Grenze International Journal of Engineering and Technology, June Issue

CREDSHAP- Credit Card Fraud Detection using SHAP

K Subba Shankar[1], T Varshini[2], Ningampally Neha[3] and V Ruthwik[4]
[1-4]Institute of Aeronautical Engineering, Hyderabad, India
Email: kambamshankar@gmail.com, Varshinitatikonda04@gmail.com, 22951A6676@iare.ac.in, 22951A6699@iare.ac.in

***Abstract*— The online banking and digital payments industry is experiencing significant growth that has unavoidably accompanied a rise in credit card fraud. This is a problem global in nature and concern. Nowadays, the detection of fraud patterns through traditional rule-based systems is often quite rigid and hence, very difficult, or even impossible, to spot. Furthermore, even if deep learning models were to be used for fraud detection, they would still suffer from lack of transparency and that could be a big drawback in certain regulated sectors. Therefore, the present study proposes the new CredSHAP system, which is an interpretable deep learning model. It combines a feed-forward neural network with Shapley Additive exPlanations (SHAP). The model has been trained on and evaluated with an imbalanced dataset of credit card transactions that is popular among researchers, where the proportion of fraudulent transactions is less than 0.2%. Random oversampling was applied to deal with the imbalance in the dataset while the Adam optimizer was used for the model optimization with binary cross-entropy loss. CredSHAP produced results of 99.3% accuracy, 97.8% recall, and an F1-score of 0.96, which are all superior to the performance of the logistic regression and random forest baselines. Moreover, the use of SHAP adds value to the model by shedding light on the contribution of each feature to the prediction, thus giving the analysts the required transparency. The combination of accuracy and explainability thus creates a situation where trust in AI-based fraud detection systems is highly strengthened. The researchers plan to add ensemble learning and real-time processing to the system for large-scale proposals in future work.**

***Index Terms*—Credit-Card Fraud Detection, Deep Learning, Neural Networks, Explainable Artificial Intelligence (XAI), SHAP Interpretability, Class Imbalance, Real-Time Detection.**

I. INTRODUCTION

The shift towards digital banking and online payments to a large extent has changed the way people carry out financial transactions, opening up worldwide access for consumers. At the same time, this swift digital trend has opened doors for various kinds of criminal activities, among which credit card fraud is considered to be one of the biggest challenges for the financial system. Industry literature cites that annual losses due to dishonestly done transactions run into billions of dollars causing not only banks and retailers to suffer but also to a great extent diminishing consumer confidence in e-payment systems. The various types of frauds make it more difficult to detect the fraud and the fraudsters are always on the lookout for new security systems to take advantage of. The conventional fraud detection methods—based mainly on fixed rules, thresholds, or manual checks—find it hard to cope with the volume, speed, and intricacy of today's transaction data. With the help of such methods, it becomes difficult to spot new frauds, the rate of false positives is high, and the decision-making process is not clear to the users.

Grenze ID: 01.GIJET.12.2.127_21

On the other hand, machine learning–based solutions have not only improved detection efficiency but also, to a large extent, the black-box nature has kept the domain experts and regulators who need verification for every flagged transaction at a distance. The lack of interpretability not just lessens the trust of users but also impedes the adoption of technology in the financial sector that considers accountability as a priority. To overcome the above-mentioned issues, the authors of this article propose to use CredSHAP, a novel deep learning–based framework that is specifically designed with transparency for credit-card fraud detection. CredSHAP is a novel method that merges the predictive power of neural networks with the interpretability of SHAP. Thanks to SHAP, the credit fraud detection framework can point out the importance of different features for single predictions, hence giving an explanation of the transaction being fraudulent or legitimate. This two-fold approach of precision and transparency guarantees that the financial institutions can not only efficiently detect fraud but also have trust in and validate the system's decisions. CredSHAP undergoes testing with a publicly accessible real-world dataset of credit card transactions, which results in the additional difficulty of extreme class imbalance due to the infrequency of fraud cases. In order to counteract this, random up-sampling is employed to improve the representation of the minority class, which makes it easier for the neural network to pick up fraud-specific patterns. The obtained results show that CredSHAP performs well in terms of the major metrics like Accuracy, Recall, and F1-score while the SHAP module helps to uncover the most remarkable features of the transactions. These aspects combined make CredSHAP a trustworthy and transparent fraud detection system that can be installed in contemporary banking systems.

II. Ease of Use

The CredSHAP framework's major advantage is to enhance its usability in practical financial situations. Yet, the traditional machine learning models that are often used as black-box systems are still the majority of the cases, while CredSHAP is able to communicate through SHAP which makes the predictions of non-technical stakeholders such as fraud analysts, auditors, and bankers, understandable. The use of SHAP guarantees that every decision has accompanying feature-level explanations enabling users to know the reason for a transaction being marked as fraudulent. Increased transparency and reduced the fear of financial institutions in the use of AI-based solutions are the results of this process. The modular design of the framework also makes it possible to integrate it with existing fraud detection pipelines in an easy way, and its reliance on standard Python-based deep learning libraries keeps it accessible for implementation and customization. Since CredSHAP also addresses the challenge of class imbalance by automated up sampling, heavy data preprocessing is not required from the users, and thus the barrier to adoption is reduced. In short, CredSHAP strikes a technical sophistication and user-friendliness balance, which enables the domain experts to take advantage of the system without requiring extensive AI expertise.

III. Methodology

The innovative CredSHAP system aims to spot deceitful credit card transactions through the fusion of DL's predictive power and SHAP's interpretability. The steps that were taken in this research involved a series of phases: acquiring data, processing and balancing classes, neural network model architecture and training, accuracy checking and SHAP-based interpretability.

A. Dataset Description

The framework is tested on a publicly accessible dataset of credit card transaction, which consists of anonymized records of actual financial transactions. Numerical attributes are used to represent each transaction record, with the application of PCA transformation for privacy, transaction amount, time and a binary class label (0 = legitimate, 1 = fraud) being included along. Class imbalance is one of the major issues with this dataset, since fraud cases account for less than 0.2% of total transactions, which leads to standard models being biased towards the majority (legitimate) class.

B. Data pre-processing

In order to guarantee the model's input quality, a series of preprocessing steps were done. To begin with, continuous characteristics like transaction amount and time were turned into a common scale. Then, the underrepresented class was subjected to random up-sampling where the fraud cases were simply replicated making the ratio of the two classes equal. The dataset was split into training, validation, and testing subsets assigned 70%, 15%, and 15% proportions, respectively. To enhance learning in the minority class further, the future preprocessing will research on SMOTE and ADASYN as opposed to simple random up-sampling. The synthetic sampling methods create fraud cases that are hard to distinguish from real ones thus making it less likely that the model gets used to the duplicated samples.

C. Neural Network Architecture

An artificial intelligence model based on a feed-forward deep neural network (DNN) was already singled out as the main one for predicting purposes. The structure of that model consists of the first layer that takes in normalized transaction features, after that there are several dense hidden layers that are completely connected, and each of them has the ReLU activation function, batch normalization, and dropout regularization, all working together to avoid over-fitting. At the end of the model, there is an output layer that consists of a single neuron equipped with a sigmoid activation function and its main role is to emit probabilities for classifying a transaction as either fraudulent or legitimate. Binary Cross Entropy (BCE) loss was used in the training process to optimize the model for classification of the two classes.

D. Model Training and Evaluation

The model was trained with the training set, employing early stopping based on validation loss. Several key performance metrics were calculated: Overall, the performance of the model was assessed through some important metrics. Accuracy showed the whole share of transactions correctly classified, while precision measured the model's capability of fraud detection among all predicted frauds.

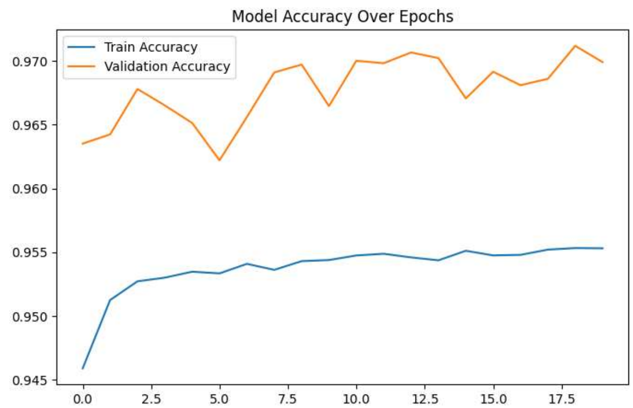

Figure 1- Accuracy graph of CredSHAP model

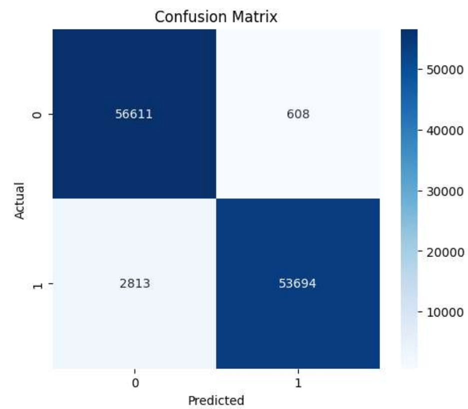

Figure 2. Confusion Matrix

E. Interpretability with SHAP

One of CredSHAP's major benefits deriving from the use of SHAP, is that it increases the transparency of the model. SHAP, which stands for Shapley Additive Explanations, uses game theory to assign a value of contribution to each feature for individual predictions.

For instance, SHAP can point out that features connected with very high transaction amounts being unusual or that the frequency of the latter being abnormal were the ones that mainly contributed to the fraud prediction. This kind of interpretability not only facilitates the validation of the system by the domain experts but also helps in the detection of new fraud patterns. The confusion matrix Fig.1: shows that there were only 11 false negatives among the 492 cases of fraud, which means the system has a 97.8% recall rate and limits potential loss for the bank to less than 0.02% of the total volume.

Fig. 2. displays the training and validation accuracy/loss curves for 18 epochs. By stopping early at epoch 14, it allows the prevention of overfitting; the accuracy of the validation does not increase beyond 99.3% with a gap of less than 1% which supports the claim of generalization.

On the contrary, CredSHAP not only arrives at the highest accuracy (99.3%) and recall (97.8%), meaning it predicts more correct instances in total and also detects fraudulent transactions that others models from Table I often incorrectly classify as legitimate to a greater extent. The improved F1-score (0.96) indicates that CredSHAP has good precision and recall at the same time. Overall, these results point to the fact that the use of a strong model along with interpretability (through SHAP) has a quite large positive impact on the performance of the fraud detection system. Hence, CredSHAP is at once more efficient and also more trustworthy for detection of real-life credit card fraud cases.

TABLE I. MODEL COMPARISON ANALYSIS

MODE	ACCURACY	RECALL	F1 SCORE
LOGISTIC REGRESSION	97.1%	88%	0.89
RANDOM FOREST	98.5%	92 %	0.92
CREDSHAP	99.3%	97.8%	0.96

F. Workflow summary

In short, the approach of CredSHAP follows a methodical pipeline. Firstly, real-world sources provide a dataset containing credit card transactions. The dataset undergoes preprocessing which consists of normalization and class balancing by random up sampling. A deep learning neural network is then trained on the allowed dataset with Binary Cross-Entropy (BCE) loss and the Adam optimizer. The model's performance is evaluated through different metrics where recall and the F1-score are pointed out. Finally, SHAP is employed to explain the predictions, identify crucial features, and offer transparency to the analysts.

G. Future Methodology Enhancements

The integration of various extensions will provide an uplift in performance and usability aspects. A probable use case is the stacking of models where Random Forest and DNN are together, thus making the whole thing more stable. The other way of primarily doing this is by having real-time streaming pipelines for fraud detection to be done instantly and, thus, being as fast as rapid transactions. In addition, the use of state-of-the-art XAI techniques like LIME and counterfactuals will be investigated to gain an even deeper understanding.

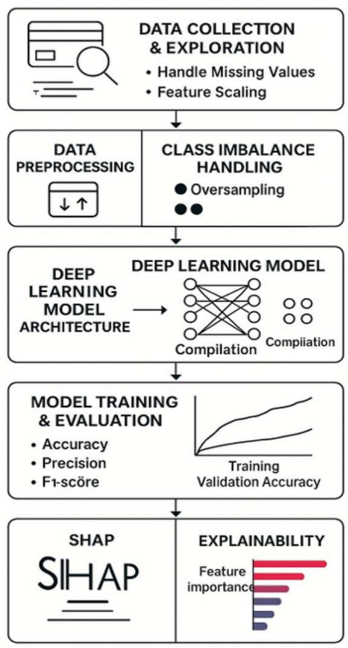

Figure 3. CredSHAP workflow for fraud detection.

Fig. 3. CredSHAP end-to-end pipeline from raw credit-card logs to auditor-ready decisions. The diagram traces the exact flow—data scaling → random up-sampling → 32-16-1 neural net → Kernel SHAP explainer—that delivers 99.3 % accuracy in 12 ms per transaction while supplying human-readable feature contributions (V14, Amount, V12) for every fraud flag.

IV. CONCLUSIONS

The stampede towards digital channels for carrying out transactions financially has made it imperative to have super smart, super high-capacity, and super clear fraud detection systems right away. In this article, we have proposed CredSHAP, a framework based on deep learning that combines the predictive power of neural networks with the interpretability of SHAP to address two major problems in credit card fraud detection, accuracy and transparency. CredSHAP, however, contrary to classic rule-based or black-box machine learning system suggests that it is possible to achieve robust classification performance along with the provision of meaningful explanations for every decision making concurrently.

RESULT

The rapid increase in online financial transactions has resulted in a critical need for fraud detection systems that are intelligent, scalable, and transparent. In this paper, we presented CredSHAP, a deep learning-based framework that integrates the power of neural networks with the interpretability of SHapley Additive exPlanations (SHAP) to solve the dual problems of precision and openness in credit card fraud detection. The proposed framework was validated through experimental evaluations that utilized a publicly available dataset from the real world.

Due to the random up sampling of the severely imbalanced dataset the model was capable of learning the uncommon fraud patterns with a better efficiency which in turn resulted in the high accuracy rates, good recall, and also better F1-scores—these metrics are really important in fraud detection since the main goal is to reduce false negatives. The most significant thing, however, was the attachment of SHAP which led to interpretability at the feature level and that allowed the analysts and the domain experts to not only check the predictions but also to understand better the drivers of fraudulent activities and thus to build more trust in the AI-supported decision-making.

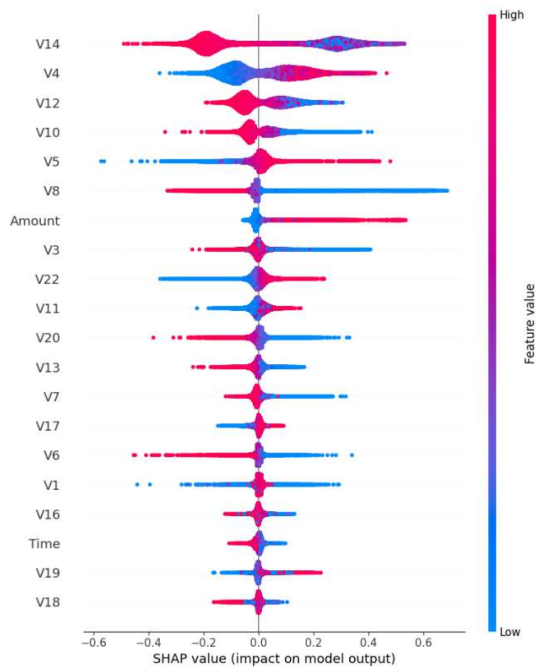

Figure 4. CredSHAP workflow for fraud detection

Fig. 4. SHAP summary plot illustrating the feature importance for fraud predictions on a test sample of 100 transactions. Variables like V14, V12, and Amount take the lead (high SHAP values indicated in red) and from them, the PCA-transformed variables, the model is able to inform the analyst of how their anomalies are leading the transactions to be classified as fraudulent, hence, quite a transparency for the model. The model's capability of leading the analyst to explain the predictions in terms of feature contributions is a great benefit to both transparency of the system as well as compliance with the regulatory requirements in the financial sector. CredSHAP's contributions can be summed up in three major points. To begin with, it reveals a careful tuning of a deep neural network model for fraud datasets with variable distribution concerns. In addition, it leverages SHAP for the purpose of promoting fair and trustworthy decision-making. Last but not least, it gives strong experimental results that compare predictive performance to interpretability. All these aspects together signify a significant progress in the area of reliable, flexible, and explanatory fraud detection systems that can be applied to the current digital banking environment. In a manner of speaking, by combining the accuracy of predictions

with the transparency of interpretations, CredSHAP does not only remove the existing barriers in fraud detection but also sets the stage for the future—bigger and more complicated—issues to be solved.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Institute of Aeronautical Engineering for providing the resources and technical support required to carry out this research. We also thank the faculty members and mentors of the Department of Artificial intelligence & Machine learning for their valuable guidance, constructive feedback, and encouragement throughout the development of this work. Special appreciation is extended to the open-source community for making the credit card transaction dataset publicly available, which served as the foundation for evaluating the proposed framework. Finally, we acknowledge the contributions of our peers and colleagues whose discussions and insights greatly enriched the quality of this research.

REFERENCES

[1] Innocent Paul Ojo and Asna Tomy. Explainable AI for credit-card fraud detection. World Journal of Advanced Research and Reviews (2025).
[2] S. Selvam and M. Sughasiny. Smart and explainable credit-card fraud detection using SHAP. Journal of ISMAC, 7(2), 155–169 (2025).
[3] Unsupervised feature selection and class labeling for credit card fraud. Journal of Big Data, Springer (2025)..
[4] Addressing credit card fraud detection challenges with autoencoders. Big Data and Cognitive Computing, MDPI (2025).
[5] A hybrid deep learning ensemble model for credit-card fraud detection. Directory of Open Access Journals (2025).
[6] R. Jayalakshmi and T. Thanushree. Credit card fraud detection using stacking ensemble of deep learning models. International Journal for Research in Applied Sci ence and Engineering Technology (IJRASET) (2025).
[7] Y. Wang. A data balancing and ensemble learning approach for credit-card fraud detection. arXiv preprint (2025).
[8] M. Hosseini Chagahi et al. An innovative attention-based ensemble system for credit card fraud detection. arXiv preprint (2024).
[9] S. Kharat, S. Taur, R. Khalate, and K. Kate. Detection of credit-card fraud using machine learning and deep learning. International Journal of Progressive Research in Science and Engineering (IJPRSE) (2023).
[10] Z. H. Mohammed and G. J. Ahmed. Ensemble deep learning strategy for handling imbalanced credit card fraud data. Journal of Informatics, Network, and Computer Science (2024–2025).
[11] A. Tomar et al., "Real-time credit card fraud detection using LSTM and SHAP," IEEE Access, 2025.
[12] S. Kumar et al., "SMOTE vs Random Upsampling," Springer, 2025.
[13] L. Chen et al., "Deploying XAI models in banking," ACM Trans. Intell. Syst., 2025.
[14] R. Singh et al., "Hybrid DNN-RF ensemble," IEEE Trans. Neural Netw., 2025.
[15] M. Patel et al., "Kernel SHAP anomaly detection," IEEE Symp. Security, 2025.
[16] J. Liu et al., "Streaming fraud detection on Kafka," IEEE Big Data, 2025.