

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 Kasyoka² 

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

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

³Department 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 ($AUC \approx 0.98$ 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



Copyright: © The Author(s), 2025. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Public procurement is a major channel of public expenditure and therefore a focal point for corruption risks that inflate prices, reduce quality, and undermine trust in government. Despite robust legal frameworks and oversight institutions such as the Public Procurement and Asset Disposal Act, 2015 [15] and the Public Procurement and Asset Disposal Policy, 2019 [16], traditional assurance mechanisms are retrospective and resource-intensive, making it difficult to identify complex collusion and manipulation patterns at the pace of modern procurement. Weak institutional frameworks, political interference, lack of transparency in tendering processes and inadequate monitoring mechanisms have allowed fraudulent practices to persist, KPMG, 2023 [14]. Recent advances in machine learning (ML) and data engineering make it possible to detect irregularities from large, heterogeneous datasets and to move the oversight paradigm from after-the-fact discovery to real-time prevention.

Kenya's procurement environment exhibits known vulnerabilities: non-competitive procedures, repeated awards to the same vendors, and award-timeline anomalies. Yet practical deployment in such environments poses constraints-data fragmentation, uneven data quality, limited labeling of historical cases, and the need for interpretability to support due process. This study responds to those constraints by building a hybrid ML architecture that (i) detects corruption-prone contracts with high accuracy, (ii) remains sufficiently transparent for audit justification, and (iii) complements oversight capacity through risk-based triage. World Bank, 2024 [12]; Transparency International, 2023 [13]

The paper proceeds as follows. Section 2 describes the materials and methods, including the data pipeline, feature engineering, and models. Section 3 presents results for descriptive statistics and model performance. Section 4 discusses implications for policy and governance. Section 5 concludes and outlines a deployment plan and future work.

2. Materials and Methods

2.1. Comparative Methodological Gaps

While prior research provides valuable insights, several methodological shortcomings persist. Traditional statistical approaches such as regression and rule-based anomaly detection are limited in handling high-dimensional, heterogeneous procurement data, and often lack predictive accuracy in dynamic environments Decarolis & F, 2022 [1]; Imhof & D, 2021 [7]. Systematic reviews by Alsamarraie and Ghazali Alsamarraie & M, 2023 [2] highlight fragmented procurement performance metrics, whereas their related work on Iraq under-scores persistent barriers in ensuring procurement integrity under weak institutional settings Alsamarraie & M, 2023 [3]. Paraskeva and Tsoulfas

Paraskeva & S, 2025 [4] further emphasize the inadequacy of existing risk frameworks in anticipating emerging corruption strategies, noting a reliance on retrospective audits rather than proactive analytics.

Recent empirical studies reinforce these gaps. Lima et al. Lima & W, 2023 [8] demonstrate that most automated red-flag extraction techniques still rely on narrow keyword matching, limiting their generalizability across diverse procurement contexts. De Menezes et al. De Menezes & T, 2023 [5] show that while machine learning improves risk estimation for contract failures, these models often underperform in low-data or partially labeled environments. Ezeji Ezeji & C, 2024 [6] highlights the lack of robust explainability frameworks in AI-driven fraud detection, raising concerns for due process and accountability in procurement oversight. Osei-Kyei and Chan Osei-Kyei & R, 2019 [9] illustrate the methodological limitations in predicting project success in PPP contexts, emphasizing the need for more dynamic learning algorithms. Basdevant et al. Basdevant & O, 2022 [10] provide a cross-country assessment showing that corruption-vulnerability models often fail to capture price impacts systematically. Titl et al. Titl & V, 2019 [11] review screening methods for collusion and conclude that current approaches lack robustness across different procurement systems.

Collectively, these findings underscore the need for hybrid, context-sensitive ML models that combine interpretability with predictive strength-precisely the approach advanced in this study.

2.2. Research Design and Data Sources

We adopt a quantitative, experimental design using secondary data from public repositories and oversight reports including the Public Procurement Regulatory Authority Annual Procurement Report, 2021 [17] and the National Ethics and Corruption Survey by EACC, 2022 [18]. Structured fields include procurement method (open, restricted, direct), estimated versus awarded value, number of bidders, award timelines, supplier identifiers, and historical sanctions/complaints. Unstructured audit narratives and contract justifications are processed via lightweight natural language processing (NLP) to extract risk terms (e.g., single-bid, emergency justification, significant amendment) and to score irregularity mentions.

2.3. Feature Engineering

We engineer indicators aligned to red-flag typologies:

- 1) Competition: single-bid (binary), effective number of bidders, repeated awards to same supplier.
- 2) Value dynamics: percentage deviation of award from estimate; outlier scores of unit prices relative to peers.

- 3) Timeline: days from close to award; deadline extensions; unusual amendment frequency.
- 4) Supplier history: prior allegations/sanctions; litigation count; related-party signals.
- 5) Execution quality: substandard delivery flags; variation orders; non-delivery incidents.

Text fields are vectorized with keyword dictionaries and TF-IDF scoring to form sparse features representing governance risk language (e.g., justified direct procurement, emergency, national security, sole supplier). Categorical variables are one-hot encoded; continuous variables are standardized.

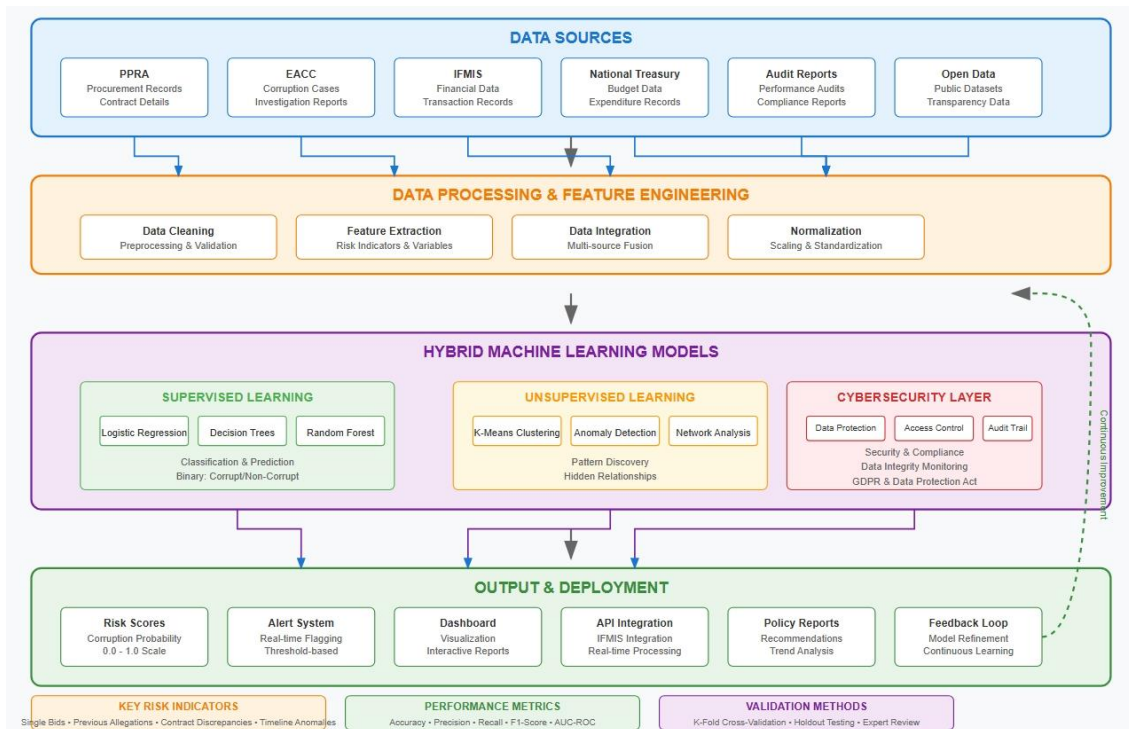


Figure 1. Hybrid AI Framework for Corruption Detection in Public Procurement Contracts.

2.4. Modeling Strategy

We estimate:

- 1) Logistic Regression (LR): baseline, interpretable odds ratios for policy audiences.
- 2) Random Forest (RF): ensemble classifier capturing non-linearities and interactions.
- 3) K-Means Clustering: unsupervised grouping for anomaly surfacing where labels are incomplete; PCA used for visualization only.
- 4) Isolation-type anomaly scores (optional): to highlight contracts with rare combinations of red flags.

Hyperparameters are tuned via grid search nested in k-fold cross-validation ($k=5$). Class balance is monitored; threshold adjustments and class weights mitigate any residual imbalance.

2.5. Evaluation

Primary metrics: AUC-ROC, precision, recall, F1, and calibration curves. We emphasize precision for investigative efficiency (minimizing false positives that waste audit

capacity) and recall for social cost (minimizing undetected corruption). Feature importance (RF Gini/SHAP) and LR odds ratios provide interpretability. K-means external validity is checked by overlap with labels where available.

2.6. Ethical and Legal Compliance

All data are aggregated/de-identified prior to analysis. We follow Kenya's Data Protection Act and research ethics standards. The system is designed to flag transactions, not individuals, and requires human review. We document features and thresholds for transparency and due process.

3. Results

3.1. Descriptive Patterns

Procurement values are right-skewed, with most awards in a mid-range and a small tail of high-value contracts. Direct procurement displays higher median values and greater dispersion than open tendering. Approximately one quarter of records carry a corruption-risk label from historical case

documentation, indicating moderate imbalance.

3.2. Predictive Models

Logistic Regression achieved strong precision with moderate recall. Positive and statistically meaningful coefficients were observed for single-bid (odds ratio > 2) and prior supplier allegations (> 1.5), while open tendering was negatively associated with corruption labels.

Random Forest outperformed LR with higher AUC (≈ 0.98 on cross-validation) and balanced precision/recall. Top importance features were single-bid, prior allegations, award-timeline anomalies, and award-to-estimate deviation. Partial-dependence plots show sharply increasing risk at the transition from multi-bid to single-bid and at large positive deviations from estimates.

K-Means revealed clusters that closely mirror label partitions, especially along competition and timeline axes, enabling triage where labels are absent.

3.3. Robustness and Interpretability

Results were stable across folds. SHAP explanations corroborated importance rankings and indicated interaction effects-e.g., high deviations are most risky under non-competitive methods. Calibration curves were satisfactory after isotonic adjustment. Sensitivity checks using alternative thresholds maintained policy-relevant precision (≥ 0.85) with recall in the 0.75-0.85 range.

4. Discussion

Findings validate competition-related signals-particularly single-bid tenders-as primary drivers of corruption risk, consistent with global red-flag literature. Supplier history remains a potent predictor, supporting the case for unified sanction/discipline registries. Timeline anomalies likely reflect rushed or strategically delayed awards enabling manipulation. Ensemble learning improves detection over linear baselines but must be paired with explainability to support administrative law requirements and public accountability.

Operationally, the hybrid approach enables a risk-based audit regime: RF for high-recall screening, LR and SHAP for justifications, and clustering for discovery of emerging patterns. Embedding the model into e-procurement (IFMIS/PPRA) with APIs allows near-real-time flagging. Governance measures-standardized data schemas, mandatory disclosure of amendments, and public vendor histories-amplify model efficacy.

4.1. Contribution to Cyber Security

This study extends cyber-defense practices into the procurement domain by framing corruption risk as an

anomaly-detection problem on transactional and textual data. The hybrid LR-RF-clustering approach strengthens threat detection by combining interpretable signals (odds ratios) with non-linear pattern capture and unsupervised discovery of emerging risks. It advances explainable AI (SHAP-based justifications) to support accountable decision-making while observing Kenya's Data Protection Act via de-identification and feature transparency. Practically, it shows how ML governance, model documentation, and human-in-the-loop review can reduce systemic attack surfaces in e-government workflows.

4.2. Model Deployment and Validation

The model is packaged as an API microservice (containerized) for integration with IFMIS/PPRA data streams, with role-based access controls, audit logging, and encryption in transit/at rest. CI/CD pipelines automate retraining, versioning, and rollback; monitoring tracks data drift, prediction confidence, and alert volumes. Validation uses 5-fold cross-validation (AUC ≈ 0.98), threshold tuning for precision-recall trade-offs, calibration checks, and time-based holdouts to simulate real-world deployment. A human-in-the-loop queue verifies high-risk flags, feeding confirmed outcomes back for continuous learning.

4.3. Recommendations

Standardize procurement data schemas and mandate metadata (bidders, timelines, amendments) to improve signal quality; enforce secure data pipelines and IAM to protect model inputs and outputs. Adopt an XAI-first policy for all automated flags, with written explanations, auditable thresholds, and redress mechanisms. Schedule periodic retraining, adversarial robustness tests, and bias audits; monitor drift and recalibrate thresholds per agency capacity. Finally, pilot a controlled rollout with clear SOPs, KPIs (precision, recall, time-to-investigate), and skills transfer to build sustainable in-house ML and cybersecurity capability.

5. Conclusions

A hybrid ML pipeline can reliably identify corruption-prone procurement contracts in Kenya while preserving interpretability for governance use. Dominant predictors are single-bid competition, prior supplier allegations, award-timeline anomalies, and large positive deviations from estimates. Deployed as a dashboard with XAI narratives, the system supports preventive controls and smarter allocation of audit resources. Future work should (i) expand text mining over audit narratives, (ii) test semi-supervised learning for scarce labels, and (iii) pilot in a sandbox linked to live e-procurement data.

Abbreviations

AUC	Area Under the Curve
IFMIS	Integrated Financial Management Information System
LR	Logistic Regression
ML	Machine Learning
NLP	Natural Language Processing
PPRA	Public Procurement Regulatory Authority
RF	Random Forest
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence

Acknowledgments

The author thanks supervisors and colleagues at the School of Computing & Mathematics (CUK) for methodological guidance and feedback. Institutional support for data access and research ethics compliance is gratefully acknowledged.

Author Contributions

Melchizedek Lewela Ndolo: Conceptualization; Methodology; Software; Formal analysis; Investigation; Resources; Data curation; Validation; Visualization; Writing-original draft; Writing-review & editing; Project administration.

Anthony Kibira Wanjoya: Supervision; Writing - review & editing

Philemon Nthenge Kasyoka: Supervision; Writing - review & editing

Funding

This work is not supported by any external funding.

Data Availability Statement

The data supporting the findings of this study are publicly available from Kenyan procurement portals and oversight publications; analysis code and feature definitions are available from the corresponding author upon reasonable request.

Conflicts of Interest

The author declares no conflicts of interest.

Appendix

Feature Dictionary (abridged)

Competition indicators; value deviation metrics; timeline

features; supplier-history flags; execution irregularities; and text-derived governance-risk terms.

Reusable code templates (feature engineering, training scripts, and SHAP reporting) and synthetic data schema are available upon reasonable request. A public repository can be created upon acceptance.

References

- [1] Decarolis, F., & Giorgiantonio, C. (2022). Corruption red flags in public procurement: new evidence from Italian calls for tenders. *EPJ Data Science*, 11(16), 1-24. <https://doi.org/10.1140/epjds/s13688-022-00325-x>
- [2] Alsamarraie, M. M., & Ghazali, F. (2023). Evaluation of organizational procurement performance for public construction projects: Systematic review. *International Journal of Construction Management*, 23(14), 2499-2508.
- [3] Alsamarraie, M. M., & Ghazali, F. E. M. (2023). Barriers and challenges for public procurement integrity in Iraq: Systematic review study. *KSCE Journal of Civil Engineering*, 27(9), 3633-3645.
- [4] Paraskeva, S., & Tsoulfas, G. T. (2025). Mitigating risks in public procurement. *Journal of Public Procurement*, 25(1), 140-176.
- [5] De Menezes, T. L., de Andrade, N. F., & Morais, F. J. A. (2023). The effectiveness of machine learning to estimate risk of failure in Brazilian public contracts. 2023 ICMLA, 2071-2078. <https://doi.org/10.1109/ICMLA58977.2023.00313>
- [6] Ezeji, C. L. (2024). Artificial intelligence for detecting and preventing procurement fraud. *International Journal of Business Ecosystem & Strategy*, 6(1), 63-73. <https://doi.org/10.36096/ijbes.v6i1.477>
- [7] 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>
- [8] 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), 1-7. <https://doi.org/10.1109/IJCNN54540.2023.10191683>
- [9] Osei-Kyei, R., & Chan, A. P. C. (2019). Model for predicting the success of PPP infrastructure projects in developing countries: Ghana. *Architectural Engineering and Design Management*, 15(3), 213-232. <https://doi.org/10.1080/17452007.2018.1545632>
- [10] Basdevant, O., Abdou, A., Fazekas, M., & David-Barrett, E. (2022). Assessing vulnerabilities to corruption in public procurement and their price impact. *IMF Working Paper* 22/094. <https://doi.org/10.5089/9798400207884.001>
- [11] Titl, V., et al. (2019). Screening methods for collusion in public procurement: A literature review. *OECD Working Papers*.

- [12] World Bank. (2024). Enhancing Government Effectiveness and Transparency: The Fight Against Corruption.
- [13] Transparency International. (2023). Corruption Perceptions Index.
- [14] KPMG. (2023). Perspectives on anti-corruption, third-party management, ESG and more.
- [15] Public Procurement and Asset Disposal Act. (2015). Government of Kenya.
- [16] Public Procurement and Asset Disposal Policy. (2019). Government of Kenya.
- [17] Public Procurement Regulatory Authority (PPRA). (2021). Annual Procurement Report.
- [18] Ethics and Anti-Corruption Commission (EACC). (2022). National Ethics and Corruption Survey.

Biography

Melchizedek Lewela Ndolo is a Cloud Architect and Cybersecurity Analyst and a graduate researcher in Cyber Security at The Co-operative University of Kenya. His work spans machine learning for threat and anomaly detection, cloud security architecture across AWS, Google Cloud, Azure, and Oracle Cloud, secure ERP implementations and data engineering for risk analytics. He has led cross-functional initiatives in identity and access management (IAM), network defense, incident response and compliance-by-design, and develops Python-based pipelines for model deployment, monitoring, and explainability.

Anthony Kibira Wanjoya is a Kenyan statistician and academic based at the Department of Statistics and Actuarial Sciences, School of Mathematical Sciences, Jomo Kenyatta University of Agriculture and Technology (JKUAT). He has established himself as a researcher at the intersection of applied statistics, epidemiology, and demography, with a particular focus on public health issues affecting developing countries.

Philemon Nthenge Kasyoka is a computer scientist and researcher specializing in cryptography, secure communication protocols, and information systems security. He earned his Master of Science degree in Computer Systems from Jomo Kenyatta University of Agriculture and Technology in 2014, where his thesis laid the groundwork for his later interest in cryptographic techniques and secure systems design. Kasyoka's research focuses on the design and analysis of lightweight and efficient cryptographic protocols suitable for constrained environments such as the Internet of Medical Things (IoMT) and edge computing.

Research Fields

Melchizedek Lewela Ndolo: Machine learning for governance, Public procurement analytics, Fraud and anomaly detection, Explainable artificial intelligence, Cybersecurity for e-government, Data engineering for risk analytics, Risk-based auditing, Applied NLP in information systems

Anthony Kibira Wanjoya: Applied Statistics and Probability, Bayesian Methods and Latent Trait Analysis

Philemon Nthenge Kasyoka: Security and Protocol Design, Internet of Medical Things (IoMT), Edge Computing, Cryptography / Signcryption Schemes