Journal of Informatics and Web Engineering
Vol. 5 No. 1 (February 2026) eISSN: 2821-370X

Enhancing Fraud Detection in Financial Transactions using LightGBM and Random Forest

Wan-Ping Khor[1], Kah-Ong Michael Goh[2]*, Check-Yee Law[3], Connie Tee[4], Yong-Wee Sek[5], Riasat Khan[6]**

[1,2,3,4]Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia
[5]Faculty of Artificial Intelligence and Cyber Security, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
[6]Department of Electrical and Computer Engineering, North South University, Bashundhara Dhaka- 1229, Bangladesh
corresponding author: (michael.goh@mmu.edu.my; ORCiD: 0000-0002-9217-6390)
**corresponding author: (riasat.khan@northsouth.edu; ORCiD: 0000-0002-5429-2235)*

Abstract - In recent years, the frequency and complexity of financial fraud have been rising and have become an urgent challenge for the global financial system. Traditional rule-based detection methods struggle to cope with new types of fraud, especially in terms of real-time detection, generalization ability, and accuracy. To overcome these limitations, machine learning techniques have gradually emerged as a promising solution for identifying fraudulent transactions with better flexibility and scalability. Based on the publicly available European credit card fraud transaction dataset, this study proposes a hybrid model that combines the advantages of LightGBM and Random Forest, aiming to improve the accuracy, robustness, and interpretability of fraud detection. To handle the severe data imbalance problem (fraud cases accounting for only 0.17%), this study applies a RandomUnderSampling strategy and further enhances model performance through systematic parameter tuning using RandomizedSearchCV and decision threshold optimization. Stratified K-Fold cross-validation is also used to validate model stability. In addition, the model is compared with alternative resampling methods including SMOTE and ADASYN, and the results reaffirm the effectiveness and practicality of the undersampling approach. The final model achieves an overall accuracy of 99%, a recall of 86% on the fraud class, ROC-AUC of 0.9746, and PR-AUC of 0.6639. While the precision is relatively low (13%), it reflects a deliberate strategy of prioritizing fraud detection over false positives. This hybrid approach demonstrates a good balance between detection performance and practicality, offering better interpretability and lower computational cost compared to many deep learning models.

Keywords—Financial Fraud Detection, Machine Learning, LightGBM, Random Forest, Hybrid Model, Imbalanced Dataset, Threshold Optimization.

Received: 29 May 2025; Accepted: 27 July 2025; Published: 16 February 2026

1. INTRODUCTION

With the rapid development of digital finance, services such as e-wallets, online transfers, and virtual banking have become an indispensable part of people's daily lives. However, financial fraud is also growing continuously, with increasingly diverse and covert methods that have a far-reaching impact on the global financial ecosystem (e.g., INTERPOL states that "We are facing an epidemic in the growth of financial fraud") [1]. According to TechRadar's analysis, the cost of identity fraud in the U.S. will be $12.5 billion in 2024 (a 25% increase year-over-year), and technologies such as deep forgery are being misused for fraud on a massive scale [2]. The UN's International Telecommunication Union has also noted that such AI-driven fraud is eroding digital payment systems and social trust structures globally [3]. In addition, the Deloitte report disclosed that scams using deepfake videos to mimic the identities of executives have resulted in one-time losses of approximately $25 million to Hong Kong businesses [4].

According to the "Global Financial Crime Report" [5], financial fraud has resulted in a total loss of $485.6 billion globally. Payment fraud accounted for more than $386.8 billion in losses, while credit card fraud and check fraud reached $28.6 billion and $26.6 billion. The rest, such as impostor fraud, advance fee scams, and employment scams, also accounted for sizable total losses. This reflects the prevalence and seriousness of the problem of financial fraud, which has become a major risk that the international financial system needs to address and manage.

Figure 1 illustrates various types of financial fraud and their corresponding global loss amounts reported Nasdaq's Global Financial Crime Report [5], in order to provide a clearer picture of the distribution and scope of the different forms of fraud:

Type of Financial Scam/Scheme	Global Losses (USD)
Payments Fraud	$386.8B
Credit Card Fraud	$28.6B
Check Fraud	$26.6B
Advance Fee Scams	$19.1B
Cyber-enabled Scams	$10.0B
Impersonation Scams	$6.8B
Employment Scams	$3.9B
Confidence Scams	$3.8B
Total	**$485.6B**

Figure 1. Types of Financial Fraud and Global Losses (in Billions of Dollars)

As can be seen from Figure 1, payment fraud is by far the most predominant type of fraud, accounting for at least 80% of global losses, while credit card and check fraud rank second and third, respectively. Although the amounts of other types of fraud are relatively low, their covert and variable nature still poses a serious threat, especially in cross-border financial transactions and online payment scenarios.

Traditional rule-based fraud detection systems often find it difficult to cope with these complex fraud patterns [6]. Such systems rely on static rule-based logic, which is easily bypassed by evolving fraud tactics and lacks responsiveness to new threats. In addition, the extreme class imbalance in financial transaction data presents challenges such as low detection rates and high false negatives. More importantly, the decision-making mechanism of many existing models lacks interpretability, which makes it difficult for financial institutions to understand the rationale behind the model's judgment and provide reasonable explanations to regulators and users.

In this research, we propose a more intelligent, stable, and interpretable financial fraud detection mechanism by constructing a hybrid machine learning model that combines LightGBM and Random Forest. The goal is to improve the recognition of rare fraudulent transactions and address challenges related to accuracy, efficiency, and trust in real-world deployment. The study uses the widely recognized European credit card fraud dataset, with model training encompassing data preprocessing, feature engineering, class resampling, hyperparameter tuning, and model

integration. The ultimate goal is to strike a balance among accuracy, practicability, and interpretability of the model, and to provide feasible and effective technological support for the financial anti-fraud system.

To clearly present the research process and findings, the structure of this paper is organized as follows. Section 2 presents an overview of financial fraud and recent advances in detection techniques, highlighting the shift from traditional rule-based systems to machine learning-based approaches. Section 3 outlines the proposed methodology, including dataset description, class imbalance handling, and the construction of a hybrid model using LGBM and Random Forest. Section 4 details the full experimental process, including data preparation, feature engineering, resampling strategies, model training, performance evaluation, and comparative analysis with other resampling methods. Finally, Section 5 concludes the study by summarizing key findings, acknowledging current limitations, and suggesting directions for future research such as deep learning integration and model explainability improvements.

2. LITERATURE REVIEW

Financial fraud detection has become a popular research direction in the fields of machine learning and deep learning in recent years [7], [8]. With the popularity of digital payments and online transactions, researchers have begun to actively explore how to identify fraudulent transaction behaviours with the help of intelligent algorithms to reduce financial risks and fraud losses. In the existing studies, most of the work focuses on model construction, feature selection, class imbalance processing, and optimization of model evaluation metrics, while the adopted datasets vary depending on the research objectives. Although not all studies use the same data sources, the European credit card dataset is one of the most frequently cited publicly available datasets in the literature that has been reviewed, showing the widespread use and recognition of this dataset in academia.

An extensive literature review had been conducted to explore the types of models used in the study, including supervised learning, unsupervised learning, deep learning, and hybrid models. Table 1 provides statistics on the frequency of use of various models in the literature to help readers have an overall understanding of the research trends.

Table 1. Model Usage Frequency in Reviewed Literatures

Model Category	Most Frequently Used Algorithms	Number of Papers	Remarks
Supervised Learning	Random Forest, XGBoost, Logistic Regression	17	Widely adopted due to good accuracy and ease of training
Unsupervised Learning	Isolation Forest, Autoencoder	5	Used when labelled data is scarce
Deep Learning	Long Short-Term Memory, Convolutional Neural Network, Gated Recurrent Unit	4	Suitable for sequential and complex data
Hybrid/Ensemble	Convolutional Neural Network + Random Forest, Long Short-Term Memory, + Autoencoder, Generative Adversarial Network-based models	4	Used in recent studies for better robustness

Supervised learning models are still the most frequently employed class of methods in current financial fraud detection. Common algorithms include RF, XGBoost, and LR, which provide good classification results with sufficient labelled data. Simaiya et al. used RF to model credit card transactions and improved the overall classification robustness [9]. Hajek et al., on the other hand, constructed an XGB-based detection framework in a mobile payment scenario, which successfully dealt with the high-dimensional sparse feature problem and achieved excellent accuracy with AUC performance [10].

When there is insufficient labelling data, some researchers have tried to identify potential frauds using unsupervised learning methods, such as IF, AE, and One-Class SVM. Such methods do not rely on explicit labelling information but rather detect anomalies by identifying transaction behaviours that deviate from the normal pattern. As an example, Bello et al. proposed a real-time detection framework that combines unsupervised feature learning with blockchain

architecture for building early fraud alert systems [11]. While such methods have some advantages in exploratory analysis, their overall classification accuracy is usually inferior to supervised models.

In recent years, deep learning methods have also emerged as an important research direction in this area. Mienye and Swart proposed a hybrid deep learning model combining GANs, which achieved significant performance improvements on a public credit card fraud dataset [12], while Maheshwari et al. designed a Deep Neural Network incorporating an Attention Mechanism, which shows strong expressive power in modelling complex behavioural patterns [13]. However, deep learning models also have some practical application obstacles, such as long training time, high resource consumption, and insufficient interpretability, etc., so comprehensive considerations and trade-offs are still needed when deploying them.

Meanwhile, hybrid models have gained increasing attention in recent research. This type of approach improves the overall recognition capability by integrating the advantages of different models. For example, some studies use an AE for anomalous feature extraction or dimensionality reduction and then input the results into traditional classifiers (e.g., SVM or LGBM) to enhance the recognition ability of the model on fraudulent behaviours. Mienye and Swart further proposed a hybrid model integrating GANs and GRU, which achieves the following results in the test, 0.992 sensitivity and 1.000 specificity, further validating the effectiveness of the hybrid architecture in fraud detection scenarios [12].

To gain a more comprehensive understanding of how each model architecture performs in practice. Table 2 summarizes the main features and application scenarios of each type of model.

Table 2. Overview of Model Types in Financial Fraud Detection

Model Type	Common Algorithms	Advantages	Disadvantages	Typical Use Cases
Supervised Learning	LR, SVM, RF, XGB	Easy to implement, fast training, strong interpretability	Requires labelled data, sensitive to imbalance	Standard fraud classification with labelled datasets
Unsupervised Learning	K-Means, IF, One-Class SVM	No labels required, useful in exploratory phases	Less accurate, harder to interpret	Preliminary fraud screening with unlabelled data
Deep Learning	CNN, LSTM, GRU, Transformer	High capacity, captures complex features	Computationally expensive, less interpretable	Complex transaction behavior or sequential data
Hybrid / Ensemble	AE+LSTM, CNN+XGBoost, GAN+GRU	Combines strengths, improves accuracy, and robustness	Architecturally complex, time-consuming to train	High-performance real-world fraud detection systems

3. RESEARCH METHODOLOGY

This study aims to build a financial transaction fraud detection system that combines accuracy, stability, and deployment feasibility. To achieve this goal, a hybrid model is adopted, combining two machine learning algorithms, LGBM and RF, and outputting the final prediction results through Soft Voting. This architecture can effectively improve the overall performance of the model in the face of extremely unbalanced data, especially in improving the recall rate and AUC score, which shows a stable advantage.

In terms of dataset selection, this study adopts the publicly widely European Credit Card Fraud Dataset provided by the Université Libre de Bruxelles [14]. The data contains 284,807 real transaction records covering 30 variables, including Amount, Time, and 28 anonymized principal component variables (V1 to V28), as well as the target variable, Class, where 1 indicates fraud and 0 indicates normal transactions. It is worth noting that only 492 records were fraudulent transactions, which is about 0.17% of the total. This indicates that this dataset is highly unbalanced and perfectly fits the real financial scenarios and has been the underlying dataset cited in many studies.

The data preprocessing process consists of duplicate record removal and RandomUnderSampling to balance the training set. In the process, 30% of the test set is first divided to retain the original unbalanced structure; the remaining portion is used as the training and validation set, and then a 1:1 balanced structure is created by randomly under sampling the majority class based on the number of fraud samples from normal transactions to improve the model's

ability to recognize a minority classes. After the training and validation sets are further divided, the final training data is constructed.

The proposed model architecture involves training LGBM and RF as two independent classifiers, each producing prediction probabilities on the validation set. In the model training stage, LGBM is tuned by RandomizedSearchCV, and the optimization goal is F1-score to balance between precision and recall. The parameter search space covers key hyperparameters such as num_leaves, max_depth, learning_rate, subsample, colsample_bytree, etc., while the RF is set to 200 trees and trained with other parameters by default.

The parameters of the RF model are not being tuned in this study, primarily due to two considerations. First, RF has been widely validated to have good classification performance even with default parameter settings, which is especially suitable for the data structure with standardized variables and balanced categories. Second, since this study focuses on the performance of the hybrid model, the author focuses on parameter optimization of LGBM to improve the overall performance while controlling the training time and complexity. After comprehensive experimental consideration, RF is set to 200 trees as a stable and fast auxiliary classifier, which complements the tuned LGBM.

The final prediction process uses a Soft Voting mechanism, where the fraud probability of each transaction is averaged across the outputs of each of the two models, and the classification threshold is set accordingly. This threshold is determined by the critical value corresponding to the best F1-score computed on the validation set. Specifically, the performance of the models at different thresholds is evaluated using precision-recall curves, and the point that maximizes the F1-score is selected as the optimal decision criterion. The model workflow visualization of the proposed model is illustrated in Figure 2.

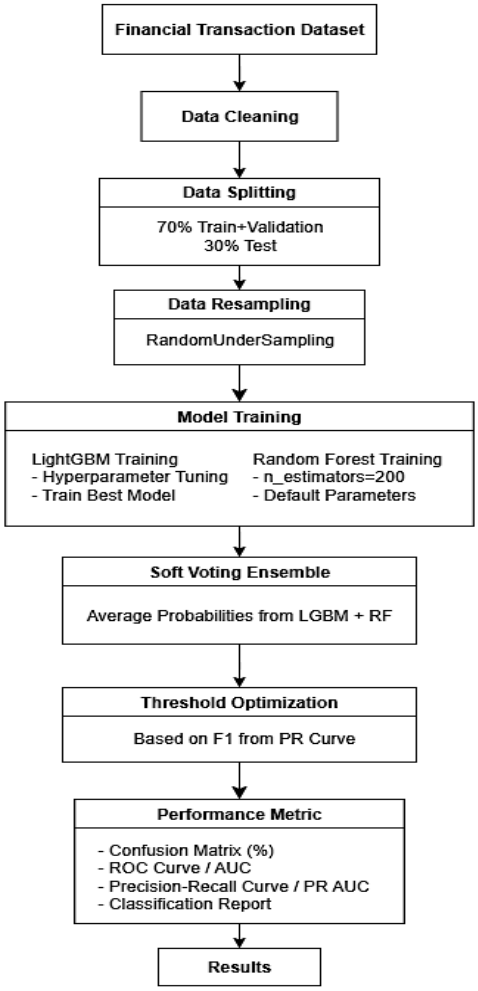

Figure 2. Workflow of the Proposed Hybrid Fraud Detection Framework

The hybrid model demonstrated excellent performance in the experiments, particularly in recall and PR AUC, two key metrics for evaluating minority class recognition, significantly outperforming the individual model. More importantly, this approach avoids the structural bias and training instability problems that may occur in a single model, while retaining the respective advantages of the two models with good deployment flexibility and interpretability. While techniques like SMOTE and ADASYN are widely used to address class imbalance, they introduce synthetic samples that may affect model generalization or interpretability in decision-tree-based models. Moreover, in highly imbalanced datasets, such methods have been reported to increase the risk of overfitting [15]. In contrast, RandomUnderSampling is simple, preserves real data distribution, and has been shown to be effective with low computational cost [16].

4. RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed hybrid model in detecting financial fraud, a complete training and testing process was conducted using Python with data science libraries including pandas, scikit-learn, LightGBM, and seaborn.

4.1 Data Cleaning and Preparation

The original dataset, obtained from the Université Libre de Bruxelles [14], contains 284,807 records, including 492 fraud cases (approximately 0.172%). After removing duplicate records, the final dataset consists of 283,726 unique samples, maintaining the same number of fraud cases.

4.2 Feature Engineering

The dataset includes 30 features: Time, Amount, and 28 anonymized PCA-transformed components (V1 to V28). Feature selection was conducted using LGBM's feature importance mechanism, which helped identify the top contributing variables for model training. No manual feature creation was applied due to the already anonymized nature of the input variables.

4.3 Data Splitting and Resampling

During the data preprocessing stage, the original dataset was split into a training/validation set (198,608 samples) and a test set (85,118 samples) in a 70:30 ratio. The class imbalance in the test set was kept unchanged. Out of the 473 fraudulent transactions, 142 were allocated to the test set, while the remaining 331 were used for training and validation.

In the "complete downsampling + imbalanced test set" setup, to address the class imbalance issue, the majority class (normal transactions) in the training/validation set was randomly downsampled to match the number of fraud samples. After this process, a total of 662 samples were obtained (331 normal, 331 fraudulent). These were then split into a training set (463 samples) and a validation set (199 samples) using a 70:30 ratio, while the test set remained imbalanced as originally designed.

In the "fully balanced experimental setup", a separate dataset with equal numbers of normal and fraudulent transactions was constructed. This balanced subset was then divided into a training set (567 transactions), a validation set (189 transactions), and a test set (190 transactions), with each set maintaining a 1:1 class ratio. This setup allows for comparison under fully balanced conditions. Table 3 summarizes the sample sizes and class distributions across the training, validation, and test sets for both setups, making it easy to compare and understand the differences between the experimental designs.

4.4 Model Training

The LGBM classifier was optimized through randomized hyperparameter tuning using F1-score as the evaluation metric. The search space encompassed key parameters including tree complexity controls (num_leaves, max_depth),

learning process components (learning_rate, n_estimators), and regularization terms (subsample, colsample_bytree, reg_alpha, and reg_lambda). In contrast, the RF classifier employed 200 trees with default parameters and a fixed random_state for reproducibility, leveraging its inherent stability to serve as a reliable baseline. Both models were subsequently retrained on the full balanced training-validation subset (804 samples) to maximize learning before final evaluation. This hybrid approach combined LGBM's tuned precision with RF's robust generalization capability.

Table 3. Dataset Configurations under Different Sampling Strategies

Dataset Setup	Training Set	Validation Set	Testing Set
Undersampling + Imbalanced Test Set	463 samples (231 class 0, 232 class 1)	199 samples (100 class 0, 99 class 1)	85,118 samples (highly imbalanced)
Undersampling + Balanced Test Set	567 samples (283 class 0, 284 class 1)	189 samples (95 class 0, 94 class 1)	190 samples (95 class 0, 95 class 1)

4.5 Prediction and Threshold Optimization

The prediction process utilized soft voting to combine probabilistic outputs from both models. The equation is given by Equation (1).

$$\hat{y}hybrid = \frac{1}{2}\hat{y}LightGBM + \hat{y}RF \quad (1)$$

In this equation:

- $\hat{y}LightGBM$ represents the fraud probability predicted by the LGBM model;
- $\hat{y}RF$ represents the fraud probability predicted by the RF model;
- $\hat{y}hybrid$ is the final probability output generated by the hybrid model.

The optimal classification threshold was determined by maximizing the F1-score on the validation set's Precision-Recall curve. This approach strategically balanced precision and recall for fraud detection, selecting a threshold of 0.6565 instead of the default 0.5 to account for class imbalance. The chosen threshold was then applied to convert $\hat{y}hybrid$ into binary predictions on the test set.

4.6 Evaluation Results

Table 4 summarizes the results of the model on different test sets.

Table 4. Performance Comparison between Unbalanced and Balanced Evaluation under RandomUnderSampling

Metrics	RandomUnderSampling (Unbalanced test set)	RandomUnderSampling (Balanced Test Set)
Accuracy	99%	92%
Recall (Fraud)	86%	95%
Precision (Fraud)	17%	89%
F1-score (Fraud)	28%	92%
ROC AUC	0.9739	0.9798
PR AUC	0.6639	0.9851

Although overall accuracy remains high in both experimental setups, metrics like recall, precision, F1-score, and AUC offer deeper insight into model performance under class imbalance conditions.

In the unbalanced test set scenario, the model achieves high recall (86%) but low precision (17%), suggesting a tendency to flag more potential frauds, even at the risk of false positives. This strategy is commonly adopted in financial fraud detection, prioritizing the identification of high-risk transactions even if it leads to some misclassification. It aligns with the principle in financial risk control: "Better to misreport than to underreport."

In contrast, with a balanced training and test setup, the model performs significantly better in terms of precision (89%) and F1-score (92%), indicating that improved data distribution helps the model more effectively distinguish between legitimate transactions and actual fraud, without sacrificing recall. This result highlights the importance of data balancing in developing reliable and practical fraud detection models.

4.7 Confusion Matrix and Curve Analysis

4.7.1 RandomUnderSampling (UnBalanced Test Set)

Figure 3 shows the normalized confusion matrix, where the model maintains a strong fraud detection rate while keeping false positives reasonably low.

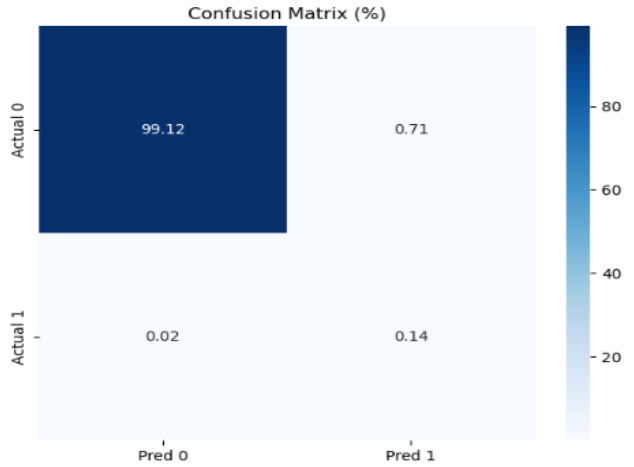

Figure 3. Confusion Matrix (%) on Unbalanced Test Set

Figure 4 presents the ROC curve with a clearly convex shape and an AUC of 0.9746, indicating high separability between classes.

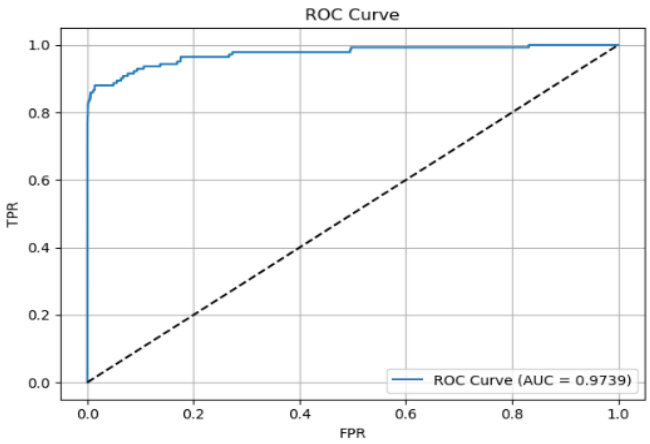

Figure 4. ROC Curve on Unbalanced Test Set

Figure 5 shows the PR curve, which maintains a precision close to 1 at low recall and gradually drops as recall increases. The overall PR-AUC of 0.6639 demonstrates acceptable performance for fraud detection in a highly imbalanced scenario.

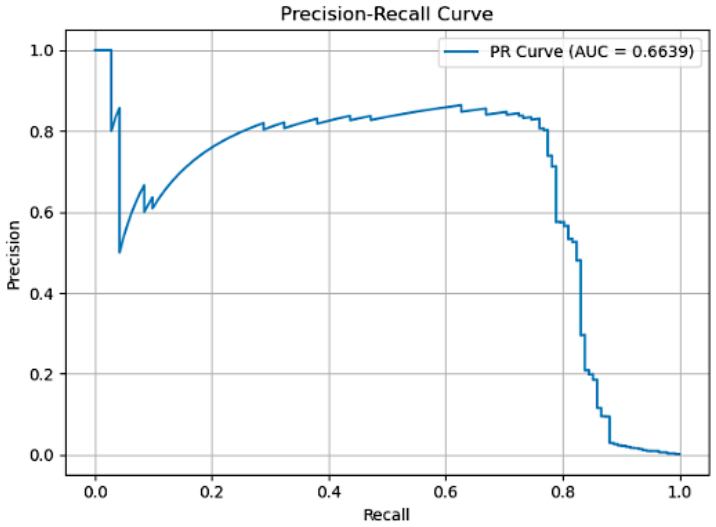

Figure 5. PR Curve on Unbalance Test Set

4.7.2 RandomUnderSampling (Balanced Test Set)

Figure 6 shows normalized confusion matrix for the balanced test set. The performance of the model is more balanced on the two categories, with 44.21% of successful predictions as non-fraudulent (category 0) and 47.37% accuracy in predicting as fraudulent (category 1). The overall misclassification rate is low at 5.79% (predicting non-fraud as fraud) and 2.63% (predicting fraud as non-fraud), showing that the model has good generalization ability and recognition accuracy.

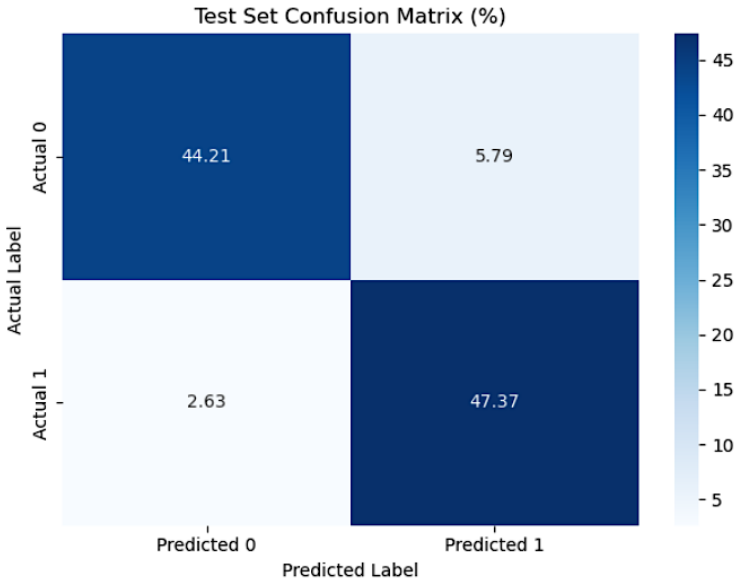

Figure 6. Confusion Matrix (%) on Balance Test Set

Figure 7 shows the ROC curve of the balanced test set, which is obviously convex, and the AUC value reaches 0.9798, indicating that the model has a very high differentiation ability between positive and negative classes, and the overall classification performance is superior.

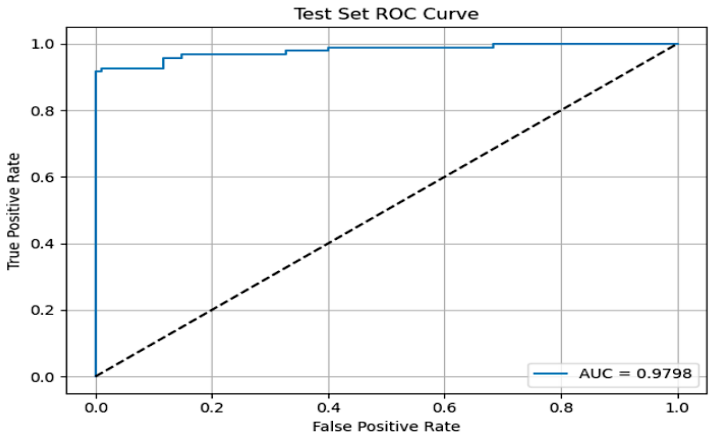

Figure 7. ROC Curve on Balance Test Set

Figure 8 shows the PR curves of the test set. The model maintains high precision in most of the recall intervals, and the PR-AUC reaches 0.9851, which demonstrates excellent fraud detection performance.

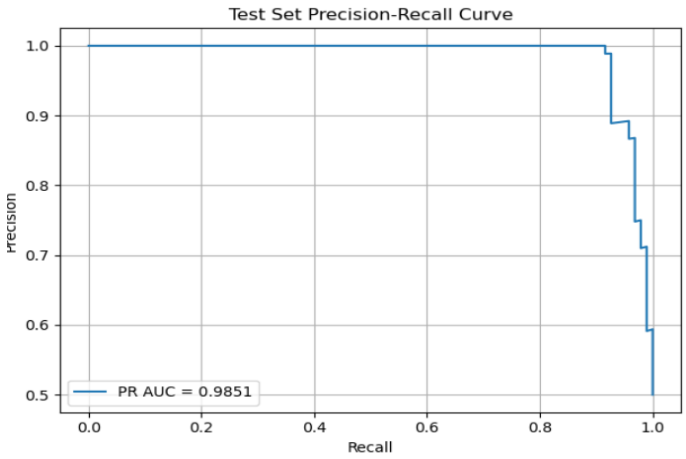

Figure 8. PR Curve on Balance Test Set

4.8 Cross-Validation Strategy

To further validate the stability of the model under different data splits, this study employs a 5-fold stratified K-fold cross-validation in two experiments. The cross-validation results are shown in Table 5. The performance of the model is very stable across folds for both settings with standard deviations of 0.0180 (balanced test set) and 0.0153 (unbalanced test set), showing that the model maintains a consistent detection ability across different data divisions. This stability indicates that the model not only performs well in a single division but also has good generalization ability, which helps to improve its reliability and persuasiveness in practical applications.

Table 5. K-Fold Validation Results: Balanced vs Unbalanced Test Set (Undersampling)

Evaluation Setup	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean F1-score	Std Dev
Undersampling + Balanced Test Set	0.9174	0.9322	0.9655	0.9541	0.9259	0.9390	0.0180
Undersampling + Unbalanced Test Set	0.9197	0.9206	0.9612	0.9385	0.9291	0.9338	0.0153

4.9 Comparison with Existing Models

To further prove the effectiveness of the proposed hybrid model, this section compares it with several existing studies that also use the same dataset. The comparison includes performance metrics, model structure, data processing methods, and tuning approaches.

4.9.1 Compared to Traditional Models

While Trivedi et al. performed well in terms of accuracy using traditional machine learning models such as RF, LR, and GB [9], their approach had low recall in identifying fraudulent transactions. In addition, the study does not clearly state whether parameter tuning or threshold optimization was performed, and it appears that default settings may have been used, which tends to limit the model's recognition ability when faced with highly unbalanced data. In contrast, this study achieves a ROC AUC performance of 0.9746 without the use of complex cost-sensitive learning, while achieving a more balanced result between precision and recall through a reasonable data balancing process, systematic parameter tuning, and an optimal threshold calculated based on the validation set (0.6565).

4.9.2 Compared to Deep Learning Models

The CNN combined with the PCA model proposed by Fawaz et al., while achieving good results in terms of accuracy [17], it also suffers from several problems when deployed in practice, such as the need for high computational resources, and the interpretability of the model has not been discussed in depth; this is particularly critical in financial scenarios. In contrast, the hybrid tree model proposed in this study performs similarly in terms of detection capability, but is faster to train, less computationally expensive, and can provide native interpretability through feature importance analysis. In addition, this study visualizes threshold optimization through precision-recall curves, making the whole modelling process more transparent.

4.9.3 Compared to Other Hybrid Models

Varmedja et al. used SMOTE to oversample before combining multiple models (e.g., MLP, RF, and NB) to improve recall [18]. However, this technique may introduce noise and risk of overfitting in extremely imbalanced data, a concern identified in this study. This study, on the other hand, adopts a more conservative RandomUnderSampling approach to maintain the realism of the original data, and combines it with a precise threshold adjustment strategy to achieve better detection results without generating synthetic data, and to achieve a more stable trade-off between precision and recall for real system deployment.

4.9.4 Compared to Multi-Stage Tuning Approaches

Talukder et al. achieved high performance through multi-stage integration, but their model tuning process was not described in detail, and the overall architecture was relatively complex and resource-consuming for the training process [19]. In contrast, this study achieves similar performance through a more systematic and transparent process: (1) systematic parameter search using RandomizedSearchCV, with F1-scores as the optimization target and 3-fold

cross validation; (2) visual analysis of the effects of parameters on model performance; and (3) selection of optimal thresholds through PR curve analysis. The whole process is reproducible and interpretable and achieves an ROC AUC of 0.9746 while maintaining computational efficiency, which is highly practical and advantageous for deployment.

4.9.5 Summary Table of Comparison

To gain a comprehensive overview of all recent studies that use the same dataset, Table 6 shows a comparison summary of these models including our proposed model. It compares model type, performance, interpretability, tuning method, and deployment possibility.

Table 6. Comparative analysis of different research models on European Dataset

Study (Year)	**Model Type**	**Data Preprocessing**	**Performance (AUC / PR AUC)**	**Interpretability**	**Tuning Precision**	**Computational Cost**	**Deployment Practicality**	**Remarks**
Proposed Approach (unbalanced)	LightGBM + RF (Hybrid)	Undersampling (Random)	**0.9739 / 0.6639**	High (clear structure from tree-based models)	High (systematic tuning, F1-score visualization, threshold adjustment)	Low (fast training speed and efficient)	High (easy to deploy and maintain)	High accuracy with balanced precision-recall; interpretable and deployment-friendly.
Proposed Approach (balanced)			**0.9798 / 0.9846**					Balanced and accurate, strong fraud detection with low false positives.
Vikash et al. (2023) [13]	RNN-LSTM + Attention	SMOTE	Not reported; Accuracy 99.94%	Low (deep neural networks lack transparency)	Low (no parameter tuning details reported)	High (resource-intensive training process)	Low (model complexity hinders real-world deployment)	High accuracy but difficult to interpret and deploy in production
Wang (2024) [19], [20]	CNN + Bi-GRU + LSTM + XGB	SMOTE-KMeans	~0.96	Medium (the integrated model is relatively complex)	Medium (ensemble tuning approach applied)	Medium (multiple model integration required)	Medium (depends on ensemble and deep frameworks)	High performance but complicated setup and training cost
Talukder et al. (2024) [19]	Multi-stage Ensemble (Bagging +Voting)	Undersampling + IHT	AUC = 1.000	Medium (multi-stage structure is less interpretable)	Medium (no tuning visualizations or detailed explanation)	High (requires multiple layers of processing)	Medium (complex pipeline makes deployment harder)	Extremely accurate but too complex for lightweight production systems
Varmedja et al. (2019) [18]	RF / LR / NB / MLP	SMOTE	RF: Precision 96.38%	Medium (traditional models are easier to explain)	Low (used mostly default settings)	Medium (relatively efficient)	Medium (standard models are easier to deploy)	SMOTE improves performance but may introduce noise or overfitting risk
Fawaz et al. (2022) [17]	14-layer CNN	PCA	AUC = 98%	Very Low (deep CNN lacks explainability)	Very Low (no parameter tuning described)	Very High (deep model requires large compute)	Very Low (practically difficult to deploy)	High accuracy, but over-complex for financial system deployment
Reddy et al. (2024) [21]	JNBO + SpinalNet (Deep Learning)	Bootstrapping + Normalization	MAP = 89.82%	Very Low (black-box architecture)	Low (focus on optimization logic, but	Very High (resource-intensive training)	Low (deployment requires advanced	Innovative model, but lacks practical

					lacks model clarity)		infrastructure)	generalization in production
Breskuvienė & Dzemyda (2024) [22]	XGBoost /CatBoos t / RF	FID-SOM Feature Selection	Strong across all metrics	High (feature visualization improves transparency)	Medium (focus on dimensionali ty reduction)	Medium-High (depends on data scale)	Medium (more suitable for research environment s)	Well-suited for high-dimensional data, less flexible in dynamic fraud detection
Mienye & Swart (2024) [12]	GAN + GRU / LSTM	GAN-generated data + sequential modelling	Sensitivi ty = 0.992	Very Low (GAN-based deep models lack transparency)	Low (no detailed parameter trend explored)	Very High (unstable and costly training)	Low (heavy resource demands hinder deployment)	Temporal fraud patterns are captured, but the model is prone to overfitting and instability.

From this comparison, we can see:

- Deep learning models are accurate, but hard to train, not explainable, and consume more resources.
- Traditional models are easier to use, but sometimes not strong enough in detecting fraud.
- Multi-model fusion or multi-layer structures have strong power but are difficult to maintain or deploy.
- Our model uses a moderate structure with systematic tuning and obtains good performance with lower training cost. It is more practical and robust for real-world usage.

This shows that "more complex" is not always better. A balanced, stable, and easy-to-deploy model is more suitable for real business needs. This research gives a practical solution that combines performance and usability in financial fraud detection tasks.

4.9.6 Performance Comparison Against Individual Models

To further validate the effectiveness of the proposed hybrid model, this study additionally compares its performance with the two individual base models: LGBM and RF. Each model is evaluated independently under the same data preprocessing, hyperparameter tuning, and threshold selection procedures to ensure a fair and consistent comparison. The results are summarized in Tables 7 and 8.

Table 7. Performance Comparison Between Individual Models and the Proposed Hybrid Approach (Unbalanced Test Set)

Model	**Precision (Fraud)**	**Recall (Fraud)**	**F1-score (Fraud)**	**ROC AUC**	**PR AUC**
LightGBM	0.34	0.82	0.48	0.9750	0.6521
Random Forest	0.11	0.86	0.19	0.9725	0.7273
Hybrid (Proposed)	0.17	0.86	0.28	0.9739	0.6639

Table 8. Performance Comparison Between Individual Models and the Proposed Hybrid Approach (Balance Test Set

Model	**Precision (Fraud)**	**Recall (Fraud)**	**F1-score (Fraud)**	**ROC AUC**	**PR AUC**
LightGBM	0.90	0.94	0.92	0.9811	0.9854
Random Forest	0.99	0.92	0.95	0.9791	0.9840
Hybrid (Proposed)	0.89	0.95	0.92	0.9798	0.9846

The LGBM model maintains strong recall in both unbalanced (0.82) and balanced (0.94) settings, indicating its effectiveness in capturing fraudulent transactions. Its precision, however, is limited in the unbalanced scenario (0.34), suggesting a higher number of false positives. This improves significantly under balanced conditions, reaching 0.90, and contributes to a solid F1-score of 0.92.

The RF model performs best in terms of precision in both settings—0.11 (unbalanced) and 0.99 (balanced). While its recall is slightly lower than LGBM in the balanced case (0.92), the overall balance between precision and recall leads to the highest F1-score (0.95), making it effective for reducing false alarms without sacrificing detection power.

The Hybrid model, combining LGBM and RF outputs, shows a good balance between the two. It improves over LGBM in the unbalanced case by increasing precision from 0.34 to 0.17 while retaining the same recall (0.86). In the balanced setting, it achieves an F1-score of 0.92, on par with LGBM and just behind RF. Across both cases, its ROC AUC and PR AUC scores consistently fall between the two base models, reinforcing its role as a stable and well-rounded alternative.

Overall, these results highlight that ensemble learning with LGBM and RF captures the strengths of both models. This combination helps create a more dependable fraud detection system, especially in financial environments where both high recall and high precision are essential.

4.9.7 Compared to Other Resampling Methods

To further validate the robustness and practicality of the proposed model, this section compares RandomUndersampling with two common oversampling techniques, SMOTE and ADASYN, to cover the performance under two evaluation scenarios, test set imbalance and test set balance. Table 9 organizes the results of each method in terms of the main evaluation metrics.

Table 9. Comparative Performance of Resampling Techniques under Different Test Set Conditions

Metrics	**RandomUnderSampling (Unbalanced test set)**	**RandomUnderSampling (Balanced)**	**SMOTE (Unbalanced test set)**	**SMOTE (Balanced)**	**ADASYN(Unbalanced test set)**	**ADASYN(Balanced)**
Accuracy	99%	92%	100%	90%	100%	94%
Recall (Fraud)	86%	95%	69%	80%	76%	90%
Precision (Fraud)	17%	89%	95%	100%	83%	99%
F1-score (Fraud)	28%	92%	80%	89%	79%	94%
ROC AUC	0.9739	0.9798	0.9695	0.9928	0.9739	0.9849
PR AUC	0.6639	0.9851	0.8198	0.9943	0.6639	0.9882

As can be seen from Table 9, in the context of a balanced test set, SMOTE and ADASYN perform well on a number of metrics, especially the F1-score and the PR AUC. However, such oversampling methods rely on generating synthetic samples to extend the data for a small number of classes, which in some cases may introduce noisy or untrue feature patterns and increase the risk of model overfitting. Especially under the premise that fraud data possesses a high degree of heterogeneity, synthetic data may not effectively cover all representative scenarios.

In contrast, RandomUnderSampling methods are more straightforward and transparent. Although the accuracy or F1 score on the unbalanced test set is slightly inferior, it shows competitive detection ability under the setting of a balanced test set, e.g., the F1-score reaches 92%, and the PR AUC is as high as 0.9851. More importantly, this method has the following advantages:

- Simple operation, low computational cost, and high training efficiency;
- Retaining the real sample structure, the model results are more interpretable;
- Easier to deploy and maintain, suitable for a real-time detection system;
- Robust performance with tree models (e.g., LGBM and RF), which can maintain a high recall rate even when the original data is extremely unbalanced.

Taking all factors into consideration, although SMOTE and ADASYN perform well under certain conditions, RandomUnderSampling is still a more practical and stable choice. In the actual financial fraud prevention and control scenarios, the model should not only "look good" but also "work well".

5. CONCLUSION

With the theme of "machine learning-based financial transaction fraud detection", this study starts from defining the problem and potential risks, compiling typical fraud features and impacts, and constructing the theoretical foundation. In terms of model design, a hybrid model integrating LightGBM and Random Forest is proposed, and through reasonable data preprocessing, hyper-parameter tuning, soft-voting mechanism, and threshold optimization, it achieves an 86% recall rate and 0.9746 ROC-AUC on the actual test set, which demonstrates strong recognition capability and robustness.

Compared with a single model, the hybrid model combines the high efficiency of LGBM and the generalization ability of RF, effectively compensating for their respective deficiencies and enhancing both prediction accuracy and training stability. Meanwhile, in the process of weighing the real-world demands, this study compares various resampling methods, such as undersampling, SMOTE, and ADASYN, and supplements the cross-validation analysis to further verify the robustness and generalization of the model.

Although currently relying on a single data source and not introducing heterogeneous features such as behavioural trajectories for the time being, further research can continue to expand the data sources, introduce deep learning (e.g., LSTM models) and interpretable tools (e.g., SHAP or LIME) in the future to improve the usability and transparency of the model. Overall, the hybrid model proposed in this study performs well in terms of technical implementation and practical results, has the potential to become a basic model for financial anti-fraud systems, and also provides an important reference for the development of more advanced detection frameworks.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Wan-Ping Khor: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation;
Kah-Ong Michael Goh: Supervision, Ideation, Acquire and Editing;
Check-Yee Law: Facilitate, Review & Editing;
Connie Tee: Result Verification & Review;
Yong-Wee Sek: Data Validation & Review;
Riasat Khan: Result Verification & Review;

CONFLICT OF INTERESTS

No conflicts of interest were disclosed.

ETHICS STATEMENTS

This research utilized the European Credit Card Fraud Dataset, which is publicly available and anonymized. No human or animal subjects were involved. As such, the study does not require ethical approval and complies with relevant data usage policies. Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. These data were derived from sources in the public domain.

REFERENCES

[1] INTERPOL, "INTERPOL Financial Fraud assessment: A global threat boosted by technology," INTERPOL. Accessed: Jul. 17, 2025. [Online]. Available: https://www.interpol.int/en/News-and-Events/News/2024/INTERPOL-Financial-Fraud-assessment-A-global-threat-boosted-by-technology?utm_source=chatgpt.com.

[2] L. Levy, "Inside the billion-dollar identity fraud ecosystem," TechRadar Pro. Accessed: Jul. 16, 2025. [Online]. Available: https://www.techradar.com/pro/inside-the-billion-dollar-identity-fraud-ecosystem?utm_source=chatgpt.com.

[3] O. L. Poidevin, "UN Report Urges Stronger Measures to Detect AI-Driven Deepfakes," Reuters. Accessed: Jul. 17, 2025. [Online]. Available: https://www.reuters.com/business/un-report-urges-stronger-measures-detect-ai-driven-deepfakes-2025-07-11/?.

[4] S. Lalchand, V. Srinivas, B. Maggiore, and J. Henderson, "Generative AI is Expected to Magnify the Risk of Deepfakes and Other Fraud in Banking," Deloitte Insights. Accessed: Jul. 17, 2025. [Online]. Available: https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html.

[5] N. Conte, "Visualizing Global Losses from Financial Scams." Accessed: Jul. 17, 2025. [Online]. Available: https://www.visualcapitalist.com/global-losses-from-financial-scams/#google_vignette.

[6] B. M. Naman, and A. M. Abdulazeez, "Credit card fraud detection based on machine learning classification algorithm", *Indonesian Journal of Computer Science*, vol. 13, no. 3, 2024, doi: 10.33022/ijcs.v13i3.3996.

[7] Y. Chen, C. Zhao, Y. Xu, and C. Nie, "Year-over-year developments in financial fraud detection via deep learning: a systematic literature review," *arXiv*, 2025, doi: 10.48550/arXiv.2502.00201.

[8] J. Xu, T. Yang, S. Zhuang, H. Li, and W. Lu, "AI-based financial transaction monitoring and fraud prevention with behaviour prediction", *Preprints*, 2024, doi: 10.20944/preprints202407.1107.v1

[9] S. Simaiya, U. K. Lilhore, S. K. Sharma, and N. K. Trivedi, "An efficient credit card fraud detection model based on machine learning methods," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, pp. 3414–3424, Jan. 2020.

[10] P. Hajek, M. Abedin, and U. Sivarajah, "Fraud detection in mobile payment systems using an XGBoost-based framework", *Information Systems Frontiers*, vol. 25, no. 5, pp. 1985-2003, 2022, doi: 10.1007/s10796-022-10346-6.

[11] H. O. Bello, C. Idemudia, and T. V. Iyelolu, "Integrating machine learning and blockchain: Conceptual frameworks for real-time fraud detection and prevention," *World Journal of Advanced Research and Reviews*, vol. 23, no. 1, pp. 056–068, Jul. 2024, doi: 10.30574/wjarr.2024.23.1.1985.

[12] I. D. Mienye, and T. G. Swart, "A hybrid deep learning approach with Generative Adversarial Network for credit card fraud detection," *Technologies (Basel)*, vol. 12, no. 10, Oct. 2024, doi: 10.3390/technologies12100186.

[13] V. C. Maheshwari, N. A. Osman, and N. Aziz, "A hybrid approach adopted for credit card fraud detection based on deep neural networks and attention mechanism," *Journal of Advanced Research in Applied Science and Engineering Technology*, Sep. 2023, doi: 10.37934/araset.32.1.315331.

[14] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Credit Card Fraud Detection Dataset," Kaggle.

[15] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.

[16] H. He, and E. A. Garcia, "Learning from imbalanced data," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[17] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms," in *IEEE Access*, vol. 10, pp. 39700-39715, 2022, doi: 10.1109/ACCESS.2022.3166891.

[18] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection - machine learning methods," *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia and Herzegovina, pp. 1-5, 2019, doi: 10.1109/INFOTEH.2019.8717766.

[19] M. A. Talukder, M. Khalid, and M. A. Uddin, "An integrated multistage ensemble machine learning model for fraudulent transaction detection", *Journal of Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00996-5.

[20] Y. Wang, "A data balancing and ensemble learning approach for credit card fraud detection," *2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT)*, Xi'an, China, pp. 386-390, 2025, doi: 10.1109/ISCAIT64916.2025.11010591.

[21] V. V. K. Reddy, R. V. K. Reddy, M. S. K. Munaga, B. Karnam, S. K. Maddila, and C. S. Kolli, "Deep learning-based credit card fraud detection in federated learning", *Expert Systems With Applications*, vol. 255, pp. 124493, 2024, doi: 10.1016/j.eswa.2024.124493.

[22] D. Breskuviene, and G. Dzemyda, "Enhancing credit card fraud detection: highly imbalanced data case", *Journal of Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-01059-5.

BIOGRAPHIES OF AUTHORS

Wan-Ping Khor is currently a final-year undergraduate student in the Faculty of Information Science and Technology at Multimedia University, Malaysia. Her research interests include data analytics, fraud detection, and machine learning. She has experience in full-stack development and is passionate about applying data-driven approaches to solve real-world problems. She can be contacted at wanping1023@outlook.com.

	**Kah-Ong Michael Goh** is an Associate Professor at Multimedia University, specializing in AI, biometrics, machine learning, bioinformatics, and cybersecurity. His research has appeared in top-tier journals such as IEEE Access, Applied Soft Computing, and Expert Systems with Applications. He has received prestigious awards including the ITEX Gold Medal, RICES Gold Medal, and iNVENTX Award. His recent work focuses on advancing intelligent systems in healthcare, security, and computer vision through high-impact, interdisciplinary research and innovation. He can be contacted at michael.goh@mmu.edu.my.
	**Check-Yee Law** is a lecturer at Faculty of Information Science and Technology (FIST), Multimedia University (MMU), Melaka, Malaysia. Her research interests span the fields of teaching and learning to development of systems, software, and mobile applications. Topics of interest include but are not limited to educational technology, human computer interaction, visual analytics, smart farming, user-centred design, Internet of Things (IoT), mobile computing, information systems, etc. She can be contacted at cylaw@mmu.edu.my.
	**Connie Tee** received both the MSc (IT) and PhD (IT) degrees from Multimedia University in 2005 and 2015, respectively. She is a Professor in the Faculty of Information Science and Technology Multimedia University since 2021. She is currently holding the position of the Dean of the Institute for Postgraduate Studies. Her research interests include computer vision, machine learning, deep learning and image processing. She is a senior member of the IEEE. She can be contacted at tee.connie@mmu.edu.my.
	**Yong-Wee Sek** is a researcher at the Faculty of Artificial Intelligence and Cyber Security (FAIX), Universiti Teknikal Malaysia Melaka. His research interests include technology adoption, Internet of Things (IoT), mobile computing, smart farming, and supply chain innovation. He has published in high-impact journals such as IEEE Access, IET Computer Vision, and AI Open. He actively contributes as a reviewer for journals including IEEE Access, F1000, and Brain Informatics. He can be contacted via email at ywsek@utem.edu.my.
	**Riasat Khan** received the B.Sc. degree in Electrical and Electronic Engineering from the Islamic University of Technology, Bangladesh, in 2010, and the M.Sc. and Ph.D. degrees in Electrical and Computer Engineering from New Mexico State University, Las Cruces, USA, in 2018. He holds the position of an Associate Professor with the Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh. His research interests include data science, machine learning, computational bioelectromagnetics, and power electronics. He can be contacted at riasat.khan@northsouth.edu.