Journal of Information Systems Engineering and Management
2025, 10(24s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Credit Card Fraud Detection using Explainable AI Methods

Dr.S. Suriya[1], RM. Sireesha[2]
[1]*Associate Professor Department of Computer Science and Engineering PSG College of Technology, Coimbatore*
ss.cse@psgtech.ac.in
[2]*PG Scholar Department of Computer Science and Engineering PSG College of Technology, Coimbatore*
23mz02@psgtech.ac.in

ARTICLE INFO

Received: 19 Dec 2024

Revised: 10 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Explainable AI (XAI) system assists users in understanding the underlying processes of AI's decision making. XAI algorithms differ from conventional AI algorithms as XAI systems highlight decision-making processes and therefore can be regarded as trustworthy. Fraud detection should be precise in credit card transactions as the volume of global transactions is enormous. Most of these transactions are legitimate but an alarming amount of them are fraudulent. Detecting these fraudulent transactions enables banks and consumers to save enormous amounts of resources that would have otherwise been spent on compensation. Tools like Watson OpenScale by companies like IBM are designed to ensure that AI models are unbiased and transparent. The proposed project relies on the use of XAI methods such as LIME and SHAP designed to identify fraud in credit card transactions. LIME understands the reason an AI model made a decision and presents that rationale in a simplified manner. SHAP illustrates the transaction features like transaction amount or location and how these elements affect the model's choice. The aid of these XAI enabled methods improves the comprehension of the automated fraud detection systems and why certain transactions were unsuccessfully authenticated. Furthermore, we need to balance this dataset using SMOTE because there could be an imbalance between Legit and fraudulent transactions. XGBoost is great for large datasets which is why we will build our predictive model with that algorithm. The project merges XAI with powerful fraud detection approaches like SMOTE and hyperparameter tuning to builds a system that can be easily manipulated for its effectiveness.

Keywords: Explainable AI, Credit Card Fraud Detection, SHAP, LIME, XGBoost

1.INTRODUCTION

The growing prevalence of credit card fraud has made it increasingly important to verify and identify fraudulent transactions. Credit card usage has exploded in the last few years and the rise of eCommerce companies have led to an increase in fraudulent attempts. Millions of people and financial institutions fall victim to credit card fraud every single year. Unfortunately, the rise of fraudulent activity has made AI one of the primary tools used for spotting and identifying these frauds. The main issue with most AI systems today is that they operate as “black boxes” which keeps the general public in complete ignorance regarding what systems were implemented to arrive at their conclusions. On one hand, these systems are important for tackling fraud but these poorly managed AI systems can erode trust. This is where the magic of Explainable AI (XAI) comes in. XAI provides a way to decipher these challenges and make AI more user and trust-friendly. When it comes to fraud, XAI is especially valuable for rate setters and other officials of banks and financial institutions that need to know the credit card transactions flagged as suspicious and why they are flagged. It is much more difficult for these officials to trust the AI or act on its recommendations if they cannot see or hear how it reached its decision. Shapley Additive Explanations (SHAP) contributes assistance by revealing which factors like the amount of the transaction or the place of the transaction affected the decision the most. This renders a logical explanation on why the fraud flag was raised. LIME or Local Interpretable Model-agnostic Explanations does this by deconstructing AI decisions into simpler granular individual transactions and clarifying what drove the model’s decision. From a multi-dimensional context, this process helps in trusting the system better but, in particular, allows for understanding of the reasoning especially when the consequences are dire, such as in finance. With directives like the GDPR pointing towards a clearer need for the ability to explain and justify actions

taken, the use of XAI SHAP and LIME in fraud detection is not to improve the technique, but to make it easier for people to comprehend and depend on it as they go about their ordinary business.

2.MOTIVATION

Source: https://www.livemint.com/industry/banking/debit-credit-card-frauds-on-rise-atm-scams-down-ncrb-11661885877307.html

This fig 1 & 2 illustrates the growing trends in financial fraud and cheating cases from 2017 to 2021. The data, sourced from the National Crime Records Bureau (NCRB), underscores the urgency for robust fraud detection mechanisms.

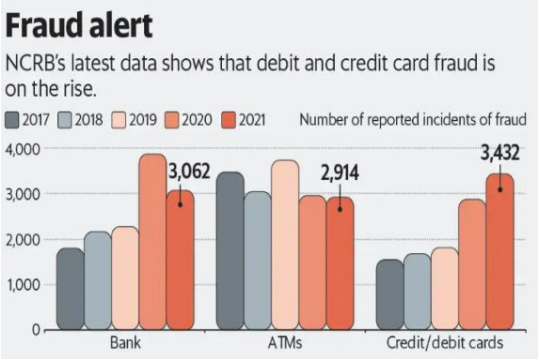

Figure 1. Trends in Reported Debit and Credit Card Fraud Incidents (2017–2021)

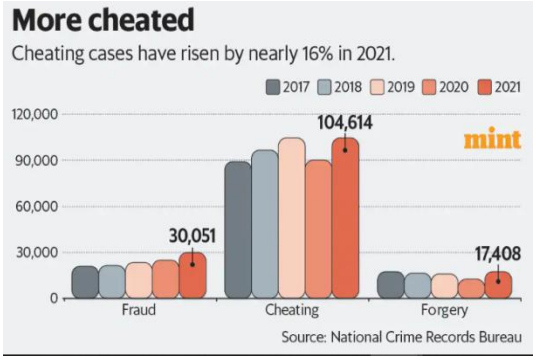

Figure 2. Yearly Trends in Fraud, Cheating, and Forgery Cases (2017-2021)

Credit and debit card fraud increased significantly, reaching a peak of 3,432 instances in 2021, as shown in Figures 1 and 2. The need for improved detection systems is highlighted by the 16% increase in cheating incidents. The alarming rise in credit card theft highlights the urgent need for innovative detection technologies that strike a compromise between efficacy and transparency.Explainable Artificial Intelligence (XAI) is revolutionary because it can identify suspicious activity and give clear justifications for transactions that are reported. This promotes consumer and financial institution trust, allows analysts to make better informed decisions, and guarantees equitable fraud detection practices. XAI allows us to respond to changing fraud patterns, create more transparent and ethical systems, and restore trust in digital financial transactions, paving the way for a safer and more secure financial environment.

3.OBJECTIVES

This research aims to develop a credit card fraud detection system based on Explainable AI (XAI) techniques. To sharpen the focus on offering explanations, SHAP and LIME are implemented. The goal of the system is to improve the transparency and trust given the methods employed in fraud detection be explaining many outcomes to end-users and financial institutions. To ensure comprehensive analysis, the model will also be tested and evaluated on other measurement metrics which include accuracy, recall, precision, F1 score, and AUC-ROC. To resolve the class imbalance problems and increase the accuracy of the model in fraudulent transaction detection, SMOTE technique will be applied. Moreover, the adoption of feature engineering methods will lead to maximized fraud detection accuracy and minimum computing resources.

4.RELATED WORK

This paper [1] develops Explainable AI (XAI) and Federated Learning (FL) for improving fraud detection in automated insurance systems without breaching data privacy. Training and testing of the model have been done with a large set of anonymous insurance transactions. Through the use of techniques such as SHAP and LIME, the AI model's rationale for its decisions is systematic and interpretable. Federated learning emboldens the model against proposals on decentralized data without the need to centralize it, protecting data privacy. The result indicates that the model is successful in detecting false insurance claims with justifications to each choice in favor of trust and adherence to data privacy laws. This represents a sound and private means of identifying fraud in the insurance sector. Graph representations of financial transactions in the dataset appearing in [2] are developed to illustrate the relationships and interactions among the entities involved. The description of GNNs as a way of assessing suspicious

activities found by pattern recognition will evaluate these graphs in order to spot frauds. The results clearly show that the graph learning approach helps improve fraud detection performance significantly compared to existing ones. This approach is of great use in large schemes with several entities and hidden relationships. Thus, this study concluded that graph learning can serve as a valuable tool for improving fraud detection in financial online systems by boosting the robustness and precision of investigations. Detecting fraudulent activities in less frequent economic transactions is what is dealt with in this research work. Databases contain transaction statistics which are highly infrequent thus rendering them harder to analyze by traditional methods. The implementation employs some familiar machine learning algorithms developed to handle low-frequency data that will recognize anomalies that suggest fraud. The authors show how this model leads to better detection accuracy and reduced false positives-a valuable solution for financial institutions to protect low-volume transactions from fraud. The author emphasis the necessity of having special approaches for those scenarios of fraud detection and offers an application-oriented solution to deal with these system-specific challenges in finance. The area of focus is discrimination of fraud in financial statements; a systematic literature review of machine learning and data mining techniques applied in various ways in this area is to be developed. The set of sample financial reports and statement records analyzed is rather diverse, ranging from one kind to another. In this regard, several models for machine learning and data mining have been applied and analyzed for purposes of detecting fraud in financial statements. These results illustrate that intelligent algorithms are more accurate in detecting financial statement fraud than conventional techniques. The other thing this research assists with is those researchers or practitioners who are interested in the current state of machine learning for financial statement fraud detection-further enrichment regarding the rise of possibilities for AI-based solutions. The final output says that mixing data mining and machine learning approaches is an upcoming solution to overcome challenges in financial statement fraud detection and prevention. [5] emphasizes on the use of AI-based unified framework for fraud detection and prevention within the US banking sector. The dataset is composed of real-world banking transaction data retrieved from different financial institutions. This combines multiple AI techniques for the detection and prevention of numerous undercover schemes. The results indicate that the framework helps in the detection of fraudulent activities and reduce financial losses. It demonstrates the feasibility of AI-driven models in applying better banking fraud detection systems, providing robust and unified measures against fraud. [6] investigated the use of explainable artificial intelligence techniques in credit card fraud detection, placing a core focus on developing interpretable models benefiting from transparent decision-making. The dataset consists of credit card transaction data from bank and other financial institutions from the USA. Implementation of XAI techniques-shap and lime-were used to throw light on Zthe decisions for the machine learning models' predictions. The result shows detection accuracy is improved and trust level heightened among the stakeholders thanks to transparent and interpretable decisions. The study concludes that explainable AI improves compliance with regulatory standards and fosters greater trust in fraud detection systems. In their paper, [7] speaks of the use of Explainable AI in card fraud detection, which is concerned mostly with the generation of interpretable models for transparent decisions. The dataset comprises transaction records from financial institutions in the USA. XAI methods, like SHAP and LIME, are used to explain the decisions or judgments made by the ML models. They prove that the improved detection accuracy led to greater trust in stakeholders by making sure that the decisions were transparent and interpretable. It was concluded in this paper that explainable AI provides conformance with regulatory compliance and with a structure of trust toward fraud detection systems. [8] sets up a model for detecting fraud in financial transactions according to principles of explainable AI in combination with machine learning and deep neural nets. The data set contains transaction records from financial institutions in the USA. The implementation uses deep neural networks and explainability methods, such as SHAP and LIME, to explain model decisions. The results show increasing detection accuracy with a decrease in false positives. The paper compares four XAI techniques—SHAP, LIME, ANCHORS, and DICE—for their ability to interpret dense neural networks for credit card fraud detection. The dataset encompasses records of credit card transactions, and uses a dense neural network for fraud detection. The results show that every XAI method possesses differential strengths regarding explanation of the model's decision. It also gives an insight into the applicability of different XAI methods followed by financial fraud detection systems. Counting fraudulent transactions, participants of the offer use novel silent approaches to sift through the massively imbalanced fraud datasets by data cleaning and unsupervised learning. The dataset consists of large-scale, highly imbalanced credit card transaction records. With implementations, iterative cleaning processes improve the quality of the data, with unsupervised learning algorithms employed to detect fraud. Results demonstrate that significant models clearly protect the degree of resistance against other traditional methods in both accuracy and speed. Credit card transaction data will be collected, and the ensemble voting machine learning model proposed in [11] will be used for

analysis of credit card fraud detection. The dataset comprises real-world credit card transactions in which there is a huge imbalance between fraudulent and genuine purchases. The methodology enhances false-positive rate and detection accuracy with voting on the different classifiers' output. The findings demonstrated a shockingly better detection performance in the analysis of unbalanced data. A clever approach for credit card fraud detection is given in [12], where improved Light Gradient Boosting Machine (LightGBM) is used. The dataset includes credit card transaction records and compromises that work with an optimization when implemented on LightGBM for performance enhancement. The findings quantitatively state that, in the end, the improved model turns out superior in terms of accuracy and speed of operation, which in itself proves the worth of relying on it as an efficient means for fraud detection. [14] explores an approach that integrates explainable artificial intelligence (XAI) with deep learning models for predicting credit card defaults with an aim of making credit scoring models more interpretable. The dataset incorporates credit card default records, and the implementation itself blends explainable artificial intelligence techniques with deep learning in order to provide more accurate predictions along with interpretability. Results testify to enhanced transparency and decision-making in credit scoring. [15] discusses the role of explainable AI (XAI) and federated learning (FL) for the enhancement of transparency and privacy in financial fraud detection systems. The dataset comprises financial transaction records, and its implementation includes XAI techniques along with federated learning in order to provide interpretable fraud detection models while ensuring data privacy. Results show that this approach increases detection accuracy while conforming to data privacy regulations, rendering it suitable for use in financial institutions.

3.1 MINDMAP OF THE RELATED WORK

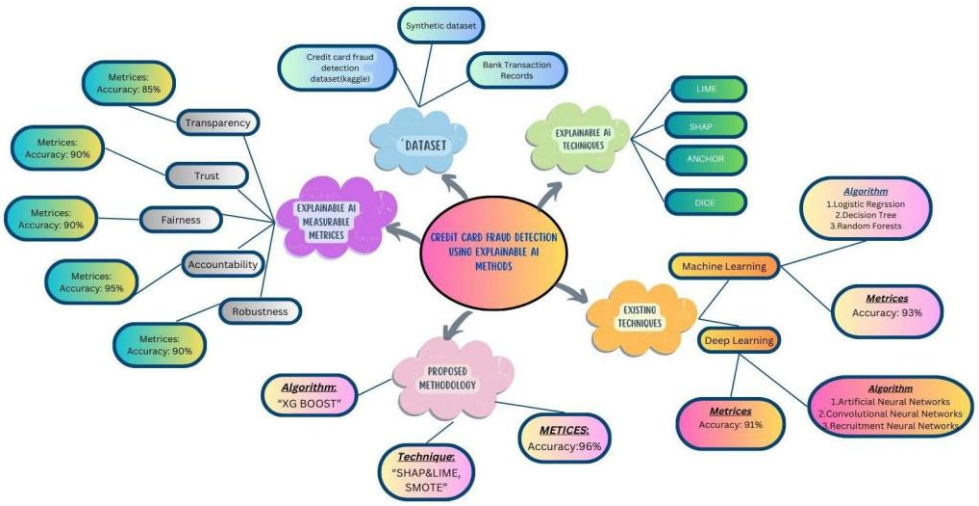

Fig 3. Mindmap of the related work

Inferences from the mindmap

The mindmap rendered in fig 3 gave rise to a few salient inferences. Specifically, they could be considered as the methods like shap, lime, anchors, dice, and suchlike to display the particular benefit and its application to the interpretation of machine learning. One is to say something about how good smote is for the rebalancing act, while further sophisticated measures might include graph learning, and hybrid models could be ingenious for the detection of advanced schemes for fraud. One such area that arose could be a potential convergence with federated learning and the blockchain approach, with the options considering how it could keep matters private and reasonably robust in the case of fraud detection. Lastly, the mindmap provides enough insights to swing between performance and interpretability, to foster compliance and instill trust among the parties involved, such as compliance officers and financial analysts. These insights provide a scaffold for designing trustworthy and efficient fraud detection systems.

4.METHODOLOGY

4.1 DATASET

Description: The dataset obtained from Kaggle has recorded credit card transactions performed by European cardholders over two days (September 2013). It consists of 284,807 transactions out of which 492 cases are marked as fraudulent, making for only about 0.172% of the dataset. The challenge of noticing fraud is mainly due to the

huge imbalance. This dataset has been prepared as an anonymized one with Principal Component Analysis (PCA), with all relevant input variables being numerical. It is most widely used for crafting machine learning models that are fairly good at exception detection.

SOURCE: KAGGLE

[41]:

	trans_date_trans_time	merchant	category	amt	city	state	lat	long	city_pop	job	dob	trans_r
0	2019-01-01 00:00:44	Heller, Gutmann and Zieme	grocery_pos	107.23	Orient	WA	48.8878	-118.2105	149	Special educational needs teacher	1978-06-21	1f76529f8574734946361c461b024
1	2019-01-01 00:00:51	Lind-Buckridge	entertainment	220.11	Malad City	ID	42.1808	-112.2620	4154	Nature conservation officer	1962-01-19	a1a22d70485983eac12b5b88dad1
2	2019-01-01 00:07:27	Kiehn Inc	grocery_pos	96.29	Grenada	CA	41.6125	-122.5258	589	Systems analyst	1945-12-21	413636e759663f264aae1819a4d4
3	2019-01-01 00:09:03	Beier-Hyatt	shopping_pos	7.77	High Rolls Mountain Park	NM	32.9396	-105.8189	899	Naval architect	1967-08-30	8a6293af5ed278dea14448ded268
4	2019-01-01 00:21:32	Bruen-Yost	misc_pos	6.85	Freedom	WY	43.0172	-111.0292	471	Education officer, museum	1967-08-02	f3c43d336e92a44fc2fb67058d594
...	...	...	...	...	...	...	...	...	...	...	...	
339602	2020-12-31 23:57:56	Schmidt-Larkin	home	12.68	Wales	AK	64.7556	-165.6723	145	Administrator, education	1939-11-09	a8310343c189e4a5b6316050d2d6b
339603	2020-12-31 23:58:04	Pouros, Walker and Spence	kids_pets	13.02	Greenview	CA	41.5403	-122.9366	308	Call centre manager	1958-09-20	bd7071fd5c9510a5594ee196368ac
339604	2020-12-31 23:59:07	Reilly and Sons	health_fitness	43.77	Luray	MO	40.4931	-91.8912	519	Town planner	1966-02-13	9b1f753c79894c9f4b71f04581835
339605	2020-12-31 23:59:15	Rau-Robel	kids_pets	86.88	Burbank	WA	46.1966	-118.9017	3684	Musician	1981-11-29	6c5b7c8add471975aa0fec023b2e8
339606	2020-12-31 23:59:24	Breitenberg LLC	travel	7.99	Mesa	ID	44.6255	-116.4493	129	Cartographer	1965-12-15	14392d723bb7737606b2700ac791b

339607 rows × 15 columns

Figure 4. Sample records of the dataset

The kaggle credit card fraud detection dataset, shown in fig. 4, summarized anonymized transactional data that was studied to look for fraudulent activities. This dataset included transaction time, amount, merchant, and its category, as well as geographic and demographic processes such as city population, job and date of birth. It offers valuable data across various categories including grocery shopping, entertainment, and travel, and supports various analytical processes. Its dataset of labeled fraudulent and legitimate transactions makes it suitable for supervised learning models for fraud detection. Moreover, the dataset is approached with rich contextual background that allows for explainable ai approaches aimed at better fraud identification and interpretation.

4.2 DATA PREPROCESSING

4.2.1 SMOTE: Smote (synthetic minority over-sampling technique) is one of a number of ways to oversample minority classes and aims to mitigate class imbalance in a dataset. Rather than duplicating existing instances, smote synthesizes new samples belonging to the minority class. It generates new data points in between the line connecting the minority data point to its nearest neighbors by interpolating them. In scenarios such as fraud detection where fraudulent transactions constitute very little in number than legal transactions, smote is indispensable. Smote consequently increases the recall and f1 score and balances the dataset while minimizing overfitting due to simple duplication of samples, which helps in detecting anomalies. Most often used in anomaly detection, healthcare, or finance, smote enhances the performance of machine learning models when applied to unbalanced datasets.

4.2.2 Key steps in SMOTE include:

1. Balancing the Dataset: The number of examples of the minority class becomes equal to the other good classes; hence class depends more on the performance of the other good classes due to generation of various classes.
2. Identifying the Minority Class: Identifying the lesser represented category, fraudulent Transactions.
3. Creating Synthetic Samples: For SMOTE, one point from the minority class is selected along with its k-nearest neighbors. A synthetic data point is generated along the line connecting one of these neighbors to this original data point.

4.2.3 Before and After SMOTE Analysis

4.2.3.1 Before SMOTE:

1. Class imbalance is significant, with non-fraudulent transactions (majority class) vastly outnumbering fraudulent ones (minority class).
2. Example from the document: "337,825 non-fraudulent transactions vs. 1,782 fraudulent transactions."

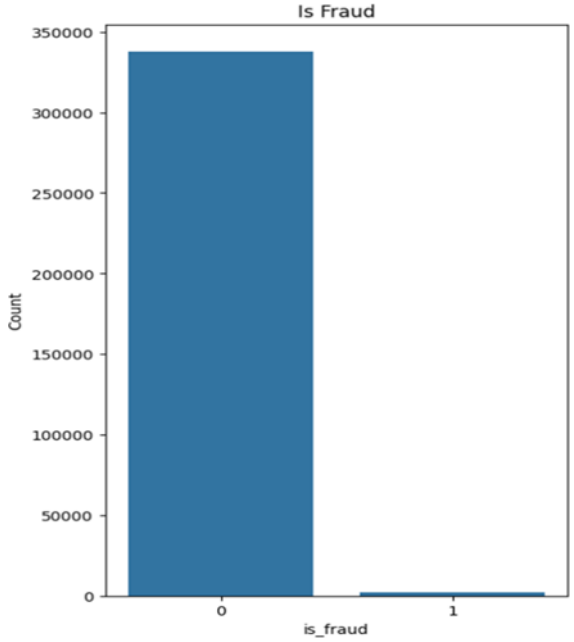

Figure 5. Before SMOTE

Figure 5 provides the class distribution of the dataset, which highlights the excessive imbalance between the majority class (non-fraudulent transactions) and minority class (fraudulent transactions). The majority of the dataset consists of 337,825 non-fraudulent classes, while only a tiny portion consists of the fraudulent class. Such an outstanding imbalance can hamper the ability of the machine learning model to identify fraudulent transactions properly because it may become biased towards predicting the majority class. This shortcoming can be overcome by utilizing techniques like SMOTE (synthetic minority oversampling technique) for achieving a balanced dataset in order to enable the model to detect instances of the minority class properly.

4.2.3.2 After SMOTE:

Through the application of Synthetic Minority Oversampling Technique, or SMOTE, the dataset has been balanced to provide approximately half the representation of the minority class and majority class. As an example, the dataset finally contains 270,270 non-fraudulent and 270,270 fraudulent transactions. This removes the bias and helps the model improve on the accurate detection of fraud. In fact, SMOTE allows the model to be less prone to overfitting for the less well-specified data for fraudulent transactions by creating synthetic samples on the minority class. SMOTE ensures that the decision boundary learned by the model does not necessarily favor the majority class, thus increasing the generalization capacity. The balanced dataset can then yield improved performance metrics, such as higher recalls and precisions for the minority class, and thus enables downstream explainability techniques to become better performers with interpretable models, bearing no biases.

4.2.4 Feature Engineering

1. Handling Imbalance: SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the dataset.
2. Feature Engineering: Anonymized features were retained, and categorical variables were label-encoded.

4.3 MODEL TRAINING

XGBoost(Extreme Gradient Boosting) is a very powerful, very efficient, and highly robust machine learning algorithm for classification and regression tasks. This algorithm is based on the gradient boosting framework. Most applications of the algorithm are characterized by developing fast models with maximum performance. The algorithm

builds an ensemble of decision trees in a sequential style such that at every step the current tree tries to correct the errors of the previous ones. To reach its final form, XGBoost performs gradient descent on the loss function, which allows it to work extremely well in iterations for arriving at a fairly accurate model.

XGBoost is a good alternative for producing models for large, complex datasets mainly due to its ability to handle missing data, regularization technique(s) to avoid overfitting, and offer parallel computation. Thus, the algorithm builds a series of decision trees, with each tree getting better at reducing the residual errors from the previous trees. It combines the predictions of these trees together to form its strong learner. With tree pruning, sparsity handling, and the estimator weighting of the classes, the XGBoost is good for imbalanced datasets. It benefits from the use of a regularized objective function that allows it to further improve performance and generalizability. We selected the XGBoost classifier for its robustness and high performance.

The model was then trained using 80% of the dataset, while the remaining 20% was used for testing the model. Using grid search, hyperparameter optimization was embraced to achieve the best combination of learning rate, maximum depth of the tree, and number of estimators, so that we would achieve optimal performance with minimal overfitting. Moreover, class imbalance was resolved with scale_pos_weight to account for uneven distribution of fraudulent vs non-fraudulent transactions.

4.3.1 Analysis of Correlation Matrix

A correlation matrix was generated to detect the correlations between the dataset attributes. Highly correlated attributes were examined and considered for exclusion to improve model efficiency. Most correlations identified by the matrix relate to the anonymized attributes like V1 through V5, along with their interaction with the amounts involved in the transactions' activities. Such analysis in its determination helped with feature selection and prioritization of important predictors for fraud detection, as depicted in Fig. 6. Correlation matrix, or relationships in correlation, is a statistical measure to establish linear relationships between any two numerical input attributes in a dataset. These values range from -1 to +1 for each pair in the correlation matrix and indicate the enzyme activity direction.

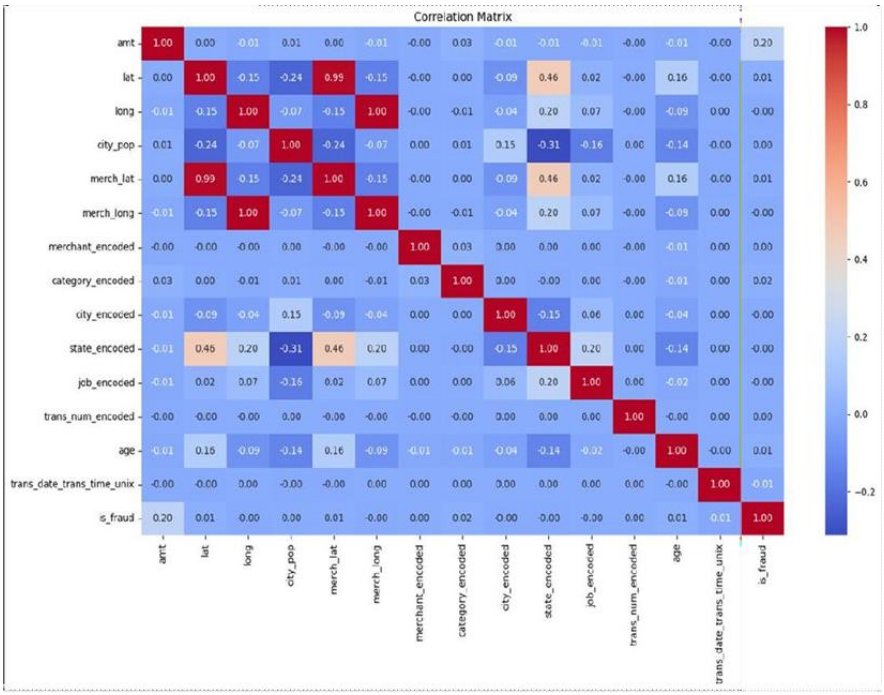

Figure 6. Analysis of Correlation Matrix

1. A value of "1" means that there is an absolute positive correlation; that is, when one variable increases, the other also increases.
2. A value of "-1" shows that there is a perfect negative correlation; that is, when one variable increases, the other decreases.
3. A value of "0" represents no linear relationship between the variables.

By analyzing the correlation matrix, features with strong correlations to fraudulent transactions were

identified, enabling more focused and effective fraud detection modeling.

4.3.2 Hyperparameter Optimization

Hyperparameters are parameter settings for a machine learning model that are specified before the training process and that cannot be trained directly from the data. Unlike model parameters-such as weights in a neural network-hyperparameters shape how the model learns and performs. Examples of hyperparameters include the learning rate, number of estimators, and maximum depth of a tree. Hyperparameters can influence the overall performance of the model: generalization capability and training efficiency. Once hyperparameters are tuned properly, there will be no risk of overfitting with respect to training data nor any underfitting with respect to unseen data. A fine-tuned model strikes a good balance between simplicity and complexity for accurate predictions on a new, real-world dataset.

The following hyperparameters were optimized during the grid search:

1. Learning Rate: This was tuned from 0.01 to 0.3 to regulate the step taken in the updates.
2. Maximum Tree Depth: The depth was explored from 3 to 10 for balance between model complexity and overfitting.
3. Number of Estimators: Fine-tuned between 100 and 500 to find the best number of boosting rounds.
4. Scale_Pos_Weight: Setting to lessen the class imbalance; the values range from 1 to 50.

Parameters	choices	Best parameter
n_estimators	50, 100, 150, 200	150
max_depth	3, 5, 7, 9	7
learning_rate	0.01, 0.1, 0.2, 0.3	0.1
subsample	0.5, 0.7, 0.9	0.7
colsample_bytree	0.3, 0.5, 0.7	0.5
gamma	0, 0.1, 0.2, 0.3	0.1
scale_pos_weight	1, 2, 5	2
min_child_weight	1, 3, 5	3
max_delta_step	0, 1, 2	1

Table 1: Highlights of the optimized hyperparameters, ensuring robust performance against class imbalance.

The selection of these hyperparameters improved both model accuracy and interpretability, ensuring robust detection of fraudulent transactions.

4.4 EXPLAINABLE AI TECHNIQUES

4.4.1 LIME (Local Interpretable Model-Agnostic Explanations)

Local Interpretable Model-agnostic Explanations-Surface Any complex machine learning model (like XGBoost), LIME gives intermediate interpretation for every single prediction made by such models by approximating them locally by a less complicated and well-interpretable model.

1. Perturbation of the Input Data Input Data

What LIME does is modify input data by perturbing it, thus forming a set of copies or modified instances surrounding the original data point. If for example the model is predicting whether a transaction is fraudulent, LIME will change features, like transaction amount, merchant location, or time of day, during which the transaction took place, thus getting perturbed instances to work with.

2.Weighting Similarity

These modified data points are then weighted according to their similarity to the original instance. The model focuses more on the perturbed examples that are closer to the original data and less on those that are farther away. This ensures the explanation is localized to the specific prediction.

3.Training a Surrogate Model

Lime, hence, returns perturbed weighted instances, to which the next step is training a surrogate model-a simple, interpretable model, usually a linear model or decision tree-using screwable models produced. This model has really acted like a surrogate for the complex model's behavior in the local neighborhood surrounding the original data point.

4. Interpreting the Surrogate Model

A surrogate model is usually simple enough that it can quite easily be understood. It gives clear explanations, like which features, for example, transaction amount or merchant location, governed the model's prediction. It uncovers the reasoning as to why a complex model made a certain prediction, like in this case, declaring a transaction fraudulent.

These are some prominent characteristics of LIME:

1. Model-agnostic: It means that LIME can be deployed on any machine learning model and be used in a number of different ways.
2. Local Explanations: It explains an individual prediction so as to give interpretations for why a particular decision is made.
3. Interpretive Explanations: LIME approximates complex models through simpler, interpretable approaches to further understand and clarify the black box models.

4.4.2. SHAP (Shapley Additive Explanations).

SHAP (SHapley Additive exPlanations) provides a framework for interpreting predictions by calculating the contribution of each feature to the model's output using concepts from cooperative game theory.

1. *SHAP SCORES: Global and Local Significance via SHAP Scores*

SHAP scores represent both global feature importance and local explanation for individual dataset predictions. Global feature importance offers a high-level picture of the features that are most significant across the dataset, whereas local explanation gives an account of how much each feature contributes to the specific prediction.

2. *SHAP Values: Feature Contributions*

SHAP values tell us how a model output varies if a particular feature is included or not in the dataset. These values represent how much a feature changes the prediction in relation to the baseline or expected value, giving information about the contribution of that feature to the model's decision.

3. *SHAP Values: Positive vs. Negative Contributions*

As per the model evaluation made above, positive SHAP values mean that the presence of some feature contributes to a greater probability of an outcome (for instance, fraud); such a feature favors fraud detection. Conversely, negative SHAP values indicate that the feature will work against the particular outcome and favor a non-fraud decision.

Key Features of SHAP

1. Cooperative Game Theory: SHAP uses the principles in cooperative game theory to fairly allocate the contribution of each feature to the prediction.
2. Global and Local Interpretability: SHAP provides global interpretation of feature importance as well as local explanation of any single prediction.
3. Impact of Features: The SHAP value explains the positive impact or negative impact given by each feature to the decision of the model.

4.5 SYSTEM ARCHITECTURE

The architecture for fraud detection integrates multiple components, including data preprocessing, model training, explainability layers, and decision-making interfaces. Real-time detection and batch processing pipelines are established to ensure system robustness.

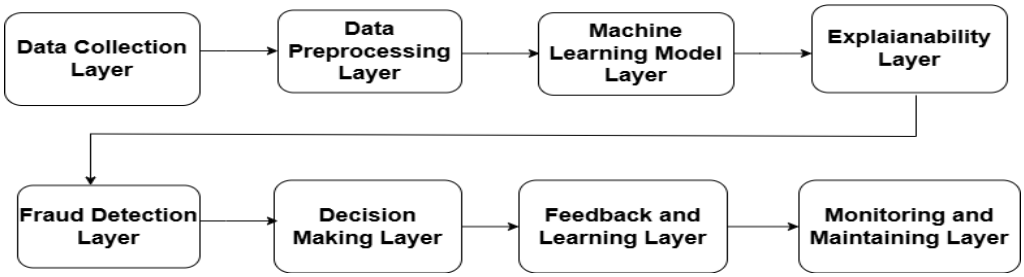

Figure 7. System architecture

1. Data Collection & Preprocessing:

Gather and clean transaction and user data, handling missing values and inconsistencies to ensure high-quality data for model training.

2. Machine Learning & Explainability:

Train fraud detection models (e.g., XGBoost) and use explainability tools like LIME and SHAP to provide insights into model predictions.

3.Fraud Detection & Decision-Making:

Deploy models for real-time and batch fraud detection, sending alerts and explanations to analysts for verification and decision-making.

4. Feedback, Monitoring & Maintenance:

Collect feedback for continuous model improvement, monitor system performance, and maintain regular updates to keep the system efficient and effective.

5.RESULTS AND ANALYSIS

5.1 MODEL PERFORMANCE

On the test dataset, XGBoost demonstrated the following performance metrics:

1. Accuracy 96.64%: the overall percentage of total predictions (fraud and legitimate) that were correctly classified.
2. Precision 13.12%: the probability of the model identifying actual fraudulent transactions among the ones predicted as fraudulent.
3. Recall 92.92%: shows how successful the model is in catching most of the fraudulent transactions.
4. F1-Score 22.99%: gives a balance between precision and recall.
5. AUC-ROC: 94.79% means very good discrimination between fraudulent and legitimate transactions.

These metrics show the model effectively discriminates between fraudulent and legitimate transactions despite the large class imbalance. The model also showed robustness, having been tested in all folds of a cross-validation, which demonstrates consistency in performance across various scenarios.

5.2 EXPLAINABILITY RESULTS

5.2.1 GLOBAL SHAP :

Global SHAP summary plots captured feature importance across the entire dataset, identifying "Transaction Amount" and "Feature V12" as the top predictors of fraudulent activity. The summary plot provided a clear visualization of how feature values influenced predictions, helping to uncover patterns like the role of transaction time in distinguishing fraud cases. These insights offer strategic value for refining fraud detection strategies and improving model performance. The plot shows which features have the greatest influence onfraud detection across all transactions in the test set. This fig 8. demonstrates a sample SHAP output for an individual transaction.

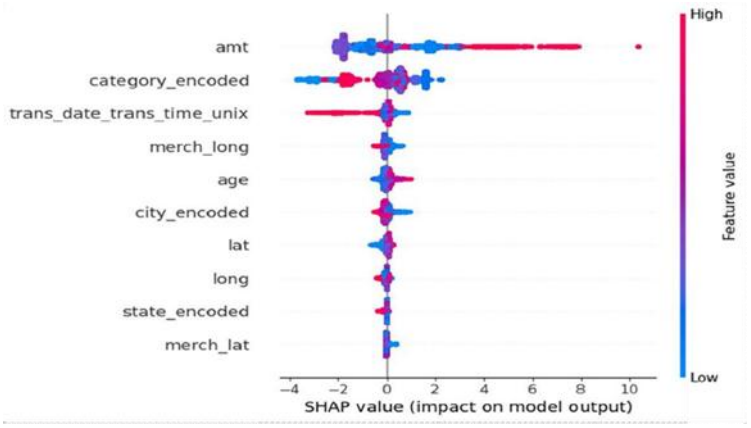

Figure 8. Global shap plot reveals feature contributions, with red indicating positive impacts and blue denoting negative influences.

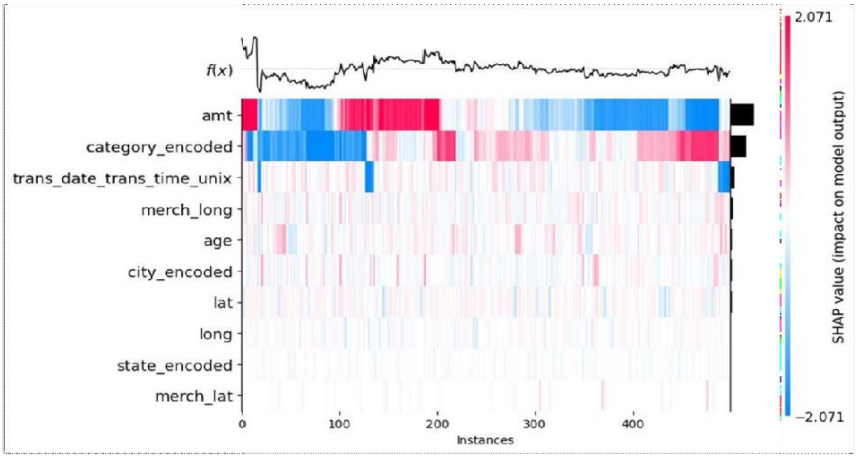

Figure 9. The plot reveals feature contributions, with red indicating positive impacts and blue denoting negative influences.

1. Feature Importance: shap summary graphs reveal the major contributing features such as "transaction amount" and "feature v12", whose effects on the model's predictions are extensive.
2. Positive and Negative impacts: red bars signify input variables with upward movements in predictions, while blue bars reflect downward movements in predictions, and the totop gives a measure of both direction and intensity of percussion by which they affect the model output.
3. Model Insights: the visualization outlines patterns such as the contribution of the transaction time in differentiating between fraud cases, which will help improve the course of model optimization.
4. Aggregated View: this summary plot serves to aggregate all shap instances and describe how they relate holistically to the contributions of the features…

5.2.2 LOCAL SHAP

The Local shap visualizations give explanations per transaction, which allows analysts to understand the contributions of each feature for every prediction. Such as above, a shap force plot for a flagged transaction would say that high transaction amount and unusual geographic location were its significant contributors to fraud classification. Model prediction verification by analysts and prioritization of cases for further scrutiny is facilitated by such insights. Local shap visualizations can provide a transaction-specific basis for decisions and accordingly allow analysts to understand the contributions of several features for each prediction. For example, a shap force plot for some flagged transactions would indicate that high transaction amounts and unusual geographic locations were significant contributors to fraud classification. Such insights will then empower an analyst to verify model predictions or prioritize cases for further investigation.

Figure 10. Demonstrates a local shap output for an individual transaction.

Local shap visualizations offered transaction-specific explanations, e.g., the force plot for instance_idx=8. Important components of this are:

- Base value (-3.728): the model average prediction over all cases serves as the initial value before contributions from features.
- Model output (f(x)) (-4.35): the model's prediction for this instance now altered by the feature values. The positive contributors: ded = 461, city_encoded = 27, age = 88.48 have increased the prediction, while the negative contributors: amt = 43.71, job_encoded = 12 have reduced it. Blue arrows indicate the lowering features of the prediction below the base value.

5.2.3 LIME: The Limetabularexplainer uses the training data (x_train.values), points to the type of job as classification (mode="classification"), and specifies the feature names (feature_names=features) and class names (class_names=['is_not_fraud','is_fraud']). Lime will explain by carving interpretable, local surrogate models around individual predictions.

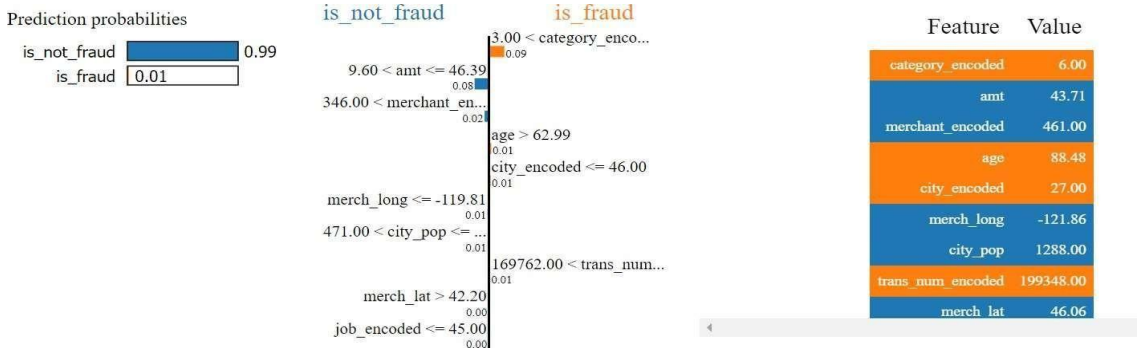

Figure 11. Demonstrates a LIME output for an individual transaction.

Lime explains predictions for specific instances, such as index 8, using test data. The explain_instance function identifies feature contributions to fraud likelihood. The outputs provide insights into the model's decision-making.

6. RANDOM CLASSIFIER AND FEATURE IMPORTANCE

A random classifier was used as the base condition and factored in predicting the performance of xgboost. The random classifier made predictions without taking into consideration any feature importance state, resulting in nearly random performance metrics and clearly signifying the need for feature-driven models for fraud detection. Feature importance was inferred from shap values and gain metrics; it was observed that the "transaction amount" and "feature v12" were the ones with maximum predictions.

6.1 Gini Index and Entropy

The Gini Index and Entropy were employed during model training to determine the best splits in decision trees. The Gini Index measures impurity in data splits, favoring nodes with homogenous groups, while Entropy quantifies the disorder of a dataset, guiding the model to make the most informative splits. These metrics ensured the XGBoost model built robust trees, enhancing prediction accuracy and interpretability.

6.2 Feature Importance and Gini Formula

The feature importance was measured using SHAP values and gain metrics. Below is a table with general features listed regarding the most influ- ential features inside the model:

Feature	Importance
amt	64.91
merch_long	5.14
Long	5.14
city_pop	5.14
merch_lat	5.14
Lat	5.14
Others	9.19

Table 2 : Most influential features in the model

Feature importance helps identify the drivers behind model predictions, enabling targeted improvements and interpretability. The Gini Index was utilized during training to determine the best splits in decision trees. The formula for the Gini Index is as follows:

Formula:

$$\textbf{Gini Index} = 1 - \Sigma (pi)^2$$

Where Pi is the proportion of instances belonging to class iii. Lower Gini values indicate purer splits, aiding robust decision tree construction.

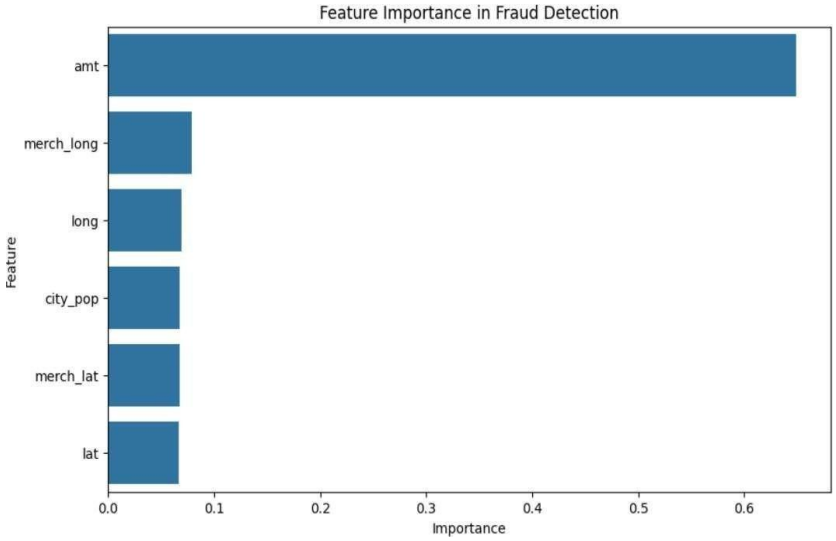

Figure 12. Feature Importance in Fraud Detection

CONCLUSION

While SHAP integrates XGBoost with its complement LIME, LiME focuses more on an understanding of specific predictions from the model. Using these two methods reduces transparency, trust, and regulatory compliance barriers in fraud detection. The usefulness of complementarity in end-user studies is illustrated, showing that explainability tools should be designed to meet the needs of different users such as compliance officers and fraud analysts. Explainable AI techniques such as LIME and SHAP make it possible to detect credit card fraud with increased transparency and reliability in predictions made by models. LIME provides information on individual predictions from the model application by being a local approximation of the model; SHAP provides consistent attribute scores based on the game theory. Applied to high-performing models such as XGBoost, these methods help identify the most important features behind the possible fraudulent transactions and, therefore, enhance interpretability while preserving performance. These properties are necessary for any financial institution regarding improving detection algorithms, rule fulfilment, and customer trust through explicit explanations about flagged transactions. Furthermore, the inclusion of XAI in fraud detection systems provides more human oversight over

automated decisions by allowing the validation of such decisions by analysts, thereby minimizing the chances of false positives and biases in the model. Combined, these now enable a much more robust and interpretable solution to fraud detection.

Conflicts of Interest

All authors declare that there is no conflict of interest.

REFRENCES

[1] N. Dhieb, H. Ghazzai, H. Besbes and Y. Massoud, "A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement," in IEEE Access, vol. 8, pp. 58546-58558, 2020, doi: 10.1109/ACCESS.2020.2983300

[2] R. Li, Z. Liu, Y. Ma, D. Yang and S. Sun, "Internet Financial Fraud Detection Based on Graph Learning," in IEEE Transactions on Computational Social Systems, vol. 10, no. 3, pp. 1394- 1401, June 2023, doi: 10.1109/TCSS.2022.3189368

[3] Z. Zhang, L. Chen, Q. Liu and P. Wang, "A Fraud Detection Method for Low-Frequency Transaction," in IEEE Access, vol. 8, pp. 25210-25220, 2020, doi: 10.1109/ACCESS.2020.2970614.

[4] M. N. Ashtiani and B. Raahemi, "Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 72504-72525, 2022, doi: 10.1109/ACCESS.2021.3096799

[5] Kotagiri, A. (2023). Mastering Fraudulent Schemes: A Unified Framework for AI-Driven US Banking Fraud Detection and Prevention. International Transactions in Artificial Intelligence, 7(7), 1–19. Retrieved from https://isjr.co.in/index.php/ITAI/article/view/197

[6] Md Rokibul Hasan, Md Sumon Gazi, & Nisha Gurung. (2024). Explainable AI in Credit Card Fraud Detection: In- terpretable Models and Transparent Decision-making for Enhanced Trust and Compliance in the USA. Journal of Computer Science and Technology Studies, 6(2), 01–12. https://doi.org/10.32996/jcsts.2024.6.2.1

[7] Sabharwal, R., Miah, S.J., Wamba, S.F. et al. Extending application of explainable artificial intelligence for managers in financial organizations. Ann Oper Res (2024). https://doi.org/10.1007/s10479-024-05825-9

[8] Sai, Chaithanya Vamshi and Das, Debashish and Elmitwally, Nouh and Elezaj, Ogerta and Islam, Md Baharul, Explainable Ai-Driven Financial Transaction Fraud Detection Using Machine Learning and Deep Neural Networks. Available at SSRN: https://ssrn.com/abstract=4439980 or http://dx.doi.org/10.2139/ssrn.4439980

[9] Raufi, B., Finnegan, C., Longo, L. (2024). A Comparative Analysis of SHAP, LIME, ANCHORS, and DICE for Interpreting a Dense Neural Network in Credit Card Fraud Detection. In: Longo, L., Lapuschkin, S., Seifert, C. (eds) Explainable Artificial Intelligence. xAI 2024. Communications in Computer and Information Science, vol 2156. Springer, Cham. https://doi.org/10.1007/978-3-031-63803-9_20

[10] Kennedy, R.K.L., Salekshahrezaee, Z., Villanustre, F. et al. Iterative cleaning and learning of big highly-imbalanced fraud data using unsupervised learning. J Big Data 10, 106 (2023). https://doi.org/10.1186/s40537-023-00750-3

[11] Chhabra, R., Goswami, S. & Ranjan, R.K. A voting ensemble machine learning basedcredit card fraud detec- tion using highly imbalance data. Multimed Tools Appl 83, 54729– 54753 (2024). https://doi.org/10.1007/s11042-023-17766-9

[12] A.A. Taha and S. J. Malebary, "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine," in IEEE Access, vol. 8, pp. 25579- 25587, 2020, doi: 10.1109/ACCESS.2020.2971354

[13] Krishnavardhan, N., Govindarajan, M. & Rao, S.V.A. An intelligent credit card fraudulent activity detection using hybrid deep learning algorithm. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-18793-w

[14] Talaat, F.M., Aljadani, A., Badawy, M. et al. Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. Neural Comput & Applic 36, 4847–4865 (2024). https://doi.org/10.1007/s00521-023-09232-2

[15] T. Awosika, R. M. Shukla and B. Pranggono, "Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection," in IEEE Access, vol. 12, pp. 64551-64560, 2024, doi: 10.1109/ACCESS.2024.3394528.