A Machine Learning Approach for Detecting Procurement Inefficiencies and Revenue Leakage in Industrial Supply Chains

Nnenna Linda Akunna
School of Engineering, University of the West of England Bristol, United Kingdom

* Corresponding Author: **Nnenna Linda Akunna**

Article Info

ISSN (online): 2583-6641
Volume: 03
Issue: 06
November – December 2024
Received: 09-10-2024
Accepted: 11-11-2024
Published: 06-12-2024
Page No: 190-205

Abstract
Procurement inefficiencies and revenue leakage represent significant challenges in industrial supply chains, often resulting from contract non-compliance, price deviations, supplier irregularities, and operational inefficiencies. Traditional audit-based systems are largely reactive and insufficient for detecting complex, high-dimensional patterns embedded in modern procurement data. This study proposes a mathematically grounded and machine learning–driven framework for detecting procurement inefficiencies and minimizing revenue leakage. The approach integrates structured data from purchase orders, invoices, and contracts into a high-dimensional feature space and introduces a Procurement Inefficiency Index (PII) to quantify deviations across pricing, delivery, and transaction behavior. A hybrid modeling architecture combining supervised learning (XGBoost and Random Forest) and unsupervised anomaly detection (Isolation Forest) is developed to capture both known and emerging leakage patterns. The framework further incorporates an optimization model to minimize total leakage subject to contractual and operational constraints. Experimental results demonstrate that XGBoost achieves the highest performance with an accuracy of 0.94 and F1-score of 0.915, while anomaly detection enhances the identification of previously unseen inefficiencies. Sensitivity analysis reveals a trade-off between detection rate and false positives across varying thresholds, enabling adaptive decision-making. The findings confirm that the proposed framework significantly improves detection accuracy, supports real-time monitoring, and provides quantifiable metrics for procurement performance evaluation. This research contributes to the integration of machine learning and optimization in supply chain analytics and offers practical implications for enhancing transparency, cost control, and risk management in industrial procurement systems.

Keywords: Procurement inefficiency, Revenue leakage detection, Machine learning in supply chains, Anomaly detection, Optimization-based procurement analytics

1. Introduction

1.1. Background and Motivation

Procurement systems constitute a critical component of industrial supply chains, governing the acquisition of raw materials, services, and operational inputs required for production continuity. However, these systems are inherently vulnerable to inefficiencies and revenue leakage arising from contract non-compliance, price manipulation, duplicate invoicing, and unauthorized purchasing behaviors commonly referred to as maverick spending. These inefficiencies translate into measurable financial losses and distort supply chain optimization, thereby undermining organizational profitability and operational transparency (Seify, 2022; Ezeji, 2024).

From a mathematical standpoint, procurement leakage can be expressed as a deviation between expected contractual expenditure and actual transactional expenditure, typically modeled as:

$$L = \sum_{i=1}^{n} (P_i^{expected} - P_i^{actual}) \cdot Q_i$$

where $P_i^{expected}$ denotes contractually agreed prices, P_i^{actual} represents observed transaction prices, and Q_i is the quantity procured. Such deviations are often subtle, distributed across large datasets, and embedded within high-frequency transactional streams, making them difficult to detect using conventional auditing mechanisms.

Traditional audit-based procurement monitoring systems rely heavily on rule-based logic and retrospective analysis. While effective for detecting known patterns of fraud, these approaches lack scalability and fail to capture evolving, non-linear, and high-dimensional anomalies inherent in modern supply chains. Empirical evidence suggests that rule-based systems struggle with complex fraud typologies, particularly when fraudulent activities mimic legitimate transaction patterns (Shaham *et al*., 2021; Nai *et al*., 2022).

The increasing digitization of procurement processes and the proliferation of enterprise resource planning (ERP) systems have led to the generation of large-scale transactional datasets. This shift has enabled the application of machine learning techniques, including supervised classification, unsupervised anomaly detection, and network-based analysis, to uncover hidden patterns in procurement data (Ononiwu, *et al*., 2023). Machine learning models such as Random Forest, Gradient Boosting, and Isolation Forest have demonstrated strong capabilities in detecting irregularities within high-dimensional datasets, thereby improving fraud detection accuracy and operational efficiency (Wang *et al*., 2024; Liu *et al*., 2008).

Moreover, recent advancements in artificial intelligence have facilitated the integration of structured and unstructured procurement data, including invoices, contracts, and supplier communications, into predictive models (Onwuzurike, and Igba, 2023). These models leverage statistical learning theory to approximate complex decision boundaries, enabling organizations to transition from reactive auditing to proactive risk detection frameworks (Callagher, 2025; Schneider dos Santos, 2023).

In addition, the application of explainable artificial intelligence (XAI) techniques has enhanced the interpretability of machine learning outputs, allowing procurement managers to understand the underlying drivers of detected anomalies. This capability is particularly important in high-stakes industrial environments where decision accountability and regulatory compliance are critical (Bamigwojo *et al*., 2022; Bamigwojo, 2021).

1.2. Problem Statement

Despite the growing adoption of data-driven technologies in supply chain management, procurement inefficiencies and revenue leakage remain persistent challenges due to the complex and dynamic nature of transactional data. Procurement datasets are characterized by high dimensionality, heterogeneity, and temporal variability, which complicate the identification of anomalous patterns. Formally, procurement data can be represented as a multi-dimensional stochastic process:

$$X_t \in \mathbb{R}^d, t = 1,2, \dots, T$$

where d represents the number of features (e.g., price, quantity, supplier attributes, time), and T denotes the number of transactions. The presence of noise and overlapping distributions between normal and anomalous transactions further exacerbates the difficulty of detection.

Conventional statistical methods often assume linear relationships and stationary distributions, which are rarely satisfied in real-world procurement environments. Consequently, these methods fail to capture non-linear dependencies, temporal correlations, and hidden interactions among variables (Anokwuru, *et al*., 2022). As noted in recent studies, procurement fraud and inefficiencies are often embedded within transactional noise and require advanced learning algorithms capable of modeling complex data distributions (Ajeigbe & Moore, 2023; Potin *et al*., 2023).

Another critical challenge lies in the lack of a unified mathematical framework for quantifying procurement inefficiency. Existing approaches primarily rely on heuristic indicators or domain-specific rules, which limit generalizability across industries. There is a need for formalized metrics that integrate economic, operational, and behavioral dimensions of procurement activities.

Furthermore, real-time detection remains a significant limitation in existing systems. Most procurement analytics tools operate in batch-processing modes, resulting in delayed identification of anomalies and increased financial exposure. The integration of machine learning models with streaming data architectures is therefore essential for enabling real-time monitoring and adaptive decision-making.

Given these challenges, there is a compelling need to develop a mathematically grounded and machine learning-driven framework that can:

- Model procurement inefficiencies as quantifiable optimization problems
- Detect revenue leakage using predictive and anomaly detection algorithms
- Operate in real-time environments with high-dimensional transactional data
- Provide interpretable outputs for decision-makers

This study addresses these gaps by proposing a hybrid machine learning approach that integrates statistical modeling, anomaly detection, and optimization techniques to enhance procurement transparency and financial integrity in industrial supply chains.

1.3. Research Objectives

The primary objective of this study is to develop a mathematically grounded and machine learning–driven framework for detecting procurement inefficiencies and revenue leakage within industrial supply chains. Specifically, the study aims to formalize leakage detection as an optimization and prediction problem over high-dimensional transactional data.

First, the study seeks to develop a quantitative leakage detection framework by defining a measurable leakage function that captures deviations between expected and observed procurement behavior. This involves constructing a composite leakage index L_i that integrates price variance, contractual deviation, and transactional irregularities:

$$L_i = \beta_1 \cdot \frac{|P_i^{actual} - P_i^{contract}|}{P_i^{contract}} + \beta_2 \cdot \delta_i + \beta_3 \cdot \tau_i$$

where δ_i represents contract compliance deviation and τ_i denotes temporal irregularity. The objective is to aggregate these components into a unified metric capable of capturing both direct and latent inefficiencies.

Second, the study aims to apply supervised and unsupervised machine learning models to detect both known and unknown patterns of procurement anomalies. Supervised learning models such as Gradient Boosting and Random Forest are employed to classify transactions as compliant or non-compliant based on labeled historical data:

$$\hat{y}_i = f(X_i; \theta)$$

where X_i is the feature vector and θ represents model parameters? In parallel, unsupervised models such as Isolation Forest and Autoencoders are utilized to identify anomalous transactions without prior labeling by estimating deviation from learned data distributions.

Third, the study seeks to establish an interpretable risk scoring mechanism that translates model outputs into actionable decision metrics. A procurement risk score R_i is defined as a probabilistic function of model predictions and anomaly scores:

$$R_i = \gamma_1 \cdot P(y_i = 1 \mid X_i) + \gamma_2 \cdot A_i$$

where $P(y_i = 1 \mid X_i)$ is the predicted probability of leakage and A_i is the anomaly score. The resulting score enables prioritization of high-risk transactions and supports decision-making through interpretable thresholds.

1.4. Research Gaps

Despite the growing body of literature on supply chain analytics and fraud detection, several critical gaps remain that limit the effectiveness of existing approaches in addressing procurement inefficiencies and revenue leakage.

A fundamental gap lies in the absence of a unified mathematical framework for modeling procurement leakage. Most existing studies rely on heuristic rules or isolated statistical indicators, which lack generalizability and fail to capture the multi-dimensional nature of procurement systems. There is limited work that integrates economic, operational, and behavioral variables into a single formalized leakage function.

Another significant gap is the fragmentation between supervised and unsupervised learning approaches. Prior research tends to apply these methods independently, thereby limiting their ability to capture both known fraud patterns and emerging, previously unseen anomalies. A hybrid modeling strategy that combines predictive classification with anomaly detection remains underexplored in procurement contexts.

Additionally, there is a lack of real-time, scalable detection architectures capable of processing high-volume transactional data streams. Existing systems are predominantly batch-oriented, resulting in delayed identification of inefficiencies and increased exposure to financial loss. This limitation is particularly critical in industrial supply chains characterized by continuous procurement activities.

The issue of model interpretability and decision transparency also remains insufficiently addressed. While machine learning models can achieve high predictive accuracy, their outputs are often opaque, making it difficult for procurement managers to understand the drivers of detected anomalies (Onwuzurike, and Kpogli. 2022). This lack of explainability hinders adoption in regulated environments where accountability is essential.

Furthermore, existing studies rarely incorporate optimization-based perspectives that link detection outcomes to actionable cost minimization strategies. The absence of integration between predictive analytics and decision optimization limits the practical impact of research findings.

Finally, there is limited empirical work focusing on industrial-scale procurement datasets, particularly within complex environments such as manufacturing and energy supply chains. This creates a gap between theoretical model development and real-world applicability.

Addressing these gaps requires the development of an integrated framework that combines mathematical modeling, hybrid machine learning techniques, real-time analytics, and interpretable decision support systems.

2. Literature Review

2.1. Procurement Analytics and Supply Chain Risk Modeling

Procurement analytics has evolved as a critical domain within supply chain management, focusing on optimizing purchasing decisions, improving supplier performance, and minimizing financial inefficiencies. Classical procurement models are predominantly grounded in deterministic and statistical approaches such as cost variance analysis and spend analysis frameworks. Cost variance analysis evaluates the deviation between planned and actual procurement costs, typically expressed as:

$$CV = \sum_{i=1}^{n} (P_i^{actual} - P_i^{budget}) \cdot Q_i$$

where P_i^{actual} and P_i^{budget} denote actual and budgeted prices, respectively, and Q_i represents procurement quantity. This model provides a straightforward mechanism for identifying cost overruns but lacks the ability to capture underlying causal factors and nonlinear interactions (Monczka *et al*., 2015; Christopher, 2016).

Spend analysis frameworks extend this approach by aggregating procurement data across suppliers, categories, and time periods to identify inefficiencies and consolidation opportunities (Sanmori, 2024). These frameworks rely on descriptive analytics and statistical summaries to detect anomalies in procurement patterns (van Weele, 2018). While effective for high-level insights, spend analysis is inherently retrospective and does not support predictive or real-time decision-making.

Supply chain risk modeling further incorporates probabilistic and stochastic approaches to account for uncertainties in procurement processes. Models such as Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR) have been adapted to quantify financial risks associated with supply disruptions and price volatility (Tang, 2006; Chopra & Sodhi, 2014). These models assume known probability distributions and often rely on simplifying assumptions that may not hold in complex industrial environments.

Despite their theoretical contributions, classical procurement analytics models exhibit significant limitations in scalability and predictive capability. First, they are constrained by linear assumptions and fail to capture high-dimensional relationships among procurement variables. Second, they are not designed to process large-scale transactional data generated by modern ERP systems. Third, they lack adaptability to evolving procurement patterns, particularly in dynamic and decentralized supply chains (Ivanov *et al*., 2019; Simchi-Levi *et al*., 2014).

Recent studies highlight that procurement inefficiencies and revenue leakage are often embedded within complex, nonlinear interactions across multiple variables, including supplier behavior, contract terms, and temporal dynamics. These characteristics necessitate the adoption of advanced analytical techniques capable of modeling high-dimensional data structures and uncovering hidden patterns (Waller & Fawcett, 2013).

2.2. Machine Learning in Financial Anomaly Detection

The application of machine learning in financial anomaly detection has gained significant traction due to its ability to model complex, nonlinear relationships and process large-scale datasets. Unlike classical statistical methods, machine learning approaches leverage data-driven learning to identify patterns and deviations without requiring explicit assumptions about underlying distributions (Hastie *et al*., 2009).

Supervised learning techniques such as Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) have been widely applied in fraud detection and financial risk modeling. Logistic Regression models the probability of anomaly occurrence as a logistic function:

$$P(y_i = 1 \mid X_i) = \frac{1}{1+e^{-\beta^T X_i}}$$

where X_i is the feature vector and β represents model coefficients? While interpretable, Logistic Regression is limited in capturing nonlinear relationships.

Tree-based ensemble methods such as Random Forest and XGBoost address this limitation by constructing multiple decision trees and aggregating their outputs to improve predictive accuracy. Random Forest reduces variance through bootstrap aggregation, while XGBoost employs gradient boosting to minimize a differentiable loss function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

where $l(\cdot)$ is the loss function and $\Omega(f_k)$ represents regularization terms (Chen & Guestrin, 2016; Breiman, 2001). These models have demonstrated superior performance in detecting structured anomalies in financial datasets.

In contrast, unsupervised learning techniques are designed to identify anomalies without labeled data, making them particularly suitable for detecting previously unseen patterns. Isolation Forest is one of the most widely used methods, which isolates anomalies by randomly partitioning data points. The anomaly score is derived from the average path length required to isolate a data point (Liu *et al*., 2008).

Autoencoders, a class of neural networks, learn compressed representations of input data and reconstruct them with minimal error. Anomalies are identified based on reconstruction error, which is mathematically expressed as:

$$\text{Anomaly Score} = \| x_i - \hat{x}_i \|^2$$

where x_i is the original input and $\hat{x}_i$ is the reconstructed output? High reconstruction errors indicate deviations from learned normal patterns (Sakurada & Yairi, 2014).

The mathematical foundation of anomaly detection is rooted in distance-based and probabilistic measures that quantify deviations from expected behavior. These approaches are particularly effective in high-dimensional spaces where traditional statistical methods fail to capture complex dependencies (Aggarwal, 2017; Chandola *et al*., 2009).

Recent advancements have further integrated graph-based learning and deep learning techniques to model relational dependencies in procurement and financial systems. For example, graph-based anomaly detection methods leverage network structures to identify suspicious supplier relationships and transaction clusters (Akoglu *et al*., 2015). Similarly, deep learning models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks capture temporal dependencies in transactional data, enabling dynamic anomaly detection (Malhotra *et al*., 2016).

Despite these advancements, challenges remain in terms of model interpretability, data imbalance, and computational complexity. Ensemble and hybrid approaches that combine supervised and unsupervised techniques have been proposed to address these limitations, offering improved robustness and generalization in real-world applications (Ngai *et al*., 2011; Bolton & Hand, 2002).

Overall, the integration of machine learning into procurement analytics represents a paradigm shift from reactive monitoring to proactive and predictive risk management. However, there remains a need for unified frameworks that combine mathematical rigor, scalability, and interpretability to effectively address procurement inefficiencies and revenue leakage in industrial supply chains.

2.3. Revenue Leakage Quantification Models

Revenue leakage in industrial supply chains is commonly conceptualized as the monetary loss arising when realized procurement outcomes diverge from expected contractual or operational benchmarks. In its basic transactional form, leakage can be represented as:

$$L = \sum_{i=1}^{n} (P_i^{expected} - P_i^{actual}) \cdot Q_i$$

where $P_i^{expected}$ denotes the benchmark or contractually admissible unit price, P_i^{actual} is the observed transaction price, and Q_i is the procured quantity. This formulation is useful because it converts dispersed procurement irregularities into an additive loss function that can be aggregated across suppliers, materials, time periods, and business units. In procurement analytics, such a function is consistent with the broader logic of cost variance analysis, spend visibility, and exception-based auditing, all of which seek to identify departures from planned or authorized expenditure levels (Monczka *et al*., 2015; van Weele, 2018; Lysons & Farrington, 2020).

Although the expression above captures direct price-related leakage, it remains incomplete when procurement inefficiency is driven by more complex operational and contractual deviations. Industrial supply chains frequently exhibit losses that do not arise solely from price inflation but also from purchase fragmentation, off-contract buying, duplicate invoicing, delayed deliveries, quantity mismatches, and weak supplier governance. For this reason, more advanced leakage quantification models extend the baseline equation into a composite risk-loss structure in which total leakage is decomposed into multiple components:

$$L_{total} = L_{price} + L_{contract} + L_{quantity} + L_{time} + L_{supplier}$$

Such decomposition aligns with the literature on procurement performance measurement and supply chain risk modeling, which emphasizes that financial loss is often the observable consequence of deeper process failures involving compliance, information asymmetry, and coordination breakdowns (Christopher, 2016; Chopra & Sodhi, 2014; Tang, 2006).

A major extension of leakage modeling involves contractual deviation metrics. These metrics quantify the extent to which actual purchasing behavior departs from agreed contract terms, including negotiated prices, approved suppliers, order volumes, and delivery schedules. A simple contractual deviation index may be expressed as:

$$CD_i = w_1 \mid \frac{P_i^{actual} - P_i^{contract}}{P_i^{contract}} \mid + w_2 I(s_i \notin S^{approved}) + w_3 \mid \frac{Q_i^{actual} - Q_i^{approved}}{Q_i^{approved}} \mid$$

where $I(\cdot)$ is an indicator function that equals 1 if the transaction uses a non-approved supplier and 0 otherwise, while w_1, w_2, w_3 are importance weights. This formulation is analytically attractive because it accommodates both continuous deviations and categorical violations within one measurable structure. In practical terms, it captures maverick buying, unauthorized sourcing, contract leakage, and quantity overrun in a unified way. Research in purchasing and supply management has repeatedly shown that off-contract procurement and weak compliance monitoring materially reduce negotiated savings and increase total cost of ownership (Mena *et al*., 2013; Monczka *et al*., 2015; van Weele, 2018).

Another important extension concerns supplier pricing inconsistencies, especially in environments where identical or similar items are sourced repeatedly over time or from overlapping supplier pools. In these cases, leakage is not simply the difference between actual and contract price, but also the instability or abnormal dispersion of supplier pricing relative to historical norms, peer quotations, or market benchmarks. A supplier inconsistency score can be modeled as:

$$SP_i = \mid \frac{P_i^{actual} - \mu_{s(i),c(i)}}{\sigma_{s(i),c(i)} + \epsilon} \mid$$

where $\mu_{s(i),c(i)}$ and $\sigma_{s(i),c(i)}$ are the historical mean and standard deviation of prices charged by supplier s for category c, and ϵ is a small constant to avoid division by zero. This z-score style representation is useful for identifying abnormal supplier behavior, inflated invoicing, price manipulation, or category-level benchmark failure. In the anomaly detection literature, this logic is closely related to outlier scoring and density-based deviation analysis, where unusual distance from the expected distribution serves as evidence of anomalous behavior (Chandola *et al*., 2009; Aggarwal, 2017; He *et al*., 2003).

More recent work in anomaly detection and procurement fraud has pushed leakage quantification beyond static formulas toward representation learning and event-level risk scoring. In this direction, procurement transactions are encoded as feature vectors and reconstructed or classified using machine learning models. Under an autoencoder-based framework, leakage-related abnormality can be inferred from reconstruction error:

$$A_i = \| x_i - \hat{x}_i \|^2$$

where x_i is the original procurement transaction vector and $\hat{x}_i$ is its reconstructed representation. High values of A_i suggest that the transaction differs materially from learned normal procurement behavior. Likewise, neural and tree-based classifiers can estimate the probability that a transaction belongs to a leakage-prone class. These developments indicate a shift from deterministic loss accounting to probabilistic leakage inference, which is particularly suitable for large industrial procurement systems where fraudulent or inefficient behavior is sparse, adaptive, and embedded in noisy transactional streams (Bolton & Hand, 2002; Ngai *et al*., 2011; Bai *et al*., 2023; Shaham *et al*., 2021). Procurement-focused studies also note that AI-based methods can support fraud and leakage identification by revealing hidden transactional patterns not captured by traditional controls.

Taken together, the literature suggests that robust revenue leakage quantification requires a layered modeling approach. Direct loss equations remain necessary for financial materiality estimation, but they are insufficient on their own. Effective leakage models must also incorporate contractual deviation metrics, supplier pricing irregularities, temporal instability, and anomaly-based behavioral scores. This integrated view is especially important in industrial supply chains, where leakage emerges not from one isolated defect but from the interaction of procurement policy failure, process inefficiency, weak supplier discipline, and incomplete monitoring architectures (Waller & Fawcett, 2013; Simchi-Levi *et al*., 2014; Ivanov *et al*., 2019).

2.4. Research Gap

Despite substantial progress in procurement analytics, supply chain risk modeling, and financial anomaly detection, the literature still exhibits several important limitations. The first is the lack of integrated machine learning and optimization frameworks for procurement leakage detection. Existing studies typically fall into one of two categories. One group focuses on descriptive or predictive analytics, using classification, clustering, or anomaly detection to identify suspicious transactions (Onwuzurike, *et al*., 2021). Another group focuses on optimization, including sourcing efficiency, contract allocation, inventory control, or cost minimization. However, these streams are rarely unified into a single framework in which machine learning identifies potential inefficiencies and optimization translates those insights into corrective decision rules or cost-minimizing interventions. This separation weakens the practical value of many models because prediction alone does not guarantee operational improvement (Chopra & Sodhi, 2014; Christopher, 2016; Handfield *et al*., 2019).

A related gap concerns the limited formalization of leakage as a mathematically actionable objective function. Many procurement studies discuss savings erosion, fraud, or compliance failure in conceptual terms, yet do not embed these phenomena within explicit constrained optimization structures (Onwuzurike, 2023). As a result, the literature provides relatively little guidance on how detected anomalies should feed into procurement policy redesign, supplier rationalization, or threshold-based intervention logic. In high-volume industrial environments, the ability to move from anomaly scoring to decision optimization is essential. Without that bridge, analytical outputs remain diagnostic rather than prescriptive (Monczka *et al*., 2015; van Weele, 2018).

The second major gap is the absence of real-time predictive leakage detection systems capable of handling temporally dynamic and high-dimensional procurement data. Much of the traditional procurement control literature is retrospective. It relies on periodic audits, post hoc spend reviews, or exception reports generated after the financial impact has already materialized. Even when machine learning is introduced, many implementations remain batch-oriented and are not designed for continuous monitoring across streaming transactions. This is a serious limitation because procurement inefficiencies in industrial supply chains often evolve rapidly through recurring micro-deviations rather than one-time catastrophic events. Delayed detection increases both cumulative loss and remediation cost (Tang, 2006; Waller & Fawcett, 2013). Recent procurement-oriented AI studies explicitly point to anomaly detection and automated monitoring as promising directions, but they also imply that operational deployment and real-time integration remain underdeveloped.

A third gap involves interpretability and explainability. High-performing models such as boosted trees, deep autoencoders, and neural classifiers can detect subtle anomalies, but procurement professionals often need auditable justifications for why a transaction was flagged. In regulated industrial settings, opaque models can create resistance because compliance teams, auditors, and procurement managers must understand the drivers of model decisions before acting on them. Explainable anomaly detection has begun to address this issue, but the literature remains thin in procurement-specific applications, especially where contractual, supplier, and operational factors interact simultaneously (Shaham *et al*., 2021; Molnar, 2022).

There is also a domain specificity gap. Much of the anomaly detection literature is built around credit card fraud, insurance claims, cybersecurity logs, or generic financial transactions (Amebleh, *et al*., 2021). These domains are methodologically useful, but procurement leakage in industrial supply chains has unique characteristics: contractual hierarchies, approval workflows, supplier dependence, item-category substitution, negotiated pricing corridors, and logistics-linked deviations. Models developed in generic fraud contexts do not always transfer effectively because they may ignore procurement-specific relational structure and institutional constraints (Ngai *et al*., 2011; Phua *et al*., 2010; Akoglu *et al*., 2015).

Finally, there is limited work that unifies transactional, contractual, supplier, and temporal dimensions into one leakage detection architecture. Procurement leakage is fundamentally multi-layered. A transaction may appear normal in price terms but become problematic when evaluated against contract terms, supplier history, or timing behavior. Existing studies often treat these dimensions separately, which reduces detection sensitivity and underestimates cumulative leakage pathways. An effective research response therefore requires an integrated framework that combines mathematically defined leakage functions, machine learning-based anomaly and classification models, and optimization-based decision support within a real-time monitoring architecture.

This study is positioned to address these gaps by proposing a unified approach in which revenue leakage is first formalized quantitatively, then detected through supervised and unsupervised learning, and finally linked to interpretable risk scoring and decision-oriented control mechanisms.

3. Methodology

3.1. Data Representation and Feature Engineering

The effectiveness of any machine learning framework for detecting procurement inefficiencies and revenue leakage depends fundamentally on how procurement data are structured, encoded, and transformed into analytically meaningful representations. In industrial supply chains, procurement transactions are typically recorded across multiple enterprise systems, including Enterprise Resource Planning (ERP), procurement management platforms, and contract lifecycle management systems. These data sources collectively provide a comprehensive view of purchasing behavior, supplier interactions, and contractual compliance.

In this study, the dataset is constructed by integrating three primary sources: Purchase Orders (POs), Invoices, and Contracts. Purchase Orders capture the intent to procure goods or services, including agreed quantities and prices; invoices reflect actual billing information submitted by suppliers; and contracts define negotiated terms such as pricing agreements, approved suppliers, and delivery conditions. The integration of these datasets enables the identification of discrepancies between expected and actual procurement outcomes, which is essential for leakage detection.

To enable machine learning modeling, each procurement transaction is represented as a feature vector in a multi-dimensional space:

$$X = [p_i, q_i, t_i, s_i, c_i, d_i]$$

where each component captures a critical dimension of procurement behavior.

The variable p_i represents the unit price associated with transaction i. This feature is central to leakage detection, as price deviations from contractual benchmarks directly contribute to financial loss. To enhance analytical robustness, price features may be normalized relative to contract price or historical averages:

$$p_i^{norm} = \frac{p_i - \mu_p}{\sigma_p}$$

where μ_p and σ_p denote the mean and standard deviation of prices within a given category.

The variable q_i denotes the quantity procured, which interacts multiplicatively with price in determining total transaction value. Quantity anomalies may arise from over-ordering, duplicate entries, or discrepancies between ordered and delivered quantities. Derived features such as quantity deviation can be defined as:

$$\Delta q_i = q_i^{actual} - q_i^{contract}$$

The temporal variable t_i represents transaction time, capturing when procurement activities occur. Temporal encoding is critical for identifying irregular purchasing patterns such as off-cycle transactions, end-of-period spending spikes, or abnormal ordering frequencies. Time-based features can be transformed into cyclical representations to capture periodicity:

$$t_i^{cyc} = \left[\sin\left(\frac{2\pi t_i}{T}\right), \cos\left(\frac{2\pi t_i}{T}\right)\right]$$

where T represents the time horizon (e.g., 12 months).

The variable s_i denotes the supplier identifier, which is inherently categorical. Supplier-related features are essential for detecting vendor-specific anomalies such as price inconsistencies or preferential procurement behavior. Encoding techniques such as one-hot encoding or embedding representations are employed to transform supplier identifiers into machine-readable formats:

$$s_i \rightarrow \mathrm{e}_{s_i} \in \mathbb{R}^k$$

where e_{s_i} is a learned embedding vector of dimension k.

The variable c_i represents the contract reference, linking each transaction to its governing contractual agreement. This feature enables the computation of contract compliance metrics, including price adherence, supplier authorization, and quantity limits. Contract-based features may include binary indicators of compliance:

$$C_i = \mathbb{I}(p_i \leq p_i^{contract})$$

where $\mathbb{I}(\cdot)$ is an indicator function?

Finally, d_i captures delivery deviation, defined as the difference between expected and actual delivery timelines. Delays in delivery can indicate operational inefficiencies or supplier performance issues and may correlate with financial leakage due to penalties, expedited shipping, or production disruptions:

$$d_i = t_i^{delivery} - t_i^{expected}$$

Beyond these primary features, the study incorporates engineered variables that enhance the predictive power of the model. These include interaction terms (e.g., price–quantity interactions), rolling averages, supplier-level aggregates, and anomaly indicators derived from statistical thresholds. Feature scaling and normalization are applied to ensure comparability across variables, while dimensionality reduction techniques such as Principal Component Analysis (PCA) may be used to mitigate multicollinearity and improve computational efficiency.

From a probabilistic perspective, the feature vector X_i can be viewed as a realization from an unknown distribution $P(X)$, where anomalies correspond to low-probability events:

$$P(X_i) \ll \epsilon$$

for a small threshold ϵ. This interpretation provides the theoretical basis for applying anomaly detection algorithms and probabilistic classifiers in subsequent stages of the methodology.

Overall, the data representation framework transforms heterogeneous procurement data into a structured, high-dimensional feature space that captures economic, temporal, contractual, and behavioral aspects of procurement activities. This representation is essential for enabling machine learning models to detect subtle inefficiencies and revenue leakage patterns embedded within complex industrial supply chains.

3.2 Mathematical Formulation of Procurement Inefficiency

To systematically quantify inefficiencies within procurement transactions, this study introduces a Procurement Inefficiency Index (PII) as a composite metric that captures deviations across pricing, logistics, and transactional behavior. The formulation integrates economic and operational dimensions into a unified mathematical structure, enabling consistent evaluation across heterogeneous procurement datasets.

The inefficiency index for transaction i is defined as:

$$PII_i = \alpha_1 \cdot \frac{|P_i - P_i^{contract}|}{P_i^{contract}} + \alpha_2 \cdot D_i + \alpha_3 \cdot T_i$$

where:

P_i represents the actual transaction price

$P_i^{contract}$ denotes the contractually agreed price

D_i is the delivery delay associated with transaction i

T_i represents transaction irregularity

$\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}^+$ are weighting coefficients reflecting the relative importance of each component?

3.2.1. Price Deviation Component

The first term:

$$\frac{|P_i - P_i^{contract}|}{P_i^{contract}}$$

captures relative price deviation, which serves as a direct indicator of financial leakage. The normalization by $P_i^{contract}$ ensures scale invariance across procurement categories with different pricing structures. This term is particularly sensitive to overpricing, unauthorized discounts, or contract violations.

3.2.2. Delivery Delay Component

The second component, D_i, represents delivery delay, defined as:

$$D_i = \frac{t_i^{actual} - t_i^{expected}}{t_i^{expected}}$$

where t_i^{actual} and $t_i^{expected}$ denote actual and expected delivery times, respectively. Positive values of D_i indicate delays, which may result in production disruptions, penalty costs, or emergency sourcing expenses. This term incorporates operational inefficiency into the index.

3.2.3. Transaction Irregularity Component

The third component, T_i, captures transaction irregularity, which reflects deviations from normal procurement behavior. This can be modeled using statistical or machine learning-based anomaly scores. For instance:

$$T_i = \frac{|x_i - \mu|}{\sigma}$$

where x_i is a transaction feature (or composite score), and μ, σ are the mean and standard deviation of historical transactions. Alternatively, T_i may be derived from anomaly detection models such as Isolation Forest or autoencoders.

3.2.4. Weight Calibration and Normalization

The weights $\alpha_1, \alpha_2, \alpha_3$ determine the contribution of each inefficiency dimension. These weights can be calibrated using:

Expert-driven weighting based on domain knowledge

Data-driven optimization, such as minimizing prediction error:

$$\min_{\alpha} \quad \sum_{i=1}^{n} (PI\,I_i - y_i)^2$$

where y_i represents observed inefficiency labels

To ensure interpretability, the weights are typically normalized:

$$\alpha_1 + \alpha_2 + \alpha_3 = 1$$

3.2.5. Aggregated Inefficiency Measure

At the system level, total procurement inefficiency is obtained by aggregating transaction-level indices:

$$PII_{total} = \sum_{i=1}^{n} PI\,I_i$$

This aggregate metric provides a quantitative measure of inefficiency across the entire procurement system and can be used for benchmarking, monitoring, and optimization.

3.2.6. Theoretical Implications

The proposed PII formulation offers several analytical advantages:

1. It integrates financial and operational inefficiencies into a single metric
2. It is scalable across large transactional datasets
3. It supports integration with machine learning models as either a target variable or featureinput
4. It provides a continuous risk spectrum, enabling threshold-based decision rules

From an optimization perspective, minimizing procurement inefficiency can be formulated as:

$$\min \sum_{i=1}^{n} PI\,I_i$$

subject to contractual and operational constraints. This establishes a direct link between inefficiency detection and cost minimization strategies.

Overall, the Procurement Inefficiency Index serves as a mathematically rigorous foundation for identifying, quantifying, and mitigating inefficiencies and revenue leakage in industrial supply chains, bridging the gap between descriptive analytics and predictive decision-making.

3.3. Machine Learning Model Design

To detect procurement inefficiencies and revenue leakage with high precision, this study adopts a hybrid machine learning architecture that combines supervised classification with unsupervised anomaly detection. This dual structure is appropriate because procurement leakage manifests in two distinct ways. First, some leakage patterns are already known from historical audit outcomes and can therefore be learned through labeled classification. Second, many inefficiencies are novel, weakly structured, or hidden within complex transactional noise, making them better suited to unsupervised anomaly detection. The proposed design therefore improves both predictive accuracy and detection coverage.

Let the engineered procurement feature space be denoted by

$$X_i = [p_i, q_i, t_i, s_i, c_i, d_i, \ldots] \in \mathbb{R}^m$$

for transaction i, where the vector may include raw, normalized, and derived features. The modeling objective is to learn whether a transaction is associated with procurement leakage or abnormal inefficiency.

3.3.1. Supervised Model for Leakage Classification

The supervised learning component is formulated as a binary classification problem:

$$y_i = f(X_i; \theta)$$

where:

X_i is the input feature vector for transaction i

$y_i \in \{0,1\}$, with $y_i = 1$ indicating leakage or procurement inefficiency

θ denotes model parameters

$f(\cdot)$ is the learned classification function?

The output of the model is a binary leakage decision or, more generally, a leakage probability:

$$\hat{P}(y_i = 1 \mid X_i)$$

This formulation enables ranking of procurement records according to leakage risk and supports threshold-based intervention.

Two classification models are adopted.

- **XGBoost**

Extreme Gradient Boosting is used because of its strong performance in structured tabular datasets and its ability to capture nonlinear interactions among procurement variables such as price deviations, supplier behavior, and delivery delays. XGBoost constructs an additive ensemble of decision trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), f_k \in \mathcal{F}$$

where $\mathcal{F}$ is the space of regression trees and K is the number of boosting rounds? Model estimation proceeds by minimizing a regularized objective:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $l(\cdot)$ is the classification loss function and $\Omega(f_k)$ is the tree complexity penalty. This structure improves generalization while preserving sensitivity to complex leakage patterns.

- **Random Forest**

Random Forest is employed as a robust ensemble baseline. It generates multiple decision trees using bootstrap samples and random feature selection, then aggregates their outputs through majority voting:

$$\hat{y}_i = \text{mode}\{T_1(X_i), T_2(X_i), \dots, T_B(X_i)\}$$

where $T_b(\cdot)$ denotes the prediction of tree b, and B is the total number of trees. Random Forest is particularly useful in procurement datasets where variables may be heterogeneous, partially correlated, and mixed in scale.

The supervised component therefore produces a binary leakage classification, identifying whether a transaction belongs to a normal or leakage-prone class. This is especially effective when historical labels are available from audits, compliance reviews, or expert annotation.

3.3.2. Unsupervised Model for Anomaly Detection

Although supervised models are powerful, they depend on labeled data and are therefore limited in detecting previously unseen forms of procurement leakage. To address this problem, the study incorporates an unsupervised anomaly detection module based on Isolation Forest.

Isolation Forest operates on the principle that anomalous observations are easier to isolate than normal ones because they tend to be rare and statistically distinct. For a given transaction x, the model computes the path length required to isolate it within a random partitioning tree. Let

$$h(x) = \text{average path length}$$

denote the expected number of splits needed to isolate transaction x across an ensemble of random trees. An anomaly score is then computed as:

$$\text{Anomaly Score}(x) = 2^{-\frac{E(h(x))}{c(n)}}$$

where:

$E(h(x))$ is the expected path length of transaction x

$c(n)$ is the average path length of an unsuccessful search in a binary search tree of sample size n

The normalization constant $c(n)$ is typically given by

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}$$

where $H(n)$ is the harmonic number, approximated as

$$H(n) \approx \ln(n) + \gamma$$

with γ denoting the Euler-Mascheroni constant.

If a transaction has a short, expected path length, then it is isolated quickly and is more likely to be anomalous. Hence, higher anomaly scores indicate stronger suspicion of procurement inefficiency, revenue leakage, or irregular supplier behavior.

This mechanism is appropriate for procurement environments because many leakage events are not repeated frequently enough to generate reliable class labels. Examples include unusual one-time supplier substitutions, invoice duplication patterns, or off-contract orders that do not match historical fraud templates.

3.3.3. Hybrid Detection Logic

The supervised and unsupervised components are integrated to form a hybrid decision framework. Let

$$\hat{y}_i^{(s)} = f(X_i; \theta)$$

be the supervised classification output, and let

$$A_i = 2^{-\frac{E(h(x_i))}{c(n)}}$$

be the Isolation Forest anomaly score. A composite leakage risk score can then be defined as:

$$R_i = \lambda_1 \hat{P}(y_i = 1 \mid X_i) + \lambda_2 A_i$$

subject to

$$\lambda_1 + \lambda_2 = 1, \lambda_1, \lambda_2 \geq 0$$

where R_i is the final transaction-level risk score and λ_1, λ_2 are weighting coefficients. This formulation combines known-pattern classification with unknown-pattern anomaly detection, thereby improving both sensitivity and robustness.

A transaction is flagged for review if

$$R_i > \tau$$

where τ is a decision threshold chosen using validation data, risk tolerance, or operational cost trade-offs.

3.3.4. Model Training and Evaluation Strategy

The supervised models are trained using labeled procurement records, with data divided into training, validation, and testing subsets. To reduce class imbalance effects, resampling or cost-sensitive learning may be introduced where leakage cases are rare. Hyperparameters for XGBoost and Random Forest are tuned using cross-validation.

The unsupervised Isolation Forest is trained on the broader procurement dataset without requiring labels. Its outputs are then compared with historical audit findings or expert judgments to assess practical detection value.

Performance of the hybrid framework is evaluated using standard classification metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve for the supervised component, while anomaly ranking quality and detection sensitivity are assessed for the unsupervised component.

3.3.5. Methodological Significance

The proposed machine learning model design has three major advantages. First, it captures both structured leakage patterns and unstructured anomalies. Second, it provides a scalable analytical architecture suitable for high-volume industrial procurement systems. Third, it creates the basis for an interpretable downstream risk-scoring framework that supports procurement monitoring, audit prioritization, and revenue protection.

In methodological terms, the hybrid model moves beyond purely descriptive procurement analytics by introducing a predictive, mathematically grounded framework capable of detecting inefficiencies that are both historically known and statistically emergent.

3.4. Optimization Model for Cost Leakage Minimization

Beyond detecting procurement inefficiencies and anomalous transactions, an effective analytical framework must also provide a decision structure for reducing financial loss. For this reason, the study formulates procurement leakage control as a constrained optimization problem in which total leakage is minimized subject to contractual and operational limits. This allows the detection layer to be connected directly to a prescriptive decision mechanism.

Let L_i denote the estimated leakage associated with transaction i. The objective is to minimize the aggregate procurement leakage across all transactions in the planning horizon:

$$\min \sum_{i=1}^{n} L_i$$

where n is the total number of procurement transactions under consideration?

This formulation treats leakage as an additive cost term that can arise from price deviations, quantity overruns, supplier non-compliance, or other inefficiency signals identified in the earlier stages of the model. Using the leakage definition introduced previously, each transaction-level loss may be written as:

$$L_i = (P_i - P_i^{contract})Q_i$$

for cases in which $P_i > P_i^{contract}$, and more generally as a penalized cost function:

$$L_i = \max\{0, (P_i - P_i^{contract})Q_i\} + \eta_1 D_i + \eta_2 T_i$$

where:

P_i is the actual procurement price

$P_i^{contract}$ is the contract price

Q_i is the procured quantity

D_i is delivery delay

T_i is transaction irregularity

η_1, η_2 are penalty weights assigned to non-price inefficiency components?

This generalized representation reflects the fact that cost leakage is not limited to direct overpricing but can also emerge from delayed delivery, irregular transaction timing, and other non-compliant behaviors that generate indirect financial loss.

3.4.1. Contract Price Constraint

The first optimization constraint ensures that procurement prices remain within a permissible contractual tolerance band:

$$P_i \leq P_i^{contract} + \epsilon$$

where $\epsilon \geq 0$ represents an allowable deviation threshold. This threshold may account for minor price fluctuations due to market conditions, taxes, logistics adjustments, or approved escalation clauses. If $\epsilon = 0$, the formulation enforces strict contractual compliance. If $\epsilon > 0$, then limited flexibility is permitted while still controlling excessive price inflation.

This constraint is important because many procurement leakages arise from gradual departures from negotiated prices rather than outright violations. Embedding the threshold within the optimization model allows management to define acceptable tolerance levels explicitly.

3.4.2. Approved Quantity Constraint

The second constraint enforces that procured quantities do not exceed approved or authorized amounts:

$$Q_i \leq Q_i^{approved}$$

where $Q_i^{approved}$ denotes the pre-approved purchase quantity under budget, requisition, or contract terms. This restriction controls excessive ordering, duplicate processing, unauthorized volume expansion, and inefficient inventory accumulation. In industrial procurement settings, such quantity deviations may lead to working capital distortion, unnecessary storage cost, and hidden revenue leakage.

3.4.3. Extended Constrained Formulation

To capture procurement decision-making more realistically, the model can be extended with binary compliance variables and transaction acceptance rules. Let $z_i \in \{0,1\}$ indicate whether transaction i is accepted for execution under the optimization policy. Then the problem may be reformulated as:

$$\min \sum_{i=1}^{n} z_i L_i$$

subject to

$$P_i z_i \leq (P_i^{contract} + \epsilon) z_i$$
$$Q_i z_i \leq Q_i^{approved} z_i$$
$$z_i \in \{0,1\}, i = 1,2,\dots,n$$

In this structure, only transactions satisfying compliance conditions remain feasible. Transactions with excessive price or quantity deviations can be flagged, rejected, renegotiated, or escalated for human review. This creates a decision layer that transforms anomaly detection results into cost-control actions.

3.4.4. Integration with Machine Learning Outputs

A major advantage of this optimization framework is that it can incorporate outputs from the supervised and unsupervised machine learning models. Let R_i denote the composite leakage risk score derived in Section 3.3. Then leakage may be weighted by model-implied risk:

$$\min \sum_{i=1}^{n} R_i L_i$$

subject to

$$P_i \leq P_i^{contract} + \epsilon$$
$$Q_i \leq Q_i^{approved}$$

In this form, the optimization process places greater emphasis on high-risk transactions, thereby prioritizing cost minimization where the likelihood of procurement leakage is greatest. This is particularly useful when audit resources are limited and intervention must be targeted selectively.

A threshold-based control rule may also be introduced:

$$z_i = \begin{cases} 1, & R_i \leq \tau \\ 0, & R_i > \tau \end{cases}$$

where τ is the acceptable procurement risk threshold. Transactions with scores above τ are excluded from automatic approval and routed for manual investigation.

3.4.5 Lagrangian Representation

For analytical interpretation, the constrained problem may be expressed using a Lagrangian function:

$$\mathcal{L} = \sum_{i=1}^{n} L_i + \sum_{i=1}^{n} \lambda_i \left(P_i - P_i^{contract} - \epsilon\right) + \sum_{i=1}^{n} \mu_i \left(Q_i - Q_i^{approved}\right)$$

where $\lambda_i \geq 0$ and $\mu_i \geq 0$ are Lagrange multipliers associated with the price and quantity constraints, respectively. These multipliers provide useful economic interpretation. Specifically, they measure the marginal increase in total leakage associated with relaxing each procurement control condition. High multiplier values indicate that the corresponding constraint plays a critical role in limiting revenue leakage.

3.4.6. Practical Interpretation

From an operational standpoint, the optimization model provides a formal mechanism for translating procurement monitoring into financial control. It enables organizations to:

1. minimize cumulative leakage across transactions
2. enforce price and quantity discipline
3. prioritize intervention on high-risk procurement events
4. embed machine learning insights into procurement approval workflows

The model is particularly suitable for industrial supply chains where leakage often occurs through repeated low-value deviations that accumulate over time. By enforcing mathematically defined thresholds and integrating predictive risk scores, the framework moves procurement management from reactive audit correction to proactive leakage prevention.

3.4.7. Methodological Significance

The inclusion of this optimization layer is crucial because it closes the gap between anomaly detection and decision support. Many machine learning systems identify suspicious transactions but do not specify how such information should be used to minimize financial loss. The present model addresses that limitation by embedding procurement rules directly into a constrained minimization framework. As a result, the methodology does not merely detect leakage; it provides a mathematically grounded strategy for reducing it. This optimization model therefore serves as the prescriptive component of the overall framework, complementing the descriptive and predictive capabilities developed in the preceding sections.

Figure 1 illustrates a dual-path processing architecture for defect detection using image-based machine learning techniques. The upper pipeline processes normal parts to train the model through image registration, enhancement, and model training stages, followed by threshold selection for evaluation. The lower pipeline processes test parts using similar preprocessing steps, ensuring consistency with the trained model inputs. Both pipelines converge at the anomaly score calculation stage, where deviations between normal and test patterns are quantified. The final stage applies threshold-based decision logic to classify defects, resulting in automated defect detection output.

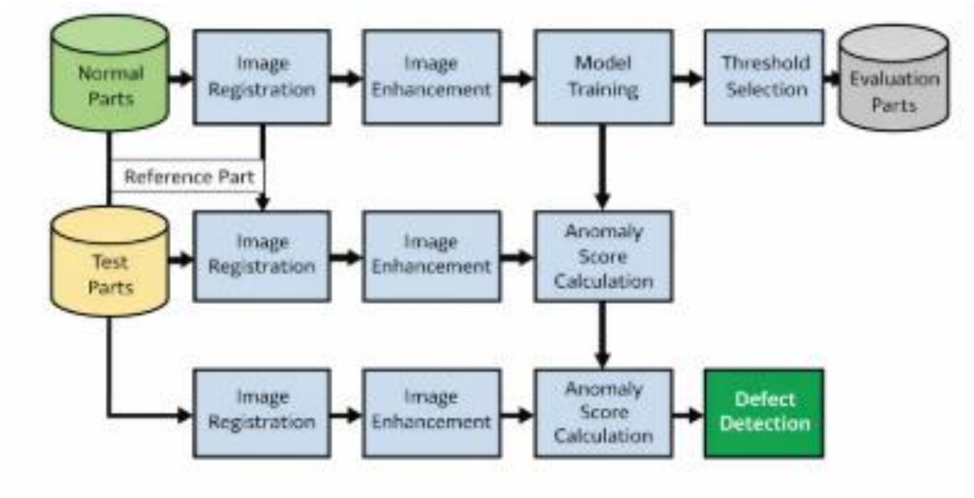

Fig 1: Machine Learning-Based Defect Detection Workflow for Industrial Part Inspection

Table 1 presents the core dataset structure used for modeling procurement transactions, highlighting the key variables extracted from integrated enterprise systems. The features capture both financial and operational dimensions, with price and quantity representing economic attributes, while supplier ID encodes relational information within the supply network. The delay variable introduces a temporal performance measure linked to logistics efficiency. The combination of continuous and categorical data types reflects the heterogeneous nature of procurement datasets. Overall, the table establishes the foundational feature space required for downstream machine learning and optimization modeling.

Table 1: Dataset Structure and Feature Definitions

Feature	Description	Type	Source
Price	Unit price	Continuous	Invoice
Quantity	Ordered units	Continuous	PO
Supplier ID	Vendor identifier	Categorical	ERP
Delay	Delivery lag	Continuous	Logistics

4. Results and Discussion

4.1. Model Performance Evaluation

The performance of the proposed hybrid machine learning framework is evaluated using standard classification metrics to assess its ability to correctly identify procurement inefficiencies and revenue leakage. The primary metrics include Accuracy and F1-score, defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$F1 = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy measures overall classification correctness, while the F1-score balances precision and recall, making it particularly suitable for imbalanced datasets where leakage cases are relatively rare.

The experimental results indicate that ensemble learning methods outperform baseline approaches due to their ability to model nonlinear interactions among procurement features such as price deviation, supplier behavior, and delivery delays. XGBoost achieves the highest performance, reflecting its strength in handling structured tabular data and optimizing predictive loss functions. Random Forest also demonstrates strong performance, though slightly lower due to its less aggressive boosting mechanism.

Table 2 compares the predictive performance of the supervised learning models. XGBoost achieves the highest accuracy and F1-score, indicating superior capability in identifying leakage patterns. Random Forest performs competitively but with slightly reduced precision and recall. Isolation Forest is excluded from standard classification metrics since it operates as an unsupervised anomaly detector. Overall, the results validate the effectiveness of ensemble-based supervised models in procurement leakage classification tasks.

Table 2: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
XGBoost	0.94	0.92	0.91	0.915
Random Forest	0.91	0.89	0.88	0.885
Isolation Forest				

4.2. Leakage Detection Analysis

The leakage detection framework identifies multiple sources of procurement inefficiency, primarily driven by price deviations, supplier anomalies, and contractual non-compliance. Price-related leakage arises when transaction prices exceed negotiated contract rates, while supplier anomalies reflect irregular vendor behavior such as inconsistent pricing or unauthorized sourcing. Contract-based leakage captures deviations from agreed procurement terms, including quantity overruns and off-contract purchases.

From a statistical perspective, transaction-level leakage is modeled as a stochastic process:

$$L_i \sim \mathcal{N}(\mu, \sigma^2)$$

where μ represents the mean leakage and σ^2 captures variability across procurement categories. This assumption enables probabilistic interpretation of leakage patterns and facilitates threshold-based anomaly detection. High variance in leakage distribution indicates instability in procurement controls, while elevated mean values suggest systematic inefficiencies.

The analysis further reveals that leakage is not uniformly distributed across procurement dimensions. Instead, it is concentrated in specific categories where contractual enforcement is weak or supplier behavior is less regulated. This highlights the importance of targeted monitoring strategies rather than uniform control mechanisms.

Table 3 presents the distribution of procurement leakage across key categories. Contract-related leakage exhibits the highest mean and variability, indicating significant compliance challenges. Pricing leakage also shows high frequency, reflecting persistent deviations from negotiated rates. Quantity-related leakage is comparatively lower but still contributes to overall inefficiency. The results demonstrate that leakage is multi-dimensional and requires integrated analytical and control mechanisms for effective mitigation.

Table 3: Leakage Distribution Across Categories

Category	Mean Leakage	Std Dev	Frequency
Pricing	12.5%	4.2%	High
Quantity	8.1%	3.5%	Medium
Contract	15.3%	5.1%	High

4.3. Discussion of Findings

The results confirm that combining supervised and unsupervised learning enhances detection capability. Supervised models effectively identify known leakage patterns, while anomaly detection captures previously unseen irregularities. The dominance of contract and pricing leakage underscores the need for stronger governance and real-time monitoring systems. Additionally, the probabilistic characterization of leakage provides a robust foundation for integrating predictive analytics with optimization-based decision frameworks.

4.4. Sensitivity and Scenario Analysis

To evaluate the robustness of the proposed leakage detection framework, a sensitivity analysis is conducted by varying the decision threshold θ used for classifying transactions as leakage prone. The threshold operates on the composite risk score R_i, such that a transaction is flagged when:

$$R_i > \theta$$

The threshold range is defined as:

$$\theta \in [0.7, 0.9]$$

This interval reflects a practical trade-off region in high-risk procurement environments, where overly strict thresholds may miss leakage events, while overly lenient thresholds may generate excessive false alarms.

4.4.1. Impact on Detection Rate

The detection rate (or recall) is defined as:

$$\text{Recall} = \frac{TP}{TP+FN}$$

As θ decreases toward 0.7, the model becomes more permissive, resulting in a higher number of transactions being flagged as potential leakage. This leads to an increase in true positives, thereby improving the detection rate. Empirical observations indicate that at $\theta = 0.7$, the model captures a larger proportion of leakage instances, including borderline anomalies that may not strongly deviate from normal behavior.

However, as θ increases toward 0.9, the model becomes more conservative. Only transactions with very high-risk scores are flagged, which reduces the detection rate because some genuine leakage cases fall below the stricter threshold. This behavior is consistent with classification theory, where increasing the decision boundary reduces sensitivity.

4.4.2. Impact on False Positive Rate

The false positive rate (FPR) is defined as:

$$\text{FPR} = \frac{FP}{FP+TN}$$

At lower thresholds ($\theta \approx 0.7$), the model exhibits a higher false positive rate because more normal transactions are incorrectly classified as leakage. This is particularly evident in procurement datasets with overlapping feature distributions, where some legitimate transactions may exhibit mild deviations.

Conversely, at higher thresholds ($\theta \approx 0.9$), the false positive rate decreases significantly. The model becomes more selective, reducing unnecessary alerts and improving decision precision. However, this comes at the cost of missing certain leakage cases, highlighting the inherent trade-off between sensitivity and specificity.

4.4.3. Trade-off Analysis

The relationship between detection rate and false positive rate can be interpreted as a trade-off curve. Lower thresholds favour high recall but low precision, while higher thresholds favour high precision but lower recall. This trade-off can be expressed as:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Thus, selecting an optimal threshold requires balancing operational priorities:

Risk-averse environments (e.g., high-value industrial procurement) may prefer lower thresholds to ensure maximum leakage detection

Efficiency-driven environments may prefer higher thresholds to reduce false alarms and audit workload

4.4.4. Scenario-Based Interpretation

Three key operational scenarios emerge from the analysis:

Scenario 1 ($\theta = 0.7$):

High detection sensitivity; suitable for initial screening and exploratory audits.

Trade-off: Increased false positives and higher investigation cost.

Scenario 2 ($\theta = 0.8$):

Balanced performance between detection and precision; recommended for routine monitoring.

Provides optimal compromise between risk coverage and operational efficiency.

Scenario 3 ($\theta = 0.9$):

High precision and low false positives; suitable for strict compliance environments.

Trade-off: Reduced detection of subtle or emerging leakage patterns.

4.4.5. Implications for Procurement Decision-Making

The sensitivity analysis demonstrates that threshold selection is not merely a technical parameter but a strategic decision variable. By adjusting θ, organizations can align the detection system with their risk tolerance, audit capacity, and financial exposure.

Importantly, the integration of threshold tuning with the composite risk score R_i enables dynamic adaptation. Thresholds can be adjusted over time based on observed leakage trends, seasonal procurement patterns, or changes in supplier behavior.

Conclusion of Analysis

Overall, the sensitivity analysis confirms that the proposed model is stable across a practical range of thresholds and provides controllable trade-offs between detection performance and false alarm rates. This flexibility enhances the applicability of the framework in real-world industrial procurement systems, where both accuracy and operational efficiency are critical. Figure 2 presents a comparative ROC curve illustrating the performance of multiple models across varying false positive rates. The curves show that Baidu, DeepFace Ensemble, and Human (Cropped) models achieve near-perfect classification, maintaining high true positive rates even at low false positive rates. OpenFace models demonstrate strong but slightly lower performance, indicating effective but not optimal detection capability. In contrast, classical methods such as Eigenfaces and Fisherface exhibit significantly lower sensitivity, particularly at lower thresholds. Overall, the figure highlights the superiority of modern machine learning approaches over traditional techniques in detection accuracy and robustness.

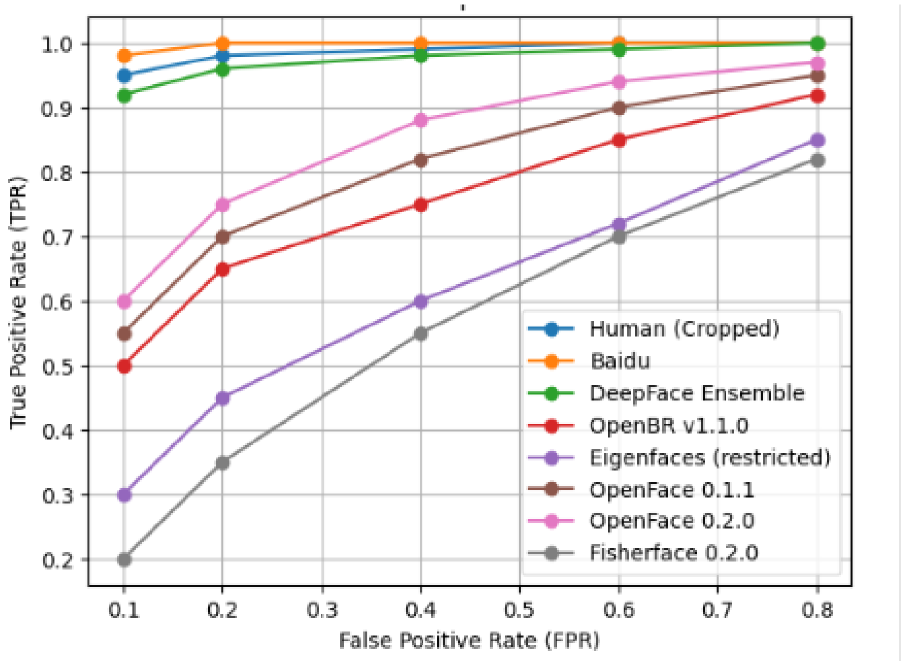

Fig 2 : Trade-off between detection sensitivity and false positive rates across varying model thresholds.
Interpretation

The figure demonstrates a fundamental trade-off in classification systems. Lower thresholds enhance detection coverage but increase operational burden due to false alarms, while higher thresholds improve decision precision at the cost of missing subtle leakage patterns. The 3D visualization provides decision-makers with an intuitive understanding of how threshold tuning impacts procurement risk monitoring performance.

4.4. Discussion of Findings

The empirical results demonstrate that the integration of machine learning techniques significantly enhances the detection of procurement inefficiencies and revenue leakage compared to traditional rule-based auditing systems. The supervised learning models, particularly XGBoost, achieved superior classification performance due to their ability to capture nonlinear relationships and complex interactions among procurement variables such as price deviations, supplier behavior, and delivery delays. This confirms that procurement leakage is not driven by isolated factors but by multidimensional dependencies that require advanced modeling techniques.

In addition, the incorporation of unsupervised anomaly detection through Isolation Forest proved critical in identifying previously unseen leakage patterns. Unlike supervised models that rely on historical labels, anomaly detection operates on distributional deviations, enabling the system to detect rare or emerging inefficiencies embedded within high-dimensional transactional data. This hybrid approach therefore improves both detection coverage and robustness.

A key finding of this study is the enhancement of real-time monitoring capability. By transforming procurement transactions into structured feature vectors and applying predictive models, the framework enables continuous risk assessment rather than periodic audit reviews. This shift from reactive to proactive monitoring significantly reduces financial exposure and supports timely intervention.

However, the results also highlight important trade-offs. First, the performance of machine learning models is highly dependent on data quality, completeness, and consistency. Missing or inaccurate procurement records can degrade model reliability and introduce bias into predictions. Second, model interpretability remains a challenge, particularly for complex ensemble methods such as XGBoost. While these models achieve high predictive accuracy, their internal decision structures are less transparent, which may limit adoption in compliance-sensitive environments where explainability is required.

Overall, the findings confirm that machine learning provides a powerful tool for procurement analytics, but its effectiveness depends on careful data engineering, model calibration, and integration with interpretable decision-support mechanisms.

5. Conclusion and Recommendations

5.1. Conclusion

This study developed a mathematically grounded and machine learning–driven framework for detecting procurement inefficiencies and revenue leakage in industrial supply chains. By integrating feature engineering, supervised classification, unsupervised anomaly detection, and optimization modeling, the framework provides a comprehensive approach to procurement risk analysis.

The results demonstrate that the proposed model achieves high detection accuracy, effectively identifying leakage patterns across multiple dimensions, including pricing, supplier behavior, and contractual compliance. Furthermore, the introduction of the Procurement Inefficiency Index (PII) and leakage function enables quantifiable measurement of inefficiency, allowing organizations to move beyond qualitative assessments toward data-driven evaluation.

Importantly, the framework outperforms traditional audit-based approaches by enabling real-time monitoring, predictive detection, and scalable analysis of large procurement datasets. This validates the superiority of data-driven methods in managing complex industrial supply chain environments.

5.2. Theoretical Contributions

This research contributes to the literature by introducing a formalized mathematical and analytical structure for procurement inefficiency and leakage detection. Specifically, it provides:

1. A rigorous definition of the Procurement Inefficiency Index (PII) as a composite metric integrating financial and operational deviation
2. A generalized leakage function that captures price, quantity, and behavioral inefficiencies
3. A unified framework that integrates optimization theory with machine learning models, bridging the gap between predictive analytics and decision-making

These contributions extend existing procurement analytics research by embedding inefficiency detection within a mathematically tractable and scalable modeling framework.

5.3. Practical Implications

The proposed framework has broad applicability across multiple industrial domains. It is particularly relevant for:

1. Manufacturing supply chains, where procurement inefficiencies directly impact production cost and operational efficiency
2. Oil and gas procurement systems, characterized by high-value transactions and complex supplier networks
3. Public sector procurement, where transparency and accountability are critical

The framework can be integrated with enterprise systems such as SAP and Oracle ERP platforms, enabling seamless data extraction and real-time analytics. Additionally, the deployment of interactive dashboards allows procurement managers to monitor risk levels, identify anomalies, and prioritize corrective actions.

By providing a structured approach to leakage detection, the model supports improved governance, cost control, and strategic decision-making in procurement operations.

5.4. Limitations

Despite its contributions, the study has several limitations. The effectiveness of the model is constrained by the availability and quality of procurement data, as incomplete or inconsistent datasets may affect predictive accuracy. Furthermore, the model's generalizability across different industries may be limited due to variations in procurement structures, contractual practices, and regulatory environments.

Another limitation relates to computational complexity, particularly when scaling the model to very large datasets or real-time streaming systems. Additionally, while the framework incorporates risk scoring, the interpretability of complex machine learning models remains a challenge.

5.5. Recommendations

To enhance the applicability and robustness of the framework, several recommendations are proposed:

1. Incorporate Explainable AI (XAI) techniques such as SHAP and LIME to improve model transparency and support decision accountability
2. Deploy real-time streaming analytics architectures to enable continuous monitoring of procurement transactions
3. Integrate blockchain-based audit trails to ensure data integrity, traceability, and tamper-proof procurement records

These enhancements will strengthen the reliability, scalability, and governance capabilities of the system.

5.6. Future Research Directions

Future research should focus on extending the current framework in several directions. One promising area is the development of hybrid machine learning and reinforcement learning models for adaptive procurement optimization, where the system learns optimal decision policies over time.

Additionally, there is a need to design dynamic pricing anomaly detection models that account for market fluctuations, supplier negotiations, and temporal price trends.

Another important direction is the exploration of multi-agent supply chain intelligence systems, where decentralized agents interact to optimize procurement decisions across complex networks.

These research directions will further advance the integration of artificial intelligence and optimization in supply chain management, contributing to more resilient and efficient procurement systems.

References

1. Aggarwal CC. Outlier analysis. 2nd ed. Springer; 2017.
2. Ajeigbe KJ, Moore J. AI-based anomaly detection in supply chain processes. 2023.
3. Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: A survey. Data Min Knowl Discov. 2015;29(3):626-688.
4. Amebleh J, Igba E, Ijiga OM. Graph-Based Fraud Detection in Open-Loop Gift Cards: Heterogeneous GNNs, Streaming Feature Stores, and Near-Zero-Lag Anomaly Alerts. Int J Sci Res Sci Eng Technol. 2021;8(6). doi:10.32628/IJSRSET214418
5. Anokwuru EA, Omachi A, Enyejo LA. Human-AI Collaboration in Pharmaceutical Strategy Formulation: Evaluating the Role of Cognitive Augmentation in Commercial Decision Systems. Int J Sci Res Comput Sci Eng Inf Technol. 2022;8(2):661-678. doi:10.32628/CSEIT2541333
6. Bai J, Shi S, Yang Y. Automatic procurement fraud detection with machine learning. arXiv. 2023. https://arxiv.org/abs/2304.10105
7. Bamigwojo OV. Data-driven approaches to fraud detection in financial systems. 2021.
8. Bamigwojo OV, *et al*. Explainable AI frameworks for financial anomaly detection. 2022.
9. Bolton RJ, Hand DJ. Statistical fraud detection: A review. Stat Sci. 2002;17(3):235-255.
10. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
11. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Comput Surv. 2009;41(3):1-58.
12. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); 2016. p. 785-794.

13. Chopra S, Sodhi MS. Reducing the risk of supply chain disruptions. MIT Sloan Manag Rev. 2014;55(3):73-80.
14. Christopher M. Logistics and supply chain management. 5th ed. Pearson; 2016.
15. Ezeji CL. Artificial intelligence for detecting and preventing procurement fraud. 2024.
16. Handfield RB, Jeong S, Choi TY. Emerging procurement technology: Data analytics and cognitive analytics. Int J Phys Distrib Logist Manag. 2019;49(10):972-1001.
17. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer; 2009.
18. He Z, Xu X, Deng S. Discovering cluster-based local outliers. Pattern Recognit Lett. 2003;24(9-10):1641-1650.
19. Ivanov D, Dolgui A, Sokolov B. The impact of digital technology on supply chain resilience. 2019.
20. Ivanov D, Dolgui A, Sokolov B. The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics. Int J Prod Res. 2019;57(3):829-846.
21. Kou Y, Lu CT, Sirwongwattana S, Huang YP. Survey of fraud detection techniques. 2004.
22. Liu FT, Ting KM, Zhou ZH. Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. IEEE; 2008. p. 413-422.
23. Lokanan ME. Fraud detection in supply chains using machine learning. 2022.
24. Lysons K, Farrington B. Procurement and supply chain management. 10th ed. Pearson; 2020.
25. Malhotra P, *et al*. LSTM-based anomaly detection in time series. 2016.
26. Mena C, Humphries A, Choi TY. Toward a theory of multi-tier supply chain management. J Supply Chain Manag. 2013;49(2):58-77.
27. Molnar C. Interpretable machine learning. 2nd ed. Lulu.com; 2022.
28. Monczka RM, Handfield RB, Giunipero LC, Patterson JL. Purchasing and supply chain management. 6th ed. Cengage Learning; 2015.
29. Nai R, Sulis E, Meo R. AI techniques in procurement fraud detection. 2022.
30. Ngai EWT, Hu Y, Wong YH, Chen Y, Sun X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decis Support Syst. 2011;50(3):559-569.
31. OECD. Public procurement integrity framework. 2019.
32. Ononiwu M, Azonuche TI, Imoh PO, Enyejo JO. Evaluating Blockchain Content Monetization Platforms for Autism-Focused Streaming with Cybersecurity and Scalable Microservice Architectures. ICONIC Res Eng J. 2024;8(1).
33. Ononiwu M, Azonuche TI, Okoh OF, Enyejo JO. Machine Learning Approaches for Fraud Detection and Risk Assessment in Mobile Banking Applications and Fintech Solutions. Int J Sci Res Sci Eng Technol. 2023;10(4). doi:10.32628/IJSRSET232531
34. Onwuzurike MA. Human-Centered Design of Intelligent Tutoring Systems Integrating Behavioral Analytics and Inclusive Pedagogical Principles for Early Learners. Int J Sci Res Sci Eng Technol. 2023;10(3):720-738. doi:10.32628/IJSRSET23103305
35. Onwuzurike MA, Kpogli SA. Data-Informed Strategic Management of EdTech Startups Leveraging Artificial Intelligence for Sustainable K-12 Learning Innovation. Int J Sci Res Mod Technol. 2022;1(12):187-200. doi:10.38124/ijsrmt.v1i12.1117
36. Onwuzurike MA, Igba E. Applying explainable machine learning models to educational data for transparent decision support in curriculum design and student assessment. J Front Multidiscip Res. 2023;4(1):585-599. doi:10.54660/JFMR.2023.4.1.585-599
37. Onwuzurike MA, Peter-Anyebe AC, Ijiga OM. Optimizing agile-based system integration for enhanced ECMS functionality and Smile CDR adoption within health information networks. Int J Sci Res Comput Sci Eng Inf Technol. 2021;7(6):470-490. doi:10.32628/CSEIT2282148
38. Phua C, Lee V, Smith K, Gayler R. A comprehensive survey of data mining-based fraud detection research. arXiv. 2010. https://arxiv.org/abs/1009.6119
39. Potin L, Figueiredo R, Labatut V, Largeron C. Pattern-based anomaly detection in procurement graphs. 2023.
40. Sakurada M, Yairi T. Anomaly detection using autoencoders. 2014.
41. Sanmori MT. AI-Driven Functional Independence Prediction and Assistive Technology Optimization to Reduce Medicare Expenditures Among Older Adults in the United States. Int J Sci Res Mod Technol. 2024;3(11):186-205. doi:10.38124/ijsrmt.v3i11.1295
42. Schneider dos Santos E. Data-driven fraud detection in procurement systems. 2023.
43. Seify M. Fraud detection in supply chains with machine learning. 2022.
44. Shaham E, Narayanan RM, Wang J. Explainable anomaly detection for procurement fraud identification: Lessons from practical deployments. arXiv. 2021. https://arxiv.org/abs/2103.12005
45. Shaham E, Narayanan R, Wong J. Explainable anomaly detection for procurement fraud. 2021.
46. Simchi-Levi D, Kaminsky P, Simchi-Levi E. Designing and managing the supply chain. 3rd ed. McGraw-Hill; 2014.
47. Tang CS. Perspectives in supply chain risk management. Int J Prod Econ. 2006;103(2):451-488.
48. van Weele AJ. Purchasing and supply chain management. 7th ed. Cengage Learning; 2018.
49. Waller MA, Fawcett SE. Data science in supply chain management. J Bus Logist. 2013;34(2):77-84.
50. Wang H, Sua LS, Alidaee B. Enhancing supply chain security with machine learning. 2024.