ISSN (Print): 1060-6076
ISSN (Online): Applied

The Journal of African Development
2026; Vol 7: Issue 1
https://www.afea-jad.com/

Ai-Based Digital Procurement Fraud Detection Framework For Public Sector Organizations

Dr. Reetika Agarwal[1], Dr. Moushumi Banerjee[2], Dr. Fatima Qasim Hasan[3], Dr. Kirti[4], Dr. Rakesh Chawla[5], Dr. Tarang Mehrotra[6], Ms. Sonal Walia[7], Ms. Shravani Mathur[8]

[1] Associate Professor, Department of Management, IILM Academy of Higher Learning, Lucknow, Uttar Pradesh, India.

Email: reetikaagarwal82@gmail.com

[2] Assistant Professor, Faculty of Hospitality Management and Catering Technology, MS Ramaiah University of Applied Sciences, Bengaluru, Karnataka, India.,Email: moushmiban1978@gmail.com

[3] Professor, School of Business, Galgotias University, Greater Noida, Uttar Pradesh, India.

Email: fatimaqasim.hasan@galgotiasuniversity.edu.in

[4] Professor, Department of Applied Science and Humanities, IIMT College of Engineering, Greater Noida, Uttar Pradesh, India.,Email: kirtiupadhyay0407@gmail.com

[5] Principal of B.Com & B.A., Department of Commerce, Integrated Academy of Management & Technology, Ghaziabad, Uttar Pradesh, India.,Email: rakesh.chawla@inmantec.edu

[6] Associate Professor & HOD of Commerce, Department of Commerce, Integrated Academy of Management & Technology, Ghaziabad, Uttar Pradesh, India.,Email: tarang.mehrotra@inmantec.edu

[7] Assistant Professor, Department of Business Administration, Chandigarh Business School of Administration, Landran, Mohali, Punjab, India.,Email: waliasonal86@gmail.com

[8] Research Scholar, Department of Electronics and Communication Engineering, School of Engineering and Technology, SAGE University, Bhopal, Madhya Pradesh, India.,Email: matshravani@gmail.com

Cite This Paper as: Dr. Reetika Agarwal, Dr. Moushumi Banerjee, Dr. Fatima Qasim Hasan, Dr. Kirti, Dr. Rakesh Chawla, Dr. Tarang Mehrotra, Ms. Sonal Walia, Ms. Shravani Mathur (2026) Ai-Based Digital Procurement Fraud Detection Framework For Public Sector OrganizationsThe Journal of African Development l, Vol.7, No.1, 1632-1647

KEYWORDS
Public procurement, AI fraud detection, anomaly detection, risk assessment, explainable AI...

ABSTRACT
Public procurement digitalization multiplies the number of transactions that can be seen, but at the same time makes the risks of fraud even bigger, faster, and more complex for control bodies to assess. This study proposes an AI-based digital procurement fraud detection system for public sector organisations, and demonstrates analytical logic using a reproducible set of benchmark data, not by 'claimed' field data. The architecture incorporates the procurement red flags, supervised classification, anomaly detection, explainability, human audit triage and governance controls. A data set comprising 20,000 procurement transactions was created based on risk patterns identified in the literature and subsequently split into training, validation and held out test partitions. The following six models were tested: logistic regression, random forest, XGBoost, isolation forest, and a weighted ensemble were compared based on precision, recall, F1, ROC-AUC, PR-AUC, calibration, and interpretability diagnostics. The performance of the supervised models was significantly superior to that of the unsupervised models for anomaly detection, and logistic regression was the best discriminator in the benchmark. The following factors were identified as risk indicators that are significant in relation to the restrictions in the procedure, single bidding, concentration of suppliers, shortened advertisement periods and changes to the contract. The results validate a risk ranking architecture, where AI can guide review without leaving the possibility for determining guilt, maintain explainability, due process, auditability, and human accountability in operational procurement oversight and responsible public administration....

.

1. INTRODUCTION

1.1 The background and research context

The public procurement of large amounts of government spending is linked to administration, competition and services, and failures of integrity have real consequences for public value. While digital procurement platforms have enhanced traceability and generated machine-readable records, they've also generated more transactions than can be manually screened. In fact, recent studies consider procurement data as the basis to identify suspicious activities, such as tendering with few procedures, single bidding, repeated supplier concentration, short tendering periods, and unusual changes in the contracts' clauses (Decarolis & Giorgiantonio, 2022; Siino et al., 2026). Data-driven approaches can transform these signals into risk scores to offer earlier intervention, and machine learning extends detection beyond fixed rules (Schneider dos Santos et al., 2025). The deployment of AI in the public sector is still unique in that regards, however, administrative decisions are required to be explainable, contestable, proportionate, and institutionally accountable (Straub et al., 2023; Sampaio et al., 2026). It is the requirements that drive the integrated framework that combines predictive accuracy with transparent oversight by design.

1.2 Problem Statement

The key issue is not that there are no procurement warning signs, but how to integrate and use the disparate indicators to make timely, defensible, and operationally useful decisions. While rule-based systems might fail to capture nonlinear interactions, opaque models can emit an alarm that is not understood or justified by auditors. Further, there are few fraud labels, imbalanced class distributions, and suspicious activity may appear like a legitimate administrative exception (Lyra et al., 2022; Westerski et al., 2021). These circumstances lead to a double risk scenario: on the one hand, harm-full transactions are not detected by weak models, on the other, false suspicion falls upon innocent suppliers and procurement officers. Public organisations therefore require a framework that prioritizes risk, clarifies the signals that make it up, facilitates human verification, and distinguishes between predictive suspicion and legal or disciplinary judgment by ensuring a framework of accountability for the review.

1.3 Research Gap and Originality

Current scholarship is partial and piecemeal. Procurement studies create corruption red flags and composite risk indicators, while machine-learning studies involve comparing classifiers, network methods or anomaly detectors (Fazekas & Kocsis, 2020; Muñoz-Cancino & Ríos, 2025). In contrast, public-administration research focuses on the concepts of trustable AI, explainability, fairness, and human control (Odilla, 2024; Zuiderwijk et al., 2021). There has been less focus on linking these streams together in a deployable analytical risk/prominent fraud decision architecture. This study aims to fill this gap by combining transaction indicators, AI comparative models, interpretable risk attribution, audit triage and governance protection measures. The novelty of its approach is to test the framework without creating institutional observations, as simulation is a repeatable representation of the coherence of internal analysis.

1.4 Research Objectives

The study aims to have three goals. First, it defines a public sector fraud detection model that involves a mix of supervised, unsupervised, and ensemble risk scoring techniques and procurement red flags. Second, it assesses the discrimination of a simulated class of fraud risk given class imbalance and uncovers the procurement properties that most affect predictions. Third, it integrates explainability, human oversight, and governance mechanisms into the workflow, ensuring that outputs from the model serve as an auditable screening signal and not as an automated accusation. These objectives are designed to evaluate whether the use of predictive analytics can reinforce procurement oversight without compromising the accountability, proportionality, transparency and verification criteria.

1.5 Research questions and/or hypotheses

The study focuses on three research questions: RQ1 - What procurement indicators are most prominent to classify the risk of frauds?; RQ2 - What are the differences between supervised, unsupervised, and ensemble approaches with an imbalanced benchmark?; and RQ3 - What can be done to turn the predictive outputs into accountable public-sector review? These are operationalized in four hypotheses and one proposition. Restrictive procedure indicators (H1); supplier concentration and network embeddedness (H2); process irregularities (H3); and superior discrimination of supervised models as compared to unsupervised models of anomaly detection (H4). It is important to note that explainability and human review are governance layers for legitimate procurement screening through AI (Proposition P1).

2. LITERATURE REVIEW

2.1 The public sector is vulnerable to fraud in a number of ways, including public sector procurement

In such a case, according to Bosio et al. (2022), formal procurement law might not match actual practice, suggesting that compliance architecture is not necessarily sufficient to ensure competitive results. This is also confirmed by a cross-national study that links low competition, low accountability, and low transparency with high corruption risk (Fazekas et al., 2016; Ferwerda et al., 2017). However, recent research indicates that single bidding, non-open bidding, shortened tender periods, and other features can be useful warning indicators but none of these alone can be a signal of fraud (Decarolis & Giorgiantonio, 2022).

The tradition of this measurement has been supported by Fazekas and Kocsis (2020), who employ the procurement records as behavioural traces that can be used to create objective risk indicators. Lisciandra et al. (2022) also demonstrate that multidimensional vulnerabilities can be captured in composite indicators, and Siino et al. (2026) test procurement risk profiles against organized crime related firm data. Together, these studies warrant data-based screening, as well as the danger of misinterpreting statistics as evidence and the need for risk prioritization—not automatic adjudication.

2.2 The digital procurement market is still evolving, and new fraud trends are emerging

Jiménez et al. (2022) show that electronic procurement could decrease the opportunities for corrupt contracting, with the impact varying depending on governance institutions and the conditions of implementation. While digitalization can be used to increase the volume of records that can be searched and standardise work processes, it can also be used to move manipulation to problems with data quality, with identity of the supplier, to coordinated bidding and to the use of exceptions in strategic terms. Falcón-Cortés et al. (2022) demonstrate that procurement behaviour can also change over political transition, highlighting the potential of the need for compliance rules that are not static in changing risk landscapes.

In their work, García Rodríguez et al. (2019) prove that machine-readable tender information can be used to help in predictive analysis of award outcomes and Potin et al. (2023) demonstrate how graph representations can be used to recover relational signals when contract attributes are missing. However, Muñoz-Cancino and Ríos (2025) propose expanding the scope of fraud screening to suspicious supplier relationships by leveraging machine learning and social network analysis. What these studies have shown is that there is a core paradox in the digital procurement environment: richer procurement data present analytical opportunities, but less comprehensive, unavailable, and manipulated data can hinder detection efforts. Data architecture, relational context and quality control are thus as important as algorithmic capacity for effective digital oversight.

2.3 Artificial Intelligence in Fraud Detection

Guida et al. (2023) trace the use of AI in procurement and find that there is significant scope for automation, prediction, classification, and decisions support, though there is less on the organizational integration. Supervised learning algorithms can also be used when it comes to detecting frauds, as they can learn to combine several weak signals and learn nonlinear relationships that fixed thresholds for red flags might miss. Random forests were created by Breiman (2001) as strong ensembles of decorrelated decision trees and scalable gradient boosting by XGBoost was contributed by Chen and Guestrin (2016). These are appealing approaches for the case of procurement risks associated with interactions between procedure, supplier, pricing, and execution variables.

Madan and Ashok (2023) demonstrate that, as well as its technical performance, public-administration AI adoption relies on institutional capacity, data readiness, and perceived legitimacy. Similarly, Straub et al. (2023) call for standards and governance across government AI applications that are coherent. Zuiderwijk and his colleagues (2021) mention a recurring need for transparency, accountability, and human-centred governance. Thus, the selection of algorithms can't be limited to accuracy. Public procurement demands performance evaluation in the setting of class imbalance, explanations of individual alerts, calibration of risk scores, and institutional processes which account for when models can impact audits, inquiries or supplier-facing decisions. When false positives have a reputational risk, these are important protections.

2.4 To identify the various approaches for detecting procurement fraud using AI

The study by Schneider dos Santos et al. (2025) uses data-driven research to synthesize findings on procurement fraud and identifies that machine learning, statistical analysis, and specialized detection methods are widely adopted in the context of collusion and favouritism. Their mapping also reveals label and data inconsistencies, as well as inconsistencies in labels, datasets, evaluation measures, and fraud definitions, which makes direct comparisons between studies difficult. Gallego et al. (2021) show the importance of early-warning modelling, rather than punishment in hindsight, and the focus is no longer on punishment, but on prioritisation for prevention.

Through real-world deployments of anomaly scoring in procurement-fraud scenarios, Westerski et al. (2021) demonstrate the need for explainability to make anomaly scores meaningful investigative signals. Velasco et al. (2021) also view fraud analytics as decision support, not decision making. Nai et al. (2023) continue public-administration analytics to address recommender systems and learning from complaints. This literature collectively indicates that the usefulness of an operation depends on three interrelated aspects—discriminative performance, interpretable evidence, and workflow integration. A model which scores well without being able to support reasoning from the reviewers is institutionally incomplete for public accountability.

2.5 Context of the public sector governance, transparency and accountability.

Predictions can cause unfair effects, even if they are not the result of unfair inputs in the data, unfair design of the model, or unfair assumptions in the institutions, or if they are used unfairly downstream, as Odilla (2024) argues. In high-stakes scenarios, Rudin (2019) also raises the question of the need of unnecessary black-box models, while Barredo Arrieta et al. (2020) present a classification of explainable-AI approaches that might reveal model reasoning without sacrificing complexity. These thoughts are relevant in situations where a procurement alert may impact reputation, investigations, and market access.

Sampaio et al. (2026) create a trustworthy-AI pipeline for public procurement, with a focus on autonomy, fairness, explainability, oversight and control. AI research in the public sector also focuses on governance and not technology-driven adoption (Wirtz et al., 2021). Sánchez-Graells (2024) points out the “dichotomy between the possibilities offered by AI-based architectures for procurement and the constraints of due process.” Governance is thus an integral component of

detection quality: a risk system must be able to document data provenance, able to be contested, maintain human authority and record the pathway from alert to administrative action.

2.6 Achievements of Science and technology

In this study, Lyra and colleagues (2022) argue that data-driven and network-aware approaches are beneficial for fraud, corruption and collusion research, but there are still some challenges in obtaining ground truth, in having a consistent definition of the concepts, and in comparability. Thus, a literature-based approach does not support a rule-based solution or a black-box solution. What's missing here is a process that connects procurement-specific red flags, comparative predictive modelling, explainability, human audit triage and public-law protections. The present study will tackle that intersection via a simulation validated architecture. It asks not if AI can be used to pronounce fraud, but if AI can rank suspicious transactions accurately enough to allow for the focus of the oversight capacity without losing sight of the importance of maintaining traceability, being able to make reviewer calls, and ensuring procedural accountability.

3. CONCEPTUAL AND ANALYTICAL FRAMEWORK

3.1 Core Constructs and Framework Dimensions

The concept of procurement fraud risk is a multidimensional screening construct not a legal finding. There are four analytical dimensions of observable signals: procedural restriction, supplier dependence, transaction irregularity and relational concentration. Non-open awards, single bidding, bidder count and advertisement duration are examples of procedural restriction. Supplier dependence reflects the repeated awards and concentration. Transaction irregularity includes the ratios of awards to estimates and changes to estimates, missing documents, split purchases, and late payment. Relational concentration is the centrality of suppliers' networks. These features are input into a predictive layer which generates a calibrated risk score and feature-level explanation. Access, logging, threshold policy, audit referral, reviewer documentation and feedback are then governed by a governance layer. Figure 1 shows this closed loop system: in which the outputs of the model are given priority in attention while human review still has control over interpretation and action..

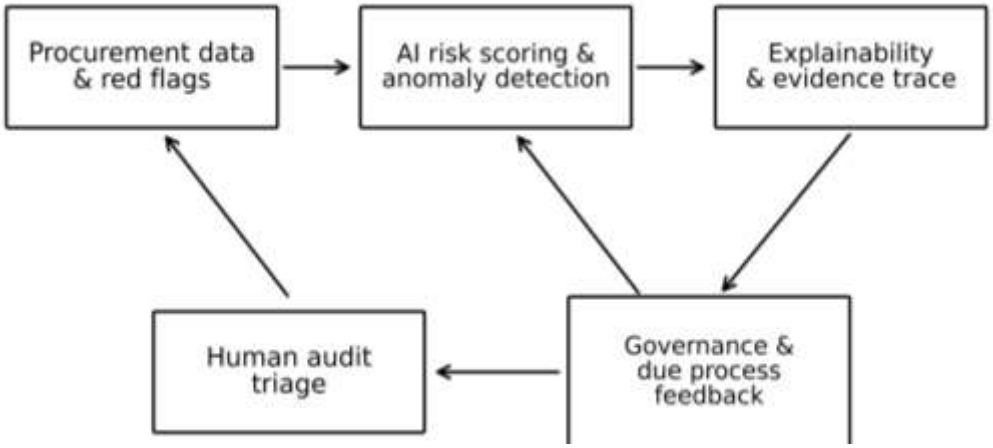

Figure 1. Conceptual architecture of the AI-based digital procurement fraud detection framework

3.2 Proposed Relationships among Framework Components

Restrictive procedures should be used sparingly, as they can lead to less competition and more chances for favouritism, and should be linked to an increase in the risk predicted. A strong supplier concentration and repetitive award may be an indicator of specialization, but can also be an indicator of entrenched relationships, therefore relational variables are modelled in conjunction to tender variables. With execution-stage irregularities, the framework has temporal evidence to update risk as contracts evolve. These heterogeneous features are converted to scores by the use of AI models and the drivers are identified by explainability. The alert that is generated is not a verdict. It goes through a human triage process involving auditor review of records, exceptions and documentation. Version control and governance approval of thresholds and models can be updated based on feedback from confirmed cases and dismissed cases. This relationship is shown in Figure 1. This is an intentional way to detract from the process of adjudications and clearly include organizational accountability in public administration contexts when doing analysis.

3.3 Research Hypotheses/Analytical Propositions

H1 suggests that the two variables are positively related (i.e., the greater the simulated fraud risk, the more restrictive the procedure). H2 predicts a positive relationship between supplier concentration and network centrality and simulated risk. H3 suggests that the information added to the execution irregularities is discriminatory information in addition to the information carried in the procedure variables. In terms of class discrimination, supervised classifiers outperform unsupervised anomaly detection, says H4. According to Proposition P1, there must be an interpretable alert, a documented

human decision, and an auditable model version, threshold, evidence, and disposition record of a legitimate operation. The assessment of H1-H4 is by a quantitative measure on the benchmark; in this study P1 is assessed analytically but not by statistical significance, but by the completeness of the framework.

4. RESEARCH METHODOLOGY

4.1 Research Design and Approach

The study is conducted following a design-science and simulation-validation method to validate the proposed framework in the absence of any actual deployment. The empirical benchmark does not represent any of the participants, agencies, suppliers or any observed procurement record. Rather, the literature-based procurement risk mechanisms are programmed in a transaction generator, and then alternative classifiers are tested under an experimental protocol that cannot be modified. This design can be used to show model behaviour without creating the evidence of that behaviour in the real world and acknowledges that the validating of effectiveness is not the same as demonstrating external effectiveness. The process outlined in figure 2 moves from data generation to data pre-processing, stratified partitioning, model training, threshold selection based on the validation set, held-out testing, explainability analysis and governance interpretation. Fixed seeds are employed with all randomized processes in order to enable the benchmark to be reproduced and independently audited.

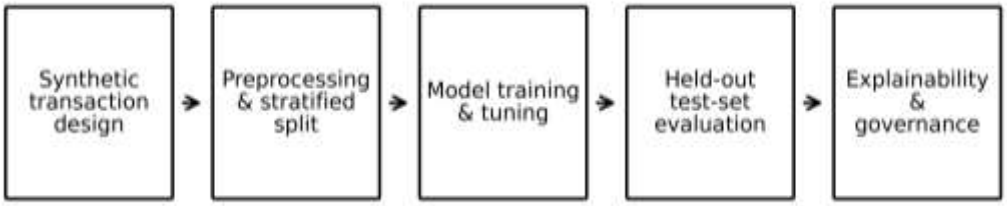

Figure 2. Simulation-validation and model-evaluation workflow

Note. No observed agency, supplier, participant, or field transaction data are used.

4.2 Research Context, Data Sources, and Sampling

The context is a generic digital public-procurement context, in which the tender, supplier, award, contract-execution and payment records can be connected on the transaction level. The benchmark has 20,000 transactions and is not associated with any jurisdiction or organization. A minority class for fraud risk was generated in a probabilistic fashion, based on interacting procedure, supplier, pricing, execution, documentation, and network conditions; prevalence was 11.27%, which was more like an imbalanced screening problem than an estimate of prevalence. Stratified sampling provided 12,000 observations for training, 4,000 for validation and 4,000 for testing. The values of partition and class distribution are reported in Table 1. The operating thresholds were selected from the validation set, but final comparative statistics were computed without using the validation set. This separation brings down the optimistic assessment that could be attributed to tuning on assessment information (He & Garcia, 2009). The adequacy of the sample size is a concern for computational stability and for minority-class representation and is not related to inferential generalization to citizens, agencies, or suppliers.

Table 1. benchmark partition and class distribution

Partition	N	positive	low-risk	Positive rate
Training	12,000	1,352	10,648	11.27%
Validation	4,000	451	3,549	11.28%
Test	4,000	451	3,549	11.28%
Total	20,000	2,254	17,746	11.27%

Note. All observations and class labels are simulated. Percentages are not estimates of real-world procurement fraud prevalence.

4.3 Data Preparation and Fraud-Related Feature Construction

Fourteen procurement behaviours were repeatedly mentioned in the corruption-risk literature (Fazekas et al., 2016; Decarolis & Giorgiantonio, 2022; Siino et al., 2026) and formed the basis for creating fourteen predictors. The following binary indicators were captured: non-open procedure; single bidding; split purchasing; missing documentation; and weekend award activity. Bidder Count, Advertisement Duration, Supplier Concentration, Repeat-award Share, Award-to-estimate ratio, Contract Modifications, Payment delay, Contract Value, Supplier network centrality were either continuous variables or count variables. Each variable is defined and explained in its role for analysis in Table 2. The generator

enforced realistic restrictions: fewer procedures led to more frequent single bidding, fewer bidders and fewer ad periods, concentration led to more frequent repeated awards, and more extended modification intensity led to more payment delay. The class was created using a logistic latent-risk function that included main effects and selected interactions before a Bernoulli draw. This mechanism leads to a probabilistic overlap between classes, thereby avoiding deterministic labels. Predictors were kept in interpretable units except for logistic regression which had standardized inputs. Observations were not excluded but rather an explicit documentation-risk feature. A lack of information can be a tell-tale sign.

Table 2. predictors and operational fraud-risk logic

Predictor	Type	Operational interpretation
Contract value	Continuous	Transaction scale; retained mainly as a control feature
Non-open procedure	Binary	Restricted procedure/discretion signal
Single bid	Binary	Competition restriction signal
Bidder count	Count	Intensity of observed competition
Advertisement days	Continuous	Time available for market participation
Supplier concentration	Continuous [0,1]	Dependence on a limited supplier set
Repeat-award share	Continuous [0,1]	Persistence of awards to recurring suppliers
Award/estimate ratio	Continuous	Potential pricing irregularity
Contract modifications	Count	Post-award change intensity
Payment delay days	Continuous	Execution/payment irregularity
Split-purchase flag	Binary	Potential threshold-avoidance behavior
Missing documentation	Binary	Opacity/data-quality warning
Weekend award	Binary	Timing irregularity
Supplier network centrality	Continuous [0,1]	Relational embeddedness/concentration

Note. Predictors encode literature-grounded warning mechanisms for simulation. No variable is treated as proof of fraud
.

4.4 AI-Based Fraud Detection Framework Development

Four modelling strategies were evaluated to prevent assuming superiority of algorithms. Logistic regression was used to obtain an interpretable linear baseline using class weighting and standardized predictors. Random forest modelled nonlinear interactions by 500 trees learned using subsampling to ensure the subsampling was balanced on the right side, using the ensemble logic introduced by Breiman (2001). It took 550 boosting rounds, shrinkage, row/column subsampling, and a positive-class weight based on the training distribution (Chen & Guestrin, 2016) for XGBoost. To provide an unsupervised comparator to rank unusual observations, without relying on fraud-risk labels, we used the isolation forest (Liu et al., 2008). To assess the impact of diversified supervised signals on operational ranking, a weighted combination of probabilities from logistic regression, random forest and XGBoost was tested. The model specifications are shown in Table 3.

In the case of model development, it was explicitly followed along the pipeline shown in figure 2. The approximate 60% training partition was used for the training of the fitted parameters. Then, the probabilities of validation were chosen in the range of candidate thresholds (0.05 to 0.95) and the threshold with the highest F1 score was selected for each supervised classification model; the same process was carried out for the isolation scores to obtain the threshold for anomaly detection. The ensemble was proposed using the weights of logistic regression (0.25), random forest (0.4), and XGBoost (0.35) and the threshold was determined with the validation data. There were no test labels used in a post hoc threshold adjustment. This separation provides an out-of-sample simulation result for this final test comparison that is not a resubstituting estimate. The framework is model-agnostic and these learners could be replaced by deployment if a public organisation validates an alternative with similar governance and performance controls.

Table 3. Model specifications and validation strategy

Model	Core specification	Threshold rule	Analytical role
Logistic regression	Class-weighted; standardized inputs	Validation F1 maximization	Interpretable linear baseline
Random forest	500 trees; min leaf 3; balanced subsampling	Validation F1 maximization	Nonlinear ensemble
XGBoost	550 rounds; depth 4; learning rate 0.035; class weighting	Validation F1 maximization	Regularized boosted trees

Model	Core specification	Threshold rule	Analytical role
Isolation forest	500 trees; contamination set from training prevalence	Validation F1 on anomaly score	Unsupervised comparator
Weighted ensemble	0.25 LR + 0.40 RF + 0.35 XGBoost	Validation F1 maximization	Diversified supervised score

Note. Hyperparameters were frozen before held-out test evaluation; thresholds were chosen using validation data only.

4.5 Model Evaluation and Validation Strategy

Evaluation included not only accuracy but also usefulness to the minority classes. The accuracy, precision, recall and F1 were reported at the operating threshold selected during validation. Ranking discrimination at every threshold was calculated using ROC-AUC; PR-AUC was also given special attention as precision-recall analysis is more informative in the presence of significant class imbalance (Saito & Rehmsmeier, 2015). To compare with the existing classification practice (Fawcett, 2006), ROC curves were kept. Probabilistic calibration was evaluated using the brier score for supervised outputs. Figures 4 and 5 illustrate the ROC and precision-recall curves, respectively, and Figure 6 assesses the calibration. For the principal supervised models, 300 bootstrap resamples of test set samples were used to compute percentile 95% confidence intervals for AUC-ROC, AUC-PR, F1, recall and precision. The classification trade-offs of false positives and false negatives were quantified by the use of confusion matrices for the selected thresholds. Decrease in average precision due to shuffling features was used to measure the model-agnostic explanation of the predictive dependence of XGBoost, which was measured using permutation importance. This further extends the concept of local explainability, which is part of the current research in XAI (Lundberg & Lee, 2017; Barredo Arrieta et al., 2020). The selection of a threshold wasn't given as a characteristic of any fitted algorithm, but rather as an operational choice.

4.6 Reliability, Robustness, and Reproducibility

The reproducibility controls comprised a fixed random seed, explicit rules for generating variables, stratified partitions, frozen hyperparameters, tuning that was only performed on the validation set, and a completely unchanged test set. Logistic regression, random forest, isolation forest, evaluation metrics, and permutation procedures were provided by scikit-learn (Pedregosa et al., 2011), while gradient boosting was provided by XGBoost. Strategies including multiple algorithm families, bootstrap intervals, threshold-specific confusion patterns and/or calibration were tested for robustness. Internal consistency simply requires that the outputs follow the specified mechanism. It doesn't mean that the same performance would result from administration records. External validation is thus, as mandatory precondition, required for operational deployment.

4.7 Ethical, privacy and public sector data considerations

There is no personal data, records, suppliers or human participants used that leads to a field level allegation. Ethical considerations are, instead, related to deployment: data access, purpose limitation, security, bias monitoring, explainability, human authorization, contestability and retention controls. Sanctions should not be based on model scores; and there must be auditable evidence of the actions taken by the reviewers as a result of any alert (Sampaio et al., 2026; Odilla, 2024).

5. ANALYSIS OF RESULTS AND DATA ANALYSIS

5.1 Descriptive and Preliminary Analysis

The minority class in the benchmark was 11.27% (17,746 low-risk and 2,254 fraud-risk). Figure 3 depicts the minority class of the benchmark. Table 1 shows that the prevalence was almost identical across all the partitions (i.e. training, validation, test) after stratification. The risk structure is overlapping as confirmed by descriptive contrasts in Table 4. For positive cases, non-open procedures accounted for 81.6 % of cases and single bidding accounted for 80.3 % of cases compared to 21.7 % of non-open procedures and 20.3 % of single bidding in low risk cases, respectively. There were also fewer bidders, shorter advertisement time, increased supplier concentration, increased award-to-estimate ratio, and more contract modifications in the case of positive-class transactions. These are associational and not estimates of real world prevalence or causal effects. They are meant to provide a “controlled” benchmark, on which combinations of procurement signals need to become separated in a model.

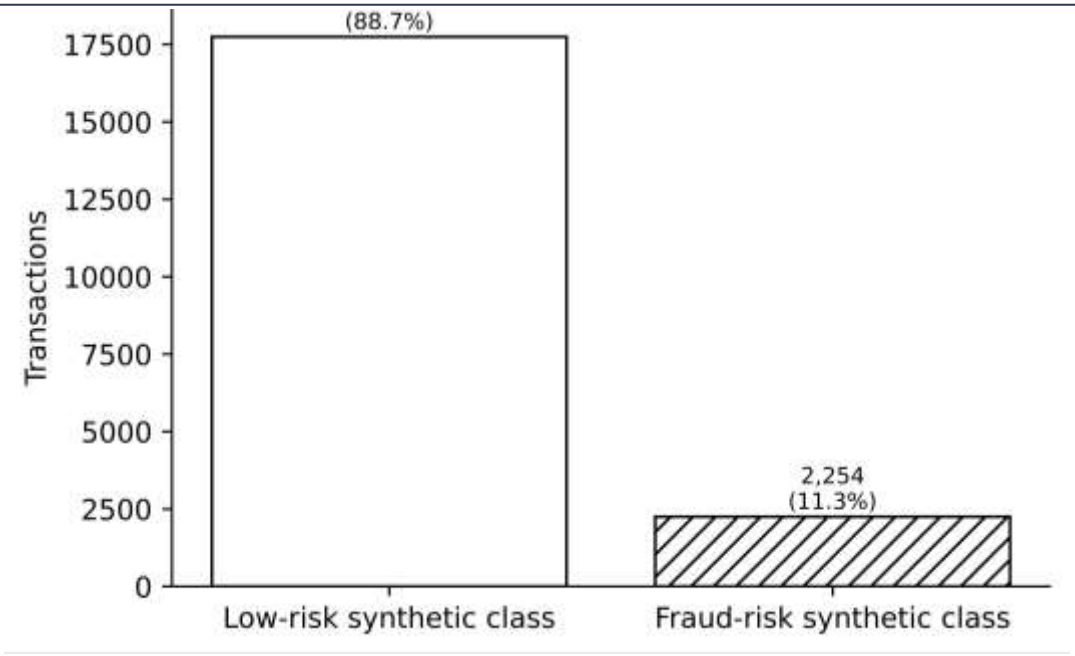

Figure 3. Class distribution in the procurement benchmark

Note. The positive class is intentionally imbalanced to represent a risk-screening problem; it is not a prevalence estimate.

Table 4. Descriptive comparison of selected procurement indicators

Indicator	Low-risk class	positive class
Non-open procedure	21.7%	81.6%
Single bid	20.3%	80.3%
Bidder count	4.075 (1.885)	2.415 (1.411)
Advertisement days	21.207 (8.823)	12.025 (7.965)
Supplier concentration	0.385 (0.182)	0.603 (0.183)
Repeat-award share	0.363 (0.168)	0.471 (0.171)
Award/estimate ratio	1.002 (0.127)	1.109 (0.122)
Contract modifications	0.880 (1.038)	1.841 (1.482)
Payment delay days	21.261 (12.756)	32.159 (15.655)
Split-purchase flag	9.1%	24.1%
Missing documentation	11.3%	26.6%
Supplier network centrality	0.345 (0.146)	0.413 (0.151)

Note. Entries are mean (SD) for continuous/count variables and percentages for binary variables. All values are simulated.

5.2 AI Fraud Detection Performance Results

The held-out test performance of the five scoring strategies is reported in Table 5. Logistic regression had the highest supervised F1 score (0.682), and the highest single model ROC-AUC (0.918); its precision was 0.686 and its recall was 0.678 at its validation selected threshold of 0.785. Random forest gave the F1 score of 0.674, ROC-AUC of 0.912 and the lowest Brier score, 0.058, which was relatively good probability calibration. XGBoost achieved F1 of 0.675 and ROC-AUC of 0.912. The weighted ensemble showed a ROC-AUC of 0.918 and an F1 of 0.681, indicating the same degree of discrimination without significantly outperforming the logistic model.

As can be seen in Figure 4, the supervised ROC curves are well clustered and significantly higher than the chance diagonal. The more challenging minority class comparison is in Figure 5 where the logistic regression achieves PR-AUC of 0.708, the ensemble 0.701, XGBoost 0.691 and random forest 0.688, well above the baseline of 0.113 for the positive class. Despite having a high recall of 0.650 for isolation forest, it was very weak with ROC-AUC of 0.830, PR-AUC of 0.394, precision of 0.324, and F1 of 0.433. The result shows that, compared to label-informed learning, unusualness alone is not as effective in recovering the simulated fraud mechanism. Importantly, the benchmark does not show that logistic regression will be superior in performance on real procurement data to complex models; it shows that added complexity should be justified in terms of incremental value, with the added value being validated. Bootstrap intervals overlaid for supervised models, suggesting a need for caution when testing the meaningfulness of differences in point estimates in simulation.

Table 5. Held-out test performance of fraud-risk scoring models

Model	Threshold	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	Brier
Logistic regression	0.785	0.929	0.686	0.678	0.682	0.918	0.708	0.108
Random forest	0.400	0.925	0.659	0.690	0.674	0.912	0.688	0.058
XGBoost	0.725	0.926	0.671	0.678	0.675	0.912	0.691	0.085
Isolation forest	0.495	0.808	0.324	0.650	0.433	0.830	0.394	—
Proposed ensemble	0.610	0.927	0.674	0.687	0.681	0.918	0.701	0.073

Note. Test N = 4,000 with 451 positives. Isolation forest uses an anomaly score, so Brier score is not reported.

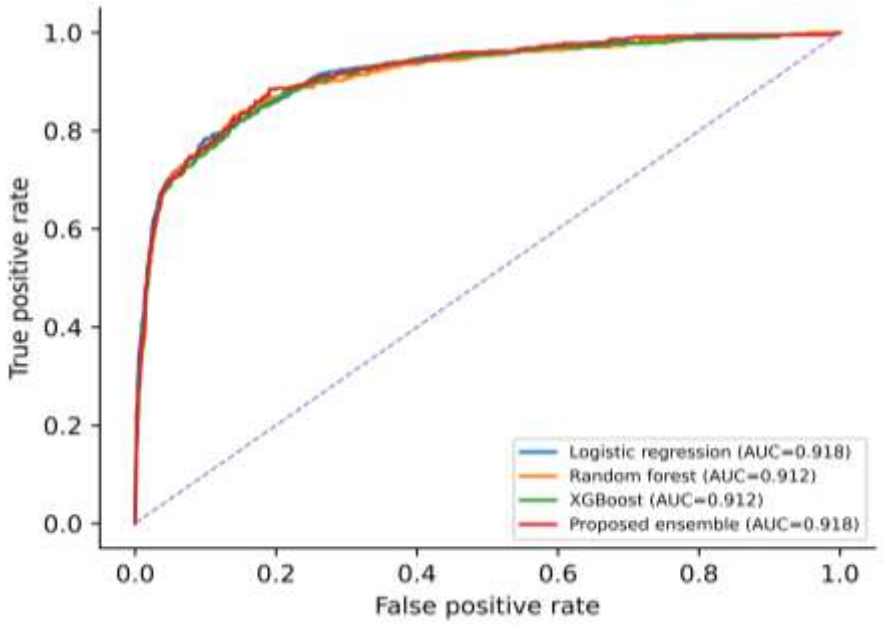

Figure 4. Receiver operating characteristic curves for supervised fraud-risk models

Note. Curves use the held-out test set; the diagonal represents chance ranking.

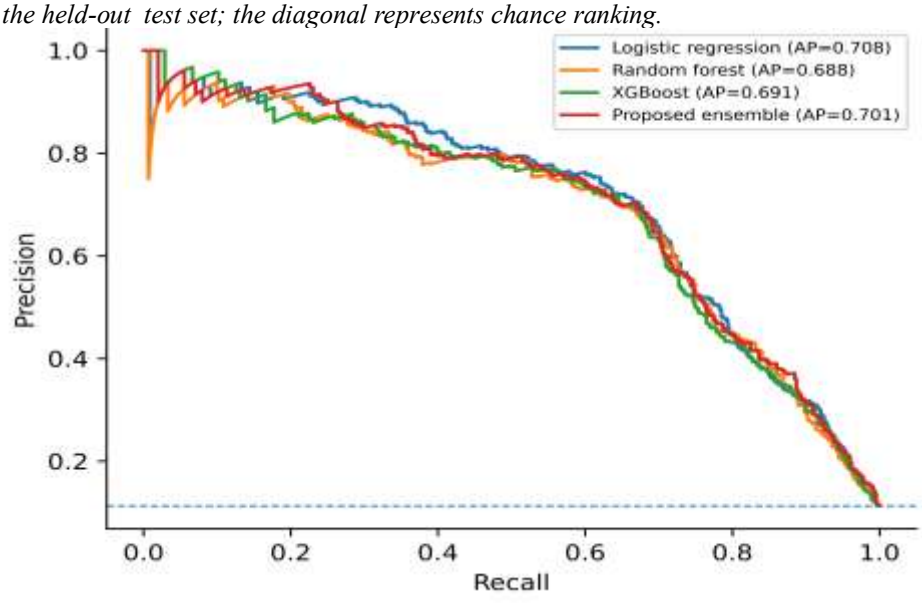

Figure 5. Precision-recall curves for supervised fraud-risk models

Note. The dashed baseline corresponds to the positive-class prevalence in the test set.

5.3 Comparative and Robustness Analysis

Robustness analysis was based on uncertainty, operating thresholds and calibration, not on a single ranking statistic. Supervised Models and their test-set confusion patterns are reported in Table 6. Logistic regression's ROC-AUC interval was 0.901-0.932, its PR-AUC interval was 0.665-0.746, and its F1 interval was 0.649-0.714. The intervals were overlapping for random forest and XGBoost, so there was no separation in the performance. Logistic regression, random forest, and XGBoost were able to identify 306 of 451 simulated positives, and 140, 161, and 150 false positives, respectively. The ensemble had 310 correct identifications and 150 false identifications.

With ranking metrics only, calibration differences are not discovered in Figure 6. Before thresholding, the random forest probabilities were less well calibrated, with a Brier score of 0.058, which was close to the observed rates; class-weighted logistic and boosting probabilities were less well calibrated. This is important for operations as the risk score can be used to determine the intensity of review rather than just to classify the transaction. Therefore, agencies should consider customizing and tracking their scores and then assign consequences of workflow actions to risk bands.

Table 6. Bootstrap robustness and threshold-specific confusion profiles

Model	ROC-AUC 95% CI	PR-AUC 95% CI	F1 95% CI	TN	FP	FN	TP
Logistic regression	0.901–0.932	0.665–0.746	0.649–0.714	3409	140	145	306
Random forest	0.897–0.928	0.647–0.730	0.638–0.708	3388	161	140	311
XGBoost	0.893–0.927	0.648–0.735	0.640–0.708	3399	150	145	306
Proposed ensemble	0.904–0.931	0.658–0.743	0.642–0.718	3399	150	141	310

Note. Percentile intervals use 300 test-set bootstrap resamples. Confusion counts use each validation-selected threshold.

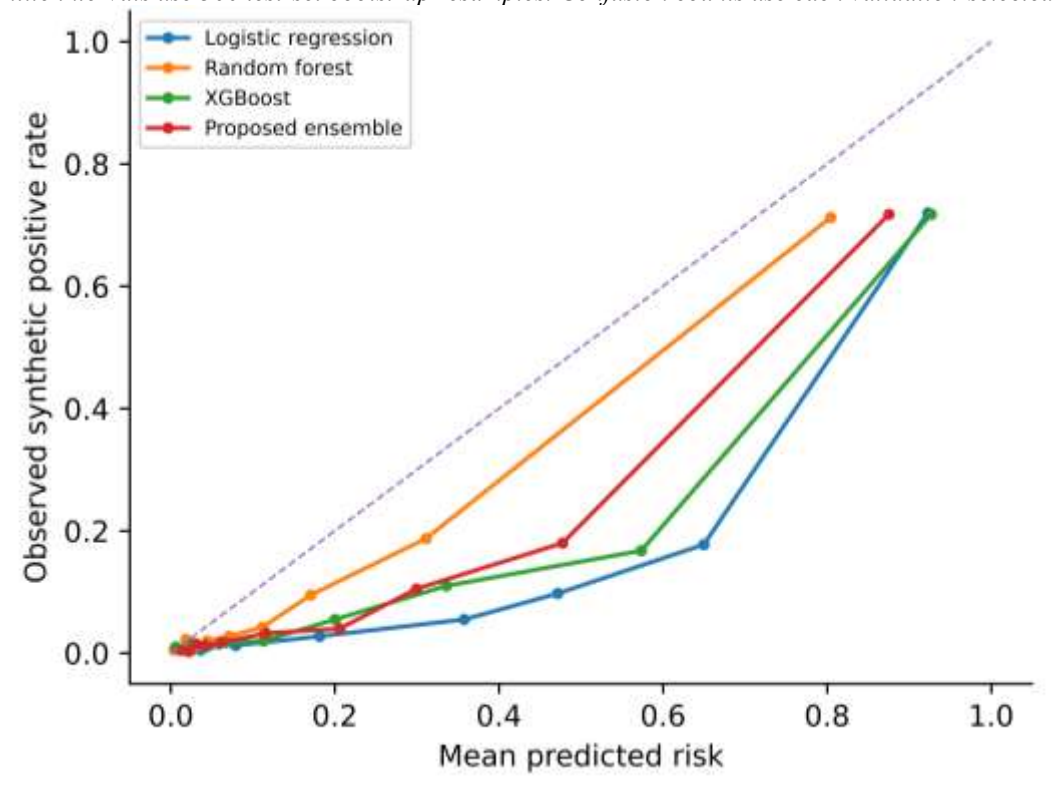

Figure 6. Calibration of supervised fraud-risk probabilities

Note. Observed rates refer only to the test labels; proximity to the diagonal indicates better calibration.

5.4 Fraud Risk Indicators and Framework-Level Results

What the nonlinear model learned from the mechanism is clarified by feature analysis. Table 7 and Figure 7 contain the values of permutation importance obtained for the XGBoost model, which is calculated as the loss of average precision when each predictor is shuffled. The biggest drop was in single bidding (-0.243), followed by the non-open procedure (-0.201). Supplier concentration accounted for 0.061, advertisement duration for 0.053, contract modifications for 0.020, supplier network centrality for 0.015, award to estimate ratio for 0.014, and repeat-award share for 0.013. The rankings correspond to the conceptual differentiation between procedure restriction, supplier dependence, irregular procedure performance and relational concentration.

The interpretation was supported by an ablation test. Only the procedure variables yielded a PR-AUC of 0.593 in the XGBoost model. Incorporating supplier concentration, repeated award, and network centrality, the PR-AUC was raised to 0.657, with the complete feature set at 0.691. Therefore, process and relational variables provided information in addition to the competition variables in the benchmark. But, don't take 'importance' as a given. The analysis is not designed to check whether a feature constitutes misconduct; rather, it is designed to determine whether the framework can be used to retrieve structured risk signals from the data generated by the data generator. Explanations should lead auditors to evidence and records in deployment, and not be a single piece of evidence for adverse action.

Table 7. Permutation importance for XGBoost fraud-risk scoring

Feature	Mean AP decrease	SD
Single bid	0.243	0.018
Non-open procedure	0.201	0.017
Supplier concentration	0.061	0.008
Advertisement days	0.053	0.009
Contract modifications	0.020	0.007
Supplier network centrality	0.015	0.004
Award/estimate ratio	0.014	0.004
Repeat-award share	0.013	0.004
Bidder count	0.008	0.003
Split-purchase flag	0.006	0.003

Note. Importance is the decline in test average precision after shuffling a feature across eight repeats; values indicate predictive dependence, not causality.

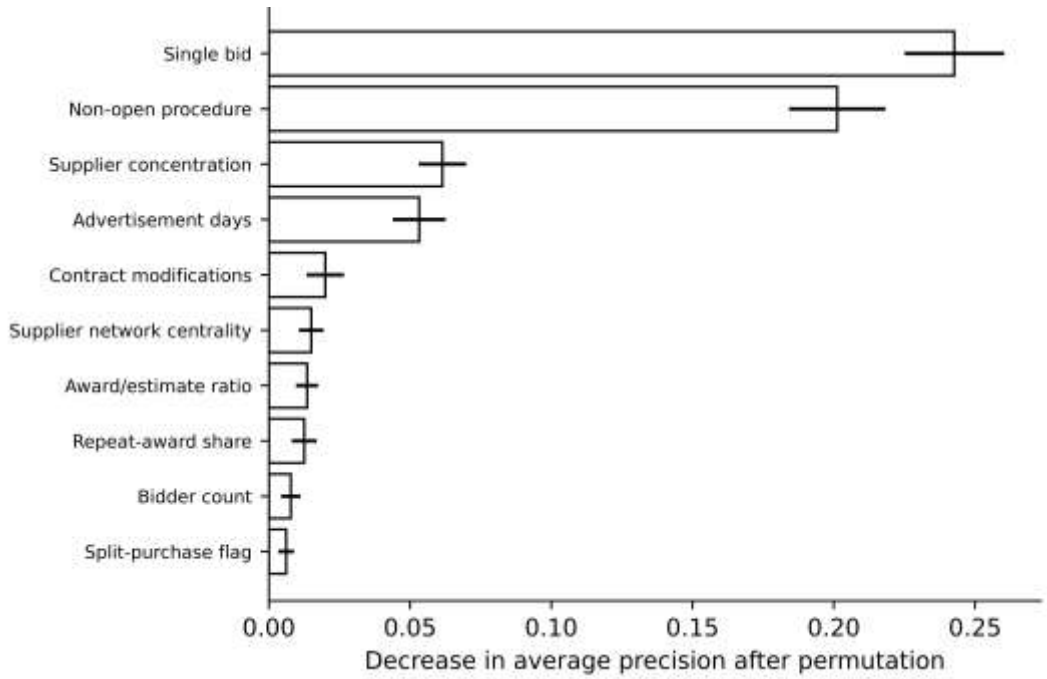

Figure 7. Permutation importance of the ten leading XGBoost predictors

Note. Error bars show the standard deviation across permutation repeats on the test set.

5.5 Hypothesis/Research Question Evaluation

The research questions, hypotheses, and evidence are summarized in Table 8. The results answer to RQ1: restrictive procedure signals are dominant in predicting, whereas the information provided by supplier, execution, and relational features is measurable. The negative class and non-open procedure and single bidding are clustered in it, which makes it a rank influential predictor and makes it supported. H2 gets support due to value being added by procedure-only modelling is not the only one, as supplier concentration and network centrality do. The PR-AUC is higher for the procedure variables (0.593) compared to the full feature set (0.691), which supports H3. Table 5 answers RQ2 and H4: all supervised methods provide higher PR-AUC and F1 than isolation forest.

The framework does not cover the RQ3, which is addressed by statistics. Proposition P1 is met with feature attribution, audit triage, reviewer documentation and governance controls for each score. This is not an indication of institutional legitimacy, it is an indication of design completeness. Agency data, stakeholder evaluations, fairness testing, legal review, and anticipated audit results are all needed to confirm in the real world. The difference safeguards the utility of screening from evidentiary overreach.

Table 8. Research question, hypothesis, and proposition evaluation

Item	Expectation	Evidence	Assessment
RQ1 / H1	Restrictive procedure signals increase risk	Non-open and single-bid indicators show the largest class contrasts and importance	Supported in simulation
RQ1 / H2	Supplier concentration and relational embeddedness increase risk	Adding supplier/network variables raises XGBoost PR-AUC from 0.593 to 0.657	Supported in simulation
RQ1 / H3	Execution/process irregularities add information beyond procedure variables	Full features raise XGBoost PR-AUC from 0.657 to 0.691	Supported in simulation
RQ2 / H4	Supervised learning discriminates better than unsupervised anomaly detection	Supervised PR-AUC 0.688–0.708 versus isolation forest 0.394	Supported in simulation
RQ3 / P1	Explainability and human review are required governance layers	Architecture links score, explanation, reviewer action, logging, and feedback	Analytically satisfied

Note. Assessments apply to the benchmark and framework logic only; they do not constitute real-world causal or legal findings.

6. DISCUSSION AND IMPLICATIONS

6.1 Interpretation of Principal Findings

The results suggest that an efficient procurement-fraud model can be viewed as a decision structure, rather than a contest to determine the most elaborate classifier. The logistic regression in the benchmark was equivalent to or more adaptable learners in discrimination and random forest provided better calibration. This trend supports one of the principles of public analytics: the complexity of a model needs to be incrementally valuable, as simpler models can be easier to explain, to validate, and to challenge (Rudin, 2019). The framework thus allows substitution of algorithms, but demands evidence related to discrimination, calibration, stability, and interpretability.

Procedural restriction and single bidding were the best predictors, which are correlated with the literature that considers constrained competition and discretionary choice of procedure as a corruption-risk marker, not a corruption-evidence indicator (Fazekas and Kocsis, 2020; Decarolis and Giorgiantonio, 2022; Siino et al., 2026). The concentration of suppliers, centrality of networks, change, and process variables enhanced discrimination as compared to procedure-only modelling. This upholds a stratified perspective of risk where suspiciousness arises out of combinations and sequences, rather than single flags. It further describes how the graph and relational approaches might bring value in case of transactional attributes which are either insufficient or are formed strategically (Potin et al., 2023; Muñoz-Cancino and Ríos, 2025).

The supervised-versus-unsupervised comparison also demonstrates that there is another purpose of anomaly detection. Isolation forest found the strange records, yet strangeness was not overlaid with the simulated mechanism of fraud-risk. In applications, hybrid systems can continue to rely on the detection of anomalies to identify new patterns, and the supervised models rank familiar risk signatures. Human verification is still needed to differentiate exceptions, incorrect data, and malpractice.

6.2 Comparison to Existing Research.

Aaraujo et al. (2024) demonstrate that transaction-level predictive screening is justified by the ability of machine-learning techniques to identify anomalous proposals in bidding. The current benchmark concurs with that orientation but places an admonition: it is not assured that the complexity is low and that the evaluation should be based on the minority-class precision, recall, and calibration rather than merely on accuracy. This aligns with the approach to imbalanced learning by He and Garcia (2009) and with the methodological argument of precision-recall analysis that Saito and Rehmsmeier (2015) present.

Sampaio et al. (2024) highlight the ethical issues related to AI-based fraud detection, whereas Sampaio et al. (2026) transform them into a reliable process of procurement. The suggested framework can be used to supplement this work as it connects the ethical control with a comparative statistical layer and a distinct human triage stage. It also aligns with Westerski et al. (2021) and Velasco et al. (2021) who locate explainable analytics as decision support. This break with the purely predictive research is then institutional: useful detection is one which is not only characterized by risk finding, but also by the generation of evidence which can be reviewed, challenged, recorded and managed.

6.3 Theoretical and Conceptual Contributions.

Theoretically, the research connects the corruption-risk measurement to responsible public-sector AI via the accountable risk ranking concept. The traditional red-flag research involves observation of the observed conditions of procurement, and the AI research involves learning patterns based on combinations of variables. Governance scholarship introduces the needs of transparency, contestability, and human authority (Zuiderwijk et al., 2021; Straub et al., 2023). A combination of these traditions creates a three-layer contribution risk indicators characterize meaningful inputs; machine learning approximates suspicion; and governance limits the potential use of scores in administrative action.

The framework also conceptually isolates predictive validity and institutional validity. High ROC-AUC or PR-AUC may indicate that a model identifies a benchmark class, but not fairness, legality, causal correctness, or procedural legitimacy. Those properties require evidence and organizational processes. This distinction assists in avoiding a category error in algorithmic governance: predictive performance as a reason enough to make consequential decisions in the state.

6.4 Implications on Managerial and Public-Sector Procurement.

To the procurement managers and audit units, this framework suggests a levelled review process. Routine controls can be applied to low-risk transactions; document verification can be carried out on medium-risk cases; and enhanced review, conflict checks, supplier-network inspection, or specialist audit can be initiated in high-risk cases. False positives and false negatives ought to have their relative cost and the capacity to review reflected in the thresholds, and not a fixed statistical threshold. The risk score, leading contributing indicators, source-record links, model version, and reviewer disposition should be shown in dashboards. Monitoring should be conducted periodically to compare the alert volumes, confirmation rates, drift, calibration, subgroup error patterns and reviewer overrides. These practices turn AI into an accusation mechanism, but with priority, to workload, keeping professional judgment intact, but making selection more systematic, documented, and available to scrutiny over time.

6.5 Policy, Governance and Accountability Implications.

Policy protections must make sure that algorithmic risk is not a corruption finding or substitute of evidentiary norms. Lawful purposes, minimum data-quality requirements, access controls, appeal or correction routes, audit logs, and independent model review should be set by the procurement authorities. The due-process issues found by Sánchez-Graells (2024) and fairness risks explored by Odilla (2024) need to be governed, not post-discovery. Public reporting may also reveal model intent, performance ranges, supervision arrangements and overall effects, but not expose detection limits that may be used to game.

7. LIMITATIONS AND FUTURE RESEARCH.

7.1 Research Limitations

The restriction is external validity. Every transaction, label, dependency, and performance are , and thus the outcomes cannot give an approximation of actual prevalence of fraud and model performance of any given public organization. The generator is based on literature-founded logic of risk but fails to replicate rules in the institution, supplier strategy, data-quality issues, or fraud pattern emergent. The benchmark also makes the temporal dependence, procurement types, geographic and jurisdictional heterogeneity, and feedback between detection and offender adaptation easier. Therefore, findings only exhibit framework Behavior.

7.2 Future Research Directions

The architecture should be validated with the future research by the use of accessible procurement records and the defensible outcome labels. Research ought to juxtapose interpretable and complex models in the audit processes, assess temporal and cross-jurisdiction transfer, and network, text, document, and graph capabilities. The analysis of fairness should look at the differences in errors without the assumption that the attribute under protection is a proxy of fraud. The researchers are to examine the trust of the reviewers, the fatigue of alerts, explanations, appeals, concept drift, adaptive adversariality, and cost sensitive thresholds. Controlled field trials might find out whether AI-assisted triage can lead to better audit yield, timeliness and consistency.

8. CONCLUSION

8.1 General Conclusions and Research Contributions.

This paper has created and simulated an AI-based digital procurement fraud detection model in the public-sector organisations. The framework integrates red flags of the procurement process, comparative machine learning, calibration, explainability, human audit triage, and governance control. In the 20,000-transaction benchmark, supervised classifiers were able to beat unsupervised anomaly detection, and logistic regression was able to compete with more complicated models. The strongest signals were restrictive procedures and single bidding, and the variables of supplier, relational and execution variables provided discriminatory information. The outcomes of this research respond to the research questions by showing an analytically sensible approach to prioritizing suspicious transactions that does not consider model outputs as evidence of misconduct. They further support the necessity to assess the public-sector AI based on transparency, accountability, and reviewability and predictive performance (Straub et al., 2023; Sampaio et al., 2026).

8.2 Final Significance of the Proposed Framework.

The importance of the framework is the fact that it puts AI as support of decisions, and not as something that enforces them. Its input is a route between the procurement-data to the risk-score, explanation, reviewer-assessment, and disposition. It dissociates predictive validity and institutional legitimacy, diminishing the risk that predictive performance itself will turn into evidentiary authority. Prior to deployment, organizations need to test models on data, test fairness and drift, calibrate thresholds to audit capacity, and govern. In such circumstances, AI will be able to enhance monitoring without compromising due process and accountability..

REFERENCE:

1. Araújo, H. R. F., Leite, P. F., Honório, J. J. C. M., Souza, I. M. L., Albuquerque, D. W., & Santos, D. F. S. (2024). Detection of anomalous proposals in governmental bidding processes: A machine learning-based approach. Anais do Latin American Symposium on Digital Government (LASDiGov), 121–132. https://doi.org/10.5753/wcge.2024.2888

2. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

3. Bosio, E., Djankov, S., Glaeser, E., & Shleifer, A. (2022). Public procurement in law and practice. American Economic Review, 112(4), 1091–1117. https://doi.org/10.1257/aer.20200738

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

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

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

7. Falcón-Cortés, A., Aldana, A., & Larralde, H. (2022). Practices of public procurement and the risk of corrupt behaviour before and after the government transition in México. EPJ Data Science, 11, 19. https://doi.org/10.1140/epjds/s13688-022-00329-7

8. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

9. Fazekas, M., & Kocsis, G. (2020). Uncovering high-level corruption: Cross-national objective corruption risk indicators using public procurement data. British Journal of Political Science, 50(1), 155–164. https://doi.org/10.1017/S0007123417000461

10. Fazekas, M., Tóth, I. J., & King, L. P. (2016). An objective corruption risk index using public procurement data. European Journal on Criminal Policy and Research, 22(3), 369–397. https://doi.org/10.1007/s10610-016-9308-z

11. Ferwerda, J., Deleanu, I. S., & Unger, B. (2017). Corruption in public procurement: Finding the right indicators. European Journal on Criminal Policy and Research, 23(2), 245–267. https://doi.org/10.1007/s10610-016-9312-3

12. Gallego, J., Rivero, G., & Martínez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. International Journal of Forecasting, 37(1), 360–377. https://doi.org/10.1016/j.ijforecast.2020.06.006

13. García Rodríguez, M. J., Rodríguez Montequín, V., Ortega Fernández, F., & Villanueva Balsera, J. M. (2019). Public procurement announcements in Spain: Regulations, data analysis, and award price estimator using machine learning. Complexity, 2019, 2360610. https://doi.org/10.1155/2019/2360610

14. Guida, M., Caniato, F., Moretto, A., & Ronchi, S. (2023). The role of artificial intelligence in the procurement process: State of the art and research agenda. Journal of Purchasing and Supply Management, 29, 100823. https://doi.org/10.1016/j.pursup.2023.100823

15. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

16. Jiménez, A., Hanoteau, J., & Barkemeyer, R. (2022). E-procurement and firm corruption to secure public contracts: The moderating role of governance institutions and supranational support. Journal of Business Research, 149, 640–650. https://doi.org/10.1016/j.jbusres.2022.05.070

17. Lisciandra, M., Milani, R., & Millemaci, E. (2022). A corruption risk indicator for public procurement. European Journal of Political Economy, 73, 102141. https://doi.org/10.1016/j.ejpoleco.2021.102141

18. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. 2008 Eighth IEEE International Conference on Data Mining, 413–422. https://doi.org/10.1109/ICDM.2008.17

19. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

20. Lyra, M. S., Damásio, B., Pinheiro, F. L., & Bação, F. (2022). Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. Applied Network Science, 7, 83. https://doi.org/10.1007/s41109-022-00523-6

21. Madan, R., & Ashok, M. (2023). AI adoption and diffusion in public administration: A systematic literature review and future research agenda. Government Information Quarterly, 40(1), 101774. https://doi.org/10.1016/j.giq.2022.101774

22. Muñoz-Cancino, R., & Ríos, S. A. (2025). Data-driven transparency: Machine learning and social network analysis for corruption detection in public procurement. Procedia Computer Science, 270, 1788–1795. https://doi.org/10.1016/j.procs.2025.09.299

23. Nai, R., Meo, R., Morina, G., & Pasteris, P. (2023). Public tenders, complaints, machine learning and recommender systems: A case study in public administration. Computer Law & Security Review, 51, 105887. https://doi.org/10.1016/j.clsr.2023.105887

24. Odilla, F. (2024). Unfairness in AI anti-corruption tools: Main drivers and consequences. Minds and Machines, 34, 28. https://doi.org/10.1007/s11023-024-09688-8

25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

26. Potin, L., Figueiredo, R., Labatut, V., & Largeron, C. (2023). Pattern mining for anomaly detection in graphs: Application to fraud in public procurement. In Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track (pp. 69–87). Springer. https://doi.org/10.1007/978-3-031-43427-3_5

27. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

28. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

29. Sampaio, I. G. B., Bernardini, F. C., & Viterbo, J. (2024). Challenges in addressing the ethical aspects of artificial intelligence to detect fraud in public procurement processes. Anais da Conferência Latino-Americana de Ética em Inteligência Artificial. https://doi.org/10.5753/laai-ethics.2024.32440

30. Sampaio, I. G. B., Andrade, E. O., Fontes, S. S. B., Bernardini, F. C., & Viterbo, J. (2026). Trustworthy AI in public administration: The PRO-Trust pipeline for ethical fraud detection in public procurement. Government Information Quarterly, 43(3), 102157. https://doi.org/10.1016/j.giq.2026.102157

31. Sánchez-Graells, A. (2024). Procurement corruption and artificial intelligence: Between the potential of enabling data architectures and the constraints of due process requirements. In S. Williams & J. Tillipman (Eds.), Routledge handbook of public procurement corruption (pp. 29–41). Routledge. https://doi.org/10.4324/9781003220374-6

32. Schneider dos Santos, E., Machado dos Santos, M., Castro, M., & Carvalho, J. T. (2025). Detection of fraud in public procurement using data-driven methods: A systematic mapping study. EPJ Data Science, 14, 52. https://doi.org/10.1140/epjds/s13688-025-00569-3

33. Siino, M., Iezzi, S., & Gara, M. (2026). Corruption risk indicators in public procurement: Definition and evaluation with organised crime data. Socio-Economic Planning Sciences, 105, 102429. https://doi.org/10.1016/j.seps.2026.102429

34. Straub, V. J., Morgan, D., Bright, J., & Margetts, H. (2023). Artificial intelligence in government: Concepts, standards, and a unified framework. Government Information Quarterly, 40(4), 101881. https://doi.org/10.1016/j.giq.2023.101881

35. Velasco, R. B., Carpanese, I., Interian, R., Paulo Neto, O. C. G., & Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. International Transactions in Operational Research, 28(1), 27–47. https://doi.org/10.1111/itor.12811

36. Westerski, A., Kanagasabai, R., Shaham, E., Narayanan, A., Wong, J., & Singh, M. (2021). Explainable anomaly detection for procurement fraud identification—lessons from practical deployments. International Transactions in Operational Research, 28(6), 3276–3302. https://doi.org/10.1111/itor.12968

37. Wirtz, B. W., Langer, P. F., & Fenner, C. (2021). Artificial intelligence in the public sector—A research agenda. International Journal of Public Administration, 44(13), 1103–1128. https://doi.org/10.1080/01900692.2021.1947319

38. Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly, 38(3), 101577. https://doi.org/10.1016/j.giq.2021.101577.