Comparative Review of Confidence and Other Evaluation Metrics in Predictive Modeling for Procurement Fraud Coalition

Saifuddin Mohd, Mohamad Taha Ijab
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, 43600 Bangi Selangor, Malaysia

***Abstract*—Procurement fraud, particularly when bidders act together through collusion or coalition schemes, remains a major threat to fair competition in public procurement. Predictive modeling has emerged as a key analytical tool for detecting such behaviors yet choosing appropriate evaluation metrics continues to be a challenge, especially with imbalanced or correlated data. This study applies a structured narrative review supported by a comparative analysis to examine commonly used evaluation metrics—Accuracy, Precision, Recall, F1-score, and AUC-ROC—in relation to the rule-based Confidence metric derived from association rule mining. The findings reveal that while traditional classification metrics are effective for general predictive tasks, they often fail to capture relational and co-occurrence patterns that characterize coalition fraud. In contrast, Confidence demonstrates higher interpretability and contextual relevance for detecting collusive behaviors among suppliers. The study highlights the potential of hybrid evaluation frameworks that combine classification and rule-based measures to improve fraud detection accuracy and explainability. This approach contributes to advancing predictive modeling, procurement analytics, and coalition detection by emphasizing metrics that balance performance, interpretability, and real-world applicability.**

***Keywords*—Procurement fraud; predictive modeling; confidence; evaluation metrics; association rule mining; coalition detection; public sector analytics**

I. Introduction

Public administration rests on a simple promise to citizens: public money must be spent properly and fairly. Public procurement sits at the heart of that promise, shaping how governments acquire goods and services at scale. Digitisation has helped by opening processes to scrutiny and producing rich data that can be analysed for fairness and efficiency. At the same time, the same openness has introduced new risks as actors learn to exploit procedural signals and digital traces for gain [1], [2].

Among those risks, collusion fraud is especially damaging. It arises when suppliers coordinate to shape outcomes through bid rigging, rotating winners, or submitting coordinated bids that blunt competition. These schemes are difficult to spot because the telltale signs rarely appear in a single tender. They live in patterns across time and across networks of bidders, which standard document checks or routine audits often miss [3].

Analytics has therefore become a practical route to detection. Predictive models sift large datasets to flag anomalies that suggest misconduct [4] The challenge is how we judge these models in a setting where fraud is rare. Accuracy, precision, recall, F1-score and AUC-ROC are widely used, yet each can mislead in imbalanced data. A model can be "accurate" by predicting almost everything as clean. These metrics also focus on isolated transactions and often ignore the relational signals that define collusion, such as repeated co-bidding or synchronized entry and withdrawal patterns [4], [5].

This study brings a complementary lens to the problem by examining Confidence from association rule mining as an evaluation and insight metric. Confidence estimates the likelihood of an event given another event. In procurement terms, it answers questions like "how often does Supplier B bid when Supplier A bids." That simple conditional view helps surface co-occurrence patterns that are consistent with coordinated behaviour and gives analysts something they can read and explain [6].

Although predictive analytics has matured, how we evaluate models for collusion specifically remains underexplored. Collusion is relational by nature. It emerges from ties, routines, and repeated interactions among bidders that traditional classification metrics do not capture well [7], [8], [9]. Ignoring those ties risks high-scoring models that miss the very behaviours of interest.

Therefore, this paper compares Confidence against common performance metrics and argue for hybrid evaluation. Classification metrics remain useful for measuring statistical performance. Confidence and related rule-based measures supply contextual signals about who appears with whom, and how often, in ways auditors and regulators can act on. The combination improves detection quality and interpretability in practice [10], [11]. This integrated focus aligns with the intent of the present manuscript, which positions Confidence alongside accuracy, precision, recall, F1-score and AUC-ROC for coalition detection in public procurement, and motivates a balanced, insight-driven assessment approach.

The rest of the paper proceeds as follows. Section II reviews literature on collusion in procurement and on evaluation metrics for predictive modelling, with emphasis on Confidence. Section III explains the methodology. Section IV discusses the comparative analysis and findings. Section V presents the synthesis and discussion and Section VI concludes and outlines directions for future research in predictive procurement analytics.

II. Background and Related Work

A. Coalition Fraud in Public Procurement

It is quite easy to deceive and change the way the government buys things. Coalition fraud is one of the most advanced and harmful types of fraud. Collusion, bid rigging, or cartel activity occurs when two or more suppliers work together to change the outcome of a procurement process. These kinds of alliances are meant to get rid of real competition. They accomplish this by bidding together over and over again, swapping the winner before the auction ends, or making bids that are very similar to each other.

These kinds of alliances are meant to get rid of real competition. They accomplish this by bidding together over and over again, swapping the winner before the auction ends, or making bids that are very similar to each other. These actions not only mess up how the market works, but they also make people lose a lot of trust and make items too expensive to acquire [12], [13]. Coalition fraud is always about connections, unlike single fraudulent activities. It takes advantage of how hard it is to grasp the different parts of the evaluation and how often bidding groups act the same way. This makes it challenging to find occurrences using typical techniques of assessment.

The Malaysia Competition Commission (MyCC) has observed that in Malaysia's public e-procurement settings, behavior that appears like a cartel is growing more sophisticated. We need to find it using improved, data-driven ways [14].

B. Evaluation Metrics in Predictive Modeling

In predictive modeling for procurement fraud detection, assessment metrics are necessary to measure both accuracy and efficacy. The most fundamental indicator is accuracy, which shows how many of the predictions were accurate. However, in datasets with significant imbalance, frequently seen in fraud detection, accuracy can be deceptive as it may overstate performance by privileging the majority class [15], [16].

Precision and recall are two indicators that provide you more information. Precision measures how many of the highlighted instances were indeed frauds. This is very important in situations when false positives might create reputational or administrative costs. On the other hand, recall evaluates how well a model can find all real fraud cases, which is very important in high-stakes situations where missing fraud can have serious consequences [17], [18]. The F1-score is a harmonic mean of accuracy and recall, which is typically used to give a fair assessment because these metrics often disagree with each other [19]. In procurement fraud detection, however, recall is frequently more important than precision since missing fraud is more damaging than overestimating its presence.

Other metrics, such as AUC-ROC, are useful since they do not depend on a certain threshold and can tell the difference between different things [20]. But since they are so abstract, they are not very good for rule-based or coalition-oriented fraud detection, where patterns that are relevant to the situation are important. Recent advancements in Explainable AI (XAI) highlight the necessity for metrics that provide transparency and interpretability, ensuring that model outputs can be successfully conveyed to policymakers, auditors, and non-technical stakeholders [21].

C. Confidence as a Metric for Coalition Detection

Given the limitations of global evaluation metrics, rule-based measures, particularly confidence from association rule mining, have garnered increasing attention. Confidence measures the conditional probability of a consequent given an antecedent, formally expressed as $P(Y|X)$. In the context of procurement fraud, this can be interpreted as the likelihood of a bidder participating or winning given the participation of another supplier, thus offering critical insight into co-occurrence behavior and potential [22], [23]. Unlike aggregate classification metrics, confidence allows for localized, pattern-specific interpretability.

This makes it particularly effective for identifying relational anomalies, such as recurrent supplier pairings or tenders with suspiciously homogeneous bidder compositions. Furthermore, confidence can be used in unsupervised settings, making it applicable even when labeled fraud data is scarce and often-cited challenge in procurement analytics. Nonetheless, confidence is not without limitations. It can be artificially inflated by low-support rules, leading to overinterpretation of spurious associations. To address this, confidence is typically used in conjunction with support (frequency of rule occurrence) and lift (statistical strength relative to random chance), forming a robust triad of rule evaluation metrics [23].

III. Methodology

In examining studies on procurement fraud and coalition detection, two methodologies for synthesizing information are commonly utilized: structured narrative reviews and systematic reviews. Each approach has various advantages and disadvantages. Systematic reviews are recognized for their methodological rigor, enhancing transparency, reproducibility, and confidence in the resultant evidence [24]. The rigor is especially crucial when looking at complex evaluation criteria for predictive modeling in procurement fraud, where the accuracy of the results is critical [25]. The strict guidelines of systematic reviews may make it harder to look into new ideas or processes, which are sometimes better handled using more flexible methods [26].

In Contrast, structured narrative reviews, on the allow researchers combine results from many different sources of literature without having to follow strict rules on the methodological guidelines [27]. This adaptability supports the formation of nuanced findings, particularly in coalition detection, where conceptual, empirical, and methodological perspectives intersect [28]. A structured narrative review is not only descriptive; it seeks to clarify linkages across contemporary studies, uncover methodological strengths, and pinpoint gaps for further study.

This study employs a structured narrative review method to examine the Confidence metric in predictive modeling for procurement fraud and compare it to traditional classification-based evaluation metrics. This technique emphasizes the study's focus on relational and rare coalition activities, requiring a balance between comprehensive synthesis and interpretative depth. The review is conducted with inclusion and exclusion

criteria, a clearly defined topic, and careful consideration of the narrative integration of results to ensure quality.

To enhance methodological transparency and replicability, a clear and systematic search strategy was developed. The review process followed four sequential phases:

- Identification: Relevant literature was identified through a comprehensive search across Scopus, Web of Science, and Google Scholar databases using defined keyword combinations. The search period covered 2010 to early 2025 to capture both foundational and emerging studies.
- Screening: Duplicate records were removed, and titles and abstracts were screened to ensure alignment with the research focus on evaluation metrics and coalition detection in procurement fraud.
- Eligibility: Full-text assessment was conducted to evaluate whether the articles met the inclusion criteria—namely, studies that (i) examined evaluation metrics used in fraud or anomaly detection, (ii) applied Confidence or association rule mining in predictive modeling, or (iii) focused on procurement, collusion, or cartel behavior detection.
- Quality Assessment: To maintain rigor, studies were appraised based on methodological soundness, clarity of metric application, and relevance to predictive modeling. Papers lacking sufficient methodological detail or relying solely on opinion-based commentary were excluded.

Exclusion criteria encompassed grey literature (e.g., blogs, white papers, or non-peer-reviewed content), studies unrelated to predictive modeling or procurement fraud, and those without measurable or interpretable performance metrics. This structured process ensured that the final body of literature was both methodologically robust and conceptually relevant. The resulting dataset of studies formed the empirical foundation for the comparative synthesis presented in later sections.

A. Research Design

The study is conceptualized as a comparative literature-based review supported by a case-informed synthesis. The objective is to evaluate the interpretability, contextual relevance, and sensitivity of common evaluation metrics: Accuracy, Precision, Recall, F1-score, and AUC-ROC against the rule-based Confidence metric. These metrics were selected due to their widespread use in fraud detection and classification modeling.

Given the exploratory nature of the study, a qualitative interpretive comparison is adopted rather than a purely statistical meta-analysis. This approach allows for incorporating insights from multiple domains (e.g., data mining, fraud analytics, public procurement) and ensures the findings remain grounded in operational and domain-specific requirements.

Limitations: It is important to acknowledge that the qualitative, narrative synthesis approach adopted in this study may limit quantitative generalizability. Findings presented here are primarily conceptual, based on expert interpretation and literature review rather than empirical validation. Therefore, future studies should consider conducting empirical analyses with large, real-world datasets to statistically validate the conceptual insights provided by this comparative assessment.

B. Data Sources and Literature Selection

The literature search was conducted across three major academic databases that is Scopus, Web of Science, and Google Scholar to ensure comprehensive coverage of both theoretical and applied studies. The search period spanned 2010 to early 2025, capturing the evolution of predictive modeling and fraud detection research.

The following keyword combinations were used to ensure search precision and breadth:

- "confidence metric" AND "fraud detection"
- "evaluation metrics" AND "predictive modeling"
- "procurement fraud" OR "coalition detection"
- "association rule mining" AND "bid rigging"

Each retrieved article underwent multi-stage screening as described in the methodology overview. Only peer-reviewed journal articles, conference papers, and technical reports were included, ensuring academic credibility.

The inclusion criteria emphasized empirical or conceptual contributions to predictive modeling, Confidence metric application, or procurement fraud detection. Studies focusing solely on algorithmic performance without interpretive or contextual analysis were excluded.

C. Analytical Framework

This study constructed a comparative framework based on three dimensions and systematically compare the metrics:

- Interpretability – the extent to which metric outputs can be meaningfully interpreted by human analysts and decision-makers;
- Sensitivity to Class Imbalance – the ability of the metric to provide reliable insights in datasets where fraud cases constitute a minority.
- Applicability to Coalition Detection – the metric's utility in modeling relational or co-occurrence patterns typical of procurement collusion.

Each metric was scored heuristically on a five-point Likert-type scale (1 = Low, 5 = High) for each dimension, drawing from synthesis of the literature and informed expert judgment. While such scoring does not provide definitive empirical quantification, it offers a structured and transparent means of articulating relative strengths and weaknesses among metrics. A common approach in methodological reviews is where direct comparisons are limited or context specific.

IV. Comparative Analysis and Findings

This section presents a comparative assessment of commonly used evaluation metrics in predictive fraud modeling, with particular emphasis on their relevance to coalition detection in procurement contexts. The metrics are examined in terms of their mathematical foundations, interpretability, sensitivity to

class imbalance, and contextual utility in identifying collusive behavior.

A. Metric Characteristics and Theoretical Basis

The principal characteristics of six commonly employed evaluation metrics: Accuracy, Precision, Recall, F1-score, AUC-ROC, and Confidence are summarised in Table I. While the first five metrics are grounded in classification theory and typically applied in supervised learning settings, Confidence originates from association rule mining and is often used in unsupervised or pattern mining contexts.

TABLE I. COMPARISON ACROSS SIX METRICS: ACCURACY, PRECISION, RECALL, F1-SCORE, AUC-ROC, AND CONFIDENCE

Metric	Mathematical Definition	Best Context	Advantages	Limitations
Accuracy	(TP + TN) / Total	Balanced datasets	Simple, standard	Biased with class imbalance
Precision	TP / (TP + FP)	High false positive cost	Positive outcome focus	Ignores false negatives
Recall	TP / (TP + FN)	High false negative cost	Captures all true positives	Ignores false positives
F1-Score	2 * (P * R) / (P + R)	Imbalanced data	Balances P and R	Hard to interpret directly
AUC-ROC	Area under the curve of TPR vs FPR	Model comparison, threshold-independent evaluation	Comprehensive discrimination measure	Lacks local interpretability
Confidence	P(Y|X)	Rule-based learning, fraud detection	Conditional, interpretable	Ignores minimum support

Accuracy, defined as the ratio of correct predictions (both positive and negative) to the total number of instances, is intuitive but problematic in imbalanced datasets. It often overstates performance by favoring the majority class, which in fraud detection typically represents non-fraudulent cases [29].

Precision and Recall offer more focused evaluations. Precision emphasizes the correctness of positive predictions (i.e., how many predicted frauds are true frauds), whereas Recall reflects the model's ability to capture all actual frauds. F1-score, the harmonic mean of Precision and Recall, provides a balanced measure, though it may lack intuitive clarity for stakeholders not versed in statistical evaluation [17].

AUC-ROC evaluates the classifier's ability to discriminate between classes across different threshold values. While it is threshold-independent, AUC lacks granularity and may obscure localized fraud patterns [20].

Confidence, expressed as P(Y|X), quantifies the conditional likelihood that event Y occurs given X. In procurement contexts, this could translate to the probability that one supplier bids or wins conditional on the participation of another. It thus enables fine-grained relational analysis and is particularly useful for uncovering suspicious co-bidding behavior that might indicate tacit collusion [22].

B. Comparative Matrix and Interpretive Scoring

A comparative assessment of six evaluation metrics commonly employed in procurement fraud detection, analyzed across two critical dimensions: interpretability, defined as the extent to which a metric can be intuitively understood and meaningfully applied by human decision makers, and imbalance sensitivity, referring to the metric's responsiveness to skewed class distributions, an inherent characteristic of fraud related datasets.

The results in Fig. 1 indicate that Confidence achieves the highest level of interpretability (score of 5) while exhibiting the lowest sensitivity to class imbalance (score of 1), positioning it as particularly advantageous for contexts where transparency and data imbalance are pressing concerns.

Conversely, accuracy demonstrates poor performance on both fronts, with low interpretability, especially at finer analytical levels, and a high susceptibility to data imbalance, thereby limiting its effectiveness in real world fraud detection scenarios. The scoring, which ranges from 1 (least favorable) to 5 (most favorable), is derived through synthesis of current literature and practitioner experience, an approach often adopted in review-based studies where direct empirical comparisons are scarce.

Confidence is rarely compared with traditional classification metrics such as Accuracy in empirical evaluations, necessitating a relative rather than absolute analytical approach. The purpose of this comparative framework is not to produce definitive rankings but to elucidate general tendencies in metric behavior, thereby informing the selection of evaluation strategies that align with the unique demands of procurement fraud analytics in imbalanced data environments.

The interpretability and imbalance sensitivity scores in Table II were derived from a synthesis of relevant literature (e.g., [29] and expert judgment based on the application of these metrics in fraud detection and procurement datasets. While not absolute, these heuristic scores facilitate a structured comparison of metric suitability in coalition fraud modelling. From the comparative analysis, it is evident that Confidence offers superior interpretability and is significantly less affected by class imbalance, an advantage that is particularly relevant in fraud detection settings where positive instances (i.e., fraud) are rare and scattered. Furthermore, its pattern-based nature enables it to surface relational anomalies, making it particularly well-suited for coalition detection.

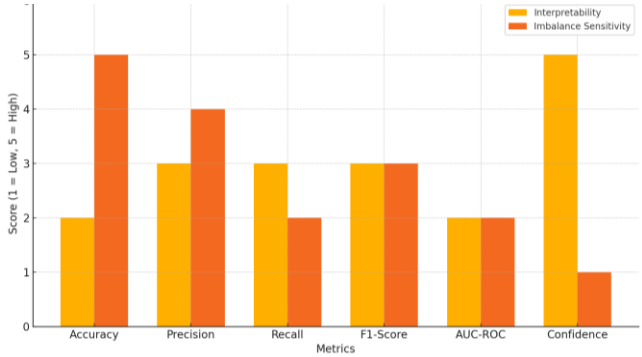

Fig. 1. Six evaluation metrics based on interpretability and imbalance sensitivity.

TABLE II. COMPARATIVE EVALUATION OF METRICS

No.	Metric	Interpretability (1–5)	Imbalance Sensitivity (1–5)	Justification
1.	Accuracy	2	5	Inflated in imbalanced datasets; does not reflect minority class detection [29]
2.	Precision	3	4	Useful to reduce false positives; appropriate in fraud context [30]
3.	Recall	3	2	Important when capturing all fraud instances; can result in high false positives [29]
4.	F1-Score	3	3	Balances precision and recall; effective for imbalanced data [31]
5.	AUC-ROC	2	2	Provides global model evaluation; lacks actionable insight for case-level interpretation [32]
6.	Confidence	5	1	Rule-based metric with high interpretability in pattern-based fraud; useful in coalition detection [22], [33]

C. Practical Insight: Rule-Based Detection

The primary advantage of Confidence lies in its ability to generate rule-level insights. For example, if a rule indicates that Supplier A and Supplier B co-bid in 90% of tenders where one is present, and both frequently win under similar conditions, this may trigger an alert. Such insight is difficult to derive from aggregate metrics like Accuracy or F1-score, which do not consider inter-entity relationships. Moreover, Confidence can be operationalized in conjunction with other rule-based measures such as support (frequency of occurrence) and lift (relative strength), enabling a multi-dimensional assessment of relational patterns. This is critical in the detection of cartel behavior, where [37] highlight the importance of using advanced analytical methods to distinguish genuine collusion from legitimate competition, a challenge made even more pressing in detecting cartel behavior where fraudulent patterns are often hidden within dense and repeated co-bidding networks.

D. Case-Based Justification

A simplified procurement case analysis drawn from Malaysian e-procurement data further underscores the utility of Confidence. These examples were derived from actual e-procurement datasets from Malaysian government agencies, with the dataset being on 2015, illustrating real-world bidding scenarios. In particular, multiple instances were identified where pairs of suppliers consistently co-bid with high Confidence scores (exceeding 80% conditional probability), strongly indicating potential collusive behavior. These relational patterns were not effectively flagged using traditional classification metrics due to their emphasis on class-labeled outcomes rather than inter-entity relationships.

The analysis revealed multiple instances where supplier pairs participated in tenders with statistically significant co-occurrence patterns. Traditional metrics failed to flag these cases due to their reliance on class-labeled outcomes, whereas Confidence-based rules identified relational anomalies worthy of further investigation. Table III illustrates the type, sensitivity to imbalanced data, interpretability, and use case in fraud detection for each metric.

TABLE III. TYPE, SENSITIVITY TO IMBALANCED DATA, INTERPRETABILITY, AND USE CASE IN FRAUD DETECTION FOR EACH METRICS

Metric	Type	Sensitivity to Imbalanced Data	Interpretability	Use Case in Fraud Detection	Author (s)
Accuracy	Overall performance	Low	High, but can be misleading	Not suitable for imbalanced datasets; may mask rare fraud cases	[15], [17]
Precision	Class-specific (Positive class)	High	High	Important to reduce false positives (e.g., avoid false alarms)	[17], [34]
Recall	Class-specific (Positive class)	High	High	Important to reduce false negatives (e.g., capture all fraud)	[17], [34]
F1-Score	Harmonic mean of precision & recall	Medium to High	Moderate	Balances precision and recall; useful in skewed data	[17], [34]
AUC-ROC	Threshold-independent	High	Moderate	Evaluates model across thresholds; compare classifier ability	[20]; [35]
Confidence	Rule-based (Association Rules)	High	High (in rule mining)	Measures rule strength; helps find fraud patterns	[23], [36]

E. Insight: Why Confidence is Useful for Coalition Detection

Confidence provides rule-level insight, which is particularly valuable for detecting procurement fraud coalitions. For example, a high confidence score for the co-occurrence of Supplier A and Supplier B (e.g., 90% confidence) suggests a strong relational pattern, which is more informative than an overall accuracy score when seeking hidden relationships. This characteristic supports early-warning systems and evidence-based procurement audits, aiding in identifying collusive behaviors that might not be apparent with aggregate metrics.

V. SYNTHESIS AND DISCUSSION

The comparative analysis presented in this review underscores the evolving challenges and opportunities in evaluating predictive models for procurement fraud detection, particularly to coalition or collusive behavior. While traditional evaluation metrics such as Accuracy, Precision, Recall, and AUC-ROC remain foundational in classification-based modeling, they are increasingly insufficient when applied to domains characterized by data imbalance, relational complexity, and pattern-level anomalies.

A. Limitations of Traditional Evaluation Metrics

A significant limitation of conventional metrics is their focus on instance-level classification performance. In procurement fraud detection, especially for coalition behaviors, the fraudulent signal often resides not in individual records but in the relationships among entities across time. For instance, Accuracy can be misleading in scenarios with extreme class imbalance, such as datasets where fraudulent tenders constitute less than 5% of all transactions by overstating model effectiveness due to the prevalence of true negatives [15], [29]. Similarly, although Precision and Recall offer greater sensitivity to the minority class, they remain inadequate in identifying the relational dynamics underlying collusion. F1-score, while useful in balancing both metrics, does not address interpretability or contextual relevance. AUC-ROC further abstracts the model's discriminatory power into a single score, which, although statistically sound, lacks the granularity needed for forensic or regulatory applications in procurement fraud.

B. Strengths of Confidence in Fraud Coalition Detection

The Confidence metric, grounded in association rule mining, provides a complementary and, in many cases, superior evaluative perspective for coalition fraud modeling. Unlike global metrics, Confidence quantifies the conditional likelihood of a co-occurrence, e.g., Supplier B submitting a bid when Supplier A does offering a highly interpretable and actionable view of potential collusion. Its unsupervised nature also makes it robust in contexts where labeled fraud data is scarce or incomplete. This is especially pertinent in the public sector, where investigations are often reactive and labels may only exist post-audit. The Confidence metric allows for early-warning signals through the detection of anomalous patterns, which can be subjected to further validation using domain expertise or additional metadata (e.g., company addresses, ownership structures, bid amounts). Importantly, Confidence should not be used in isolation. Its key limitation is the potential inflation of its value in low-support scenarios, where rare co-occurrences may appear statistically significant due to limited data. This issue can be mitigated by jointly analyzing support (absolute rule frequency) and lift (proportional deviation from independence), both of which enhance the reliability of Confidence-based findings [23].

C. Practical and Policy Implications

From a practical standpoint, integrating Confidence-based analytics into existing procurement monitoring systems could substantially enhance auditors' capabilities to identify relational anomalies indicative of collusive behavior. Policymakers and regulatory bodies, such as the Malaysia Competition Commission (MyCC), would benefit from adopting these interpretable analytics tools to improve transparency and proactively flag suspicious tenders. Implementing Confidence-based metrics can also inform risk-based auditing strategies and continuous supplier monitoring, ultimately promoting accountability and integrity in public procurement processes.

D. Implications for Practice and Policy

From a policy standpoint, the integration of Confidence into procurement analytics tools can inform risk-based auditing, automatic flagging of suspicious tenders, and ongoing supplier monitoring. As governments increase investment in e-procurement systems, embedding such interpretable analytics directly into workflows can improve the efficacy of oversight bodies like the Malaysia Competition Commission (MyCC) or anti-corruption agencies. For practitioners, the findings suggest a shift from solely performance-driven evaluation (i.e., how well does the model classify?) to insight-driven evaluation (i.e., what can the model explain about fraud structures?). This is aligned with broader trends in explainable AI (XAI) and human-centric data science.

E. Advancement Beyond Existing Studies

The comparative framework presented in this study extends prior research on evaluation metrics for fraud detection in three main ways. First, it introduces a cross-domain synthesis that integrates traditional classification metrics with rule-based interpretability measures such as Confidence an area rarely analyzed jointly in previous works. This dual perspective bridges the gap between performance evaluation and relationship discovery, enabling more meaningful assessments of models used in detecting collusion.

Second, by applying Confidence within the context of coalition fraud, the study advances the discussion from general fraud classification toward relational analytics, emphasizing inter-supplier dependencies that are typically overlooked by conventional measures like AUC-ROC or F1-score.

Third, the heuristic comparative matrix developed in this paper provides a transparent and structured tool for evaluating metrics along interpretability and class imbalance dimensions, which can be adopted by researchers and regulators for systematic assessment of fraud detection models.

Collectively, these contributions represent a methodological advancement over prior works that treated evaluation purely as a statistical exercise. The present study positions metric selection as both a technical and interpretive decision, thereby aligning predictive modeling more closely with real-world auditing and policy needs.

VI. Conclusion and Future Work

This review has critically examined the comparative utility of traditional classification-based evaluation metrics and the Confidence metric within the context of predictive modeling for procurement fraud, with a particular focus on coalition detection. The analysis demonstrates that while standard metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC provide valuable insights for general model validation, they are insufficient when the modeling objective involves detecting rare, relational, and pattern-based fraud behaviors such as supplier collusion.

The Confidence metric, rooted in association rule mining, has been shown to offer superior interpretability, resilience to class imbalance, and applicability in unsupervised or semi-supervised settings. Its capacity to uncover rule-level relationships makes it particularly effective for modeling fraud coalitions, especially in scenarios where labeled data are scarce or incomplete. Despite its limitations, such as sensitivity to low-support rules, Confidence can be effectively calibrated using complementary metrics like support and lift. Given the unique strengths of both metric types, this paper advocates for the

adoption of hybrid evaluation frameworks that integrate classification metrics with rule-based measures. Such frameworks not only enhance model robustness but also contribute to explainability, operational relevance, and stakeholder trust, key considerations in public sector analytics and regulatory oversight.

Future work in procurement fraud analytics should focus on developing more intelligent and interpretable evaluation strategies to enhance both detection accuracy and real-world applicability. One key direction is the creation of composite scoring systems that integrate metrics such as confidence, support, lift, and conventional classification indicators (e.g., precision, recall, F1-score), enabling a more balanced trade-off between predictive performance and interpretability. Additionally, graph-based learning techniques, particularly graph neural networks (GNNs), offer promise in modeling coalition behavior by capturing structured relationships among suppliers; these models should be evaluated using hybrid metrics that reflect both statistical accuracy and semantic relevance.

Embedding confidence-based alerts into real-time procurement systems can further support dynamic risk monitoring and early intervention, aligning predictive insights with operational workflows. Furthermore, empirical validation using real procurement data, including blind evaluations by domain experts, is essential to assess the practical effectiveness and trustworthiness of confidence-driven rule discovery methods. Finally, conducting cross-national comparisons of procurement fraud detection frameworks will help generalize findings and identify best practices suited to varying legal and regulatory environments. As public procurement systems continue to evolve toward digitalization, the capacity to detect and explain collusive patterns will be central to ensuring transparency and preserving market integrity. This review contributes to that objective by outlining a roadmap for advancing interpretable, actionable, and context-sensitive evaluation of predictive models in fraud detection.

This study identified potential future initiatives to enhance the detection of procurement fraud, particularly in cases of collaboration. The thoughts are split into two groups: short-term and long-term.

A. Short-Term Priorities

First, it is a good idea to test the Confidence measure with more real data from purchases. A lot of articles solely utilize synthetic or example data. We can find out if Confidence can truly assist in uncovering collusion between suppliers if we utilize genuine tender data from a government or commercial system.

Second, researchers might be able to mix diverse measures. For example, the F1-score or Confidence with precision. This can help the model be both right and easy to grasp. In real life, people like auditors need to explain how the outcome came about, not just give a figure.

B. Long-Term Goals

Not all countries have the same system and data. Researchers ought to examine the functioning of Confidence across various countries. One strategy may work well in Malaysia, but it may not work the same way in other countries.

Finally, create a hybrid system that uses both rule-based and categorization metrics. This can make the model stronger and easier for others to comprehend and utilize. Auditors or officers can understand why the pattern is there and not merely follow the score.

Each metric was scored heuristically on a five-point Likert-type scale (1 = Low, 5 = High) for each dimension, drawing from synthesis of the literature and informed expert judgment. While such scoring does not provide definitive empirical quantification, it offers a structured and transparent means of articulating relative strengths and weaknesses among metrics. A common approach in methodological reviews is where direct comparisons are limited or context specific.

REFERENCES

[1] M. Kamal and A. Tohom, "Likelihood Rating of Fraud Risk in Government Procurement," The International Journal of Business Review (The Jobs Review), 2019, doi: 10.17509/tjr.v2i1.17903.

[2] D. M. Putri, D. Syariati, and A. Ahmad, "Fraud Prevention in the Perspective of Probity Audit (The Case Study of University X)," El Muhasaba Jurnal Akuntansi (E-Journal), 2021, doi: 10.18860/em.v12i2.12392.

[3] N. W. Rustiarini, T. Sutrisno, N. Nurkholis, and W. Andayani, "Why People Commit Public Procurement Fraud? The Fraud Diamond View," Journal of Public Procurement, 2019, doi: 10.1108/jopp-02-2019-0012.

[4] M. Gunasegaran, R. Basiruddin, and A. M. Rizal, "Detecting and Preventing Fraud in E-Procurement of Public Sector: A Review, Synthesis and Opportunities for Future Research," International Journal of Academic Research in Business and Social Sciences, 2023, doi: 10.6007/ijarbss/v13-i1/15970.

[5] N. Anggraeni, L. H. Husnan, and E. Pituringsih, "School Procurement Information System (SIPLah) as Moderating of Determinant of Fraud in Goods/Services Procurement," European Journal of Theoretical and Applied Sciences, 2023, doi: 10.59324/ejtas.2023.1(6).62.

[6] S. Sutoyo, M. F. Nugroho, and W. Windyastuti, "E-Procurement Implementation, Internal Control, and Organizational Commitment to Fraud Prevention Procurement of Goods and Servicesin Device Organizations Area (DOA) City Yogyakarta," Journal of Economics Finance and Management Studies, 2023, doi: 10.47191/jefms/v6-i5-54.

[7] N. W. Rustiarini, S. Sutrisno, N. Nurkholis, and W. Andayani, "Fraud Triangle in Public Procurement: Evidence From Indonesia," J Financ Crime, 2019, doi: 10.1108/jfc-11-2018-0121.

[8] A. P. Wicaksono, D. Urumsah, and F. Asmui, "The Implementation of E-Procurement System: Indonesia Evidence," SHS Web of Conferences, 2017, doi: 10.1051/shsconf/20173410004.

[9] Y. B. Adi and A. Rohman, "Determinants of Fraud Prevention of Procurement of Goods and Services in Government Agency," Jak (Jurnal Akuntansi) Kajian Ilmiah Akuntansi, 2023, doi: 10.30656/jak.v10i2.5498.

[10] E. Amwoha, "Constituents That Affect the Implementation of Sustainable Public Procurement in Kenyan Public Universities a Case of Technical University of Kenya," strategicjournals.com, 2015, doi: 10.61426/sjbcm.v2i1.78.

[11] N. Kumar and K. Ganguly, "Non-Financial E-Procurement Performance Measures," International Journal of Productivity and Performance Management, 2020, doi: 10.1108/ijppm-07-2019-0353.

[12] World Bank, "Common Red Flags of Fraud and Corruption in Procurement." [Online]. Available: https://documents1.worldbank.org/curated/en/223241573576857116/pdf/Warning-Signs-of-Fraud-and-Corruption-in-Procurement.pdf

[13] OECD, "ALGORITHMS AND COLLUSION - Competition policy in the digital age," 2017.

[14] Malaysia Competition Commission (MyCC), "Bid Rigging," 2022. Accessed: Jun. 13, 2023. [Online]. Available: https://www.mycc.gov.my/sites/default/files/pdf/newsroom/Bid%20Rigging%20BI_outline_0_0.pdf

[15] N. V Chawla, "Data mining for imbalanced datasets: An overview," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds., Springer, 2005, pp. 853–867. doi: 10.1007/0-387-25465-X_40.

[16] J. D. Kelleher, B. Mac Namee, and A. D'arcy, Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press, 2020.

[17] V. Lopez, A. Fernandez, S. Garcia, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf Sci (N Y), vol. 250, pp. 113–141, 2013, doi: 10.1016/j.ins.2013.07.007.

[18] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," 2022.

[19] S. T. Boppiniti, "Machine learning for predictive analytics: Enhancing data-driven decision-making across industries," International Journal of Sustainable Development in Computing Science, vol. 1, no. 3, p. 13, 2019.

[20] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit Lett, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[21] M. Bowles, Machine Learning with Spark and Python: Essential Techniques for Predictive Analytics. John Wiley & Sons, 2019.

[22] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago, Chile: Morgan Kaufmann, 1994, pp. 487–499. [Online]. Available: https://www.vldb.org/conf/1994/P487.PDF

[23] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 2nd ed. Pearson, 2019.

[24] A. P. Siddaway, A. M. Wood, and L. V Hedges, "How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses," Annu Rev Psychol, vol. 70, no. 1, pp. 747–770, 2019.

[25] M. Motevalli, "Comparative analysis of systematic, scoping, umbrella, and narrative reviews in clinical research: critical considerations and future directions," Int J Clin Pract, vol. 2025, no. 1, p. 9929300, 2025.

[26] M. Pautasso, "The structure and conduct of a narrative literature review," A guide to the scientific career: Virtues, communication, research and academic writing, pp. 299–310, 2019.

[27] J. Sukhera, "Narrative reviews: flexible, rigorous, and practical," J Grad Med Educ, vol. 14, no. 4, pp. 414–417, 2022.

[28] E. Toomey et al., "Focusing on fidelity: narrative review and recommendations for improving intervention fidelity within trials of health behaviour change interventions," Health Psychol Behav Med, vol. 8, no. 1, pp. 132–151, 2020.

[29] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans Knowl Data Eng, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[30] R. Saia and S. Carta, "Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks," Future Generation Computer Systems, vol. 93, pp. 18–32, 2019.

[31] M. T. Islam, M. A. Rahman, M. T. R. Mazumder, and S. H. Shourov, "Comparative Analysis of Neural Network Architectures for Medical Image Classification: Evaluating Performance Across Diverse Models," American Journal of Advanced Technology and Engineering Solutions, vol. 4, no. 01, pp. 1–42, 2024.

[32] I. Michael, "Evaluation Metrics and Benchmarking for Predictive Accuracy," Jun. 2025.

[33] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable artificial intelligence in cybersecurity: A survey," Ieee Access, vol. 10, pp. 93575–93600, 2022.

[34] S. Dutta, P. Bhattacharya, and L. Dey, "An ensemble approach for early detection of banking fraud using soft computing techniques," Journal of Intelligent & Fuzzy Systems, vol. 31, no. 5, pp. 2689–2700, 2016, doi: 10.3233/IFS-162229.

[35] E. Lopez-Rojas and S. Axelsson, "Money laundering detection using synthetic data," in 2012 European Intelligence and Security Informatics Conference (EISIC), IEEE, 2012, pp. 252–259. doi: 10.1109/EISIC.2012.53.

[36] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Association for Computing Machinery, 1993, pp. 207–216. doi: 10.1145/170035.170072.

[37] C. Carbone, F. Calderoni, and M. Jofre, "Bid-rigging in public procurement: cartel strategies and bidding patterns," Crime Law Soc Change, vol. 82, no. 2, pp. 249–281, 2024.