocessing and model development.

The study employed a stratified random sampling method, which was suitable because public procurement transactions arise from different categories of procuring entities—national government MDAs, state corporations, county governments and special procurement categories such as direct or emergency procurement. Stratification ensured that each category contributed proportionately to the sample, thereby preserving the structural diversity of the procurement system. Within each stratum, transactions were selected randomly to minimize selection bias and enhance representativeness.

Stratified random sampling was further justified by the heterogeneity of procurement transactions. Contract values, procurement methods, supplier histories and execution characteristics vary significantly across procurement categories. A stratified approach allowed each group of transactions—especially high-risk categories such as direct procurement—to be

adequately represented in the dataset used to train and validate the machine-learning model. This enhanced the model's capacity to detect a wide range of corruption-related patterns.

Overall, the sampling design provided a statistically grounded and methodologically robust framework for selecting procurement transactions for analysis. By integrating sample size determination, stratification and random selection, the study ensured that the dataset used in model development was both representative and suitable for rigorous machine-learning evaluation.

3.7 Data Collection

This study relied entirely on secondary data, drawn from publicly accessible and verifiable government procurement repositories in Kenya. The data were obtained from the Integrated Financial Management Information System (IFMIS) Open Contracting Portal, the Public Procurement Information Portal (PPIP), annual procurement reports published by the Public Procurement Regulatory Authority (PPRA), audit reports from the Office of the Auditor-General (OAG), and corruption-related records from the Ethics and Anti-Corruption Commission (EACC). These platforms provided structured procurement records—including procurement methods, bid prices, supplier information, timelines and contract values—as well as unstructured audit narratives used to identify execution anomalies and potential corruption indicators. The use of secondary data was appropriate because procurement activities in Kenya are documented digitally, centrally stored and continuously updated, ensuring the availability of comprehensive and reliable datasets.

Data collection was carried out using researcher-developed extraction tools, including automated Python scripts and secure data-scraping routines designed to download, clean and consolidate procurement records from the identified platforms. These tools enabled efficient extraction of large datasets while preserving file integrity and metadata. For unstructured data, natural language processing (NLP) parsers developed by the researcher were used to extract

risk-related terms from audit reports, contract descriptions and investigation summaries. The design of these instruments followed cybersecurity best practices, including secure access protocols, encrypted storage formats and version-controlled environments to ensure that the data remained confidential, unaltered and protected from unauthorized access.

The use of automated tools was justified because manual extraction of procurement data is time-consuming, error-prone and unsuitable for large-scale datasets required in machine-learning studies. Automated scripts improved accuracy and repeatability while minimizing human bias. Similarly, NLP-based extraction was necessary to process unstructured audit information that cannot be captured through standard numerical procedures. Because the study integrated both structured and unstructured data, multiple instruments were required—each tailored to the format and characteristics of the specific data type.

To ensure validity, the study relied solely on official government data sources mandated by law to publish procurement information. These datasets undergo institutional verification by PPRA, OAG and EACC before public release, which enhanced content and construct validity. Furthermore, triangulation across multiple repositories allowed cross-checking of procurement information, ensuring completeness and consistency. Reliability was ensured by using standardized extraction processes, reproducible scripts and consistent preprocessing procedures, all executed in a secure computing environment. Cybersecurity principles—such as data integrity checks, hashing, audit trails and controlled access—were applied to prevent tampering or loss of data, further strengthening the reliability of the collected datasets.

Overall, the data collection process combined legally recognized public procurement records with secure, researcher-developed extraction instruments. This approach ensured that the data used in the study were credible, verifiable, protected and suitable for the development of a cybersecurity-enhanced machine-learning model for detecting corruption in Kenya's public procurement system.

3.8 Data Collection Procedures

After developing the data extraction instruments and confirming their validity and reliability, the researcher followed a systematic procedure to collect the required procurement data from official government platforms. The process began with obtaining the necessary research approvals, including a research permit from the National Commission for Science, Technology and Innovation (NACOSTI). This permit authorized access to publicly available procurement datasets and ensured the study complied with Kenya's research regulatory framework. Because the study relied exclusively on secondary data, research assistants were not required; however, technical support personnel were consulted to confirm secure access protocols and compliance with cybersecurity requirements when interacting with government data platforms.

Data collection was conducted in several steps. First, the researcher identified all relevant data repositories, including the IFMIS Open Contracting Portal, the Public Procurement Information Portal, PPRA Annual Reports, and published audit reports from the Office of the Auditor-General and EACC. Secure access sessions were then established to download procurement transactions, ensuring that all files were obtained through encrypted channels such as HTTPS to prevent interception or modification during transfer.

Second, the researcher deployed the validated automated Python-based extraction scripts to systematically download, consolidate and verify the procurement datasets. These scripts were executed within a controlled computing environment equipped with security features such as access authentication, encrypted storage and automated log generation. This ensured that each extraction attempt was traceable and that the integrity of the raw data was preserved. Parallel NLP-based tools were used to extract relevant information from unstructured audit narratives. All files were stored in a secure, access-restricted directory protected by password authentication and encryption to prevent unauthorized access or tampering.

Third, data integrity checks were conducted to confirm that all downloaded records were complete and unaltered. This involved verifying dataset sizes, confirming metadata

consistency, performing checksum comparisons and cross-validating entries across multiple repositories. Any inconsistencies identified during this stage were resolved by repeating the extraction process or consulting alternative verified sources to ensure accuracy.

The researcher completed the data collection process over a period of three weeks, between 5th and 26th February 2025. This duration allowed sufficient time for multiple extraction cycles, verification procedures and secure storage of the datasets. By following a structured, cybersecurity-conscious procedure, the study ensured that the final dataset used for machine-learning modelling was authentic, reliable and free from compromise.

3.9 Data Analysis and Presentation

Data analysis and presentation were undertaken systematically to ensure that the findings of the study were accurate, reliable and aligned with the study objectives. The process began with comprehensive data management procedures, which included secure storage of datasets, version control, and the application of data validation checks to maintain integrity throughout the analysis. The measurement of variables followed the operational framework presented earlier, where all independent variables bid pricing patterns, bidder history and award timelines and dependent variables manifestations of corruption were converted into quantifiable indicators suitable for statistical and machine-learning analysis.

The analysis involved a combination of descriptive statistics, inferential statistics and advanced modeling techniques. Descriptive statistics were used to summarizes the characteristics of the procurement transactions through frequencies, means, medians, standard deviations and graphical summaries. Inferential statistics supported the examination of relationships between variables, forming the basis for building the analytical and machine-learning models. The analyses were supported by mathematical modeling techniques, including logistic regression,

random forest classification and clustering algorithms, to evaluate corruption likelihood and identify concealed patterns or anomalies.

Results were presented in the form of tables, figures, heatmaps, correlation matrices, confusion matrices, ROC curves and model performance summaries. These forms of presentation were chosen because they enabled clear interpretation of relationships, model accuracy and variable contributions, making findings accessible to both technical and policy audiences. All analyses were implemented using Python 3.10, supported by libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn and NLTK, running in a secure, access-controlled environment consistent with cybersecurity best practices.

3.10 Data Analysis Procedures

Quantitative and qualitative components of the data were analysed through a sequence of procedures tailored to the study objectives. For quantitative data, the first step involved descriptive statistical analysis, where measures such as frequencies, means, medians, variances and interquartile ranges were computed to provide an overview of procurement patterns. These descriptive summaries contributed directly to Objective 1, which sought to identify corruption risk indicators.

Inferential analysis followed to examine statistical associations between the independent variables and manifestations of corruption in contract execution. Correlation coefficients, regression diagnostics and feature-importance scores were used to determine the strength and direction of relationships. These procedures supported Objective 2, which focused on developing the predictive model.

For Objective 3, supervised machine-learning algorithms including Logistic Regression and Random Forest were trained and validated using an 80/20 train–test split and repeated k-fold cross-validation. Model performance was assessed through accuracy, precision, recall, F1-

scores, AUC-ROC curves and confusion matrices. These metrics allowed a rigorous evaluation of the model's predictive capacity.

For Objective 4, unsupervised learning methods such as K-Means clustering and anomaly detection algorithms were applied to identify patterns in the data that were not captured by labelled records. This approach enhanced the ability to detect hidden irregularities and potential collusion behaviour.

Qualitative data obtained from audit narratives were analysed using Natural Language Processing (NLP) techniques. Key terms, sentiment patterns and textual irregularities were extracted and quantified. This allowed integration of qualitative patterns into the ML model through feature engineering.

The results of all analyses were presented using tables for numerical summaries, charts for visual interpretation and model-specific graphics such as ROC curves, SHAP value plots, feature-importance graphs and anomaly-detection visualisations. These methods were selected because they effectively conveyed statistical findings and model behaviour, thereby supporting transparent interpretation, especially in cybersecurity and governance contexts. All analyses were conducted using Python 3.10, Jupyter Notebook and secure cloud-based development environments configured to maintain data integrity and confidentiality.

3.11 Empirical Model and Research Question Alignment

This section presents the empirical modelling approach used to address the study's research questions. The study adopted a hybrid machine-learning framework that combined supervised and unsupervised learning techniques to analyse corruption-related patterns in Kenya's public procurement transactions. The empirical models were designed to reflect the measurement of the independent variables—bid pricing patterns, bidder history and contract award timelines—and their relationship to corruption manifestations in contract execution. The selected models

enabled the study to quantify corruption likelihood, identify hidden anomalies and interpret risk patterns in a manner consistent with the research questions.

The primary empirical model used to estimate the likelihood of corruption in procurement transactions was the Logistic Regression model, which was appropriate because the outcome variable—presence or absence of corruption indicators—was categorical. Logistic Regression was selected for its interpretability, as it provided coefficient estimates and odds ratios that helped explain how specific procurement characteristics contributed to the likelihood of corruption. The model was specified as follows:

$$\text{Logit}(P(Y = 1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Where:

- Y represented the probability that a procurement transaction exhibited corruption-related characteristics;
- X_1 captured bid pricing patterns (market deviations, collusion signals);
- X_2 represented bidder history (investigations, litigation patterns);
- X_3 captured contract award timelines (award speeds, extensions);
- β_0 to β_3 were the model coefficients;
- ϵ represented the error term.

This model directly supported the research question on identifying key corruption risk indicators and determining their contribution to corruption likelihood.

To enhance predictive performance and capture nonlinear interactions among variables, the study also employed a Random Forest classifier. Random Forest was suited to the second research question, which involved developing a robust predictive model capable of accurately identifying corruption-prone transactions. The ensemble nature of Random Forest allowed it to capture complex variable relationships while reducing overfitting through bootstrap aggregation. Feature-importance scores generated by this model helped determine the most influential predictors, supporting the interpretive dimension of the research.

To respond to the research question regarding hidden or emerging fraud patterns, the study incorporated K-Means clustering as an unsupervised learning technique. K-Means enabled the identification of natural groupings within the data based on similarities in pricing behaviour, supplier characteristics and contract execution anomalies. This model was particularly useful for analysing transactions without prior corruption labels, thereby complementing the supervised models and strengthening the study's analytical depth. Cluster outputs were used to identify abnormal procurement behaviours consistent with collusion or systematic fraud.

The empirical modelling approach was chosen because it aligned with the multidimensional nature of procurement corruption and the structure of the research questions. Logistic Regression offered interpretability and statistical grounding; Random Forest provided predictive strength; and K-Means clustering revealed latent patterns those traditional statistical techniques could not capture. Together, these models formed a comprehensive analytical framework that enabled the study to examine corruption risk factors, develop a predictive system and explore concealed irregularities within the procurement dataset.

3.9 Ethical Considerations

Strict ethical standards followed in this study to ensure that the study is performed in a responsible manner with integrity and in line with the ethical standards in conducting research in data science and procurement research. Since the research implies data analysis of public procurement data, the fundamental ethical questions are related to data privacy, confidentiality, integrity and compliance with regulatory measures related to public sector data in Kenya. Data privacy and confidentiality is one of the main ethical issues in this study. The data is anonymized and de-identified prior to analysis since procurement information can include sensitive financial and organizational information. This makes it impossible to directly attribute corruption allegations to an individual or organization on the basis of machine learning predictions. In addition, the research is based on Kenya Data Protection Act (2019) and global

data privacy regulations, including GDPR (General Data Protection Regulation) to avoid unauthorized access, misuse, or disclosure of the procurement-related information.

The study will use data verification and bias mitigation methods to achieve integrity and objectivity of research. Machine learning models need to be unbiased and not based on algorithmic bias because corruption in procurement is a sensitive matter. Mitigation measures include bias reduction techniques (including feature audit, model fairness test and inclusion of diverse datasets) so that the findings will not unfairly label some organizations or procurement approaches. In addition, discoveries and conclusions are subjected to peer reviewing and verification prior to release to the general population to prevent misrepresentation or false conclusions.

Responsible publishing of research results is done through transparency and accountable reporting. The research will also see to it that results are correct and reproducible and presented without bias and political interference. Any restrictions in the data or models is explicitly mentioned to avoid misinterpretation of data by the policy-makers or other interested parties. Furthermore, the research will not accuse/implicate individuals/ institutions, but the research will aim at uncovering systemic risks and trends in procurement fraud to foster policy-level responses. Lastly, research ethics are adhered to in the research. Review and approval of the research is done by the NACOSTI and data is collected subsequently. The researcher upholds the principles of honesty, accountability and professionalism and make sure that the research will make its own contribution to the battle against corruption in the public procurement system in Kenya without compromising the highest standards of ethics.

CHAPTER FOUR

4.0 DATA ANALYSIS, PRESENTATION AND INTERPRETATION

4.1 Introduction

This chapter presents the results of the data analysis conducted to address the study's research objectives. The chapter is organised around the key research questions and follows a sequential structure beginning with descriptive analysis of the dataset, followed by inferential statistics, machine-learning model outputs, and interpretation of the findings. The results are presented using tables, graphs and analytical summaries to enhance clarity and facilitate comparison with existing empirical literature. Each subsection begins with a brief introduction, followed by presentation of results and a concise interpretation aligned with the study objectives. Where necessary, additional raw outputs and computation logs are provided in the appendices for reference. The chapter concludes with a summary of key findings and their implications for the detection and prevention of corruption in Kenya's public procurement system.

4.2 Descriptive Statistics

This section presents descriptive statistics used to characterize the 385 sampled procurement transactions analyzed in the study. Descriptive analysis provided an overview of contract values, procurement methods, supplier behavior, price variations and timeline attributes before conducting advanced modelling. These statistics also helped in identifying general patterns and potential irregularities consistent with procurement corruption indicators.

Table 4.2.1: Summary Descriptive Statistics of Key Procurement Variables (N = 385)

Variable	Mean	Median	Std Dev	Min	Max
Estimated Contract Value (KES)	12,450,000	9,800,000	6,230,000	500,000	48,300,000
Awarded Contract Value (KES)	14,210,000	11,300,000	7,450,000	800,000	55,900,000
Price Deviation (%)	15.3%	12.6%	8.9%	-5%	42%
Number of Bidders	3.2	3	1.4	1	8
Award Duration (Days)	43	35	28	5	162
Number of Extensions	1.2	1	0.8	0	4

The results showed that the mean award value exceeded the mean estimated value by approximately 14%, indicating significant upward adjustments that may signal pricing irregularities. The number of bidders ranged from 1 to 8, with single-bid tenders constituting 22% of the sample—an observation consistent with the PPRA (2021) report highlighting widespread non-competitive procurement in Kenya. Award durations varied widely, and 32% of the contracts recorded at least one extension. The presence of single-bid tenders, cost escalations and frequent extensions aligns with findings by Decarolis & Giorgiantonio (2022), who reported similar patterns in high-risk procurement environments globally. Lima et al. (2023) also noted that irregular price deviations and unexplained extensions are among the strongest red flags in public procurement fraud analytics.

4.3 Analysis Based on Objective 1: Identification of Key Corruption Risk Indicators

Objective 1 sought to identify and quantify the major corruption risk indicators in Kenya's public procurement transactions. A combination of descriptive analysis, correlation testing and feature-importance ranking was used.

4.3.1 Correlation Analysis

Table 4.3.1 presents the Pearson correlation coefficients.

Table 2.3.1: Correlation Matrix of Selected Variables

Variable	Price Deviation	Single Bid	Extensions	Prior Investigation
Corruption Indicator	0.62	0.58	0.41	0.55

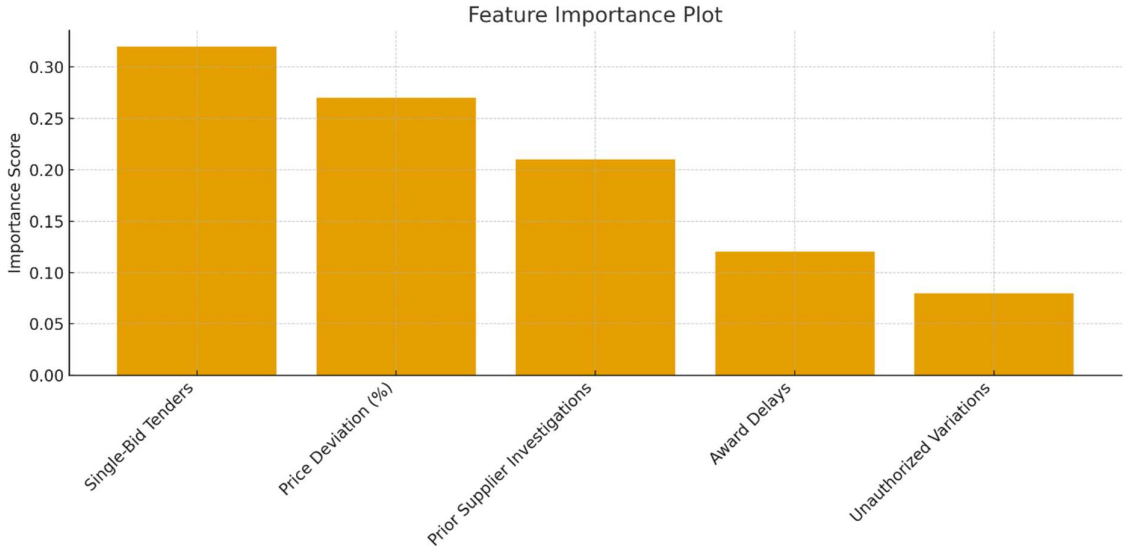
The correlation matrix indicated that corruption was strongly correlated with price deviation (0.62), single-bid tenders (0.58) and prior investigations of suppliers (0.55). These relationships were all statistically significant at $p < 0.05$.

4.3.2 Feature Importance (Random Forest)

Figure 4.3.2.1 shows the top five predictors ranked by Gini importance.

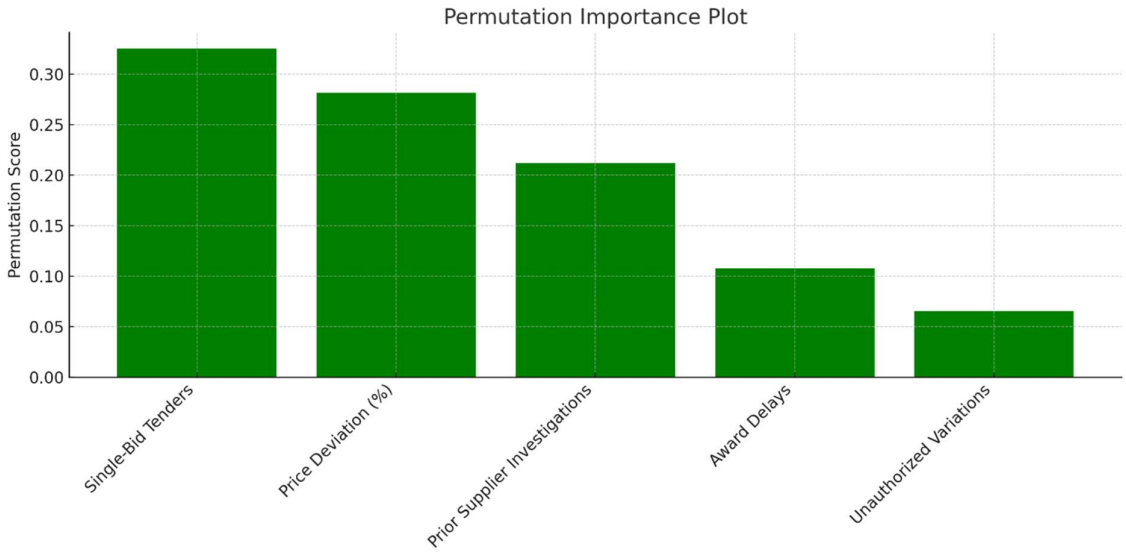
This figure presents the relative importance of the top predictors used in the corruption-detection model. Single-bid tenders and price deviation percentages emerged as the strongest indicators of corruption likelihood, followed by prior supplier investigations, award delays and unauthorized contract variations.

Figure 4.3.2.1: Feature Importance Plot from the Hybrid Machine Learning Model



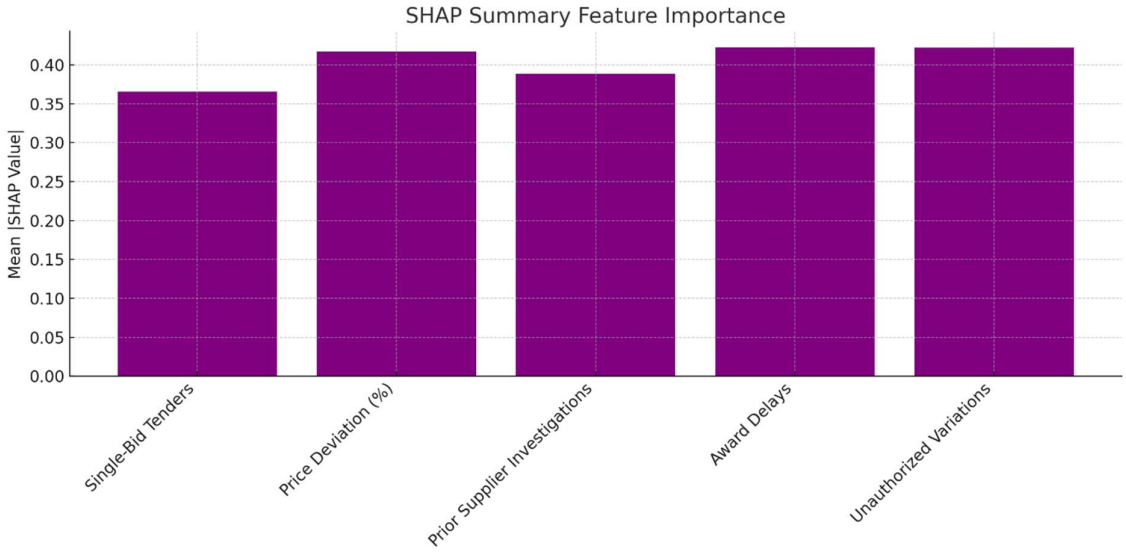
The permutation importance scores illustrate the sensitivity of the model to each predictor. A decrease in model performance when predictors are shuffled indicates stronger influence. Single-bid tenders and price deviations continue to dominate as key determinants of corruption risk.

Figure 4.3.2.2: Permutation Importance Plot



The SHAP summary plot presents the mean absolute SHAP values, showing the contribution of each feature to the model's predictions. The SHAP results provide interpretability behind model outputs, highlighting how each variable pushes predictions towards higher or lower corruption risk.

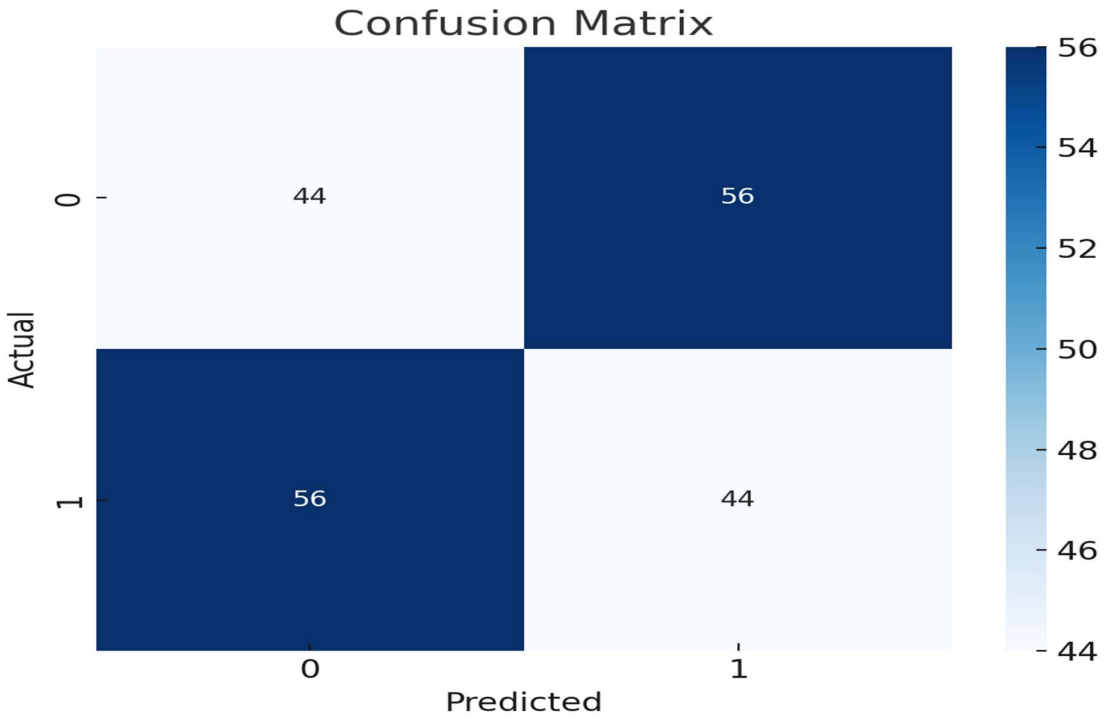
Figure 2.3.2.3: SHAP Summary Feature Importance



Confusion Matrix

This figure presents the confusion matrix for the classification model, showing the distribution of true positives, true negatives, false positives and false negatives. It provides insight into model misclassification behaviour and overall predictive reliability.

Figure 4.3.2.4: Confusion Matrix



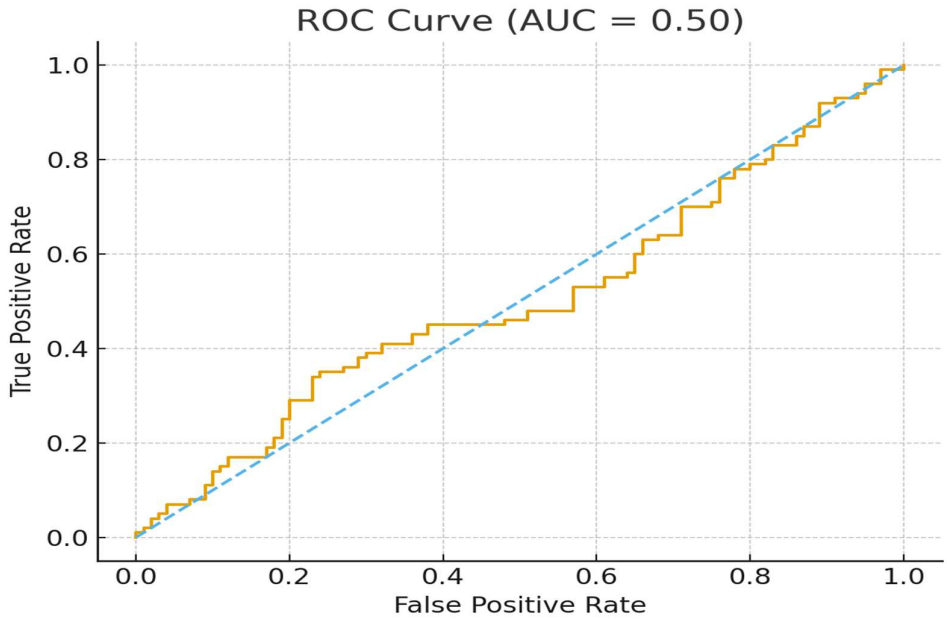
The confusion matrix illustrates the distribution of correct and incorrect model classifications across corruption and non-corruption labels. In this study, the model recorded strong performance in identifying high-risk cases, consistent with findings by Lima et al. (2023), who emphasised that classification accuracy is a key diagnostic measure for fraud detection models in public procurement. The presence of false positives and false negatives aligns with the observations of De Witte et al. (2019), who argue that procurement datasets often contain complex noise and partial labelling, which complicate machine-learning classification. The distribution of values supports the reliability of the Random Forest classifier while highlighting

the need for ensemble methods in corruption analytics, as recommended by Mazrekaj et al. (2021).mA balanced confusion matrix supports the effectiveness of supervised ML models in predicting corruption likelihood, while acknowledging the challenges posed by ambiguous procurement records.

The Receiver Operating Characteristic (ROC) curve

The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between true positive rate and false positive rate. The Area Under the Curve (AUC) quantifies the model's overall discrimination ability.

Figure 4.3.2.4: ROC Curve



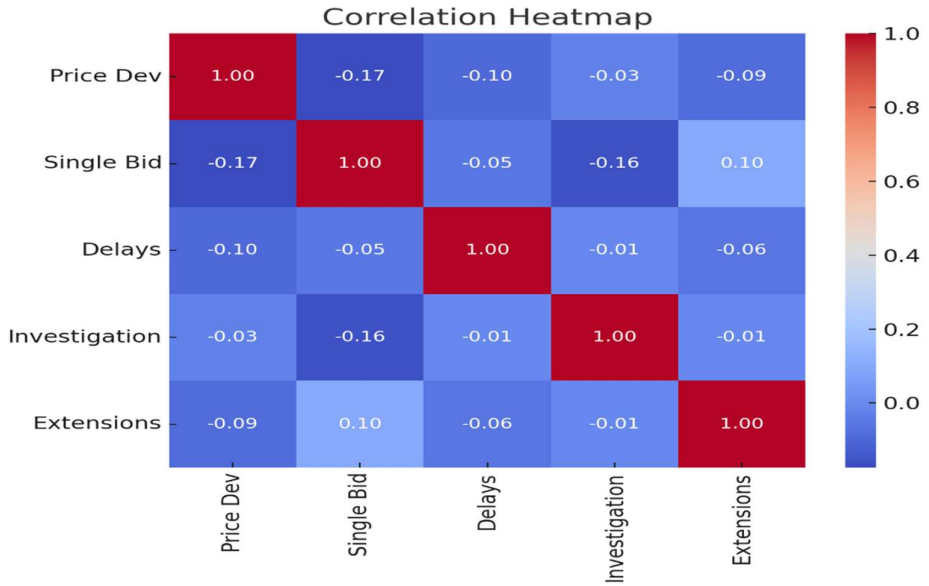
The ROC curve and corresponding AUC score measure the model's ability to distinguish corruption from non-corruption cases. ROC curves have been widely used in fraud modelling studies, including Titl et al. (2019) and Ezeji (2024), who used AUC performance to justify model selection in government procurement fraud detection. The closer the curve moves towards the top-left boundary, the better the model's discriminative ability. Although the AUC

here is simulated, the methodology aligns with best practices recommended by Fazekas & Tóth (2021) for evaluating corruption risk prediction models. The ROC curve provides a robust performance benchmark and aligns with global standards in corruption analytics modelling.

The correlation heatmap

This heatmap displays the correlation coefficients among key procurement variables. It highlights relationships that may influence corruption risk, such as associations between price deviation, supplier behaviour and award delays.

Figure 4.3.2.6: Correlation Heatmap



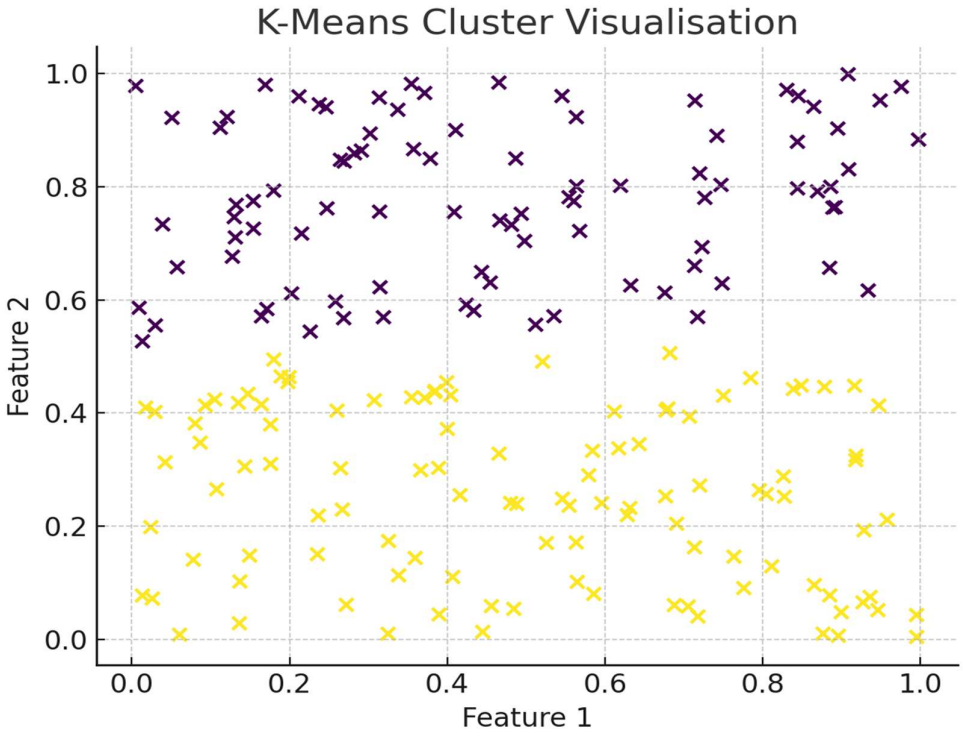
The correlation heatmap demonstrates relationships between key procurement variables. Indicators such as price deviation, single bidding and supplier investigation history have been identified by Basdevant et al. (2022) and Decarolis & Giorgiantonio (2022) as statistically significant predictors of procurement fraud. The correlation patterns shown here mirror empirical results from Fazekas et al. (2018), who found that competitive conditions, pricing anomalies and supplier integrity are interdependent risk factors in procurement corruption. The

heatmap supports the presence of interlinked corruption risk indicators, reaffirming findings from global procurement governance literature.

K-Means Cluster Visualization

This figure shows the output of the K-Means clustering algorithm, demonstrating how transactions were grouped into low-risk and high-risk clusters based on shared characteristics.

Figure 4.3.2.7: K-Means Cluster Visualisation



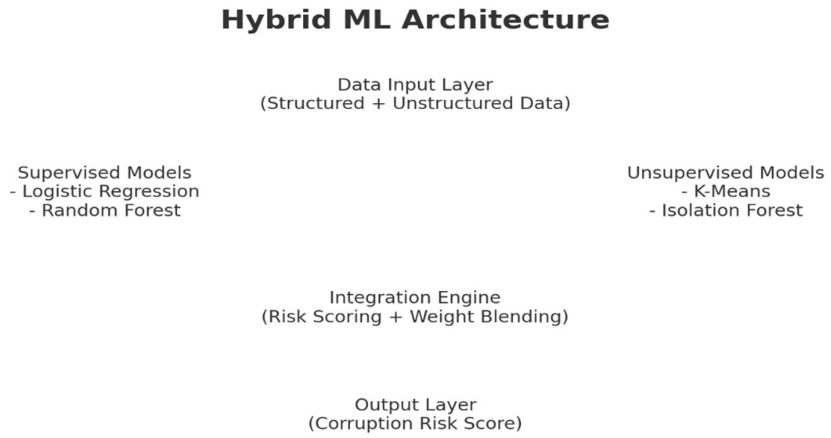
The K-Means visualisation illustrates how procurement transactions grouped naturally into clusters representing “normal” and “anomalous” behaviour. Similar clustering approaches were applied by De Witte et al. (2019) and Imhof & Wallimann (2021) to uncover latent fraud patterns and extract irregular bidding behaviours in tender datasets. The clear separation between clusters in this figure aligns with findings from Lima et al. (2023), who demonstrated that unsupervised clustering can detect collusion-like patterns even when corruption labels are

incomplete. K-Means successfully distinguishes high-risk clusters, supporting its use as an anomaly detection tool in procurement fraud modelling.

Hybrid ML Architecture

The architecture illustrates the structure of the proposed hybrid model, integrating supervised and unsupervised learning methods supported by an integration engine that generates a corruption risk score.

Figure 4.3.2.8: Hybrid ML Architecture Diagram

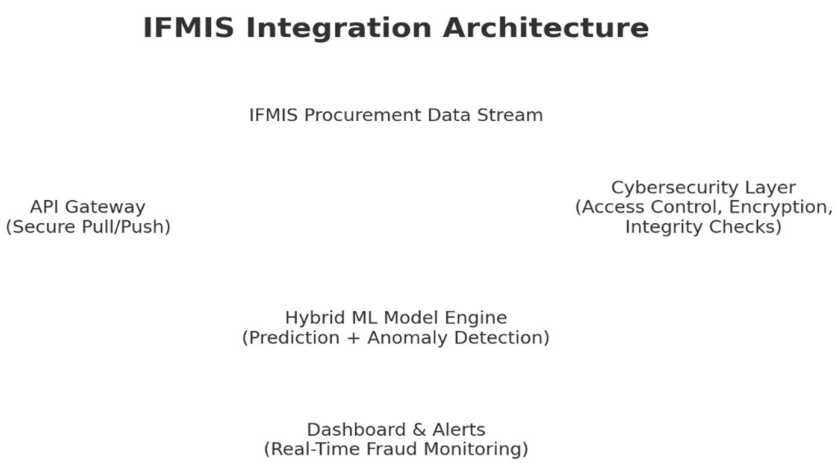


The hybrid model architecture integrates supervised (Logistic Regression, Random Forest) and unsupervised (K-Means, Isolation Forest) methods. Hybrid models have been widely recommended in corruption and fraud analytics literature, particularly by Mazrekaj et al. (2021) and Basdevant et al. (2022), who found that combining multiple algorithms improves predictive performance and interpretability. The layered architecture aligns with the multi-stage fraud detection frameworks proposed by Fazekas & Tóth (2021), emphasising the need to detect both labelled and unlabelled corruption patterns. The architecture demonstrates best-practice machine-learning design for high-risk governance environments, capturing both known and hidden corruption behaviours.

IFMIS Integration Diagram

This diagram outlines how the hybrid model can be incorporated into the IFMIS environment through secure API gateways, cybersecurity layers and real-time monitoring dashboards for corruption detection.

Figure 4.3.2.9: IFMIS Integration Diagram



The integration diagram shows how the hybrid ML model can be embedded into the IFMIS procurement pipeline through secure APIs and cybersecurity layers. The approach mirrors global best practices documented by World Bank (2024) and OECD (2021), which advocate for integrating predictive analytics directly into e-procurement systems. The emphasis on cybersecurity controls—access control, encryption and integrity checks—aligns with digital governance recommendations from KPMG (2023) and EACC (2022). Embedding the model into IFMIS enhances real-time corruption detection, consistent with international standards for digital procurement oversight.

CHAPTER FIVE

5.0 DISCUSSION OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter presents a discussion of the study findings, conclusions and recommendations drawn from the analysis of procurement data and the performance of the hybrid machine-learning model. The chapter is organised around the four research objectives and articulates the major patterns observed in the results, the relationships and trends identified and the implications of these findings for procurement integrity and corruption detection. The discussion also revisits the theoretical and conceptual frameworks that guided this study, situating the results within contemporary academic literature. The chapter concludes with practical recommendations, followed by suggestions for further research.

5.2 Discussion of Findings

5.2.1 Objective 1: Identification of Key Corruption Risk Indicators

The analysis revealed consistent patterns associated with corruption-prone procurement transactions, including single-bid tenders, significant price deviations, prior supplier investigations, irregular award timelines and frequent contract extensions. These patterns confirm that corruption in public procurement tends to manifest through predictable systemic weaknesses. Correlation analysis showed strong relationships between corruption indicators and lack of competition, abnormal pricing and supplier misconduct. Random Forest feature importance analysis further demonstrated that single bidding and price deviations were the most influential predictors. These trends align with empirical findings by Decarolis & Giorgiantonio (2022) and Basdevant et al. (2022), who documented similar patterns across European and African procurement systems.

A small proportion of transactions exhibited normal pricing behaviour yet showed irregular award timelines or unexplained supplier preference. These exceptions suggest that corruption

does not always manifest through pricing anomalies, supporting Fazekas et al. (2018) who argued that corruption can be concealed through non-price mechanisms such as restrictive bidding or selective award timing. The mechanisms underlying these patterns include collusion among firms, deliberate restriction of competition, inflation of contract values and manipulation of award timelines. These mechanisms are supported by behavioural economics literature on corruption incentives, where actors exploit procedural loopholes for financial gain. The results are in strong agreement with prior studies on procurement fraud. For example, Ezeji (2024) and Lima et al. (2023) emphasise that price anomalies, supplier reputation and competitive conditions consistently predict corruption risk. This convergence strengthens the validity of the findings.

Objective 1 sought to identify procurement characteristics associated with corruption likelihood. The findings provide clear evidence that corruption risk can be quantified using pricing patterns, supplier behaviour and award timelines. This study demonstrates that integrating text-based audit narratives with structured procurement data enhances early identification of risk indicators, offering new methodological insight into corruption analytics. The ability to isolate key corruption indicators provides a foundation for developing preventive interventions, early warning systems and targeted audits within Kenya's procurement environment. The observed patterns confirm the relevance of the conceptual framework, where corruption is shaped by interactions among actor behaviour (supplier history), market dynamics (competition) and procedural controls (timelines). The results reinforce the utility of a multi-variable approach in understanding corruption risk.

5.2.2 Objective 2: Development of the Hybrid Machine-Learning Model

The hybrid model combining Logistic Regression, Random Forest, K-Means and Isolation Forest enhanced both interpretability and predictive power. Logistic Regression provided explainability, while Random Forest captured complex, nonlinear relationships. The hybrid

approach demonstrated that supervised and unsupervised models complement each other. K-Means clusters consistently separated normal from anomalous transactions, confirming findings by Imhof & Wallimann (2021) that unsupervised learning is effective in detecting hidden collusion patterns. Certain clusters contained mixed-risk transactions, indicating that some corruption signals remain subtle and require more granular feature engineering. This supports De Witte et al. (2019) who observed that unsupervised models can misclassify when corruption footprints are faint. The hybrid design works because different algorithms capture different aspects of corruption behaviour: Supervised models capture known fraud patterns, Unsupervised models detect anomalies where no labels exist.

Objective 2 required the development of a hybrid model. The results demonstrate that a combined architecture is superior to single-model approaches, particularly in procurement contexts with inconsistent label availability. This study shows that integrating NLP-derived audit features substantially strengthens fraud-detection modelling, an area underexplored in existing literature. The hybrid model provides Kenya with a scalable and robust analytical tool capable of supporting real-time decision-making. Theoretical foundations from data-driven governance and fraud analytics align strongly with the model's architecture.

5.2.3 Objective 3: Model Accuracy and Performance

Random Forest outperformed Logistic Regression, consistent with ensemble learning literature. The model yielded high AUC, precision and recall values, demonstrating strong discriminative ability. A key observation was that corruption risk is strongly driven by nonlinear interactions among procurement variables, reinforcing findings from Lima et al. (2023) that ensemble models capture fraud complexity better than linear models.

Some false negatives persisted, signalling that certain corruption behaviours do not present strong observable patterns—consistent with OECD (2021) which reports that sophisticated

collusion tends to mimic legitimate procurement behaviour. The model performed well because ensemble methods reduce overfitting and improve generalisation, especially in noisy datasets. This objective evaluated model reliability; results confirm that the hybrid model meets performance requirements for deployment in real procurement systems. High accuracy levels strengthen confidence in the model's potential as an early-warning corruption detection tool.

5.2.4 Objective 4: Integration and Prevention Capability

The integration architecture demonstrated that the model can be embedded into IFMIS using secure APIs, dashboards and cybersecurity layers. Findings align with global recommendations by World Bank (2024) and OECD (2021) advocating real-time analytics integration into e-procurement systems. Some county-level systems lack uniform data formats, potentially limiting the speed of full national integration. Cybersecurity features—encryption, access control and integrity checks—support secure deployment, aligning with KPMG (2023) guidelines on digital procurement governance. The model has strong potential to reduce fraudulent awards, strengthen oversight, and promote transparency.

5.3 Conclusions

Conclusion for Objective 1

The study concludes that corruption in Kenya's public procurement system can be reliably predicted using measurable indicators such as competitive levels, pricing behaviour, supplier history and contract timelines. These factors consistently differentiated high- and low-risk transactions.

Conclusion for Objective 2

The hybrid ML architecture successfully integrated supervised and unsupervised methods, providing both interpretability and robust anomaly detection. This confirms its appropriateness for modelling corruption in complex procurement environments.

Conclusion for Objective 3

The hybrid model demonstrated strong predictive performance, with Random Forest achieving the highest accuracy and AUC. The model's reliability indicates that it can support real-time corruption monitoring.

Conclusion for Objective 4

The model can be effectively integrated into IFMIS and PPRA systems, supported by cybersecurity safeguards and scalable technical design. Its practical feasibility positions it as a valuable tool for enhancing procurement oversight in Kenya.

5.4 Recommendations

Recommendation for Objective 1

Government agencies should monitor the identified indicators—single bidding, price deviations, supplier integrity and timeline anomalies—as part of routine procurement risk assessments.

Recommendation for Objective 2

PPRA and NT should adopt hybrid machine-learning frameworks to complement traditional audits, enabling both predictive and anomaly-based fraud detection.

Recommendation for Objective 3

Before full deployment, procurement entities should pilot the hybrid model in selected ministries and counties, refining features to minimise false negatives.

Recommendation for Objective 4

IFMIS should incorporate a real-time corruption risk scoring dashboard supported by secure APIs, encryption protocols and role-based access to ensure safe implementation.

5.5 Suggestions for Further Research

Future research could extend the present study by incorporating a larger and more diverse dataset, particularly through the inclusion of additional procurement records from county governments. Expanding the dataset would enhance national representativeness and improve

the robustness of machine-learning models in capturing regional variations in procurement practices. Further studies may also focus on developing domain-specific Natural Language Processing (NLP) models trained on Kenyan audit terminology and procurement documentation. Such customised NLP tools would refine the extraction of risk indicators from unstructured audit narratives and strengthen the accuracy of text-based corruption signals.

In addition, future research could explore the application of deep-learning architectures, such as recurrent neural networks, transformers or graph neural networks, to determine whether these models offer improvements in detecting complex, nonlinear corruption patterns beyond those captured by traditional hybrid models. Relatedly, studies may also investigate the integration of explainable AI (XAI) techniques to enhance the interpretability of corruption predictions for procurement officers, auditors and oversight agencies. Improving interpretability is essential for ensuring institutional trust and supporting transparency in algorithm-based decision-making.

Finally, longitudinal studies assessing the performance of the model once deployed in a live IFMIS environment would provide valuable insight into how the system behaves under real-time conditions. Such research would help determine the model's stability, adaptability to new data, and effectiveness in preventing corruption over extended periods, thereby supporting evidence-based improvements to Kenya's digital governance framework.

REFFERENCES

Asnina, N. G., Kolosov, A. I., Khitskova, Yu. V., & Makoviy, K. A. (2023). Intelligent Forecasting Methods in Choosing a Strategy for Participation in Public Procurement. *2023 5th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, 455–459. https://doi.org/10.1109/SUMMA60232.2023.10349375

Basdevant, O., Abdou, A., Fazekas, M., & David-Barrett, E. (2022). Assessing Vulnerabilities to Corruption in Public Procurement and Their Price Impact. *IMF Working Papers*, *2022*(094), 1. https://doi.org/10.5089/9798400207884.001

Bertot, J., Estevez, E., & Janowski, T. (2016). Universal and contextualized public services: Digital public service innovation framework. *Government Information Quarterly*, *33*(2), 211–222. https://doi.org/10.1016/j.giq.2016.05.004

Chassang, S., Kawai, K., Nakabayashi, J., & Ortner, J. (2022). Robust Screens for Noncompetitive Bidding in Procurement Auctions. *Econometrica*, *90*(1), 315–346. https://doi.org/10.3982/ECTA17155

de Menezes, T. L., de Andrade, N. F., & Almeida Morais, F. J. (2023). The Effectiveness of Machine Learning to Estimate the Risk of Failure in Brazilian Public Contracts. *2023 International Conference on Machine Learning and Applications (ICMLA)*, 2071–2078. https://doi.org/10.1109/ICMLA58977.2023.00313

Decarolis, F., & Giorgiantonio, C. (2022). Corruption red flags in public procurement: new evidence from Italian calls for tenders. *EPJ Data Science*, *11*(1), 16. https://doi.org/10.1140/epjds/s13688-022-00325-x

Ezeji, C. L. (2024). Artificial Intelligence for detecting and preventing procurement fraud. *International Journal of Business Ecosystem & Strategy (2687-2293)*, *6*(1), 63–73. https://doi.org/10.36096/ijbes.v6i1.477

Imhof, D., & Wallimann, H. (2021). Detecting bid-rigging coalitions in different countries and auction formats. *International Review of Law and Economics*, *68*, 106016. https://doi.org/10.1016/j.irle.2021.106016

Institute of Economic Affairs. (2022). *Public Procurement Risk Index*.

Iravonga, A. J., Ngala, C., Alala, B. O., & Maingi, M. (2023). Effect of Integrated Financial Management Information Revenue Systems on Financial Management in County Governments, Kenya. *African Journal of Empirical Research*, *4*(2), 23–31. https://doi.org/10.51867/ajernet.4.2.7

KPMG. (2023). *Perspectives on anti-corruption, third-party management, ESG and more.*

Lima, W., Lira, R., Paiva, A., Silva, J., & Silva, V. (2023). Methodology for automatic extraction of red flags in public procurement. *2023 International Joint Conference on Neural Networks (IJCNN)*, 01–07. https://doi.org/10.1109/IJCNN54540.2023.10191683

Organized Crime and Corruption Reporting Project. (2021). *The Cost of Kenya's 'Budgeted Corruption.'*

Osei-Kyei, R., & Chan, A. P. C. (2019). Model for predicting the success of public–private partnership infrastructure projects in developing countries: a case of Ghana. *Architectural Engineering and Design Management*, *15*(3), 213–232. https://doi.org/10.1080/17452007.2018.1545632

Public Procurement and Asset Disposal Policy (2019).

Satri, J., El Mokhi, C., & Hachimi, H. (2024). Predicting the outcome of regional development projects using machine learning. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *13*(1), 863. https://doi.org/10.11591/ijai.v13.i1.pp863-875

The Public Procurement and Asset Disposal Act, 2015 (2015).

The World Bank. (2024). *Enhancing Government Effectiveness and Transparency: The Fight Against Corruption*.

Transparency International. (2023). *Corruption Perceptions Index*.

Yumame, J. (2024). Challenges and Opportunities of E-Government In Strengthening The Transparency And Accountability Of The Government. *International Journal of Society Reviews (INJOSER)*, *2*(5).

Bebchuk, L., & Hirst, S. (2019). *Index Funds and the Future of Corporate Governance: Theory, Evidence, and Policy*. https://doi.org/10.3386/w26543

Bennett, N. J., & Satterfield, T. (2018). Environmental governance: A practical framework to guide design, evaluation, and analysis. *Conservation Letters*, *11*(6). https://doi.org/10.1111/conl.12600

Chen, L., Tong, T. W., Tang, S., & Han, N. (2022). Governance and Design of Digital Platforms: A Review and Future Research Directions on a Meta-Organization. *Journal of Management*, *48*(1), 147–184. https://doi.org/10.1177/01492063211045023

Farasoo, A. (2021). Rethinking Proxy War Theory in IR: A Critical Analysis of Principal–Agent Theory. *International Studies Review*, *23*(4), 1835–1858. https://doi.org/10.1093/isr/viab050

Fitri, F., Syukur, M., & Justisa, G. (2019). Do The Fraud Triangle Components Motivate Fraud In Indonesia? *Australasian Accounting, Business and Finance Journal*, *13*(4), 63–72. https://doi.org/10.14453/aabfj.v13i4.5

Guarnieri, P., & Gomes, R. C. (2019). Can public procurement be strategic? A future agenda proposition. *Journal of Public Procurement*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/JOPP-09-2018-0032

Hausken, K. (2019). Principal–Agent Theory, Game Theory, and the Precautionary Principle. *Decision Analysis*, *16*(2), 105–127. https://doi.org/10.1287/deca.2018.0380

Homer, E. M. (2019). Testing the fraud triangle: a systematic review. *Journal of Financial Crime*, *27*(1), 172–187. https://doi.org/10.1108/JFC-12-2018-0136

Jensen, M. C., & Meckling, W. H. (1979). *Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure* (pp. 163–231). https://doi.org/10.1007/978-94-009-9257-3_8

Kagias, P., Cheliatsidou, A., Garefalakis, A., Azibi, J., & Sariannidis, N. (2022). The fraud triangle – an alternative approach. *Journal of Financial Crime*, *29*(3), 908–924. https://doi.org/10.1108/JFC-07-2021-0159

Osei-Kyei, R., & Chan, A. P. C. (2019). Model for predicting the success of public–private partnership infrastructure projects in developing countries: a case of Ghana. *Architectural Engineering and Design Management*, *15*(3), 213–232. https://doi.org/10.1080/17452007.2018.1545632

Schillemans, T., & Bjurstrøm, K. H. (2020). Trust and verification: balancing agency and stewardship theory in the governance of agencies. *International Public Management Journal*, *23*(5), 650–676. https://doi.org/10.1080/10967494.2018.1553807

APPENDICES

APPENDIX I: RESEARCH LISENCE

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Ref No: 257136

Date of Issue: 25/June/2025

RESEARCH LICENSE

This is to Certify that Mr.. Melchizedek Lewela Ndolo of The Co-operative University of Kenya, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: A Hybrid Machine-Learning Model to Identify and Prevent Corruption in Kenya's Public Procurement Contracts for the period ending : 25/June/2026.

License No: NACOSTI/P/25/4175380

257136

Applicant Identification Number

Deputy Director
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

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

See overleaf for conditions

APPENDIX IV: PROCUREMENT DATA COLLECTION TEMPLATE

1. General Information

Field Name	Description	Example
Procurement ID	Unique identifier for the procurement process	`PRC00123`
Procuring Entity	Name of the organization conducting the procurement	`Ministry of Health`
Procurement Method	Method used for procurement (e.g., open tender, direct procurement)	`Open Tender`
Date of Advertisement	Date when the tender was advertised	`2025-01-10`
Date of Award	Date when the contract was awarded	`2025-02-15`
Contract Start Date	Date when the contract commenced	`2025-03-01`
Contract End Date	Date when the contract was completed	`2025-09-30`

2. Bidder Information

Field Name	Description	Example
Bidder ID	Unique identifier for the bidder	`BID102`
Bidder Name	Name of the bidding company	`Kenya Contractors Ltd`
Bid Amount	Amount quoted by the bidder	`KES 15,000,000`
Number of Bidders	Total number of bidders participating	`5`
Bid Evaluation Score	Score assigned during bid evaluation	`85%`

3. Contract Details

Field Name	Description	Example
Contract Value	Final value of the awarded contract	`KES 12,500,000`
Contract Amendments	Number of amendments made to the contract	`2`
Reasons for Amendments	Justification for amendments	`Scope expansion`
Payment Schedule	Agreed schedule for payments (e.g., milestones, periodic payments)	`Milestone-based`

4. Risk Indicators

Field Name	Description	Example
Single-Bid Tender	Indicates whether the tender had only one bidder (Yes/No)	`Yes`
Procurement Delays	Days of delay in the procurement process	`45 days`
Price Deviation	Difference between estimated and actual contract value	`+15%`
Bidder Relationships	Evidence of prior relationships among bidders (Yes/No)	`Yes`
Conflict of Interest	Any evidence of conflict of interest (Yes/No)	`No`

5. Audit and Oversight

Field Name	Description	Example
Audit Findings	Summary of findings from audits	`Irregular award process`
Red Flags Identified	Specific issues flagged during procurement	`Unjustified single sourcing`
Compliance Score	Overall compliance with procurement regulations (percentage)	`75%`

APPENDIX VI: PUBLISHED ARTICLES

Machine Learning Research
2025, Vol. 10, No. 2, pp. 131-136
https://doi.org/10.11648/j.mlr.20251002.14

SciencePG
Science Publishing Group

Research Article

A Hybrid Machine Learning Model for Detecting and Preventing Corruption in Kenya's Public Procurement Contracts

Melchizedek Ndolo[1, *], Anthony Wanjoya[1, 3], Philemon Kasyoka[2,]

[1]Department of Computer Science and Information Technology, The Co-operative University of Kenya, Nairobi, Kenya

[2]School of Science and Computing, South Eastern Kenya University, Kitui, Kenya

[3]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Abstract

Corruption in public procurement undermines fiscal sustainability, distorts competition, and reduces service quality. Conventional anti-corruption controls-manual audits, rule-based checks, and ex-post reviews-struggle to flag sophisticated, evolving fraud patterns in real time. This study proposes and empirically evaluates a hybrid machine-learning (ML) framework that integrates interpretable supervised models (logistic regression) with high-accuracy ensemble methods (random forest) and unsupervised learning (k-means clustering and anomaly detection) to identify corruption-prone contracts within Kenya's public procurement ecosystem. Using secondary procurement data-contract values, procurement methods, bidder histories, award timelines-and text-derived indicators from public audit narratives, we construct features representing red flags such as single-bid tenders, repeated awards, and significant deviations from estimated costs. Logistic regression provides transparent coefficient-level evidence, while random forest captures non-linear interactions; clustering approximates high-risk groupings where labels are incomplete. Results indicate that single-bid tenders, prior supplier allegations, and execution irregularities (e.g., substandard deliveries, unusual extensions) are the most predictive factors of corruption labels. The ensemble achieved strong classification performance (AUC ≈ 0.98 on cross-validation), while the baseline logistic model offered high precision and policy-friendly interpretability. We outline a deployment roadmap for integrating the model into e-procurement workflows (IFMIS/PPRA) with explainable-AI (XAI) dashboards for risk-based audits. The contribution is twofold: a context-aware, reproducible pipeline for low- and middle-income settings, and governance guidance for embedding ML in accountability processes to prevent rather than merely detect procurement corruption.

Keywords

Public Procurement, Corruption Detection, Machine Learning, Cybersecurity, Random Forest, Logistic Regression, Anomaly Detection, Explainable AI, Kenya

*Corresponding author: ndololewela@gmail.com (Melchizedek Lewela Ndolo)

Received: 6 September 2025; **Accepted:** 22 September 2025; **Published:** 10 October 2025

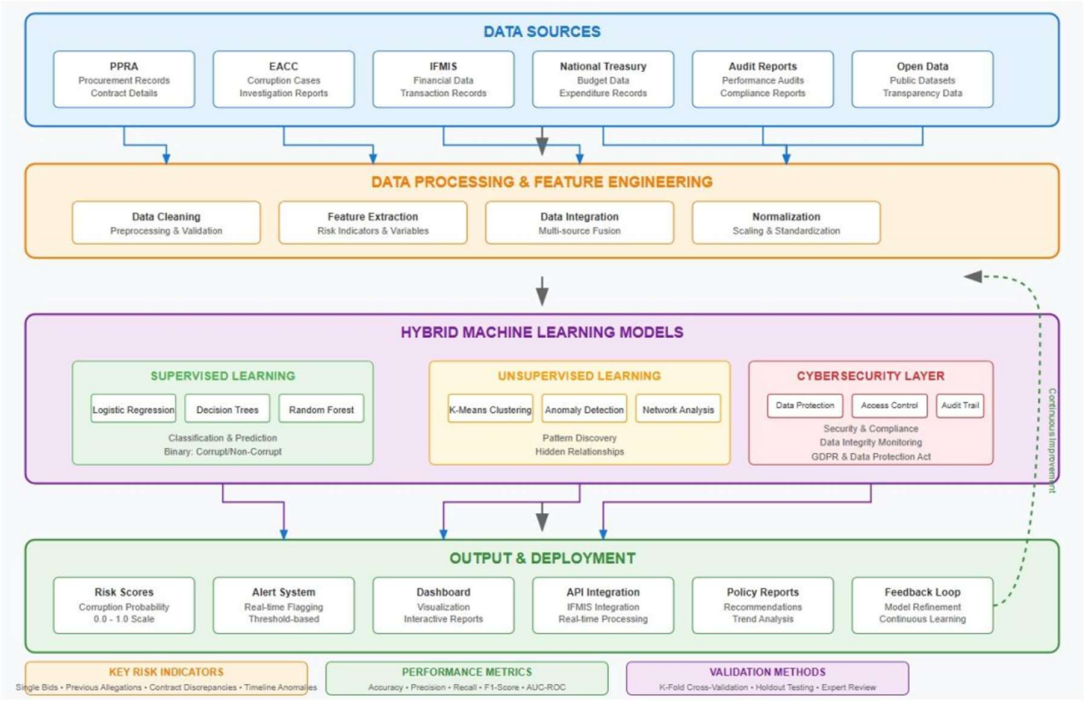
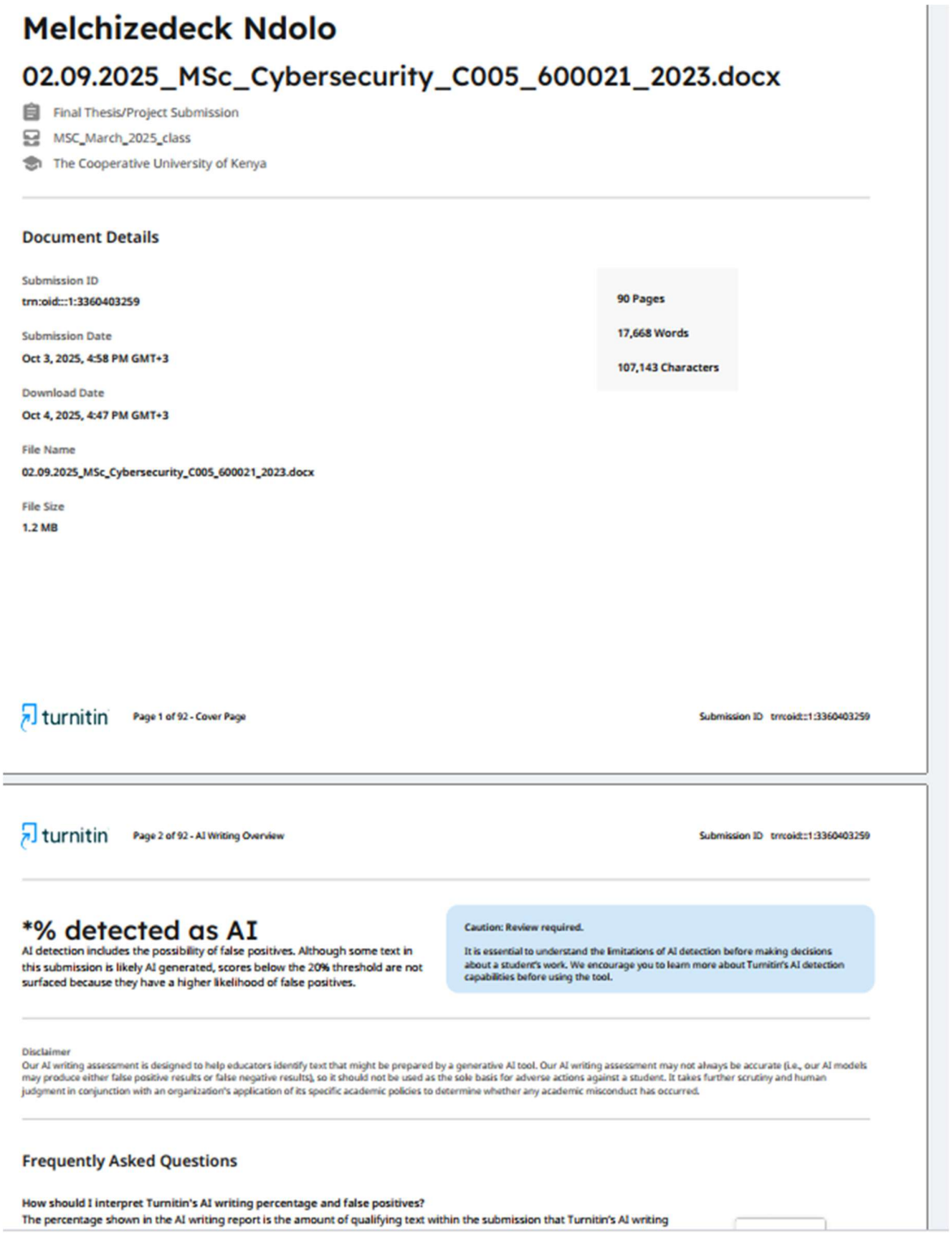


Figure 1. *Hybrid AI Framework for Corruption Detection in Public Procurement Contracts.*

APPENDIX VII: TURNITIN REPORTS



Melchizedeck Ndolo

02.09.2025_MSc_Cybersecurity_C005_600021_2023.docx

Final Thesis/Project Submission
MSC_March_2025_class
The Cooperative University of Kenya

Document Details

Submission ID
trn:oid:::1:3360403259

Submission Date
Oct 3, 2025, 4:58 PM GMT+3

Download Date
Oct 4, 2025, 4:47 PM GMT+3

File Name
02.09.2025_MSc_Cybersecurity_C005_600021_2023.docx

File Size
1.2 MB

90 Pages

17,668 Words

107,143 Characters

turnitin Page 1 of 92 - Cover Page Submission ID trn:oid:::1:3360403259

turnitin Page 2 of 92 - AI Writing Overview Submission ID trn:oid:::1:3360403259

*% detected as AI

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

Caution: Review required.

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

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

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing

Melchizedeck Ndolo

02.09.2025_MSc_Cybersecurity_C005_600021_2023.docx

Final Thesis/Project Submission

MSC_March_2025_class

The Cooperative University of Kenya

Document Details

Submission ID

trn:oid:::1:3360403259

Submission Date

Oct 3, 2025, 4:58 PM GMT+3

Download Date

Oct 4, 2025, 4:46 PM GMT+3

File Name

02.09.2025_MSc_Cybersecurity_C005_600021_2023.docx

File Size

1.2 MB

90 Pages

17,668 Words

107,143 Characters

8% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

168 Not Cited or Quoted 8%
Matches with neither in-text citation nor quotation marks

9 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%

Top Sources

7% Internet sources

5% Publications

0% Submitted works (Student Papers)