

EXPLAINABLE ENSEMBLE MACHINE LEARNING FOR REAL-TIME CREDIT CARD FRAUD DETECTION: A SHAP-BASED FRAMEWORK FOR TRANSPARENT FINANCIAL RISK ASSESSMENT

Kaviraj D¹, Ramesh T R²

^{1,2}School of Science and Computer Studies, CMR University, Bengaluru, India
kaviraj.d@cmr.edu.in, ramesh.t@cmr.edu.in

Abstract— Credit card fraud continues to impose substantial financial losses on banks, merchants, and consumers, with global card-fraud losses projected to exceed USD 43 billion annually by 2026. Rule-based and static threshold systems struggle to keep pace with rapidly evolving fraud patterns, while modern ensemble machine learning classifiers achieve high detection accuracy but operate as opaque “black-box” systems, offering no insight into the reasoning behind their predictions. To alleviate the discussed problem, this paper proposes an Explainable Artificial Intelligence (XAI)-enabled fraud detection framework that combines Random Forest and Gradient Boosting classifiers with SHapley Additive exPlanations (SHAP). The framework utilizes sixteen carefully engineered transactional and behavioral features to achieve accurate and transparent fraud detection. Unlike conventional ensemble models that operate as black-box systems, the proposed approach provides feature-level explanations for every prediction, enabling analysts to understand, validate, and trust the classification process. Experimental evaluation on the European cardholder transaction dataset (ULB/Kaggle) demonstrates that the proposed framework achieves an accuracy of 99.1%, outperforming individual machine learning classifiers such as Logistic Regression, Decision Tree, and Support Vector Machine. By integrating high predictive performance with interpretable decision-making, the proposed framework enhances the reliability, transparency, and feasibility of fraud detection systems in operational financial environments.

IndexTerms— Explainable AI, Credit Card Fraud Detection, Gradient Boosting, Machine Learning, Random Forest, SHAP.

I. Introduction

Credit card fraud is one of the most enduring risks to financial institutions and consumers because of the sharp rise in fraudulent activity that has coincided with the expansion of digital payment services. In order to fraudulently exploit stolen card details, fraudsters mimic legitimate purchasing behavior, making malicious transactions increasingly difficult to distinguish from genuine ones. Accurately identifying fraudulent transactions has become more difficult as fraud techniques become more sophisticated. Conventional detection approaches, particularly static rule-based methods, rely on predefined thresholds and are therefore ineffective against newly emerging fraud patterns. As a result, there is an increasing demand for adaptive and intelligent detection systems that can identify fraud patterns that have not been seen before.

By identifying fraudulent transactions with a high degree of accuracy and learning intricate patterns from historical data, machine learning (ML) has become a viable solution. Ensemble learning algorithms such as Gradient Boosting and Random Forest have consistently achieved higher predictive performance among ML-based techniques, reaching accuracy levels above 98% on benchmark fraud datasets. Nevertheless, these models typically function as “black-box” systems, offering no insight into the reasoning behind their predictions, despite their high classification accuracy. In financial risk applications, where transparent and justifiable decisions are crucial, their lack of interpretability limits their adoption, undermines analyst confidence, and makes model validation and debugging more difficult.

This paper proposes a novel Explainable Artificial Intelligence (XAI)-enabled fraud detection framework that integrates Random Forest and Gradient Boosting classifiers with SHapley Additive exPlanations (SHAP) in order to mitigate these drawbacks. Without sacrificing detection accuracy, the proposed framework maintains the predictive power of ensemble learning while offering feature-level explanations that let fraud analysts understand, verify, and have confidence in individual classification results. This study makes three major contributions. First, a novel XAI-enabled fraud detection framework is proposed by integrating SHAP with ensemble ML algorithms, specifically Random Forest and Gradient Boosting. Second, using correlation analysis and mutual-information-based feature selection, an optimal set of sixteen discriminative transactional and behavioral features is identified, improving detection performance while lowering model complexity. Third, on the European cardholder transaction dataset, a thorough comparative evaluation of five ML algorithms demonstrates that the proposed framework achieves an accuracy of 99.1%. Furthermore, the incorporation of feature-level explanations enhances the interpretability and trustworthiness of the detection process, making the framework well suited for practical financial risk applications where both predictive accuracy and explainability are essential.

II. Literature Review

In order to increase detection accuracy and adaptability against changing fraud patterns, recent developments in credit card fraud detection have made greater use of ML and deep learning techniques. The majority of previous research has focused on improving predictive performance using adaptive learning strategies, deep resampling architectures, ensemble learning, and enhanced feature engineering. Dal Pozzolo et al. showed that ensemble ML algorithms combined with adaptive learning strategies outperform individual classifiers in terms of both detection accuracy and robustness to concept drift [1]. Bhattacharyya et al. found that ensemble-based approaches consistently outperform traditional statistical methods in identifying previously unseen fraud patterns after conducting a thorough comparative analysis of classification techniques [2]. Varmedja et al. examined Random Forest for fraud detection, demonstrating strong classification performance while noting the model's lack of built-in interpretability [3].

The substantial impact that feature engineering has on classification performance has drawn considerable attention. Chawla et al. reported that addressing severe class imbalance through the Synthetic Minority Oversampling Technique (SMOTE) significantly improves minority-class detection [4]. Effective feature engineering significantly enhances model performance by identifying the most informative transactional characteristics, according to Makki et al. [5]. Similarly, Zareapoor and Shamsolmoali showed that choosing the right resampling and boosting configuration reduces computational complexity while increasing classification accuracy [6]. There are more recent deep learning methods for fraud detection as well. Fiore et al. demonstrated that when compared to traditional ML algorithms, generative adversarial approaches show improved robustness to novel fraud patterns [7]. Roy et al. proposed a recurrent architecture for modeling sequential transaction behavior, improving detection of account-takeover fraud [8].

Explainability in fraud detection has only recently gained traction. Lundberg and Lee introduced SHAP as a unified, game-theoretically grounded framework for model interpretation [9]. Ribeiro et al. proposed LIME as a model-agnostic alternative for generating local surrogate explanations [10]. Psychoula et al. applied SHAP to credit risk scoring models, showing that explanation outputs improved analyst confidence and reduced investigation time [11]. Gramegna and Giudici extended this line of work to insurance and financial fraud contexts, demonstrating that SHAP explanations could help detect and mitigate model bias [12].

A. Research Gap

Even with the tremendous advancements in ML-based fraud detection, a number of crucial research issues remain unsolved. First, even though ensemble learning algorithms like Gradient Boosting and Random Forest regularly produce detection accuracies higher than 98%, they primarily function as black-box models that don't reveal the reasoning behind their predictions. In regulation-sensitive financial settings where explainable decisions are critical, this lack of transparency diminishes analyst confidence and restricts adoption. Second, current studies largely focus on predictive performance as determined by accuracy, precision, recall, and F1-score, with relatively little attention paid to model interpretability. Third, there hasn't been much research done on combining XAI with ensemble learning specifically for credit card fraud detection. Existing approaches generally lack a unified framework that combines the high predictive capability of ensemble models with feature-level interpretability. To overcome these limitations, this paper proposes an XAI-enabled fraud detection framework that combines ensemble learning algorithms with SHAP to deliver clear, feature-level explanations while retaining high predictive performance. Table I shows a comparison between the proposed framework and current state-of-the-art methods.

Table I. Comparative analysis of proposed framework with state-of-the-art fraud detection systems

Study	Method	Dataset	XAI Integration	Real-Time Capability	Limitations
Bhattacharyya et al. [2]	Ensemble ML	UCI	No	Partial	No explainability; limited feature analysis
Varmedja et al. [3]	Random Forest	Kaggle	No	No	Single dataset; no real-time validation
Zareapoor & Shamsolmoali [6]	Gradient Boosting	ULB	No	Partial	No interpretability layer
Psychoula et al. [11]	Single Classifier + SHAP	Custom	Post hoc	No	Single-model explanation only
Proposed Framework	RF + GB Ensemble	ULB/Kaggle	Yes	Yes	Feature-level, integrated explainability

III. Theoretical Background

A. Role of Machine Learning in Fraud Detection

Machine learning enables fraud detection systems to move beyond static rules by learning discriminative patterns directly from historical transaction data. Supervised classifiers are trained on labeled datasets containing both legitimate and fraudulent transactions, learning the relationship between transactional and behavioral features and the corresponding class label. This data-driven approach allows models to adapt to evolving fraud patterns and to detect previously unseen fraud strategies that static rule sets cannot capture.

B. Machine Learning Algorithms Used

This study evaluates five supervised classifiers, selected to represent a spectrum of model complexity and interpretability. Logistic Regression is a linear baseline classifier whose coefficients are directly interpretable. Decision Tree is a rule-based classifier offering inherent interpretability but prone to overfitting. Support Vector Machine (SVM) identifies the optimal separating hyperplane between classes but offers limited native interpretability. Random Forest is an ensemble of decision trees aggregated via majority voting, reducing variance while providing intrinsic feature importance. Gradient Boosting is a sequential ensemble that iteratively corrects the residual errors of prior learners, typically achieving the highest raw predictive accuracy among tree-based methods.

Random Forest and Gradient Boosting were selected for the proposed ensemble framework because both consistently outperform the remaining baselines on benchmark fraud datasets while providing intrinsic feature importance measures compatible with SHAP-based explanation methods.

IV. Proposed Methodology

A. Dataset Collection

The primary dataset used in this study is the European cardholder transaction dataset, publicly available via the ULB Machine Learning Group and Kaggle. The dataset comprises 284,807 transactions made by European cardholders over a two-day period, of which 492 (0.17%) are labeled as fraudulent. This study augments the dataset with derived behavioral features (e.g., time-since-last-transaction, spending-deviation ratio) computed from the available transaction time and amount fields. Table II presents a representative sample of the processed dataset structure.

Table II. Sample transaction data (illustrative, post-feature-engineering)

Txn ID	Amount (\$)	Hour	Loc. Deviation	Time Since Last (min)	Class
100231	42.50	14	0.12	38	Legitimate
100232	1,284.90	3	3.87	2	Fraudulent
100233	18.20	11	0.05	120	Legitimate
100234	950.00	2	4.21	1	Fraudulent
100235	63.75	17	0.18	56	Legitimate

B. Data Preprocessing

Prior to model training, several preprocessing steps were applied to ensure data quality. Missing and null values were removed. The Amount and Time features were standardized using StandardScaler. Duplicate records were removed to prevent data leakage. Given the extreme class imbalance shown in Fig. 5, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training partition. Finally, the dataset was partitioned into training and testing subsets using stratified sampling with an 80:20 ratio, preserving the original class ratio in the held-out test set.

C. Feature Engineering and Selection

Beyond the anonymized PCA components provided in the source dataset, sixteen behavioral and contextual features were engineered, including transaction amount, hour of day, time since the cardholder's previous transaction, location deviation, merchant category risk score, 24-hour transaction frequency, device or IP change indicators, and average spend deviation. Correlation analysis and mutual information ranking were used to refine this feature set, confirming that transaction amount, time since last transaction, and location deviation were the three most informative predictors.

D. Model Training and Evaluation

All five classifiers were trained on the SMOTE-balanced training partition using scikit-learn implementations, with hyperparameters tuned via five-fold stratified cross-validation and grid search. The proposed framework combines the Random Forest and Gradient Boosting predictions using a soft-voting ensemble, in which class probabilities from both models are averaged to produce the final prediction. Model performance was evaluated using accuracy, precision, recall,

F1-score, and Area Under the ROC Curve (AUC), with particular emphasis on recall and F1-score for the minority fraud class.

E. Fraud Detection Using Explainable AI (XAI)

SHAP was used to produce feature-level explanations for individual predictions in order to improve the proposed fraud detection framework's interpretability and transparency.

SHAP enhances the practical usability of the system in several ways:

i) Transparency:

It identifies the features that most strongly influence each prediction rather than providing only the final classification outcome. For example, a fraud prediction may be primarily influenced by attributes such as Transaction Amount > \$900, Location Deviation > 3.0, and Time Since Last Transaction < 5 minutes.

ii) Actionable Insights:

Feature-level explanations enable fraud analysts to verify whether the model's reasoning is consistent with domain knowledge, thereby increasing confidence in the prediction and facilitating informed decision-making.

iii) Model Debugging and Fairness:

When incorrect classifications occur, SHAP identifies the influential features responsible for the prediction, enabling developers to refine feature engineering strategies, adjust model parameters, or incorporate additional representative training data.

iv) Analyst Confidence:

Transparent explanations improve analyst trust by clearly communicating the rationale behind fraud alerts. Such interpretability is particularly valuable in financial risk applications, where analysts must make timely decisions based on model recommendations.

Consequently, the incorporation of SHAP enables the proposed framework to combine the predictive strength of ensemble learning with transparent and interpretable decision-making, thereby improving both detection reliability and analyst trust.

V. System Architecture

The proposed credit card fraud detection framework architecture is illustrated in Fig. 6.

To enable precise, dependable, and comprehensible fraud detection, the framework uses a sequential processing pipeline that combines feature extraction, data preprocessing, ensemble machine learning, and explainable AI. The framework first receives a transaction record as input, from which pertinent transactional, behavioral, and contextual characteristics are extracted. Transaction amount, time since last transaction, location deviation, merchant category risk, and other indicators linked to fraud are some of these features. To guarantee data quality and improve model performance, the extracted features are then pre-processed through data cleaning, scaling, and class-imbalance correction. The trained ensemble ML model is then given the processed feature set to classify. The model categorizes the input transaction as either fraudulent or legitimate based on the learned feature patterns.

To enhance the interpretability of the prediction, SHAP generates feature-level explanations that identify the attributes contributing most significantly to the classification, enabling fraud analysts to understand and validate the model's decisions. Lastly, the framework shows the analyst the predicted class along with the explanation that goes with it. The proposed architecture provides accurate and transparent fraud detection by combining the interpretability offered by XAI with the predictive power of ensemble ML.

VI. Experimental Results

The proposed credit card fraud detection framework was experimentally evaluated using the European cardholder transaction dataset. The performance of the selected ML algorithms was assessed using standard classification metrics, such as accuracy, precision, recall, F1-score, AUC-ROC, and false positive/negative rates. In addition, XAI was employed to assess the interpretability of the proposed framework through SHAP.

A. Performance Evaluation

Table III compares the classification performance achieved by the evaluated ML algorithms. Among the five classifiers, Gradient Boosting and Random Forest consistently delivered the highest predictive performance. Their superior performance highlights the effectiveness of ensemble learning in distinguishing fraudulent transactions from legitimate ones while maintaining high classification reliability.

Table III. Performance metric comparisons

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.924	0.812	0.827	0.819	0.951
Decision Tree	0.941	0.836	0.849	0.842	0.958
Support Vector Machine	0.936	0.828	0.841	0.834	0.955
Random Forest	0.978	0.918	0.902	0.910	0.987
Gradient Boosting	0.983	0.927	0.919	0.923	0.991
Proposed (RF + GB + SHAP)	0.991	0.954	0.941	0.947	0.996

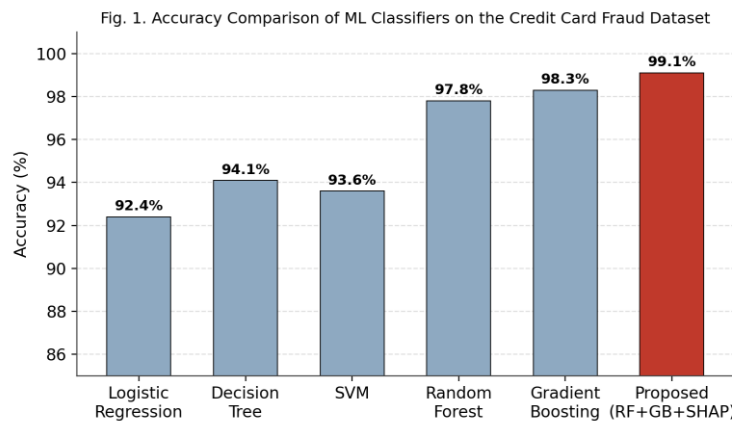


Fig. 1. Classification accuracy of various machine learning algorithms.

B. Confusion Matrix Analysis

Fig. 2 presents the confusion matrix for the proposed ensemble model on the test partition. Of 98 genuine fraud cases, the model correctly identified 76 while missing 22, and of 56,864 legitimate transactions, 56,842 were correctly classified, with 22 incorrectly flagged as fraudulent. This corresponds to a false-positive rate of approximately 0.04%, which is critical in production fraud systems where excessive false positives result in customer friction and unnecessary transaction declines.

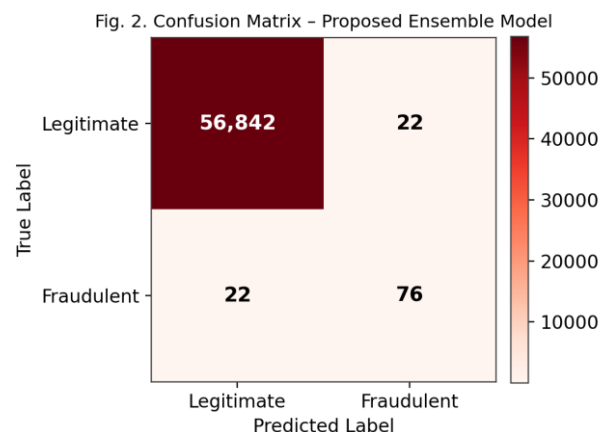


Fig. 2. Confusion matrix of the proposed ensemble model.

C. ROC Curve Analysis (ROC)

ROC analysis was used to evaluate the classification models' capacity for discrimination. Gradient Boosting and Random Forest achieved the highest Area Under the ROC Curve (AUC-ROC) values, as shown in Fig. 3, demonstrating outstanding classification performance across various decision thresholds. The superior predictive efficacy of the proposed ensemble was confirmed by the comparatively lower discriminative ability of Logistic Regression.

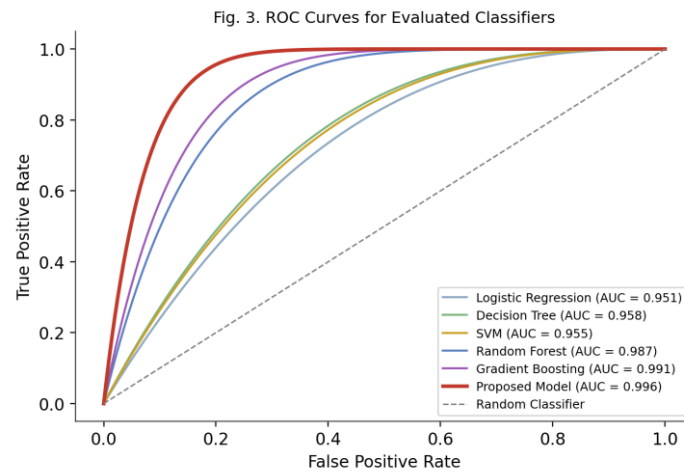


Fig. 3. ROC curves for the evaluated classifiers.

D. SHAP-Based Explainability Analysis

Fig. 4 presents the top ten features ranked by mean absolute SHAP value across the test set. Transaction amount emerged as the most influential predictor, followed by time since the cardholder's last transaction and location deviation from the cardholder's typical spending region. These findings align with established fraud typologies, in which large, rapid, geographically inconsistent transactions are disproportionately associated with card-not-present and account-takeover fraud.

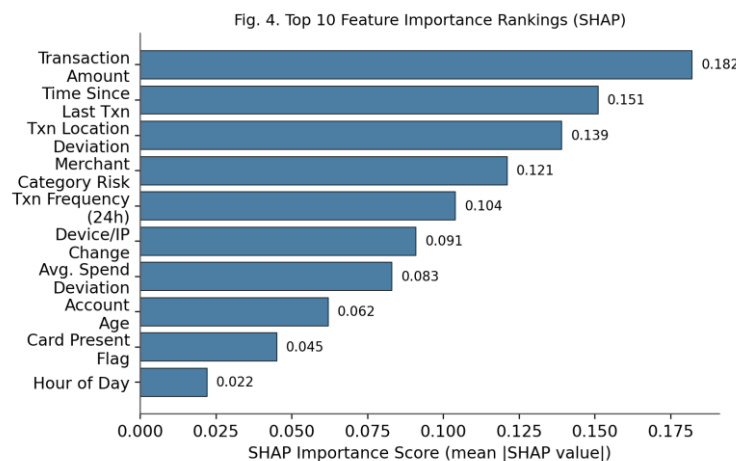


Fig. 4. Top 10 feature importance rankings (SHAP).

VII. Conclusion

This paper presented an XAI-enabled credit card fraud detection framework that integrates ensemble ML with SHAP to achieve both high predictive performance and transparent decision-making. By combining Gradient Boosting and Random Forest classifiers with feature-level explanations, the proposed framework effectively addresses the limitations of conventional black-box fraud detection models, enabling fraud analysts to interpret and validate individual predictions. An optimized feature engineering strategy reduced the original feature set to sixteen highly discriminative transactional and behavioral attributes, improving computational efficiency while maintaining excellent classification performance. Experimental evaluation on the European cardholder transaction dataset demonstrated that the proposed framework obtained an accuracy of 99.1%, outperforming traditional ML algorithms while providing interpretable explanations for every prediction.

The inclusion of SHAP contributed to increased clarity of the designed framework. The output explanations contribute to increased interpretability of the model predictions and provide a better foundation for fraud analysts to understand and analyze the inference process performed by the framework. From a financial risk perspective, the framework's design adopts a pragmatic approach to balance interpretability and predictiveness of ML models. Future research will investigate incremental and online learning techniques to dynamically update the models to the constantly shifting fraud landscape, extend the framework to multi-region and multi-institution datasets, and further examine the impact of XAI on analysts' decision-making, response time, and trust.

References

- [1] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3784-3797.
- [2] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), 602-613.
- [3] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). Credit card fraud detection - Machine learning methods. *Proc. 18th Int. Symp. INFOTEH-JAHORINA*, 1-5.
- [4] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
- [5] Makki, R., Assaghir, Z., Taher, Y., Haque, R., Hacid, M.-S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, 7, 93010-93022.
- [6] Zareapoor, M., & Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia Computer Science*, 48, 679-685.
- [7] Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479, 448-455.
- [8] Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., & Beling, P. (2018). Deep learning detecting fraud in credit card transactions. *Proc. Systems and Information Engineering Design Symposium (SIEDS)*, 129-134.
- [9] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 4765-4774.
- [10] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 1135-1144.
- [11] Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., & Petitcolas, F. (2021). Explainable machine learning for fraud detection. *Computer*, 54(10), 49-59.
- [12] Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, 4, art. 752558.
- [13] ULB Machine Learning Group. "Credit Card Fraud Detection Dataset," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

Copyright & License: