The Application of Machine Learning in Financial Fraud Analysis

Binyang Xu

University of Hong Kong, Pok Fu Lam Road 999077, Hong Kong
xby0266@connect.hku.hk

Abstract. Against the backdrop of the increasingly severe problem of financial fraud, this study is dedicated to constructing an efficient fraud detection model. First, exploratory data analysis is employed to analyze financial data. Analyses of univariate and multivariate variables are carried out to gain insights into the distribution and correlation characteristics of different features. Subsequently, data preprocessing is performed on continuous and categorical variables to improve data quality and usability. On this basis, machine learning models such as logistic regression, decision trees, random forests, and Support Vector Machine (SVM) are utilized for modeling. After the model training is completed, a series of model metrics, such as accuracy, recall, and F1-score, are used to comprehensively evaluate the prediction performance of the models. The experimental results demonstrate that different models exhibit varying advantages and limitations in the task of financial fraud prediction, providing important references for subsequent model optimization and practical applications. The findings of this study are expected to assist financial institutions in enhancing their fraud prevention capabilities and effectively reducing losses caused by fraud risks.

Keywords: Machine Learning, Decision Tree, Random Forest, SVM, Recall

1 Introduction

The escalating severity of financial fraud has posed critical threats to the global financial system, urging the need for innovative detection approaches. With digital transactions (usually referring to the total amount, scale, or number of transactions) growing at an annual rate of 25%, traditional rule-based methods—such as fixed-threshold determination and expert experience rules—have become inadequate to address the dynamic complexity of fraudulent patterns, including synthetic identity fraud and cross-border transaction anomalies. The mutation cycle of such new-type frauds has been shortened to within 45 days, highlighting the urgency of dynamic modeling with machine learning.

Previous studies have laid the foundation for the application of machine learning in fraud detection. Tarique Ameer et al emphasized that Exploratory Data Analysis (EDA) can reveal critical associations between features [1]. Nishikanta Mohanty et al validated the effectiveness of the SMOTE technique in balancing imbalanced datasets [2]. In their experiments, after applying the SMOTE technique to different amounts of data, the F1-

A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_21

score of fraud prediction tasks continuously increased with the rise in data percentage, eventually increasing from the initial 57.5% without SMOTE processing to 83.5%, which significantly enhanced the ability to detect rare fraudulent events. Bouhannache Mohammed et al demonstrated that machine learning models such as random forests can perform excellently in high-dimensional financial data [3]. Their research showed that due to the ensemble learning capability of random forests, which enables them to capture complex patterns, the performance on the test set can reach 0.91, indicating a good fit of the model. Jiwon Chung et al highlighted that recall is a core metric in fraud detection [4]. They pointed out that an extremely high recall of at least 0.9362 was achieved in four tested fraud datasets, which allowed the corresponding models to identify fraudulent credit cards efficiently and accurately. However, existing studies remain limited in optimizing hyperparameters for real-time applications and integrating deep learning to address evolving fraud schemes.

Kaggle's IEEE-CIS Fraud Detection dataset has three key traits: marked class imbalance (needing SMOTE), 101 diverse variables, and prominent nonlinear feature associations, making it suitable for machine learning (ML) models. This study employs exploratory data analysis on the distributions of TransactionDT and TransactionAmt, data preprocessing with median and mode for different variables, and SMOTE for balancing the 8:2 training-test dataset. Four machine learning models—logistic, decision trees, random forests, and Support Vector Machine (SVM) —are evaluated using accuracy, recall, and F1-score to construct a machine learning model, particularly by leveraging random forests' feature importance ranking to provide actionable insights for financial institutions. The research aims to enhance practical fraud prevention capabilities and reduce fraudulent losses.

2 Methodology

2.1 Dataset

The IEEE - CIS Fraud Detection dataset on Kaggle originates from a fraud detection competition jointly held by the IEEE Computational Intelligence Society (IEEE - CIS) and Vesta Corporation, a leading payment service company [5]. The competition aims to enhance the accuracy of fraud prevention systems. After preprocessing, this dataset has 101 dimensions, containing various types of variables, including TDT (transaction time), TA (TA), and so on.

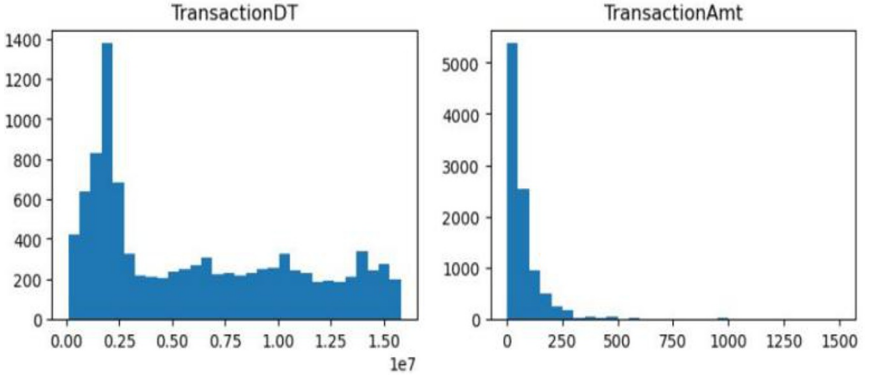
2.2 Exploratory Data Analysis (EDA)

Univariate Analysis and Bivariate Analysis. The univariate analysis here focuses on the distribution characteristics of individual variables, while the bivariate analysis explores the associations between variables and the target variable.

Multivariate Analysis. In the multivariate analysis, the correlation coefficient matrix is obtained by calculating the correlations between variables, and visualized using a correlation heatmap.

In the univariate analysis, the study focuses on two continuous variables, Transaction Date Time (TDT) and Transaction Amount (TA). Transaction ID, as a unique identifier for transactions, is merely used to distinguish different transaction events, has no practical analytical significance, and will not have a significant impact on the experimental results, so it is not included in the in-depth research scope. TDT measures the time interval of transactions in seconds, and TA represents the TA; both are continuous variables.

Fig. 1 (a) shows that the data distribution of TDT is right-skewed. Specifically, its frequency reaches a peak at approximately 0.25 seconds, and after exceeding this value, it gradually decreases and becomes stable. For TA, Fig. 1 (b) clearly shows that most data points are concentrated in the range of 0 to 250, with a particularly dense distribution between 0 and 100. Only a small number of data points are distributed around 1000, which can be regarded as outliers. It is worth noting that in the context of fraud detection, such outliers may contain key clues for identifying fraudulent behaviors. Because fraudsters sometimes use large transactions to quickly obtain illegal benefits or adopt special amount settings to evade detection, the existence of these outliers cannot be ignored [1, 6].

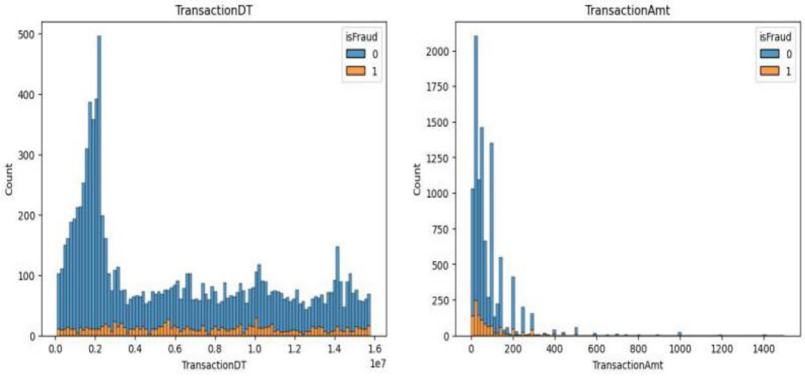


(a) (b)

Fig. 1. The distribution of variable, (a) The distribution of TDT; (b) The distribution of TA (Photo/Picture credit: Original).

Regarding TDT, Fig. 2 reveals significant differences in the distribution between the two different fraud status categories. In non-fraud cases, the distribution is right-skewed; while in fraud cases, the distribution is relatively uniform. This phenomenon indicates that there may be an association between TDT and the possibility of fraud occurrence. For example, fraudulent behaviors are more likely to occur under certain specific time distribution patterns, which provides a noteworthy feature for the subsequent fraud detection model. However, when comparing the distribution plots of TA, no obvious differences are observed between the two groups (fraud and non-fraud). It can thus be inferred that, compared with TDT, the impact of TA on the probability of

fraud may not be significant. But further observation of the charts finds that lower TAs may be associated with an increase in the probability of fraudulent activities.



(a) (b)

Fig. 2. The distribution of variables in fraud or not, (a) The distribution of TDT; (b) The distribution of TA (Photo/Picture credit: Original).

From the Fig. 3, it can be observed that except for TA, Card1, Card2, and feature C1 (Count-based features, whose specific business meanings have not been disclosed due to privacy or commercial confidentiality considerations, same as C2 and so on), which have strong correlations with other count-based variables, most other variables have no significant associations. This suggests that there may be information redundancy among these variables, characterized by strong correlations. Consequently, in the subsequent modeling process, dimensionality reduction or feature selection may be necessary to enhance the model's efficiency and accuracy. In addition, it is particularly worth noting here that variables with relatively high correlation with the target variable "isFraud"; Fig. 3 shows that it has a strong relationship with, for example, "TDT". This further verifies the conclusion found in the bivariate analysis that TDT is related to the possibility of fraud.

2.3 Data Preprocessing

Here, data preprocessing is specifically carried out in three parts. Firstly, the dataset contains 101 variables, covering integer, string, and logical types, with a large number of empty strings and "NA" values. To facilitate subsequent research, empty strings are first converted to "NA", and column vectors where missing values exceed one-third of the total observations are removed. During the analysis of individual variables, it is found that the standard deviations of "C5" and "C9" are zero, meaning all values in these columns are identical. Therefore, it is considered that they cannot provide useful information for the model, so they are eliminated. For missing values, appropriate imputation is required, and different imputation strategies are adopted here based on variable types. For continuous variables with significant skewness, median imputation is

used because the median is less sensitive to outliers. For categorical and logical variables, due to the high proportion of missing values, mean or median substitution is inappropriate, so mode imputation is employed. It can be observed that there are quite a few outliers in the data, so variables with excessively low correlation need to be selectively removed.

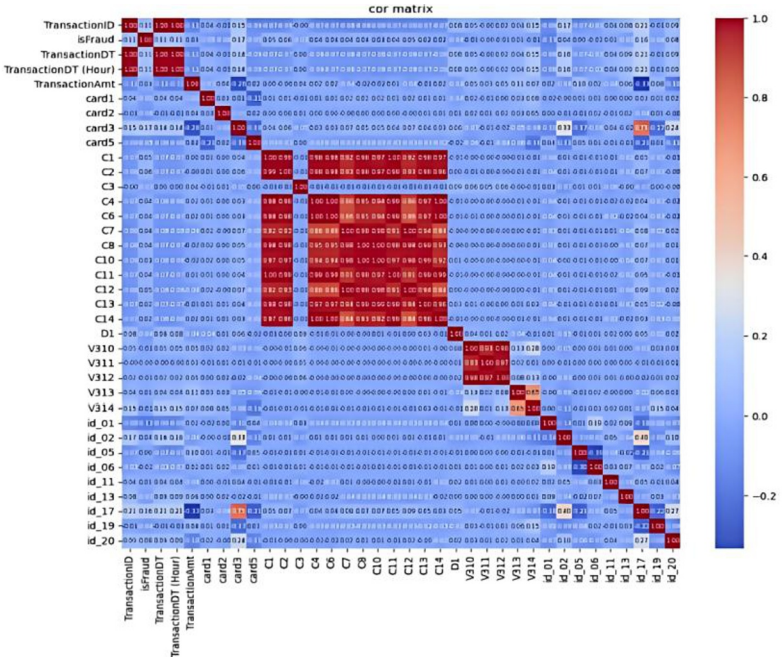


Fig. 3. The correlation matrix of features (Photo/Picture credit: Original).

2.4 Model Building and Selection

The model construction process involved in this paper requires selecting appropriate machine learning models based on data characteristics, computing resources, and business scenarios. To address the common issue of imbalanced data in fraud detection, linear models such as logistic regression and linear Support Vector Machines (SVM) often adopt the Synthetic Minority Oversampling Technique (SMOTE) to alleviate data imbalance [2]. Among them, logistic regression maps linearly combined features to the 0-1 interval based on the logistic function to predict the probability of an event [7]. It has clear principles and strong interpretability, which can intuitively present the influence weight of each feature on the fraud probability, facilitating the understanding of the role of different factors in fraud detection. Moreover, it has a low computational cost and can operate efficiently even when the data scale is large. Linear SVM divides different categories by finding the optimal hyperplane, which has a good classification effect on linearly separable data, stable performance in processing high-dimensional data, strong generalization ability, and certain robustness to noisy data [8]. For high-dimensional and non-linear data, the non-linear models adopted in this paper also play

an important role. Decision trees divide data step by step through a series of judgments on features to form a tree-like structure. They can automatically handle the interactive relationships between features, are easy to visualize, can clearly show the decision-making process, and help understand which features are more critical for fraud detection. Random forests (RF) consist of multiple decision trees, which reduce the risk of overfitting through ensemble learning. They can handle high-dimensional data, are insensitive to missing values and outliers, have strong generalization ability, and can effectively capture complex patterns in data for fraud detection. Radial basis function kernel SVM maps low-dimensional data to high-dimensional space through kernel functions to solve non-linear classification problems. It performs excellently in processing data with strong non-linear relationships, can better capture the implicit patterns in data, and improves the accuracy of fraud detection. Finally, this paper can use some visualization results to further analyze data characteristics and determine a model suitable for financial fraud analysis according to the various performances of the final model. In this process, recall and F1-score are more important indicators. Recall can reflect the model's ability to identify fraudulent transactions and avoid missed detections; F1-score comprehensively considers precision and recall, which can evaluate model performance more comprehensively.

In financial fraud analysis, RF stands out among various machine learning models as an ideal choice for handling high-dimensional and non-linear data, thanks to its multiple advantages. Firstly, by integrating multiple decision trees, they can automatically capture complex non-linear relationships between variables without the need for manual feature engineering. Secondly, the random selection of feature subsets during each split reduces the risk of overfitting in high-dimensional spaces and enables the evaluation of feature importance, thereby quantifying the contribution of each feature to fraud detection. This capability helps domain experts prioritize high-impact features, enhancing the interpretability and compliance of the model. Moreover, their bootstrap sampling mechanism inherently addresses the issue of class imbalance in financial data, improving the detection ability for rare fraudulent transactions. In addition, the ranking of feature importance facilitates business interpretation and provides excellent model interpretability. Compared with linear methods such as logistic regression and support vector machines, as well as single decision trees, RF can better balance predictive performance [9, 10].

3 Result

3.1 Metric

In terms of model indicators, the performance of different models varies: Among linear models, the accuracy of Logistic Regression is 0.7875, the recall is 0.7143, and the F1-score is 0.4295; the accuracy of linear SVM is 0.7955, the recall is 0.7188, and the F1-score is 0.4405. The accuracy and recall of the two are relatively close, with linear SVM being slightly higher, but their F1-scores are both low. This indicates that when dealing with imbalanced data, such as fraud detection, their comprehensive effect in identifying minority classes (fraud samples) is unsatisfactory.

However, among non-linear models, the accuracy of Decision Tree is 0.8595, the recall is 0.8348, and the F1-score is 0.5710; the accuracy of Random Forest is 0.8925, the recall is 0.8125, and the F1-score is 0.6287; the accuracy of SVM with RBF kernel is 0.8695, the recall is 0.6830, and the F1-score is 0.5397 [9, 10].

Random Forest performs the best in terms of accuracy and F1-score, indicating that its overall predictive performance and comprehensive effect in identifying minority classes are better. This is consistent with its advantages, such as being able to capture complex non-linear relationships and reduce the risk of overfitting, but its recall is slightly lower than that of Decision Tree.

In terms of recall, although Decision Tree has the highest recall, meaning it can identify more fraud samples, the overall recall is almost the same as that of Random Forest. Moreover, its accuracy and F1-score are not as good as those of Random Forest, and there may be a certain risk of overfitting.

SVM with RBF kernel has a moderate level of accuracy, but its recall is low, its ability to identify fraud samples is relatively weak, and its F1-score is also low, so its comprehensive performance is average.

In conclusion, Random Forest has the best balance among various indicators and is more suitable for fraud detection scenarios; Decision Tree has advantages in recall, but its overall comprehensive performance is slightly inferior; linear models and SVM with RBF kernel perform weakly in comprehensive indicators, especially the low F1-score of linear models, which may make it difficult to effectively identify fraudulent behaviors in fraud detection. However, the specific selection still needs to be based on the focus of the actual business on different indicators. If more emphasis is placed on identifying all fraud samples (such as reducing missed detections), Decision Tree may be more appropriate; if pursuing overall predictive effect and comprehensive performance, Random Forest is the better choice (Table 1).

Table 1. Evaluation metrics

	Model	Accuracy	Recall	F1-score
Linear	Logistic:	0.7875	0.7143	0.4298
	SVM (linear):	0.7955	0.7188	0.4405
Non Linear	Decision tree:	0.8595	0.8348	0.5710
	Random forest:	0.8925	0.8125	0.6287
	SVM (rbf):	0.8695	0.6830	0.5397

3.2 Visualization

Regarding the most important features, as can be seen from Fig. 4, “C1” is the most crucial one. (It should be noted that the original dataset does not provide descriptions of the relevant features, and here it is treated as a count variable.) In addition, the previously mentioned features, TDT and TA, also play a key role in the performance of this model. Moreover, the use of card3 may also have a certain effect on the analysis of financial fraud.

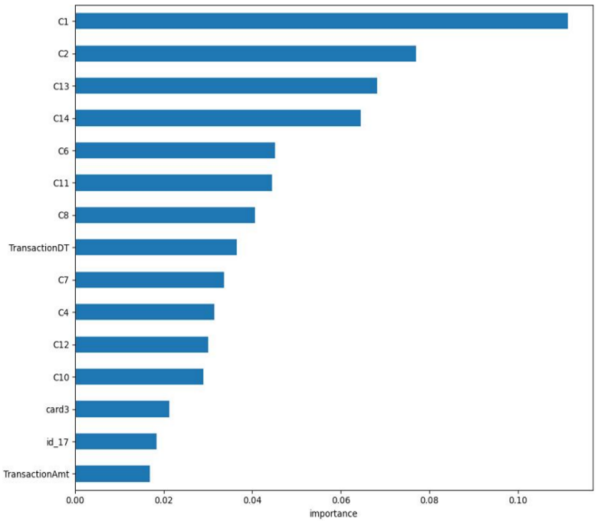


Fig. 4. The importance of random forest (Photo/Picture credit: Original).

4 Discussion

In this study, aiming at the complex task of financial fraud detection, the effectiveness of various key technologies and analytical methods has been explored in depth.

The SMOTE technique has shown excellent performance in handling data imbalance issues, which is highly consistent with the research conclusions put forward by Nishikanta Mohanty et al. In financial data, fraud samples usually account for a small proportion and belong to the minority class, so models tend to be biased towards the majority class. The SMOTE technique expands the number of fraud samples by synthesizing minority class samples, effectively improving the classification performance as shown in Table 2. For example, in relevant experiments, after applying the SMOTE technique, the recognition rate of models for fraud samples has been significantly improved, alleviating the model bias problem caused by data imbalance and enabling models to have better generalization ability when dealing with minority class samples.

Table 2. Evaluation metrics with/without SMOTE

	Model	Recall	F1-score
WITHOUT	Logistic:	0.2723	0.3935
	SVM (linear):	0.1830	0.2982
	SVM (rbf)	0.0938	0.1687
WITH	Logistic:	0.7143	0.4298
	SVM (linear):	0.7188	0.4405
	SVM (rbf):	0.6830	0.5397

Random Forest has shown remarkable advantages in processing high-dimensional and nonlinear data, which strongly confirms the viewpoint of Nishikanta Mohanty [1].

By integrating multiple decision trees, it can automatically mine complex nonlinear relationships between variables without tedious manual feature engineering. The random selection of feature subsets during each split greatly reduces the risk of overfitting in high-dimensional spaces. At the same time, Random Forest can also conduct a quantitative evaluation of feature importance, providing a key basis for an in-depth understanding of fraud factors. This helps domain experts focus on key features, improve the interpretability of the model, make the model results easier to understand and apply to actual business scenarios, and better meet compliance requirements.

In addition, recall plays a core role in fraud detection, which is in line with the view put forward by Jiwon Chung et al that the ability to identify minority classes should be prioritized in risk-sensitive scenarios [9]. A high recall rate means that it can more effectively capture potential fraud transactions and greatly reduce the risk of missed detections. For financial institutions, a missed detection of a fraud transaction may lead to huge financial losses and reputational damage, so models with high recall rates have irreplaceable value in practical business. In this paper, the relatively low performance of indicators such as recall and F1-score is related to multiple factors, including data, features, and models, such as data imbalance, poor feature quality, and insufficient model adaptability. In response to this, the following three improvement measures can be adopted: For addressing data issues, ADASYN is used to handle imbalanced data, supplemented by edge cases, and data is enhanced through methods such as feature perturbation; In terms of improving feature quality, features are screened with the help of indicators like Pearson correlation coefficient, and key features are mined from perspectives such as transaction frequency in combination with business logic; When optimizing model selection, parameters of XGBoost are adjusted to control complexity, and the histogram algorithm of LightGBM is utilized along with setting the class_weight parameter to focus on minority samples [11].

5 Conclusion

Although the dataset has been balanced using SMOTE, the F1 score of the models remains low, potentially due to the relatively small dataset size compared to typical financial fraud analysis datasets. Comparative analysis of metrics shows that non-linear models outperform linear models in both accuracy and recall. Model parameter optimization remains questionable, as suboptimal settings may constrain model performance. Furthermore, despite SMOTE processing, complex class overlap or noise in the data may still interfere with the identification of minority-class samples. Comprehensively, Random Forest demonstrates the best overall performance, achieving the highest accuracy and F1 score, with recall similar to decision trees and significantly higher than other models. This advantage stems from its algorithmic characteristics: as an ensemble learning model, it captures non-linear relationships and complex patterns by integrating multiple decision trees, outperforming linear models constrained by linearity assumptions in expressing non-linear patterns.

Future model optimization can be carried out by combining deep learning and grid parameter tuning: in terms of deep learning, Convolutional Neural Networks (CNNs)

can be used to capture local patterns in transaction data, Long Short-Term Memory (LSTM) networks to mine temporal dependency relationships, autoencoders to build anomaly detection models, or graph neural networks to analyze the topological features of transaction networks for identifying gang fraud. Grid parameter tuning requires systematic optimization of hyperparameters for traditional models and deep learning models. It should be combined with frameworks like Bayesian optimization and Ray Tune to improve parameter tuning efficiency. Meanwhile, the Stacking fusion strategy should be adopted to integrate the advantages of traditional and deep learning models, construct high-order interactive features by combining automatic feature engineering tools and domain knowledge, and use techniques such as focal loss to handle imbalanced data, to ultimately enhance the model's ability to recognize complex fraud patterns.

References

1. Ameer, T., Valilai, O.F.: Predictive exploratory data analysis of shopfloor CNC machine operation through a machine learning model. Journal of Open Innovation: Technology, Market, and Complexity 11(2), 100559 (2025)
2. Mohanty, N., Behera, B.K., Ferrie, C., Dash, P.: A quantum approach to synthetic minority oversampling technique (SMOTE). Quantum Machine Intelligence 7(1), 38 (2025)
3. Mohammed, B., Hamza, C.: A robust estimation of blasting-induced flyrock using machine learning decision tree algorithms: Random Forest, Gradient Boosting Machine, and XGBoost. Mining, Metallurgy & Exploration 42(3), 1–16 (2025)
4. Chung, J., Lee, K.: Credit card fraud detection: An improved strategy for high recall using KNN, LDA, and linear regression. Sensors 23(18) (2023)
5. Kaggle: IEEE-CIS Fraud Detection. https://www.kaggle.com/competitions/ieee-fraud-detection/data, last accessed 2025/07/25
6. Rahmanparast, A., Milani, M., Camci, M., Karakoyun, Y., Acikgoz, O., Dalkilic, A.S.: A comprehensive method for exploratory data analysis and preprocessing the ASHRAE database for machine learning. Applied Thermal Engineering 273, 126556 (2025)
7. Haghighat, M., Choupani, A.A., Afshar, F., Zakeri, S.: Assessing utility theory and logistic regression models to predict drivers' stop/go behavior considering random taste variations. International Journal of Data Science and Analytics (prepublish), 1–17 (2025)
8. Fatemi, S.P., Derakhshanfard, N., Rashidjafari, F., Ghaffari, A.: Distributed data storage using decision tree models and support vector machines in the Internet of Things. Sustainable Computing: Informatics and Systems 46, 101134 (2025)
9. Mohammad, A., Adewale, L.F., Zakariya, A.Y.: Liu regression after random forest for prediction and modeling in high dimension. Journal of Chemometrics 36(4) (2022)
10. Prasetiyo, B., Alamsyah, M., Muslim, M.A., Baroroh, N.: Evaluation performance recall and F2 score of credit card fraud detection unbalanced dataset using SMOTE oversampling technique. Journal of Physics: Conference Series 1918(4) (2021)
11. Hu, J., Zhang, Y., Zhang, H.: Hybrid optimization and deep learning for enhancing accuracy in fraud detection using big data techniques. Peer-to-Peer Networking and Applications 18(4), 179 (2025)