Perbanas International Conference on Economics, Business, Management, Accounting and IT (PROFICIENT) 2026
"Integrity-Driven Innovation: Reimagining Business, Finance, and Digital Transformation for Sustainable Impact"

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Application of Random Forest Algorithm to Detect Financial Transaction Fraud Based on Machine Learning

Sri Wahyuni Aprianti[1], Yoga Armando[2], Mercurius Broto Legowo[3]

[1,2,3] Data Science Study Program, Faculty of Information Technology, Perbanas Institute

[a)]Corresponding author : sri.wahyuni06@perbanas.id,
[b)]yoga.armando08@perbanas.id,
[c)]mercurius@perbanas.id

Abstract - *The increasing number of digital financial transactions through online banking, e-wallets, and e-commerce platforms has provided convenience but also introduced new risks in the form of financial fraud. As the volume and complexity of transaction data continue to grow, fraud patterns become more diverse and increasingly difficult to detect manually. One of the main challenges in fraud detection is the limitation of conventional systems that are not able to identify suspicious transactions quickly and accurately. This issue occurs due to the imbalance between normal transaction data and fraudulent transaction data, as well as continuously evolving fraud patterns. As a result, manual detection processes become difficult and may lead to financial losses for financial institutions and service users. This study proposes the application of the Random Forest algorithm based on Machine Learning to detect financial transactions that are potentially fraudulent. The Random Forest algorithm is selected because it can handle large datasets, reduce the risk of overfitting, and provide stable classification performance. The objective of this research is to detect financial transaction fraud using the Random Forest method. The expected outcome is a model that can effectively identify suspicious transactions with minimal errors and improve the security of digital financial transactions.*

Keywords: Financial Fraud Detection, Random Forest, Machine Learning, Digital Financial Transactions, Fraud Detection System

INTRODUCTION

The rapid growth of digital financial transactions through online banking, e-wallets, and e-commerce platforms has transformed the way people conduct financial activities. According to the, digital payment systems continue to expand globally, reflecting a shift toward a cashless society. In addition, the (Edeh & Raj, 2026) highlights the increasing adoption of digital financial services as part of global financial inclusion efforts. One of the main challenges in digital financial systems is the increasing complexity of fraud patterns. Recent studies indicate that data imbalance significantly affects model performance and increases the risk of misclassification (Zaman & Chamidy, 2024). One of the most widely used algorithms is Random Forest, which is capable of handling large datasets and identifying complex patterns. Recent research shows that machine learning methods, particularly ensemble approaches, provide strong performance in handling complex and imbalanced datasets (Sarker, 2021).

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Several recent studies have demonstrated that Random Forest performs well in fraud detection. Research by (Zaman & Chamidy, 2024) shows that machine learning approaches, particularly ensemble methods, are effective in detecting fraudulent transactions. In addition, (Sarker, 2021) highlights the advantages of machine learning in improving fraud detection accuracy. Recent research by (Khalid et al., 2024) indicates that ensemble-based machine learning algorithms remain a reliable approach for detecting fraud patterns in financial datasets.

LITERATURE RIVIEW

Financial Transaction Fraud

Financial transaction fraud refers to illegal activities aimed at obtaining unauthorized financial benefits through system manipulation or misuse of data in digital transactions. Recent studies indicate that in many financial systems, fraudulent transactions account for less than 1% of total data, posing significant challenges in training detection (Agara et al., 2026).

Machine Learning in Fraud Detection

Machine Learning (ML) is widely used in fraud detection because it can analyze large amounts of data and identify complex patterns. ML approaches also help improve fraud detection accuracy and address challenges such as data imbalance and complex financial data (Compagnino et al., 2025).

Random Forest Algorithm

Random Forest is an ensemble learning algorithm widely used in fraud detection due to its ability to produce stable and accurate predictions. Recent studies indicate that ensemble-based algorithms, including Random Forest, demonstrate superior performance in detecting complex fraud patterns and can improve classification accuracy in financial transaction data (Boulieris et al., 2024).

Data Preprocessing and Feature Engineering

Data preprocessing is a crucial step in improving data quality before applying Machine Learning models. Research shows that the quality of preprocessing and feature engineering significantly affects model performance, especially in fraud detection cases involving complex and imbalanced data. Without proper preprocessing, models tend to produce less accurate predictions (Anant et al., 2025).

Model Evaluation in Fraud Detection

Model evaluation in fraud detection should use metrics such as precision, recall, and F1-score because fraud datasets are highly imbalanced. Recall is especially important to measure the model's ability to detect fraudulent transactions and reduce financial losses. Recent studies also emphasize the importance of appropriate evaluation metrics to avoid bias in fraud detection systems (Kazeem, n.d.).

Previous Studies

Various recent studies indicate that the use of Machine Learning, particularly ensemble methods such as Random Forest, provides effective results in detecting financial transaction fraud. Additionally, studies highlight

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

that the main challenges in fraud detection include limited data availability, evolving fraud patterns, and the need for adaptive and real-time detection systems (Boulieris et al., 2024).

Literature Synthesis

Based on the reviewed literature, financial transaction fraud detection is a complex problem that requires advanced Machine Learning approaches. The Random Forest algorithm has proven effective in handling large, imbalanced datasets and capturing non-linear patterns. The success of the model depends on the quality of data preprocessing, feature engineering, and appropriate evaluation metrics.

Conceptual Framework

The conceptual framework of this study describes the process of detecting financial transaction fraud using the Machine Learning approach with the Random Forest algorithm. Finally, the model performance is evaluated using accuracy, precision, recall, and F1-score metrics to measure the effectiveness of fraud detection.

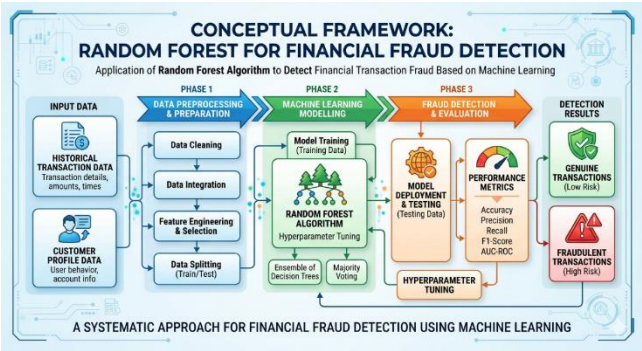

FIGURE 1. Conceptual Framework of Fraud Detection Using Random Forest

As shown in Figure 1, the conceptual framework illustrates the relationship between input data, data processing stages, feature engineering, model development, and evaluation. It provides a structured overview of how raw transaction data is transformed into meaningful insights through a machine learning pipeline, applied algorithm, and the expected outcomes in detecting fraudulent transactions accurately and efficiently.

RESEARCH METHODOLOGY

Research Design

This study adopts a quantitative research approach by utilizing Machine Learning techniques to classify financial transactions into fraud and non-fraud categories. Recent studies indicate that Machine Learning approaches significantly improve fraud detection performance compared to traditional rule-based systems, particularly in handling large and complex financial data (Credit et al., 2022)(Dastidar et al., 2024).

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Dataset

The dataset used in this study consists of financial transaction records that include attributes such as transaction amount, transaction time, user behavior, and other relevant variables. Previous research indicates that fraud datasets often contain less than 1% fraudulent transactions, making classification tasks more complex (Dastidar et al., 2024).

Data Preprocessing

Data preprocessing is a crucial step in preparing raw data for Machine Learning modeling. Effective preprocessing has been shown to significantly enhance the accuracy and robustness of fraud detection models (Btoush et al., 2026).

Feature Engineering

Feature engineering is the process of transforming raw data into meaningful features that can improve model performance. In this study, new features are generated based on transaction patterns and user behavior. These features include transaction frequency within a specific time frame, average transaction amount, time intervals between transactions, and behavioral patterns of users.

Model Implementation

Feature engineering aims to provide additional information that helps the model better distinguish between fraudulent and legitimate transactions. Research indicates that feature engineering plays a critical role in improving classification accuracy, especially in complex datasets such as financial transactions (Raymond et al., 2024). Recent studies highlight that ensemble learning methods, including Random Forest, consistently achieve high performance in detecting fraudulent transactions (Boulieris et al., 2024).

Evaluation Metrics

Model evaluation is conducted to assess the effectiveness of the proposed fraud detection system. Due to the nature of fraud detection problems, relying solely on accuracy can be misleading. Therefore, recall and F1-score are considered more important in evaluating model performance. Recent research emphasizes the importance of using multiple evaluation metrics to ensure reliable and unbiased model assessment (Sujon et al., 2025).

RESULTS AND DISCUSSION

System Implementation Overview

The fraud detection system is built using the Python programming language on the Anaconda and Jupyter Notebook ecosystems. The main libraries used are Pandas and NumPy for data manipulation, Scikit-learn for machine learning modeling, and Matplotlib and Seaborn for visual analysis.

TABLE 1. Libraries Used in the Research

Library	Function
Pandas	Data manipulation and preprocessing
NumPy	Numerical Computation

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Matplotlib	Data visualization
Seaborn	Visualization enhancement
Scikitlearn	Mechine learning modeling and evaluation

Dataset Input and Exploration

The transaction dataset in Excel format was imported using Pandas' built-in functions to examine its initial structure. Exploration using the head(), info(), and describe() methods was carried out to map the characteristics of the attributes such as data types, descriptive statistics, and to detect the presence of missing data.

TABLE 2. Initial Dataset Exploration Process

Process	Purpose
head()	Display the first rows of the dataset
info()	Display dataset structure and data types
describe()	Display statistical summary
isnull().sum()	Check missing values

Based on the exploratory data visualization, several key columns were identified such as transaction amount, merchant category, device type, authentication method, and fraud label that are ready to be sent to the cleaning stage.

FIGURE 2. Distribution of Fraud and Non-Fraud Transactions

As shown in figure 2, the distribution of transaction classes within the dataset. The visualization shows that non-fraudulent transactions dominate the dataset, while fraudulent transactions represent a significantly smaller proportion. This imbalance condition is common in fraud detection research and justifies the use of the class_weight='balanced' parameter in the Random Forest model to reduce classification bias toward the majority class.

Data Cleaning, Feature Selection, and Splitting

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Before the modeling stage, data cleaning was performed to improve dataset quality and ensure reliable machine learning performance. This process included duplicate checking, duplicate removal, and handling missing values to maintain data consistency throughout the analysis process.

TABLE 3. Data Cleaning Activities

Activity	Method Used
Duplicate Checking	duplicated()
Duplicate Removal	drop_duplicates()
Missing Value Handling	SimpleImputer(strategy='most_frequent')

Based on the cleaning process, no significant redundancy was found in the dataset. Missing values were successfully handled using the most frequent value approach, allowing the dataset to be prepared effectively for the feature selection and model training stages.

Feature Selection

Feature selection was conducted to retain only relevant attributes for fraud prediction. Unique identifiers such as user ID and transaction ID were removed because they do not contribute meaningful predictive information and may potentially cause data leakage. The remaining features, including transaction-related and behavioral attributes, were used as input variables for the Random Forest model.

Data Splitting

The dataset was divided into training and testing sets using the `train_test_split` function from Scikit-learn. A split ratio of 70:30 was applied, where 70% of the data was used for training and 30% for model evaluation. The parameter `random_state=42` was implemented to ensure experiment reproducibility and consistent results across multiple runs.

Random Forest Model Training and Prediction

The model is initialized with tree depth parameter restrictions to prevent the model from memorizing patterns from the training data (overfitting). The parameters are configured specifically: n_estimators=100, max_depth=10, min_samples_split=10, min_samples_leaf=5, and the weight balancer class_weight='balanced'. After the training is completed with the model.fit() command, 30% of the test data is used to obtain a binary classification prediction label.

Random Forest Feature Importance

Before evaluating the overall performance of the model, feature importance analysis was conducted to identify which attributes contributed the most to fraud detection. The Random Forest algorithm automatically measures the influence of each feature during the tree-building process, allowing important variables to be ranked based on their predictive contribution.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

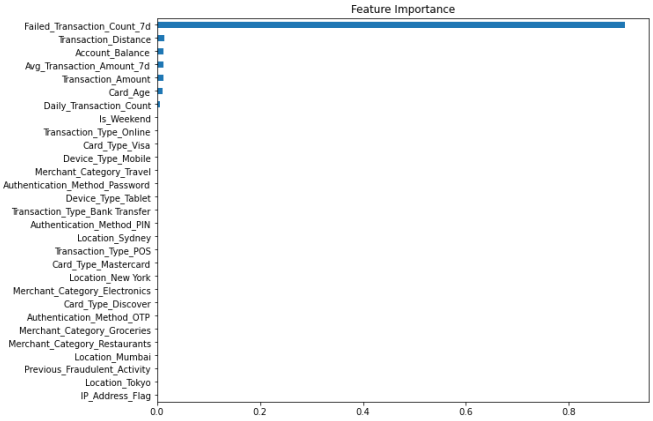

FIGURE 3. Feature Importance

After the training process, several features were identified as having strong influence on fraud classification results. Transaction amount, authentication method, device type, and merchant category showed the highest importance scores, indicating that transaction behavior patterns play a significant role in distinguishing fraudulent and legitimate transactions.

Model Evaluation Results

At this stage, the model performance was evaluated using several classification metrics including accuracy, precision, recall, and F1-score.The following code snippet shows the evaluation process.

TABLE 4. Model Evaluation Results

Metric	Value
Accuracy	0.87772
Precission	1.0000
Recall	0.6178
F1-score	0.7638

Confusion Matrix

To further analyze the classification results, a confusion matrix was generated using the testing dataset. This visualization helps evaluate how well the model distinguishes between fraudulent and legitimate transactions by comparing actual and predicted classes.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

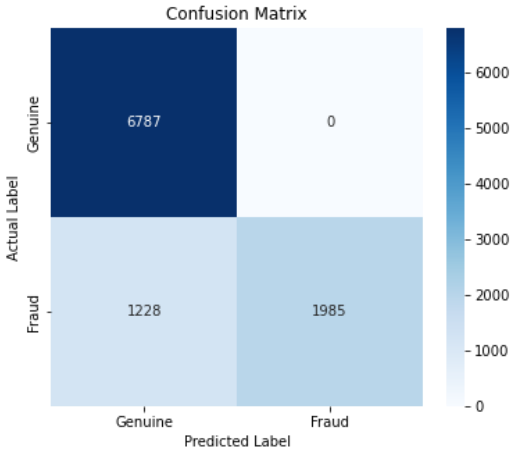

FIGURE 4. Confusion Matrix of Fraud Detection Results

As shown in figure 4, the confusion matrix shows that the Random Forest model successfully classified most legitimate transactions correctly while still detecting a substantial number of fraud cases. Although several fraudulent transactions were not identified, the model demonstrated strong capability in minimizing false positive predictions, which is important in fraud detection systems.

Classification Report

To provide a more comprehensive evaluation of the Random Forest model, a classification report visualization was generated. This visualization summarizes the main evaluation metrics, including accuracy, precision, recall, and F1-score, to measure the effectiveness of the fraud detection model.

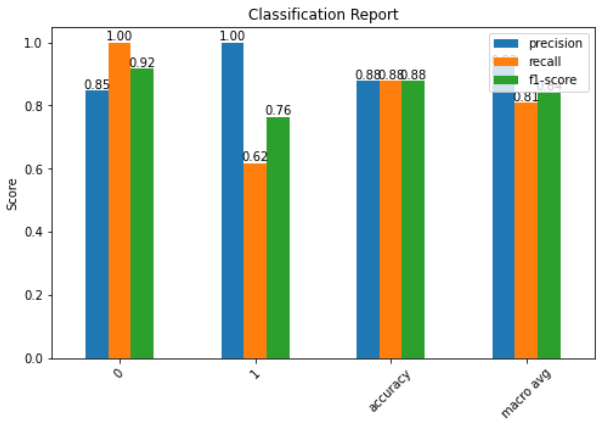

FIGURE 5. Classification Report

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

As shown in figure 4, the classification report shows that the Random Forest model achieved strong overall performance with an accuracy value of 87.77%. The precision score of 1.0000 indicates that all predicted fraud transactions were classified correctly, while the recall value suggests that several fraud cases were still undetected. These results indicate that the model performs reliably in minimizing false positive predictions despite the challenges of imbalanced fraud transaction data.

Discussion

The results of this study demonstrate that the Random Forest algorithm can be effectively applied to detect fraudulent financial transactions within a machine learning-based classification system. The model achieved an accuracy of 87.72%, indicating strong capability in classifying both fraudulent and genuine transaction patterns. Feature importance analysis further revealed that behavioral transaction features, such as transaction frequency, failed transactions, and abnormal account behavior, contributed significantly to the fraud detection process.

Future research may improve performance by applying resampling techniques such as SMOTE, hyperparameter optimization, and comparisons with algorithms such as XGBoost, LightGBM, or Neural Network using real-world transaction datasets. Overall, the findings confirm that Random Forest is a reliable algorithm for fraud detection due to its strong classification capability and ability to minimize false positive predictions.

CONCLUSIONS

This study successfully implemented the Random Forest algorithm to detect fraudulent financial transactions using a machine learning approach. The research process included dataset exploration, preprocessing, feature engineering, model training, prediction, and evaluation using Python libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn. The evaluation results showed strong classification performance with an accuracy of 87.72%, precision of 1.0000, recall of 0.6178, and F1-score of 0.7638. These findings indicate that the model could classify most transactions correctly and produce highly reliable fraud predictions with very low false positive rates. However, some fraud cases were still not detected due to the imbalanced dataset, where fraudulent transactions represented only a small portion of the data.

The confusion matrix and classification report confirmed that the model performed very well in identifying genuine transactions while maintaining good fraud detection capability. Feature importance analysis also revealed that transaction behavior variables, such as transaction activity patterns, failed transactions, and abnormal account behavior, significantly contributed to fraud detection performance. In addition, preprocessing and feature engineering stages played an important role in improving the stability and reliability of the model. Overall, the findings confirm that Random Forest is an effective and reliable approach for detecting financial transaction fraud in digital financial systems. For future research, it is recommended to use real-world transaction datasets, apply balancing techniques such as SMOTE, and compare Random Forest with other advanced machine learning algorithms to further improve fraud detection performance.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

REFERENCES

Agara, D. T., Omotoso, S. S., & Alabi, O. J. (2026). *Machine Learning Models for Multi Sector Fraud Detection in Public and Private Financial Systems*. *Ml*.

Anant, P., Shweta, G., Ghandat, S., & Ghorpade, S. (2025). *Fraud Detection Using Machine Learning*. *11*(05), 184–190.

Boulieris, P., Pavlopoulos, J., Xenos, A., & Vassalos, V. (2024). Fraud detection with natural language processing. *Machine Learning*, *113*(8), 5087–5108. https://doi.org/10.1007/s10994-023-06354-5

Btoush, E., Kobbaey, T., & Tamimi, H. (2026). *Machine Learning-Based Cyber Fraud Detection : A Comparative Study of Resampling Methods for Imbalanced Credit Card Data*. 1–34.

Compagnino, A. A., Maruccia, Y., Cavuoti, S., Riccio, G., Tutone, A., Crupi, R., & Pagliaro, A. (2025). *An Introduction to Machine Learning Methods for Fraud Detection*. *Ml*, 1–32.

Credit, E., Fraud, C., & Model, D. (2022). *Enhanced Credit Card Fraud Detection Model Using Machine Learning*.

Dastidar, K. G., Caelen, O., & Granitzer, M. (2024). Machine Learning Methods for Credit Card Fraud Detection : A Survey. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2024.3487298

Edeh, M., & Raj, A. (2026). *Digital Connectivity and Financial Inclusion : An Empirical Cross-Country Analysis Using the Global Findex 2025 Database Digital Connectivity and Financial Inclusion : An Empirical Cross-Country Analysis Using the Global Findex 2025 Database*. 0–21. https://doi.org/10.20944/preprints202603.2167.v1

Kazeem, O. (n.d.). *FRAUD DETECTION USING MACHINE*.

Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., & Adejoh, J. (2024). *Enhancing Credit Card Fraud Detection : An Ensemble Machine Learning Approach*.

Raymond, J., Joseph, O., Joseph, S., & Iseal, S. (2024). *Financial Fraud Detection Feature Engineering Techniques for Enhanced*. *Ml*.

Sarker, I. H. (2021). Machine Learning : Algorithms , Real - World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Sujon, K. M., Hassan, R., Choi, K., & Samad, A. (2025). *empirical evidence from advanced statistics , ML , and XAI for evaluating business predictive models*.

Zaman, S., & Chamidy, T. (2024). Bibliometric Analysis and Visualization of Machine Learning-Based Credit Card Fraud Detection. *2024 International Conference on Information Technology Research and Innovation (ICITRI)*, 236–241. https://doi.org/10.1109/ICITRI62858.2024.10699070