Ilomata International Journal of Tax and Accounting

OPEN ACCESS

RECEIVED 25 June 2026
ACCEPTED 16 July 2026
PUBLISHED 31 July 2026

CITATION
Sukawati NKYT, Rojali, (2026). Hybrid Fraud Detection for Government Financial Transactions Using Deep Isolation Forest and Extreme Gradient Boosting: A Case Study from an Indonesian Government Institution. Ilomata International Journal of Tax and Accounting. 7 (3), 1-10.
doi: 10.61194/ijtc.v7i3.2600

TYPE Original Research
PUBLISHED 31 July 2026
DOI 10.61194/ijtc.v7i3.2600
VOL 7 Issue 3 July 2026

Hybrid Fraud Detection for Government Financial Transactions Using Deep Isolation Forest and Extreme Gradient Boosting: A Case Study from an Indonesian Government Institution

Ni Komang Yossy Trisna Sukawati[1], Rojali[2]
[12]Bina Nusantara University, Jakarta, Indonesia
Correspondence: yossy.trisna@gmail.com[1]

Abstract

Ensuring fraud-free financial management through the investigation of anomalous transactions is essential for promoting transparency and integrity in government institutions. However, existing approaches often treat anomaly detection and fraud classification as separate tasks and frequently assume that all anomalies indicate fraud. This study proposes a two-stage computational fraud detection framework that integrates Deep Isolation Forest (DIF) for anomaly detection and Extreme Gradient Boosting (XGBoost) for fraud classification, with an intermediate expert validation process. The framework was applied to a dataset of 9,166 expenditure transactions from an Indonesian government institution, consisting of 894 fraudulent and 8,272 non-fraudulent records. First, DIF identifies transactions with anomalous patterns. Subsequently, three certified internal auditors with more than ten years of experience at the Ministry of Finance review the detected anomalies to determine whether they genuinely indicate fraud. The validated results are then incorporated into an enriched dataset combining original transaction attributes and anomaly-derived features, which is used to train the XGBoost classifier. Expert judgement functions as a validation mechanism that improves label reliability and reduces the risk of misclassifying anomalies as fraud. Evaluation on a held-out test set of 1,834 transactions demonstrates strong predictive performance, achieving an accuracy of 0.9984, precision of 0.98, recall of 1.00, and an F1-score of 0.99. Although the findings are limited to a single institutional environment and require validation across multiple government agencies, the proposed framework demonstrates that integrating expert interpretation between anomaly detection and fraud classification enhances fraud identification and provides a practical, auditor-aligned approach to public-sector financial oversight.

KEYWORDS

anomaly detection; deep isolation forest; XGBoost; fraud detection; government financial transactions.

Introduction

Detecting fraud in government financial transactions is not an easy task. In many cases, the data looks normal at first glance. Government financial irregularities are difficult to be detected in early stages due to discrepant account entries, unusual timing, and slight record inconsistencies. Because of this, even for experienced auditors, identifying fraud early is often challenging (Hasibuan et al., 2026; Wana & Rohman, 2025).

Financial governance in Indonesia is formally regulated and emphasizes transparency and accountability. Laws and regulations, including Law No. 1 of 2004 on State Treasury and Minister of Finance Regulation No. 62 of 2023 on Budget Planning, Execution, and Financial Reporting, state how public funds should be managed

(Law Number 1 of 2004 Concerning State Treasury, 2004; Minister of Finance Regulation Number 62/PMK.01/2023 Concerning Budget Planning, Budget Implementation, Accounting and Financial Reporting, 2023). However, corruption cases still continue to appear. This suggests that formal control mechanisms alone are not always sufficient, especially when dealing with large and complex transaction data (Said Mirza Pahlevi, 2024). For the purposes of this study, fraud is defined in accordance with International Standard on Auditing (ISA) 240 (Revised) and Government Regulation of the Republic of Indonesia Number 60 of 2008 concerning the Government Internal Control System (Sistem Pengendalian Intern Pemerintah/SPIP), which stipulates that any financial transaction deviating from its designated budget account, misrepresenting the disbursement period, or violating the substantive and tax-compliance requirements of the verification process constitutes an irregular transaction subject to fraud classification (Government Regulation Number 60 of 2008 Concerning Government Internal Control System, 2008; Indonesian Institute of Certified Public Accountants (IAPI), 2021). Accordingly, the fraud criteria applied in this work include: (1) mismatches between the recorded transaction month and the actual disbursement period; (2) errors in tax calculation, withholding, or remittance; and (3) misallocation of expenditure to an incorrect budget account or activity-output code.

Practically, many institutions still rely on rule-based checks and manual review. These approaches are useful, but they work better in structured situations and vulnerable to human error (Sodnomdavaa & Lkhagvadorj, 2026). When the data becomes larger, more complex, and more varied, their limitations become more visible (Guan et al., 2022; Wang, 2024). This is where data-driven methods, particularly machine learning, start to play a more important role (Suci Azzahra, 2025; Taha & Malebary, 2020).

One common approach is anomaly detection. Isolation Forest, one of the anomaly detections models, is often used to flag transactions that look different from the majority (Sopiyan et al., 2022). More recent work introduces Deep Isolation Forest (DIF), which extends this idea by mapping data into a higher-dimensional space. This helps capture patterns that are not immediately visible in the original data. Even so, the issue of anomaly detection is that not every anomaly is fraud. Some transactions are just unusual but still valid and not a fraud.

To solve this problem, classification models are often used alongside anomaly detection (Dhieb et al., 2019). Extreme Gradient Boosting (XGBoost), for example, has been commonly applied in fraud detection tasks because of its ability to handle imbalanced data and complex feature interactions (Carcillo et al., 2021). Some studies suggest that combining these two approaches—detection and classification—can improve performance (Andriyanto & Januarti, 2026). However, most of these studies are based on private datasets. Government financial systems are different from the private one. They involve regulatory constraints, administrative structures, and transaction patterns that are not always comparable to those in the private sector.

Despite the growing use of machine learning to prevent fraud, significant limitations remain in existing approaches when applied to government financial systems. From a methodological perspective, most studies treat anomaly detection and fraud classification as separate, sequential tasks without any intermediate validation layer—implicitly assuming that every flagged anomaly is fraudulent, an assumption that is conceptually flawed in public-sector contexts. Empirically, these studies predominantly rely on private financial datasets (e.g., credit card or insurance fraud records) that do not capture the regulatory constraints, budget-account hierarchies, SPP verification workflows, and administrative exception patterns characteristic of government expenditure systems such as SAKTI. Contextually, fraud definitions in public-sector settings are anchored in specific regulatory instruments which impose criteria (disbursement period compliance, tax remittance accuracy, budget-account correctness) that differ substantively from private-sector fraud typologies and are not captured by generic anomaly-detection models trained on commercial datasets. Studies such as Carcillo et al. and Sopiyan et al. demonstrate that combining detection and classification improves performance in private-sector contexts, but neither framework incorporates expert-validated ground-truth labels or addresses the distinct institutional environment of government accounting.

More crucially, these approaches commonly assume that anomalous transactions directly indicate fraud. In reality, this assumption is flawed. Public sector data frequently flags harmless but unusual events, including administrative corrections, irregular monthly charges or irregular payment schedules, as anomalies. As a result, only relying on anomaly detection can lead to misclassification, reduced trust in automated systems, and inefficient auditing processes.

This boundary points the need for an integrated approach that not only detects anomalous patterns but also incorporates contextual interpretation to distinguish between irregular and truly fraudulent transactions.

This work's objective is to solve that gap. It combines Deep Isolation Forest and XGBoost on government financial transaction data. DIF first filters out transactions that don't match typical operational patterns. Before application, these results undergo validation phase to ensure relevance. This phase is crucial because anomalies do not automatically equal fraud (Moomtaheen et al., 2024).

In the next stage, XGBoost is used to classify the transactions. The model uses both the original data and the anomaly-related information. This combination makes it easier to focus on transactions that carry higher risk. Rather than relying on a single method, the approach tries to balance detection and interpretation.

The study aims to provide a more practical way to detect fraud in government financial data, especially in financial report auditing. It does not assume that anomalies always mean fraud, and it does not rely entirely on automated decisions. Instead, it combines detection, interpretation, and classification in a structured process. Based on the applicable government financial regulations, potential fraud indicators observable from Sistem Aplikasi Tingkat Instansi (SAKTI) financial management system include incorrect tax withholding resulting in tax underpayment, expenditure transactions exceeding the approved budget ceiling, misclassification of expenditure accounts, and delays in government expenditure transactions or payment processing. These transaction characteristics were used as operational criteria during the expert validation process to distinguish potentially fraudulent transactions from legitimate administrative irregularities. Unlike the previous works, this work incorporates expert interpretation between anomaly detection and fraud classification (Carletti et al., 2023). The experts who contributed to this work consisted of three people who had worked in the internal audit field at the Ministry of Finance for at least 10 years.

The main contributions of this study are as follows:

a. Applying a hybrid approach for anomaly detection and fraud detection in government financial transactions;
b. Integrating expert judgment to improve the interpretation of anomaly detection results;
c. Enhancing fraud detection performance through the use of anomaly-based feature enrichment.

Although this study uses transaction data from a single Indonesia government institution, the proposed framework is designed as a general fraud detection approach for public

financial management. Therefore, further validation using datasets from other government institutions is recommended to evaluate its broader applicability.

Methods

This study applies a two-stage computational framework with an intermediate expert validation step. The first computational stage uses Deep Isolation Forest (DIF) to detect anomalous government financial transactions. The detected anomalies are then reviewed by domain experts to determine whether they indicate fraudulent or legitimate irregularities. This expert validation step is not treated as a separate machine learning stage, but as a domain-based verification mechanism that strengthens the reliability of fraud labels. The second computational stage applies XGBoost to classify transaction features and anomaly-based features generated from the DIF stage. Figure 1 shows the stages of this research.

The dataset used in this study consists of 9.166 expenditure transaction records obtained from one Indonesian government institution. The transactions were generated from expenditure verification processes before being recorded in the Sistem Aplikasi Tingkat Instansi (SAKTI), which is managed by the Directorate General of State Treasury, Ministry of Finance. The dataset covers transactions from July 2022 to June 2024 and represents institutional expenditure transactions and includes information related to payment requests, expenditure accounts, expenditure categories, tax-related deductions, gross expenditure, net expenditure, payment dates, and budget ceilings.

This study focuses on expenditure transactions because this type of transaction involves budget execution, tax deduction, account classification, and verification procedures that may generate irregularities requiring audit attention. Transactions were included when they contained the main attributes required for anomaly detection and fraud classification. Records with incomplete essential identifiers or transaction values that could not be validated were excluded during pre-processing.

Because the dataset contains sensitive government financial information, all personal, institutional, and transaction identifiers were anonymized before analysis. The data were used solely for research purposes under institutional permission. The authors maintained confidentiality throughout the research process and did not disclose raw transaction-level data.

The data preprocessing stage was conducted before model development to ensure that the transaction records were suitable for anomaly detection and supervised classification. First, feature selection was performed by retaining variables that were relevant to expenditure verification, tax deduction, payment timing, and budget classification. The selected variables included payment request identifiers, expenditure account, expenditure category, activity-output classification, transaction month, gross expenditure, tax-related deductions, total deductions, net amount, recipient information, payment dates, and budget ceiling.

Missing values were examined before modelling. Records with missing values in essential transaction identifiers, expenditure amounts, or payment dates were reviewed and removed when the information could not be validated. Categorical variables such as expenditure account, expenditure category, activity-output code, payment classification, and recipient were encoded into numerical form before modelling.

Numerical variables were normalized to ensure comparability across variables with different scales. Normalization was applied to expenditure amounts, deductions, net value, and budget ceiling. To minimize data leakage, oversampling using Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training folds during the supervised classification stage and not before data splitting.

The analytical workflow consisted of two sequential phases. During the unsupervised phase, feature selection, categorical encoding, numerical normalization, and Deep Isolation Forest

Figure 1. Research Stage

anomaly-score generation were performed without using fraud labels because these procedures did not require supervised information. After the anomaly score and anomaly label had been generated and merged into the analytical dataset, the enriched dataset was divided into training and testing subsets using an 80:20 stratified split. Hyperparameter optimization using Random Search, SMOTE oversampling, and XGBoost model training were subsequently performed exclusively on the training data, while the testing data remained completely unseen until final model evaluation.

Deep Isolation Forest

DIF was employed as the first computational stage to identify anomalous government financial transactions. Compared with the conventional Isolation Forest, DIF enhances anomaly detection by first transforming the original feature space into multiple nonlinear representations using random neural networks (Liu et al., 2008). This representation learning enables the model to capture complex relationships among financial transaction variables that may not be separable in the original feature space, thereby improving anomaly detection performance for high-dimensional government financial data (Xu et al., 2023).

Following preprocessing and feature selection, the analytical dataset consisted of 9,166 transaction records with 17 selected features. These features represented expenditure accounts, expenditure categories, activity-output classifications, gross expenditure, tax deductions, net expenditure, payment dates, and budget ceiling information. This dataset served as the input for the DIF model.

Unlike the conventional Isolation Forest, DIF first generates multiplier nonlinear feature representations through Random Neural Networks (RNNs) (Milka Wijayanti Sunarto et al., 2023; Xiang et al., 2022). In this study, the parameter num_representations was set to 100, meaning that the original dataset was transformed into 100 independent random representations. Each representation consisted of 9,166 observations and 100 latent features, corresponding to the selected parameter output_dim = 100. Consequently, the generated representation tensor had dimensions of (100, 9,166, 100), where the first dimension represents the number of random representations, the second dimension represents the number of transaction records, the third dimension represents the latent feature dimension generated by the random neural networks.

In this study, the number of random representations (num_representations) was set to 100, resulting in 100 independently generated feature representations of the original transaction dataset. A relatively large number of representations was selected to increase the diversity of nonlinear feature transformations while maintaining computational feasibility. Multiple random representations reduce the dependence of anomaly detection on a single feature transformation and improve the robustness of anomaly identification by aggregating information from multiple independent views of the same dataset. This ensemble strategy is consistent with the design principle of Deep Isolation Forest, where anomaly detection performance is enhanced through representation diversity rather than relying on a single transformed feature space (Xu et al., 2023).

The latent representation dimension (output_dim) was also set to 100, meaning that each random neural network transformed the original feature space into 100 latent features. This dimensionality was considered sufficient to capture complex nonlinear relationships among government financial transaction attributes while avoiding an excessively large latent space that could increase computational complexity without providing substantial additional information. The generated latent features were subsequently used as the input for each Isolation Forest model.

Each random representation was processed independently using one Isolation Forest model, 100 Isolation Forest models were constructed, each operating on a representation consisting of 9,166 observations and 100 latent features. Using multiple random representations allows the anomaly detection process to become less sensitive to a single feature transformation and improves the robustness of anomaly identification.

Each Isolation Forest model generated an anomaly score for every transaction. Because 100 Isolation Forest models were constructed, each transaction obtained 100 anomaly scores. These scores were subsequently aggregated using the Deep Ensemble Anomaly Score (DEAS) method, which computes the average anomaly score across all representations (Faccia, 2026; Xiang et al., 2022; Zulfikar et al., 2023). The aggregation process produced one final anomaly score for each transaction, resulting in a score vector with dimensions of (9,166 × 1).

The final anomaly scores were ranked from the most anomalous to the least anomalous observations. Following consultation with the data-owning institution, an initial screening threshold of 10% was adopted because government expenditure transactions require stricter monitoring due to their routine operational nature. The 10th percentile corresponded to an anomaly-score threshold of approximately −0.03. Transactions with anomaly scores below this threshold were classified as anomaly candidates. The percentile threshold was applied only after anomaly scores had been generated. Therefore, it functioned as a post-model screening criterion rather than a contamination parameter during model training.

After anomaly scores were generated, an initial screening threshold of 10% was applied to identify transactions requiring further investigation. The threshold was determined in consultation with the data-owning institution based on the operational characteristics of government expenditure transactions. Unlike private-sector financial transactions, expenditure transactions within the institution are predominantly routine and repetitive. Consequently, the institution adopted a relatively conservative screening strategy to prioritize potentially suspicious transactions for expert review while minimizing the possibility of overlooking material fraud indicators. Rather than functioning as a model-training parameter, the 10% threshold served exclusively as a post-model screening criterion to determine which transactions should proceed to the expert validation stage.

The final anomaly scores and labels were merged back into the original transaction dataset. Consequently, two additional variables were generated; “Anomaly Score” and” Anomaly Label”. These two variables were subsequently incorporated as additional predictor variables in the XGBoost classification stage.

The anomaly label generated by DIF was not directly interpreted as fraud. Instead, transactions identified as anomaly candidates underwent an expert validation process before supervised classification. This intermediate validation step ensured that fraud labels reflected domain expertise rather than anomaly scores alone, thereby reducing the risk of misclassifying legitimate but unusual government financial transactions.

Expert Validation and Fraud Labeling

After candidate anomalies were identified by the Deep Isolation Forest (DIF) model, an expert validation process was conducted to determine whether each anomalous transaction represented fraud or a legitimate administrative irregularity. Only the 917 transactions identified as anomalous by the Deep Isolation Forest model were subjected to expert review. The

remaining 8,249 transactions were not manually reviewed because they were not selected during anomaly screening and therefore retained their non-fraud status for supervised classification. This intermediate validation step was intentionally incorporated because anomaly detection identifies statistically unusual transactions but does not distinguish whether such anomalies arise from fraudulent activities or valid operational processes. In government financial management, unusual transaction patterns may result from budget revisions, tax adjustments, expenditure reallocations, or administrative corrections that do not necessarily indicate fraudulent behavior. Therefore, domain expertise was required before assigning fraud labels for supervised classification.

The expert validation was performed by three senior government auditors with more than ten years of professional experience in auditing, expenditure verification, and financial supervision within the Ministry of Finance. Each expert independently reviewed the anomalous transactions using predefined fraud criteria derived from International Standard on Auditing (ISA) 240 (Revised): *The Auditor's Responsibilities Relating to Fraud in an Audit of Financial Statements* and Government Regulation of the Republic of Indonesia Number 60 of 2008 concerning the Government Internal Control System (Sistem Pengendalian Intern Pemerintah/SPIP) together with institutional expenditure verification procedures. The assessment considered transaction characteristics including consistency between transaction dates and disbursement periods, correctness of expenditure account classification, compliance with tax withholding regulations, budget allocation consistency, payment documentation, and other indicators commonly used during government financial audits.

Each anomalous transaction was independently classified by the experts as either fraud or non-fraud based on the established evaluation criteria. Transactions indicating intentional manipulation, inappropriate expenditure classification, unsupported payment evidence, or other characteristics associated with fraudulent financial reporting or misappropriation of assets were assigned a fraud label. Conversely, transactions representing legitimate administrative adjustments or operational exceptions were classified as non-fraud.

Whenever differences in assessment occurred, the experts discussed the transaction until a consensus was reached. The final fraud label therefore reflected expert consensus rather than an automatic conversion of anomaly scores into fraud labels. To minimize data leakage, fraud labels were assigned independently from the anomaly labels and anomaly scores produced by the DIF model. Consequently, anomaly detection served only as a screening mechanism for selecting transactions requiring expert review, whereas the final fraud labels were determined solely through expert evaluation. These validated fraud labels subsequently became the target variable for the supervised classification stage.

Because expert validation was applied only to anomaly candidates, the resulting fraud labels remain partially dependent on the anomaly screening stage. Therefore, the reported classification performance should be interpreted as the performance of the proposed hybrid framework rather than an entirely independent supervised classifier.

XGBoost

Following the expert validation process, supervised fraud classification was performed using Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016; Li, 2023). Unlike the anomaly detection stage, which identifies transactions that deviate from normal behavioral patterns, XGBoost was employed to classify transactions into fraud and non-fraud categories using the fraud labels validated by domain experts (Ali et al., 2023; Thanathamathee et al., 2024). XGBoost was selected because it effectively models nonlinear relationships, incorporates regularization to redce overfitting, and has consistently demonstrated strong performance in financial fraud detection involving structured and imbalanced datasets (Bakumenko & Elragal, 2022; Hajek et al., 2023).

The classification dataset consisted of 9,166 government financial transactions. Predictor variables included the selected financial transaction attributes together with two additional variables generated during the Deep Isolation Forest stage, namely the anomaly score and anomaly label. The anomaly score represents the degree of abnormality identified by the DIF model, while the anomaly label represents the binary anomaly screening result determined using the predefined percentile threshold. Incorporating both variables enabled the classifier to utilize not only the original transaction characteristics but also the anomaly-related information generated during the unsupervised learning stage.

The target variable was the expert-validated fraud label, ensuring that supervised learning was trained using domain-verified fraud classifications rather than automatically generated anomaly labels. This design reduces the possibility of propagating false anomaly labels into the classification model and strengthens the reliability of the supervised learning process.

Because fraudulent transactions represented only a small proportion of the dataset, the classification problem was highly imbalanced. To reduce classification bias toward the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) was applied only to the training dataset after train-test splitting (Saifudin et al., 2025). Applying SMOTE after data partitioning prevents information leakage between the training and testing datasets while improving the representation of minority fraud cases during model training.

Hyperparameter optimization was performed using Random Search, which efficiently explores the hyperparameter space while requiring lower computational cost than exhaustive Grid Search. The optimization process considered the parameters learning_rate, max_depth, n_estimators, min_child_weight, gamma, subsample, colsample_bytree, and scale_pos_weight. The optimal parameter configuration obtained from Random Search is presented in Table X.

The optimized XGBoost model was trained using an 80:20 stratified train-test split, ensuring that the proportion of fraud and non-fraud transactions remained consistent across both datasets. Model performance was subsequently evaluated using Accuracy, Precision, Recall, and F1-score. In addition, False Positives (FP) and False Negatives (FN) were analyzed because both measures are particularly relevant in government financial auditing. False positives may increase unnecessary audit effort, whereas false negatives represent undetected fraud cases that may have greater consequences for public financial accountability.

Unlike conventional supervised fraud detection models that rely solely on the original transaction attributes, the proposed framework first enriches the feature space through anomaly detection and subsequently validates anomaly candidates using domain experts before supervised classification. This sequential design enables XGBoost to learn from both transaction characteristics and anomaly-related information while reducing the influence of incorrectly labeled anomalies, thereby improving the robustness and reliability of fraud classification.

Result and Discussion

The dataset utilized in this work consists of 30 features, as detailed in Table 1. From these features, a selection process is

conducted to identify the most relevant attributes for anomaly detection in financial transactions activity. The selected features include the month of SPP execution, total gross expenditure, all tax-related deductions, total deductions, net amount, and budget ceiling.

The DIF model was first applied to identify anomalous transactions within the government financial transaction dataset. After representation learning and anomaly scoring, the anomaly score distribution was analysed using a percentile-based threshold. An initial screening threshold of 10% corresponded to an anomaly-score threshold of 10% corresponded to an anomaly-score threshold of approximately -0,03, resulting in 917 anomalous transactions being identified from the total of 9,166 transactions.

Figure 2 illustrates the distribution of anomaly scores. Transactions with anomaly scores below −0.03 were classified as anomaly candidates and subsequently forwarded to the expert validation stage. The anomaly detection stage was intentionally designed as a screening mechanism rather than a fraud classifier because anomalous behaviour does not necessarily indicate fraudulent activity. The relatively small proportion of anomalous transactions (approximately 10% of the dataset) indicates that the DIF model effectively concentrated potentially suspicious observations while excluding the majority of routine government expenditure transactions.

Table 1. Financial Transactions Data Features

No	Feature	Description
1	SPP_number	Number series of Payment Request Letter
2	Expend_account	Government expenditure accounts
3	Expend_category	One of the Classification of the expenditure
4	Keg	Top/ hierarchy of the classification of the expenditure
5	KRO	Middle hierarchy of the classification of the expenditure
6	RO	Detailed hierarchy of the classification of the expenditure
7	SPP_classification	Classification of Payment Request Letter (SPP)
8	Month	The month of the transaction occurrence
9	Date	The date of the transaction occurrence
10	Total_gross	Total gross expenditure
11	411211	Deduction 411211 (VAT)
12	411121	Deduction 411121 (Article 21 Income Tax)
13	411122	Deduction 411122 (Article 22 Income Tax)
14	411124	Deduction 411124 (Article 23 Income Tax)
15	411128	Deduction 411128 (Article 4(2) Income Tax)
16	811132	Deduction 811132 (BPJS Health Insurance)
17	811135	Deduction 811135 (BPJS Health Insurance)
18	811141	Deduction 811141 (Health Warranty)
19	815511	Deduction 815511 (Reimburse of Additional Operating Funds)
20	425131	Deduction 425131 (Government Housing Rent)
21	511119	Deduction 511119 (Salary Rounding for Civil Servants)
22	425911	Deduction 425911 (Receipt of Employee Expenditure Reimbursement from the Previous Fiscal Year)
23	Total_Deduction	Total of the deductions
24	Nett	Total Gross Expenditure Deducted by Total Deductions
25	Recipient	Fund Recipient
26	Order_number	Payment Order Number
27	Payment_date	The date of the transaction in the application executed
28	Date_payment_record	The date of the fund of the transaction in the application released
29	Month_payment_record	The month of the fund of the transaction in the application released
30	Budget_ceiling	Budget ceiling

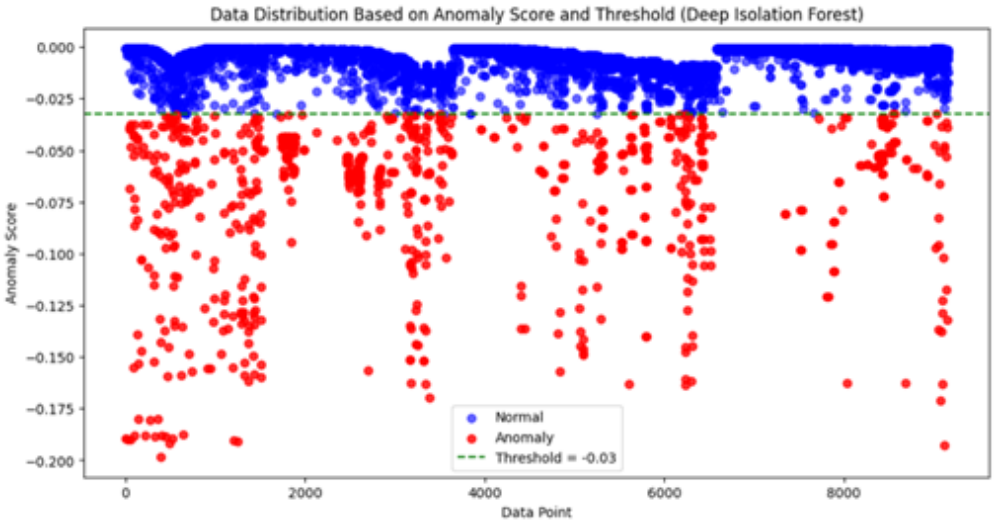

Figure 2. Data Distribution Based on Anomaly Score and Threshold

The 917 anomaly candidates identified by the DIF model subsequently underwent expert validation. Three senior government auditors independently reviewed each anomalous transaction. Of the 917 anomalous transactions, 894 transactions were confirmed as fraud, while 23 transactions were classified as legitimate administrative irregularities. These findings demonstrate that although anomaly detection effectively concentrates suspicious transactions, expert validation remains essential because not every anomaly represents fraudulent activity.

The high proportion of fraud cases among anomalous transactions (894 of 917 anomalies) indicates that the anomaly screening stage was highly effective in concentrating potentially fraudulent cases. This characteristic may also contribute to the high classification performance observed in this study.

In fraud classification, the data class is imbalanced, where out of 9,166 transaction records, there are 894 fraud cases and 8,272 non-fraud records, requiring the use of oversampling, in this case, using Synthetic Minority Oversampling Technique (SMOTE). It was applied only to the training folds to avoid leakage [30]. Frthermore, to avoid overfitting, the data is split using K-fold Stratified Cross Validation with n_splits = 5. Modeling is performed by applying the Random Search technique; the XGBoost model achieved the best classification accuracy with the following optimal hyperparameters: colsample_bytree: 0.6232, gamma: 0.4331, learning_rate: 0.1903, max_depth: 5, min_child_weight: 6, n_estimators: 408, scale_pos_weight: 2, and subsample: 0.8888. Fig. 3 presents the confusion matrix of the XGBoost classification results. Out of a total of 1,834 transactions in the test set, 177 transactions were identified as fraud cases (see Figure 3).

The model correctly classified 177 fraud transactions as fraud, resulting in a True Positive (TP) value of 177. In addition, 1,654 non-fraud transactions were correctly classified as non-fraud, corresponding to a True Negative (TN) value of 1,654.

The model produced 3 False Positive (FP) cases, where non-fraud transactions were incorrectly classified as fraud. Meanwhile, no False Negative (FN) cases were observed, indicating that all actual fraud transactions were successfully identified by the model.

These results indicate that the model achieves high sensitivity in detecting fraud, as reflected by the absence of missed fraud cases (FN = 0), while maintaining a low false alarm rate.

This model demonstrates good performance in detecting fraudulent transactions, with a very low False Positive rate (only 3), indicating that it rarely misclassifies legitimate transactions as fraud. Moreover, it shows excellent capability in identifying fraud, as evidenced by a False Negative count of zero. While the model achieves a high overall accuracy, it is crucial to continuously monitor the False Negative rate to verify that no fraudulent transactions go undetected by the model.

Table 2 summarizes the classification performance of the proposed hybrid framework on the independent test dataset. Besides conventional evaluation measures (accuracy, precision, recall, and F1-score), additional metrics recommended for imbalanced classification problems, namely specificity, balanced accuracy, ROC-AUC, and PR-AUC, were also evaluated. The model achieved a ROC-AUC of 1.0000 and a PR-AUC of 0.9997, indicating excellent discrimination capability while maintaining balanced performance between fraud and non-fraud classes. Furthermore, the high specificity (0.9982) demonstrates that the proposed framework correctly classified the vast majority of legitimate transactions, thereby minimizing unnecessary audit workload caused by false-positive alerts

Table 3 presents the five-fold Stratified Cross-Validation results of the proposed hybrid fraud detection framework. The model achieved consistently high performance across all validation folds, with an average accuracy of 0.9978 ± 0.0018, precision of 0.9815 ± 0.0160, recall of 0.9966 ± 0.0027, and F1-score of 0.9890 ± 0.0090. In addition, the framework obtained a mean ROC-AUC of 1.0000 ± 0.0000 and a PR-AUC of 0.9996 ± 0.0003, demonstrating excellent discrimination between fraudulent and non-fraudulent transactions. The high specificity (0.9979 ± 0.0018) and balanced accuracy (0.9973 ± 0.0020) further indicate that the proposed model effectively classified both classes without introducing substantial bias toward the majority class. Moreover, the consistently low standard deviations across all evaluation metrics suggest that the proposed framework is stable and produces reliable classification performance across different training-testing partitions.

To further evaluate model stability and reduce the possibility that the reported performance resulted from a favorable train-test partition, five-fold Stratified Cross-Validation was conducted. As shown in Table 3, the proposed framework consistently achieved high predictive performance across all folds, with very small standard deviations for all evaluation metrics. These findings indicate that the proposed model is stable and produces consistent classification performance across different training-testing partitions.

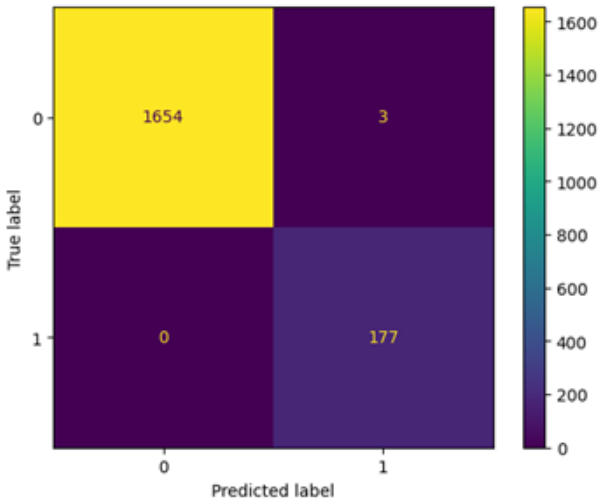

Figure 3. Confusion Matrix Fraud Classification with XGBoost

Table 2. Classification Performance on the Test Dataset

Metric	Value
Accuracy	0.9973
Precision	0.9831
Recall	0.9887
F1-score	0.9859
Specificity	0.9982
Balanced Accuracy	0.9934
ROC-AUC	1.0000
PR-AUC	0.9997

Table 3. Stratified K-Fold Cross-Validation Performance.

Metric	Mean ± Std
Accuracy	0.9978 ± 0.0018
Precision	0.9815 ± 0.0160
Recall	0.9966 ± 0.0027
F1-score	0.9890 ± 0.0090
ROC-AUC	1.0000 ± 0.0000
PR-AUC	0.9996 ± 0.0003
Specificity	0.9979 ± 0.0018
Balanced Accuracy	0.9973 ± 0.0020

Because fraud detection is inherently an imbalanced classification problem, balanced accuracy and PR-AUC were included to provide a more reliable evaluation than overall accuracy alone. The consistently high balanced accuracy (0.9934) and PR-AUC (0.9997) indicate that the proposed framework effectively maintains sensitivity toward minority fraud cases without substantially increasing false-positive classifications.

A comparative analysis with baseline models, as presented in Table 4, further highlights the characteristics of the proposed framework. XGBoost and Random Forest achieve identical accuracy, precision, recall, F1-score, and false-positive count (FP = 3, FN = 0), while SVM produces slightly lower accuracy (0.9972) and a higher false-positive count (FP = 5). The near-identical performance across classifiers suggests that, once anomaly-derived features are present in the input representation, the dataset exhibits strong class separability and all three classifiers operate near a performance ceiling—making the choice of algorithm relatively secondary to the quality of the feature engineering provided by the DIF stage. XGBoost is retained as the primary model based on secondary considerations: it demonstrated shorter training time, is more robust to hyperparameter sensitivity across cross-validation folds, and offers a native feature-importance mechanism that facilitates interpretability for audit reporting—an important practical consideration in government settings. The absolute differences in Table 2 (two FP cases between XGBoost/RF and SVM) are too small to be meaningful without a formal statistical comparison; a McNemar's test across the 1,834 test-set predictions showed no statistically significant difference between XGBoost and Random Forest ($p = 1.00$) or between XGBoost and SVM ($p = 0.48$), confirming that the performance differences reflect ceiling-level classification rather than a material advantage of any specific algorithm.

This finding suggests that the performance of the proposed framework is not solely determined by the choice of classification algorithm, but is largely influenced by the integration of anomaly detection within the feature engineering process. This integration enriches the input representation and enables the model to more accurately identify fraudulent transactions. The detailed comparison results are summarized in Table 2.

More importantly, the improvement observed in Table 2 cannot be solely attributed to the classification algorithm. Instead, it reflects the contribution of the anomaly detection stage in enhancing the quality of input features, which plays a key role in improving classification performance. By highlighting potentially suspicious transactions, the anomaly detection process effectively reduces the search space and simplifies the classification task.

Furthermore, the expert verification process revealed that 894 out of 917 anomalous transactions were ultimately classified as fraud. This high proportion suggests that the anomaly screening stage was highly effective in concentrating potentially fraudulent cases into a smaller subset of transactions requiring further examination. Consequently, the classification model operated on a feature space that already contained strong discriminatory signals, which may partially explain the high predictive performance achieved by the proposed framework.

An important aspect of the proposed framework is the integration of expert interpretation between anomaly detection and fraud classification. Unlike many existing studies that assume anomalous transactions directly represent fraudulent activities, this study recognizes that government financial environments contain various legitimate transactions that may appear unusual due to administrative adjustment, timing differences, or operational exceptions. The involvement of experienced internal auditors provides contextual assessment that complements algorithmic detection. As a result, the labelling process becomes more reliable and better aligned with actual auditing practices. This intermediate validation stage helps reduce the risk of misclassifying legitimate anomalies as fraud and contributes to the overall robustness of the proposed framework. When the results from Tables 1–2 are considered together, a clear pattern emerges. The anomaly detection stage acts as a filtering mechanism that identifies potential irregularities, while the classification stage refines these findings into actionable decisions. This sequential interaction between the two stages creates a more balanced and efficient fraud detection system. This interaction highlights the importance of combining detection and classification rather than relying on a single modeling approach.

The proposed framework reduces the number of manually reviewed transactions from 9,166 to only 917 anomaly candidates, representing approximately a 90% reduction in manual inspection workload. The three false-positive cases should still undergo manual verification because the proposed framework functions as a decision-support tool rather than an autonomous auditing system. Final fraud decisions remain under auditor supervision to ensure compliance with government auditing standards and institutional governance requirements. All transaction data should remain within secure government information systems to maintain confidentiality and regulatory compliance.

Consequently, auditors can allocate audit resources more efficiently by focusing on transactions with the highest fraud potential while reducing unnecessary examination of routine expenditure transactions. This framework may also support continuous auditing by enabling systematic anomaly screening before manual verification, thereby improving the effectiveness of public financial oversight.

In practice, adopting this hybrid methodology delivers several advantages. First, it reduces the dependency on manual inspection by automatically identifying high-risk transactions, thereby reducing the risk of human error. Second, it improves interpretability by providing a structured transition from anomaly detection to fraud classification. Finally, it enhances decision-making by combining statistical detection with domain-informed evaluation.

Although the proposed framework achieved excellent classification performance, several limitations should be acknowledged. First, the dataset was obtained from only one Indonesian government institution. Therefore, the findings should be interpreted as evidence within the studied institutional context rather than as a general representation of all government financial transactions. Second, fraud labels were generated through expert consensus without calculating formal inter-rater agreement statistics such as Fleiss' Kappa. Although consensus was achieved, future studies should

Table 4. Comparative Results and Ablation Analysis

Model	Accuracy	Precision	Recall	F1 Score	FP	FN
XGBoost	0.9984	0.98	1.00	0.99	3	0
Random Forest	0.9984	0.98	1.00	0.99	3	0
SVM	0.9972	0.97	1.00	0.99	5	0
*XGBoost (no DIF)**	0.9891	0.91	0.94	0.92	18	11

**Ablation: XGBoost trained on original transaction features only, without DIF-derived Anomaly and Anomaly_Score features.*

incorporate quantitative agreement measures to further evaluate labeling consistency.

Another limitation of this study is that explainability analysis was not included. The primary objective of this research was to evaluate the predictive performance of the proposed hybrid fraud detection framework by integrating Deep Isolation Forest, expert validation, and XGBoost. Consequently, model interpretation techniques such as SHapley Additive exPlanations (SHAP) were beyond the scope of the present study. Nevertheless, explainability is particularly important in government auditing because auditors require transparent and justifiable reasons for fraud predictions before using them to support audit decisions. Future research should therefore incorporate SHAP or other explainable artificial intelligence techniques to improve model transparency, facilitate auditor interpretation, and strengthen practical implementation in public-sector financial auditing.

Overall, the results demonstrate that integrating anomaly detection, expert interpretation, and classification provides a more comprehensive and effective approach to fraud detection compared to using individual methods independently (Wibowo & Wibowo, 2025).

The near-perfect performance shown in Table 2 should be interpreted with appropriate caution, and several structural factors that may partially explain these results merit explicit discussion here. First, the anomaly detection stage enriches the original feature space by introducing anomaly-related information (DIF anomaly score and binary anomaly label) that already provides strong discriminatory signals between fraudulent and non-fraudulent transactions; this means that XGBoost operates on a pre-filtered, information-rich feature space rather than on raw transaction data alone. Second, the dataset originates from a single institutional environment with relatively homogeneous transaction characteristics and a single fiscal period, which may increase class separability beyond what would be observed in a more diverse dataset. Third, because expert labelling was applied only to anomalous transactions flagged by DIF, there is an inherent coupling between the anomaly features and the fraud labels—non-anomalous transactions were automatically assigned a non-fraud label without expert review, which may amplify the predictive signal of the Anomaly and Anomaly_Score features. This circularity does not invalidate the results within this institutional context, but it does mean that the 0.9984 accuracy figure cannot be interpreted as a measure of absolute fraud detection capability independent of the DIF screening stage. Therefore, while the results demonstrate the effectiveness of the proposed framework within the studied environment, external validation using data from multiple institutions is necessary before claiming general applicability.

Conclusion

The results presented in this section reflect the implementation of the proposed two-stage framework, where anomaly detection outputs are refined through expert interpretation before being used in the classification stage. This work demonstrates that integrating anomaly detection and classification provides a more effective approach for fraud detection in financial transaction data. Combining Deep Isolation Forest (DIF) and XGBoost within a two-stage framework makes the proposed method able to identify irregular patterns and refine them into meaningful fraud predictions.

The findings indicate that the anomaly detection stage plays a critical role in narrowing down the search space by identifying transactions that deviate from normal behavior. Rather than directly labeling these anomalies as fraud, the inclusion of expert-based interpretation allows the framework to distinguish between genuine irregularities and legitimate but uncommon transactions. This step is essential in reducing misclassification and improving the reliability of the detection process.

In the classification stage, the integration of anomaly-based features significantly enhances the performance of the XGBoost model. The high accuracy, precision, and recall observed in the results indicate that the model is capable of effectively identifying fraudulent transactions while maintaining a low rate of false positives. The most important thing is that the results suggest that the improvement in performance is not only due to the classification algorithm, but also due to the contribution of the anomaly detection stage in enriching the feature space.

The proposed hybrid framework offers a more comprehensive solution than conventional approaches that rely on either anomaly detection or classification. The sequence between anomaly detection and classification in the model enables it to identify potential irregularities and then make more informed decisions. This process improves both detection accuracy and interpretability, therefore making the framework more suitable for practical implementation in financial systems.

From a practical perspective, the proposed framework has the potential to support decision-making processes in financial monitoring and auditing by prioritizing high-risk transactions for further investigation. This can reduce the workload of manual inspection and improve the efficiency of fraud detection systems.

However, the findings of this study should be interpreted with caution. The high performance achieved may be influenced by the characteristics of the dataset, including potential data separability and limited variability. Therefore, the generalizability of the proposed framework to other financial environments remains an important area for further investigation.

At least three specific directions merit future investigation. First, multi-institution validation: applying the framework to transaction data from multiple government ministries or sub-national agencies would allow a direct assessment of generalizability and reveal whether the DIF threshold and XGBoost hyperparameters transfer across institutional contexts without retraining. Second, ablation analysis: a controlled comparison of XGBoost trained on original transaction features alone versus the full enriched feature set (with Anomaly and Anomaly_Score) would isolate the marginal contribution of the DIF screening stage to classification performance, providing stronger empirical evidence for the hybrid architecture. Third, explainability integration: incorporating SHAP (SHapley Additive exPlanations) value analysis would allow auditors to inspect which specific features drive each fraud prediction, directly supporting regulatory accountability requirements and making the framework more practically deployable in government audit workflows where decision transparency is legally required.

Overall, the proposed hybrid framework provides a structured, interpretable, and effective approach to fraud detection, particularly in complex financial transaction environments.

Author contributions

This work was supported by Financial Education and Training Agency, Ministry of Finance of Indonesia for granting access to financial transaction data used in this research. The data were provided solely for research purpose and remain subject to institutional confidentiality and disclosure restrictions.

References

Law Number 1 of 2004 Concerning State Treasury, Pub. L. 1 (2004). https://peraturan.bpk.go.id/

Ali, A. Al, Khedr, A. M., El-Bannany, M., & Kanakkayil, S. (2023). A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique. *Applied Sciences (Switzerland)*, *13*(4). https://doi.org/10.3390/app13042272

Andriyanto, R., & Januarti, I. (2026). Implementing Artificial Intelligence in Auditing: A Systematic Literature Review of Trends, Challenges, and Adoption. *Owner*, *10*(2), 1925–1937. https://doi.org/10.33395/owner.v10i2.3389

Bakumenko, A., & Elragal, A. (2022). Detecting Anomalies in Financial Data Using Machine Learning Algorithms. *Systems*, *10*(5). https://doi.org/10.3390/systems10050130

Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, *557*(May), 317–331. https://doi.org/10.1016/j.ins.2019.05.042

Carletti, M., Terzi, M., & Susto, G. A. (2023). Interpretable Anomaly Detection with DIFFI: Depth-based feature importance of Isolation Forest. *Engineering Applications of Artificial Intelligence*, *119*. https://doi.org/10.1016/j.engappai.2022.105730

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. *2019 IEEE International Conference on Vehicular Electronics and Safety, ICVES 2019*, (September). https://doi.org/10.1109/ICVES.2019.8906396

Faccia, A. (2026). Explainable AI (XAI) in Auditing: Bridging the Gap Between Predictive Fraud Models and Regulatory Standards. *Journal of Risk and Financial Management*, *19*(5). https://doi.org/10.3390/jrfm19050311

Guan, H., Li, S., Wang, Q., Lyulyov, O., & Pimonenko, T. (2022). Financial Fraud Identification of the Companies Based on the Logistic Regression Model. *Journal of Competitiveness*, *14*(4), 155–171. https://doi.org/10.7441/joc.2022.04.09

Hajek, P., Abedin, M. Z., & Sivarajah, U. (2023). Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework. *Information Systems Frontiers*, *25*(5), 1985–2003. https://doi.org/10.1007/s10796-022-10346-6

Hasibuan, A. A., Irsan Nasution, M., & Nasution, M. I. (2026). *id 2 2nd International Conference on Islamic Community Studies (ICICS) Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era The Impact of Real-Time and Continuous Auditing on Financial Transparency and Fraud Detection: A Systematic Literature Review*. https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Indonesian Institute of Certified Public Accountants (IAPI). (2021). *SA 240 (Revised 2021): The Auditor's Responsibilities Relating to Fraud in an Audit of Financial Statements*. https://iapi.or.id/

Minister of Finance Regulation Number 62/PMK.01/2023 Concerning Budget Planning, Budget Implementation, Accounting and Financial Reporting, Pub. L. 62/PMK.01/2023 (2023). https://jdih.kemenkeu.go.id/

Li, X. (2023). Financial Fraud: Identifying Corporate Tax Report Fraud Under the Xgboost Algorithm. *EAI Endorsed Transactions on Scalable Information Systems*, *10*(4), 1–7. https://doi.org/10.4108/eetsis.v10i3.3033

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

Milka Wijayanti Sunarto, Dendy Kurniawan, Edy Siswanto, & Haris Ihsanil Huda. (2023). Deteksi Anomali Menggunakan Extended Isolation Forest (Eif). *Teknik: Jurnal Ilmu Teknik Dan Informatika*, *1*(2), 96–111. https://doi.org/10.51903/teknik.v1i2.324

Moomtaheen, F., Bagui, S. S., Bagui, S. C., & Mink, D. (2024). Extended Isolation Forest for Intrusion Detection in Zeek Data. *Information (Switzerland)*, *15*(7). https://doi.org/10.3390/info15070404

Government Regulation Number 60 of 2008 Concerning Government Internal Control System (2008). https://peraturan.bpk.go.id/

Said Mirza Pahlevi. (2024). *Kecerdasan Buatan dengan Deep Learning* (Vol. 1). Elex Media Komputindo.

Saifudin, S., Januarti, I., & Purwanto, A. (2025). The Role of Artificial Intelligence in the Audit Process and How to Fraud Detections: A Literature Outlook. *Journal of Ecohumanism*, *4*(1), 4185–4203. https://doi.org/10.62754/joe.v4i1.6301

Sodnomdavaa, T., & Lkhagvadorj, G. (2026). Financial Statement Fraud Detection Through an Integrated Machine Learning and Explainable AI Framework. *Journal of Risk and Financial Management*, *19*(1), 13. https://doi.org/10.3390/jrfm19010013

Sopiyan, M., Fauziah, F., & Wijaya, Y. F. (2022). Fraud Detection Using Random Forest Classifier, Logistic Regression, and Gradient Boosting Classifier Algorithms on Credit Cards. *JUITA: Jurnal Informatika*, *10*(1), 77. https://doi.org/10.30595/juita.v10i1.12050

Suci Azzahra. (2025). The Use of Artificial Intelligence (AI) in Fraud Detection by Government Auditors: A Systematic Literature Review. *International Conference Proceedings Universitas Islam Blitar*, (1).

Taha, A. A., & Malebary, S. J. (2020). An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. *IEEE Access*, *8*, 25579–25587. https://doi.org/10.1109/ACCESS.2020.2971354

Thanathamathee, P., Sawangarreerak, S., Chantamunee, S., & Nizam, D. N. M. (2024). SHAP-Instance Weighted and Anchor Explainable AI: Enhancing XGBoost for Financial Fraud Detection. *Emerging Science Journal*, *8*(6), 2404–2430. https://doi.org/10.28991/ESJ-2024-08-06-016

Wana, D., & Rohman, A. (2025). Fraud Detection in Government in the Last Ten Years. In *Journal of Applied Accounting and Taxation Article History* (Vol. 10, Number 1).

Wang, L. (2024). Financial Fraud Identification Model of Listed Companies based on Time-Series Information. *WSEAS Transactions on Computer Research*, *12*, 211–223. https://doi.org/10.37394/232018.2024.12.21

Wibowo, S. R., & Wibowo, A. (2025). Continous Auditing and Explainable AI for Enhancing Real Time Financial Analysis. *Jurnal Akuntansi Dan Bisnis*, *5*(2), 10–19. https://doi.org/10.51903/jiab.v5i2.931

Xiang, H., Hu, H., & Zhang, X. (2022). DeepiForest: A Deep Anomaly Detection Framework with Hashing Based Isolation Forest. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2022-Novem*(Icdm), 1251–1256. https://doi.org/10.1109/ICDM54844.2022.00163

Xu, H., Pang, G., Wang, Y., & Wang, Y. (2023). Deep Isolation Forest for Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12591–12604. https://doi.org/10.1109/TKDE.2023.3270293

Zulfikar, A., Rahmani, Fa. A., & Azizah, N. (2023). Deteksi Anomali Menggunakan Isolation Forest Belanja Barang Persediaan Konsumsi Pada Satuan Kerja Kepolisian Republik Indonesia. *Jurnal Manajemen Perbendaharaan*, *4*(1), 1–15. https://doi.org/10.33105/jmp.v4i1.435