Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.34999

Application of Explainable Ensemble Learning in Credit Card Fraud Detection

Siyu Wang

College of Information Engineering, Sichuan Agricultural University, Yaan, China
wsy1258052959@outlook.com

Abstract. Credit card fraud is a global problem, which causes huge economic losses every year and reduces consumers' trust in the digital transactions. Therefore, it is very important to detect the credit card fraud. Nowadays, traditional machine learning models have limited detection of increasingly complex fraud means and often has black box characteristics, which lack explainability and bring obstacles to financial institutions. However, the current research on the credit card fraud detection still lacks explainability. Therefore, this paper proposes a stacking model integrating XGBoost, LightGBM and MLP, and compares the stacking model with a single model. The results show that the stacking model has the best performance. Finally, Explainable AI (SHAP and LIME) is used to identify the complex features of the stacking model and interpret the output. This study provides a feasible scheme and explanation for credit card fraud detection in the contemporary digital economy, and effectively reduces financial losses.

Keywords: Credit card fraud, Explainable AI, Stacking model, Machine learning

1. Introduction

Nowadays, credit card transactions are very common in online banking, especially in electronic payments. But this kind of transaction has potential risks. Fraudsters use complex fraud methods to carry out illegal transactions, which bring serious economic losses to consumers and enterprises, while traditional fraud detection methods are difficult to deal with the changing fraud methods. In order to deal with this kind of fraud, it is very important to use machine learning to detect fraud.

In recent years, credit card fraud detection has attracted significant research attention. In credit card fraud detection, basic machine learning models such as Logistic Regression, SVM and Random Forest have been widely used. Srivastava et al. used multiple machine learning models and tried the hybrid combination method to comprehensively compare and analyze the detection capabilities of each model [1]. Iqbal et al. discussed the fraud prediction performance by implementing and comparing Random Forest and XGBoost models [2]. However, the accuracy of the basic model in fraud detection is often limited, so some research mostly uses hybrid models or relatively novel models for detection. Nguyen and others have effectively improved the accuracy of model detection based on the fusion of CatBoost and DNN [3]. Shaymaa et al. introduced a cutting-edge FS strategy algorithm and used the BBO algorithm to randomly adjust positions, significantly improving classification accuracy, reducing wrongly classified transactions and maximizing correctly classified

DOI: 10.54254/2755-2721/2026.34999

transactions [4]. Mienye et al. not only captured anomalies at the feature level, but also identified irregular patterns at the structure level by integrating Autoencoder, Graph Attention Network and XGBoost, so as to achieve stable and reliable prediction results [5]. Mienye and Sun proposed a method combining deep learning integration and data resampling, which verified the effectiveness of the integration strategy on unbalanced data [6]. Although the existing research on fraud detection models has been relatively mature, the explainability of the model is limited. In practical applications, financial institutions must prove the rationality of their decisions [7].

Therefore, this paper proposes a stacking model combined with Explainable AI (SHAP and LIME) to improve the model detection performance and interpret the model output from the global and local levels, providing a reliable and transparent detection system for practical applications.

2. Methodology

2.1. Data set

2.1.1. Data description

This paper selects the European credit card transaction dataset, which contains 284,807 credit card transaction records, including 492 fraudulent transactions, accounting for 0.172%. The features include transaction time, transaction amount and 28 anonymous features obtained by PCA transformation. The data were divided into 80% training set and 20% test set.

2.1.2. Data oversampling

Because the dataset has few fraud categories, it is extremely unbalanced. In order to deal with the imbalanced categories of the dataset, this study uses the SMOTE oversampling strategy [8]. Unlike random oversampling, SMOTE uses linear interpolation to construct new samples between the k-nearest neighbors of a few minority class samples. This avoid over fitting phenomenon caused by directly copying samples, while retaining the local data structure of the minority class. As a result, the model can more fully learn the characteristic distribution of fraud samples in the training process, and improve the recognition ability of the minority class.

2.2. Prediction model

This paper uses a stacking integrated model to predict. Stacking is a two-layer integration strategy. The first layer is independently trained by multiple base learners, and the second layer is fused by meta learners with the output of the base learner as the input, so as to integrate the advantages of each base learner and reduce the bias and variance of a single model. If the base learning machine is diverse, it can provide additional benefits of the meta-learning machine. Therefore, this study selected three heterogeneous base learners: XGBoost, LightGBM and MLP. XGBoost and LightGBM are gradient boosting tree models, which are good at capturing nonlinear interactive and sparse patterns of features. However, the growth strategies of the two model trees are different. XGBoost uses hierarchical growth, uses a pre-sorting algorithm and histogram to approximate and efficiently find the optimal splitting point, and has built-in regularization items to control the complexity of the model. LightGBM uses leaf growth and accelerates training through gradient-driven unilateral sampling and exclusive feature bundling. The combination of the two can effectively complement each other, balancing accuracy and generalization ability. As a feedforward neural network, MLP can learn the high-order abstract representation of features through multi-layer

DOI: 10.54254/2755-2721/2026.34999

nonlinear transformation, capture the smooth nonlinear mapping and global trend that the tree model is difficult to learn, and increase the prediction ability of the model. In the meta learner part, the standard logistic regression is selected, and the prediction probability of the three base learners is taken as the input. L2 regularization is used to prevent overfitting, and the final fraud detection result is output. In this paper, the prediction accuracy and overall performance of the stacking model will be compared with that of the single XGBoost, LightGBM and MLP models, so as to prove the accuracy and superiority of the stacking model [9].

2.3. Explainable methods

Explainable methods are divided into global and local levels. The global interpretation uses SHAP, which calculates how each feature marginally influences the model output based on the Shapley value in game theory [10]. In this study, SHAP is used to analyze the overall impact of characteristics on fraud prediction, identify the key features that make the largest contribution to identifying fraud, help understand the overall decision logic of the stacking integrated model, and provide data support for the formulation of risk control rules. The local interpretation uses LIME, which visually displays the judgment basis of a single transaction by fitting an Explainable local linear model in the neighborhood of the prediction sample. In this study, LIME analyzes the positive or negative contribution of each feature to the fraud probability of the transaction one by one for the transaction samples that are judged to be fraudulent, so as to help business personnel understand the reason why the transaction is intercepted in practical application. When the two are used together, they confirm each other in the global mode and local details, meet the global transparency requirements of regulatory compliance and the interpretation needs of business personnel, and enhance the integrity and credibility of the Explainable model.

2.4. Evaluation metrics

To comprehensively measure the performance of the model, the following indicators are used in this paper:

AUC-ROC measures the overall ability of the model to distinguish between normal transactions and fraudulent transactions, reflecting the comprehensive discrimination ability of the model under different threshold settings. The value range is 0 to 1. A value closer to 1 indicates a stronger capability of the model to discriminate between positive and negative classes.

As the harmonic mean of precision and recall, the F1-Score is wellsuited for credit card fraud scenarios with a wide proportion of positive and negative samples, preventing threshold distortion that may arise from relying on a single metric. Due to the high imbalance of this dataset, the single dependence on accuracy or recall rate may lead to threshold deviation. F1-Score provides a comprehensive basis for selecting the optimal threshold for the model by balancing the relationship between the two, ensuring that fraudulent transactions are detected as much as possible while controlling false detection.

The accuracy rate reflects the correct proportion of the overall forecast and visually shows the overall discrimination ability of the model. However, it may be dominated by most categories under extremely unbalanced data, which needs to be combined with other indicators for comprehensive judgment.

The precision rate measures the proportion of transactions predicted as fraud that are actually fraudulent to control the cost of false positives. In this project, high accuracy means that fewer

DOI: 10.54254/2755-2721/2026.34999

normal transactions are misjudged as fraud, so this is one of the key indicators to evaluate the practicability of the model.

Recall rate indicates the percentage of actual fraud cases that are successfully detected and controls the risk of missing reports. High recall means that more real fraudulent transactions are successfully intercepted.

3. Results

3.1. Model performance

Table 1. Evaluation indicators of different models in credit card fraud detection

Model	AUC	F1-Score	Accuracy	Precision	Recall
XGBoost	0.9833	0.7692	0.9991	0.6911	0.8673
LightGBM	0.9844	0.8300	0.9994	0.8137	0.8469
MLP	0.9571	0.8182	0.9994	0.8100	0.8265
Stacking	0.9864	0.8497	0.9995	0.8632	0.8367

As shown in Table 1, the stacking model has the best value among the four indicators of accuracy, precision, F1-Score and AUC. Although the recall value is slightly lower, the overall indicator is the best, showing the stability and feasibility of fraud detection in the face of unbalanced data sets.

3.2. Confusion matrix

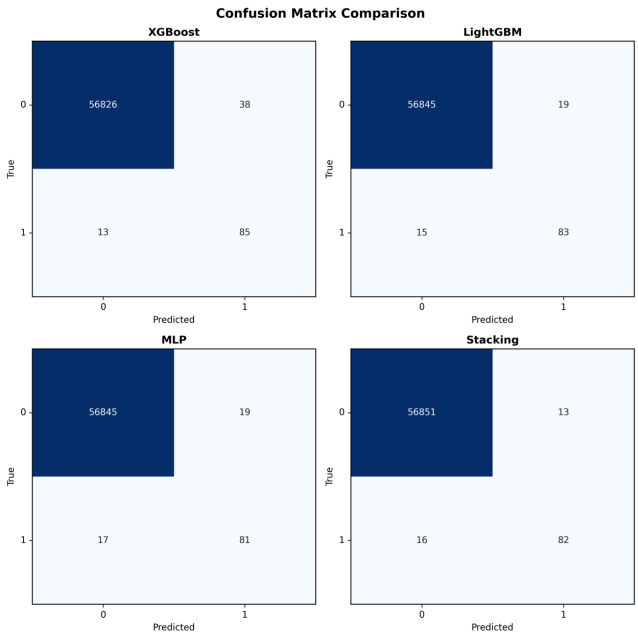

Figure 1. Confusion matrix of different models

The classification details of each model on the test set are shown in the figure, which includes the quantity distribution of true negative TN (identifying normal transactions), false positive FP

DOI: 10.54254/2755-2721/2026.34999

(misclassifying normal as fraud), false negative FN (misclassifying fraud as normal) and true positive TP (identifying fraudulent transactions). As can be seen from Figure 1, the stacking model has the highest TN value and the lowest FP value; that is, it has the best ability to identify normal transactions and misjudge fraud. In terms of FN and TP values, the stacking model is not significantly different from the other three models. Therefore, from the perspective of the confusion matrix, the stacking model has the optimal detection ability and the ability to control false positives, and the comprehensive performance is the best.

3.3. ROC curve

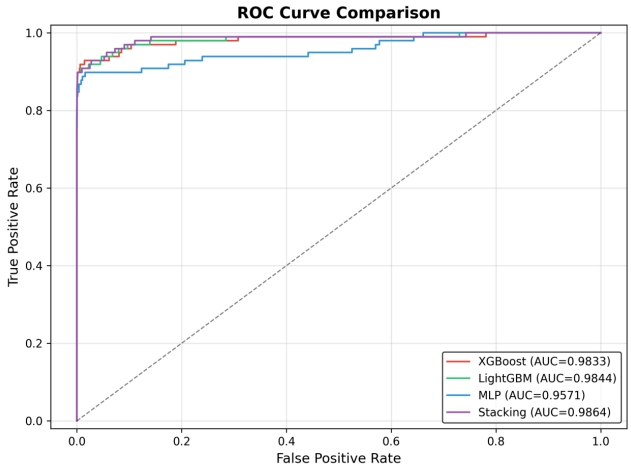

Figure 2. ROC comprehensive curve

The ROC curves of four models illustrate the balance they strike between the true positive rate and the false positive rate. As shown in Figure 2, the stacking model achieves the highest AUC value among all models, which means that the model can maintain high sensitivity and specificity within a wide range of thresholds.

3.4. Explainable results using SHAP and LIME

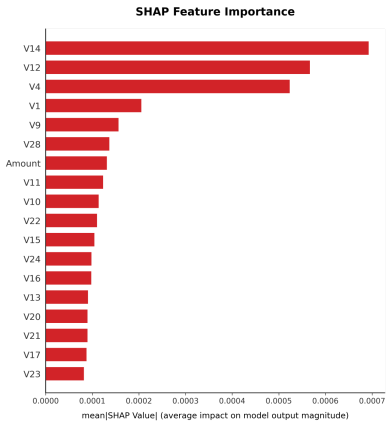

Figure 3: Local SHAP interpretation

Figure 3 shows the average marginal contribution of each feature to the fraud prediction of the model. The results show that V14, V12 and V4 are the three most important features of the model,

DOI: 10.54254/2755-2721/2026.34999

and their average absolute SHAP values are significantly higher than those of other features, which constitute the core basis of fraud identification. The low SHAP values of V20, V21, V17, V23 and other features indicate that these features have little impact on the average output of the model. They are not the key variables to identify fraud at the global level, but it does not mean that they are completely ineffective. In some specific cases, fraud alarms may still be triggered due to abnormal values of these features.

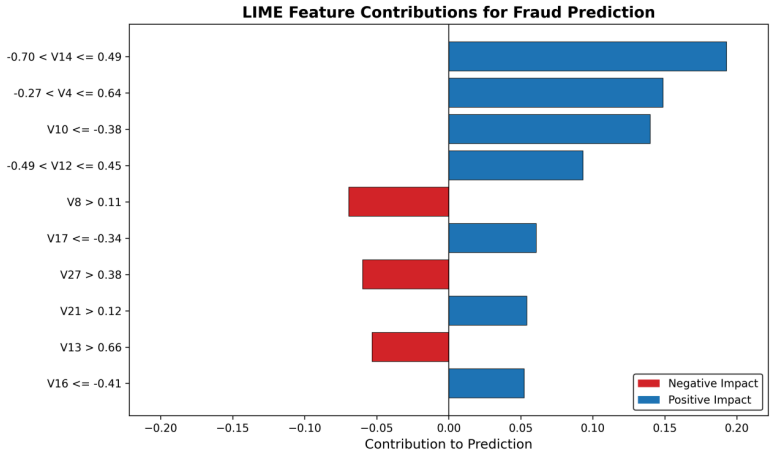

Figure 4. Explanation of LIME features

Figure 4 shows the effect of each feature on the fraud prediction probability of the sample. The blue bar represents the contribution of positive support to fraud prediction, and the red bar represents the contribution of negative inhibition to fraud prediction. It can be seen from the figure 4 that V14, V4, V10 and other features contribute the most to fraud prediction, while V8, V27, V13 and other features play a negative role in fraud prediction.

4. Discussion

The stacking model and Explainable method constructed in this paper perform well in credit card fraud detection. Compared with the existing research, this paper combines the advantages of different types of models by integrating XGBoost, LightGBM and MLP, and reveals the role of sample characteristics and single transactions in fraud detection by using an explainability method, which makes up for the lack of explainability in current fraud detection research.

5. Conclusion

In addressing credit card fraud detection, this paper proposes a stacking integrated model integrating XGBoost, LightGBM and MLP, and interprets the model at the global and local levels by combining the explainability methods of SHAP and LIME, and constructs a set of high-performance transparent fraud detection systems. Experimental results show that the stacking model is superior to the single model in accuracy, precision, F1-score, and AUC, showing stable and superior model performance on highly unbalanced data sets, achieving high detection accuracy, effectively reducing false positives and unnecessary manual review. In terms of explainability, the global analysis of SHAP reveals the dominant role of V14, V23, V4 and other features in fraud prediction, while the local interpretation of LIME provides an intuitive feature contribution for a single suspicious transaction. The combination of the two provides a transparent and credible financial decision-making basis for

DOI: 10.54254/2755-2721/2026.34999

financial institutions. This study not only significantly improves the classification accuracy and anti-interference ability of the credit card fraud detection system, but also meets the actual needs of financial institutions for model transparency in compliance review and risk communication through explainability, providing a feasible scheme for constructing practical credit card fraud identification systems in the context of digital finance.

Although this model significantly enhances the ability of fraud detection, it still has some limitations. Future research can take into account the dynamic evolution characteristics of fraud patterns and add an incremental learning framework to the research to enable the model to achieve adaptive concept drift. At the same time, other Explainable methods can be tried to improve the explainability of the model. In addition, a more lightweight and efficient detection model can be explored to reduce the computational delay in real-time monitoring scenarios.

References

[1] Srivastava, A., Srivastava, A., Tewatia, S., & Sharma, V. S. (2026). Credit card fraud detection using machine learning. In Mathematics and logics in computer science (ICMLCS 2025). Springer.
[2] Iqbal, S., Awan, K. M., Kamal, S., & Rehman, Z. U. (2025). Interpretable ensemble learning models for credit card fraud detection. Applied Sciences, 15(22), 12073.
[3] Nguyen, N., et al. (2022). A proposed model for card fraud detection based on CatBoost and deep neural network. IEEE Access, 10, 96852–96861.
[4] Sorour, S. E., AlBarrak, K. M., Abohany, A. A., & Abd El-Mageed, A. A. (2024). Credit card fraud detection using the brown bear optimization algorithm. Alexandria Engineering Journal, 104, 171–192.
[5] Mienye, I. D., Esenogho, E., & Modisane, C. (2025). Detecting imbalanced credit card fraud via hybrid graph attention and variational autoencoder ensembles. AppliedMath, 5(4), 131.
[6] Mienye, I. D., & Sun, Y. (2023). A deep learning ensemble with data resampling for credit card fraud detection. IEEE Access, 11, 30628–30638.
[7] Martins, T., de Almeida, A. M., Cardoso, E., & Nunes, L. (2024). Explainable artificial intelligence (XAI): A systematic literature review on taxonomies and applications in finance. IEEE Access, 12, 618–629.
[8] Breskuviene, D., & Dzemyda, G. (2024). Enhancing credit card fraud detection: Highly imbalanced data case. Journal of Big Data, 11, 182.
[9] Gupta, R. K., et al. (2025). Enhanced framework for credit card fraud detection using robust feature selection and a stacking ensemble model approach. Results in Engineering, 26, 105084.
[10] Keerthana, C. S., Nalluri, S. C., Muskaan, S., & Sadagopan, P. (2025). Explainable AI in credit card fraud detection: SHAP and LIME for machine learning models. In 2025 10th International Conference on Signal Processing and Communication (ICSC) (pp. 387–392). IEEE.