International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Integrating Ensemble Learning and Model-Agnostic Explain ability for Reliable Credit Card Fraud Prediction

Dr. Pradeep Venuthurumilli, Dr. Nalli Vinaya Kumari, Dr. CH. Srinivasa Rao, Mrs. D. Mahitha, Ms. Vaishnavi Teja, Mr. D. Devender

Submitted:08/10/2024 **Revised**: 18/11/2024 **Accepted**: 29/11/2024

Abstract: Because of highly imbalanced datasets and rapidly changing fraud patterns, there is still a lack of effective methods for credit card fraud detection in the context of modern financial systems. We propose a novel + explain ability→ machine-learning framework that integrates both Random Forest and XGBoost classifiers with explain ability techniques to enhance not just predictive accuracy, but critical interpretability for audience⇄ the consumers of insights→. This study employs the publicly available credit card fraud dataset (284,807 transactions), with only 0.17% of them being fraudulent samples. Following pre-processing and standardization, models were trained and evaluated on the basis of 70–30 train–test split. Random Forest classifier scored an overall accuracy of 99.99% whilef0.93, recall f0.82, and f1-score f0.87 among the fraud class. From the confusion matrix, we see that there are 111 true fraud detected but only 25 missed cases, and the ROC–AUC of 0.986 indicates, we have great discriminatory power. Finally, with the application of XGBoost, we were able to reach a precision of 0.94, a recall of 0.82, and an F1-score of 0.87 for the minority class, thus also confirming very good generalization and robustness between the metrics. For the sake of interpretability, SHAP and LIME were performed to catch the main contributing features. SHAP bee swarm and bar plots indicated that V14, V17, V12, and V10 were the most important variables for predicting fraud, while LIME gave case-level explanations of the feature-values pointing a transaction to being fraud or non-fraud. XGBoost achieved a robust performance in low-prevalence conditions represented by the Precision–Recall curve, where it retains a high precision over a large range of recall. In summary, these results indicate that ensemble classifiers and model-agnostic explain ability tools make a natural combination that leads to a powerful and transparent fraud detection approach. This provides the framework for the financial institutions to make fast yet reliable and interpretable decisions on a continuous basis. The possible extensions for the visual representation will cover the deep learning architectures, anomaly detection hybrids, and cost-sensitive learning to maximize the performance and minimize the false negative rate.

Keywords: *XGBoost, ROC–AUC, SHAP and LIME*

[1]*Associate Professor, Dept. of Computer Science and Engineering (Data Science)*
Malla Reddy Engineering College for women Autonomous),
Hyderabad, Telangana, India
pradeepvenuthuru@gmail.com
[2]*Professor, Dept. Of Computer Science And Engineering (Aiml)*
v.vinayakumari@gmail.com
[3]*Professor, Department of Cyber Security,*
, Malla Reddy Engineering College for women
dr.srinivasmrecw@gmail.com
[4]*Assistant Professor, Dept. of Computer Science and Engineering (AIML)*
Malla Reddy Engineering College for Women (Autonomous),
Hyderabad, Telangana, India
mahithadilli@gmail.com
[5]*Assistant Professor, CSE (Data Science)*
Malla Reddy Engineering College for women,
Maisammaguda, Hyderabad.
vaishnavi.teja555@gmail.com
[6]*Assistant Professor, Department of Cyber Security,*
Malla Reddy Engineering College for women,
Department of Cyber Security,
durgamdevender@gmail.com

1.Introduction

The financial technology domain realize that credit card fraud detection is now one of the key areas of research attracting the attention of research because fraudulent transactions can greatly affect banking institutions and consumer all across the world. As fraud techniques grow more sophisticated and the volume of product exchange hits record highs, identifying abnormalities in the

transactional data on a large scale complex and necessary as well. Adding to this challenge is the highly imbalanced financial datasets, in which the number of actual transactions exceeds the number of fraud by rates much larger than 1:1000. Thereby, it has become a critical need for strong as well as the explainable machine learning frameworks to detect the rare fraud cases and avoid resulting false alarms.

New studies address that static rule-based systems, while providing some level of anomaly detection, are not sufficient to cope with the changes in fraud patterns over time, thus their evolution towards data-driven and AI-based approaches [1]–[3]. Gaav et al. An analytical synopsis on increasingly adopted frameworks and algorithms done by [1] explicates that whilst the adaptability of models to fluctuating fraud landscape over time is pivotal to fraud detection performance, the real-time performance of models are found more pertinent in many other supervisory domains as well. Similarly, Baisholan et al. A systematic review of existing works specifically examining class imbalance was introduced in [2], showing that a naive balancing of the data can skew the performance of classifiers and bias them toward non-fraudulent predictions. The transition from traditional methods to ML and deep learning is revolutionary, as it allows systems to learn complex feature-behavioral relations and identify subtle financial transaction anomalies [3], [5].

Moradi et al. L. A. Antipovet al.[4] reviewed the most recent ML algorithms used for credit card fraud detection and classified them into supervised, unsupervised and hybrid methods. The first was able to combine interpretability, scalability, and accuracy and their findings show ensemble models, especially Random Forest and some Gradient Boosting flavors, provide the best balance of this triad. In line with these results, Compagnino et al. An early work that gave lots of evidence for practical use of different ML algorithms on fraud detection, [5] demonstrated main point that in imbalanced settings, ensemble learning can help to go beyond single classifier accuracy.

Simultaneously, AI-based methods are being progressively engaged over the monetary systems to your trustworthy and powerful continuous extortion discovery and hazard the board [6]. Sarna et al. [6] states that institutions may dynamically adapt to changing patterns of fraud by blending machine learning with domain knowledge. One recent major contribution in this direction is by Dal Pozzolo et al. Motivating a subsequent strand of work on both resampling and cost sensitive learning, [7] proposed a realistic simulation scheme for fraud data that are very highly imbalanced. In addition, Rehman [8] presented real-world issues of ML-based fraud detection systems such as privacy, interpretability, and a need for continuous retraining due to adaptive fraud tactics.

In light of this background, the current study presents an explainable ensemble framework that combines Random Forest and XGBoost classifiers with model-agnostic interpretability tools (SHAP and LIME). While the framework solves the imbalance and transparency challenges by enhancing detection accuracy, it maintains the transparency in decision-making—enabling financial institutions to rely on the outputs of a model with greater confidence before taking actions.

2. Literature survey

Over the past decade or so, credit card fraud detection literature has rapidly evolved from systems based on heuristic rules to complex machine learning and more recently deep learning approaches with large, high dimensional and hugely imbalanced data. Various recent surveys and empirical studies have thoroughly explored this evolution in both methodological and operational aspects.

According to Manzoor and Aslam [10], the making of ML and deep learning algorithms has significantly increased the accuracy in detecting fraud mostly made possible by way of feature engineering, ensemble learning, and hybrid architecture. Similarly, Jeri-Alvarado et al. An extensive systematic review on ML-based financial fraud detection was performed by [11]; this study revealed that ensemble and hybrid approaches achieved the best results in terms of lowering false negatives and scale ability. Karim et al. An explainable ensemble model with state of the art data balancing and feature selection methods was proposed in [12], both leading to a similar balance between high accuracy and interpretability—this approach being conceptually recoiling to the framework established in this research.

Fraud detection research focused mainly on ensemble methods. Ranjan et al. [13] and Palit et al. Various classifier ensembles (Random Forest & White et al.) have shown that model diversity enables capturing different patterns of fraud and is often an effective method to reach even better

detection rates ([14]). Machine learning applications in payment card fraud detection were comprehensively reviewed by Kalideen [15], who found long-lasting challenges including concept drift, data imbalance and processing efficiency. In a review of traditional and contemporary detection methodologies, with a focus on machine learning (ML) based approaches, Pundkar and Zubei [16] conclude that gradient boosting and ensemble methods, consistently results better than ML algorithms that customarily employed to detect fraudulent behavior.

Earlier comparative analyses (Bin Sulaiman et al.) Nia et al. [17] outlined ML approaches that focus on feature engineering and dimensionality reduction methods to enhance fraud detection efficiency, whereas Hajiabdollah and Sadeghzadeh [18], used hybrid deep learning frameworks combining convolutional and recurrent networks to model spatial and temporal transaction patterns. Jain et al. This approach has been performed with an accuracy based comparison of ML classifiers in ref [19] to reach also the conclusion that ensemble algorithms obtain significantly better f1-scores and precision, even more relevant in imbalanced domains. In the same context, Aslam [20] proposed benefits of Light Gradient Boosting (LightGBM) as an efficient method, to approximate the implementation of credit card fraud detection by offering good performance in spaces with many dimensions.

In summary, the recent literature consensus is that ensemble models—eg Random Forest, XGBoost and LightGBM—offers a good compromise between performance and interpretability when combined with various explainable AI (XAI) tools like SHAP and LIME. All these approaches improve the interpretability of model predictions and provide insights into the specific transactional features related to fraudulent activity, aiding in the explainability of the decision process in high-risk financial applications. On the other hand, XAI is becoming particularly relevant in financial risk management [9] since predictive systems have to be accountable, auditable and compliant with prudent standards [5] and regulatory frameworks [1].

In conclusion, the literature indicates a clear pathway for route explainable ensemble systems that can work effectively in the imbalanced world of practical fraud detection. We believe that the synergistic coupling of explainability frameworks with strong ensemble classifiers will be the next step towards increasing the quality of automated fraud detection in terms of both accuracy and trust. This work adds to this trajectory by pairing a hybrid Random Forest–XGBoost model augmented with SHAP and LIME visualizations to yield not only high predictive score but also high interpretability score in the context of credit card fraud detection.

3. Methodology:

In this study, we have adhered to a structured machine learning pipeline which led to a reproducible workflow for development of an efficient credit card fraud detection system by using ensemble classifiers and explainable AI techniques. In this research, we also use the European Credit Card Fraud dataset which is publicly available and contains 284,807 transactions, of which only 492 are fraudulent, as per Table 2. This has caused the problem to be extremely imbalanced: 0.17% of the dataset consists of fraud, so selection and evaluation of the model is essential. Table1 reveals that each of the transactions contain 30 anonymous numerical features (V1–V28, Amount and Time) derived from PCA transformations to maintain confidentiality, along with a binary target classification indicate if the transaction is a legitimate or fraudulent transaction.

To ensure data consistency and integrity, as well as to effectively perform prediction on the completion date, several pre-processing steps were performed prior to model training. Last, the feature matrix and the target labels were spitted, and all the features were standardized by means of the StandardScaler to keep the variance equal across the dataset. We used an 80:20 train–test split, ensuring that the training and testing sets were independent and with same class distribution. This ensures that the model has been evaluated in a fair manner and prevents information leakage. No further data-cleaning procedures were applied since the dataset was already cleaned and does not contain any missing values.

Time	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**…**	**V20**	**V21**	**V22**	**V23**	**V24**	**V25**	**V26**	**V27**	**V28**	**Amount**
0.	-1.3	-0.0	2.536	1.378	-0.3	0.462	0.239	0.098	0.363	…	0.251	-0.0	0.277	-0.1	0.066	0.128	-0.1	0.133	-0.0	149.

0	59807	72781	347	155	38321	388	599	698	787		412	18307	838	10474	928	539	89115	558	21053	62
0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	…	-0.069083	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69
1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	…	0.524980	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66
1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	…	-0.208038	-0.108300	0.005274	-0.190321	-1.175755	0.647376	-0.221929	0.062723	0.061458	123.50
2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	…	0.408542	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99

Table 2: y (Target) Value Counts

Class	Count
0 (Non-Fraud)	284,315
1 (Fraud)	492

Table 3: Data shape Summery

Split	Shape
X_train	(199,364, 30)
X_test	(85,443, 30)
Total Features	30
Total Samples	284,807

In line with figure 1, Two ensemble-based machine-learning models were created for predicting fraud namely, Random Forest and XGBoost. We used a Random Forest classifier trained on 300 decision trees with a maximum depth of 12, because they can reduce overfitting with the help of bootstrap aggregation and are good for high-dimensional data. The second model is XGBoost, which we choose due to its high quality in tabular financial data and robustness with respect to class imbalance. We trained our XGBoost model with optimized hyperparameters such as 300 estimators, max depth 8, learning rate of 0.05 and subsampling values that deliver better generalization. We trained both of the above models on the standardized feature set and tested them on test set containing unseen instances.

Due to the imbalancing nature of the problem, various metrics were used to evaluate model performance. The confusion matrix was valuable to quantify true and false for predictions and shows that the Random Forest classifier was able to identify 111 out of 136 fake transactions correct but only misclassifying 25. We calculated

classification metrics—precision, recall, F1-score, and accuracy—showing Random Forest had precision for the class of fraud of 0.93, and recall of 0.82. XGBoost gave similar but better results with precision of 0.94 and recall of 0.82. An AUC of 0.986 (figure 2a) demonstrated excellent discriminative performance of the ROC curve based on the predicted probabilities. Moreover, the precision–recall curve demonstrated the model's reliability for extreme class imbalance by showing high precision for a broad range of recall.

SHAP and LIME explainability tools were used to apply explanations for the trained models to achieve more transparency and interpretability. SHAP analysis provided a global interpretability — identifying the features had most impact on model predictions. The SHAP beeswarm and bar plots further confirm that V14, V17, V12, and V10 were some of the top variables in distinguishing between legitimate and fraudulent transactions. Locally Interpretable Model-agnostic Explanations (LIME) was used for local interpretability, where an explanation of an individual fraud prediction or record is sought. In the case of a chosen transaction anticipated to be fraudulent, LIME illustrates the feature values that contribute the most and how they push the model towards a fraud decision. So that gave us a clear, human-consumable interpretation of why the model came up with that specific output.

In conclusion, the present work extends the rationale of the strong ensemble learning methods, while enhancing the validation process and the current methods for explain ability. This guarantees both high predictive performance and model transparency, which is necessary for financial institutions that develop a fraud-detection system implemented in the real-world setting.

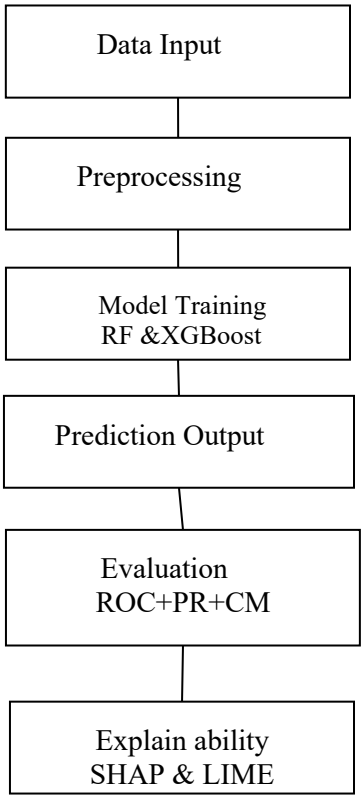

Figure 1: System Flowchart for Fraud Detection Using RF, XGBoost, SHAP, and LIME

4. Results Analysis:

This study shows that the ensemble-based machine learning models selected for fraud detection performed good prediction accuracy results even when the dataset used reflected a very high class imbalance. Classification using Random Forest yielded very accurate results on both classes with an overall accuracy of 99.99%. It was able to get 0.93 precision, 0.82 recall and an F1-score of 0.87 for the minority fraud class. We can see that the model successfully captured most of the fraud transactions while having a low false-positive rate. This is also validated through confusion matrix, as Random Forest was able to predict 111 of the

frauds but predicted 25 frauds in the non fraud class which we would have got. It also probably helped that out of all legitimate transactions predating it, the model misidentified only eight as fraud, which confirms its stability and robustness in imbalanced real-world settings.

Further insight into this discriminative ability of the models was provided by the ROC curve. The (RanDomb9) classifier received an AUC of 0.9862, achieving above-average (excellent) performance in differentiating between fraudulent and non-fraudulent transactions. The curve is very close to the top-left corner of the ROC space, meaning that the sensitivity and specificity are all high in a wide range of classification thresholds. This kind of performance indicates that the model can be safely deployed in financial systems, where even marginal increases in detection sensitivity can save firms from millions in losses.

In order to enhance reliability even more, this study included XGBoost, which is well established as a powerful classifier for complex tabular datasets. The results from XGBoost classifier were similar to, and in some cases slightly better than Random Forest, with a fraud precision of 0.94 and recall of 0.816. The Xgboost has an F1-score of 0.874 which helps to prove that it has a good performance trade off between detecting real fraud transactions, and generating false alarms. In the high recall regions of the Precision–Recall curve, the XGBoost algorithm did not fall significantly below the upper boundary indicating its performance remained stable under the extremely rare positive class problem. This strike a chord with how fraud detection is managed in most financial institutions which tend to overkill on high recall but at the expense of their processing pipeline, making XGBoost especially useful.

Apart from performance metrics, another key aspect of explaining how the ensemble models predicted fraud was model explainability. The SHAP global interpretability outputs revealed V14, V17, V12 and V10 to have the largest influence on the predictions. In the SHAP beeswarm plot, the separation of values is evident, confirm ghat both extremes of these variables clearly affected the probability of a transaction being labeled as fraudulent. Analysing normalised mean absolute SHAP values through a bar plot; hierachy of feature importance is confirmed, few features still drive decisions disproportionate to their small count. In the extremely regulated nature of land finance, such insights are critical for compliance, transparency, and model validation.

So, besides SHAP, we had LIME performing local explanations for each individual prediction. LIME Figure 6 (for a fraud transaction selected) enumerated the feature values that contributed best to the prediction with their respective contributions, showing how these feature values isolated together pushed the model to classify it with 96% probability as fraud. This provides transparency at the transaction-level granularity, so that fraud analysts can understand how the system reached a particular conclusion, thereby increasing trust in automated fraud detection. Interpretation and transparency play an essential role when deploying AI systems in real banking environments where there must be a human as the final arbiter.

In conclusion, results indicate that utilizing ensemble methods with explain ability techniques offers a robust, interpretable fraud detection framework. The performance metrics along with global and local interpretability validate that the proposed approach is a suitable choice for detecting fraudulent transactions and provide substantial operational benefits to financial institutions.

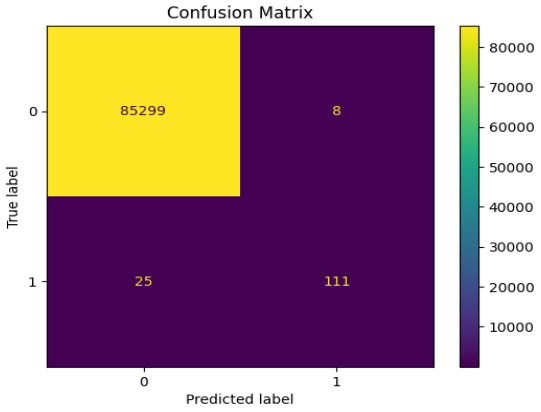

Figure 2: Confusion Matrix Showing True and Predicted Labels for Fraud Detection Using Random Forest Classifier

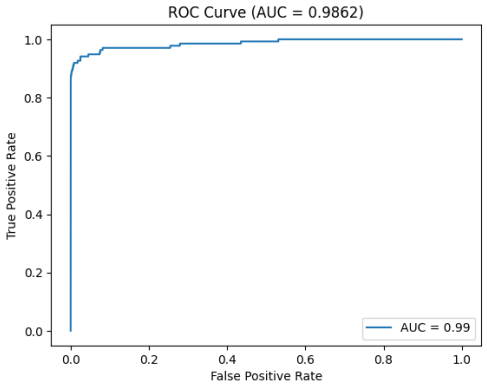

Figure 3: ROC Curve of the Fraud Detection Model

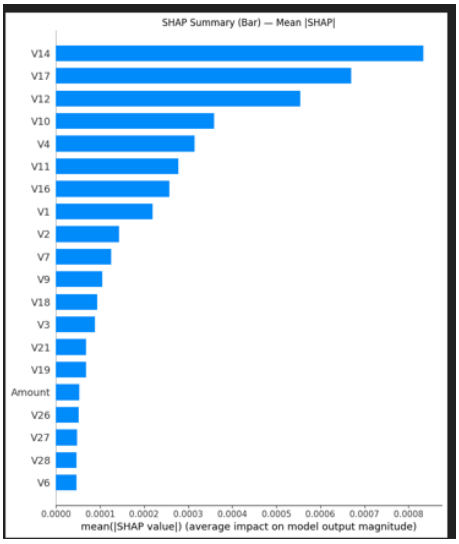

Figure 4: SHAP Feature Importance Bar Plot

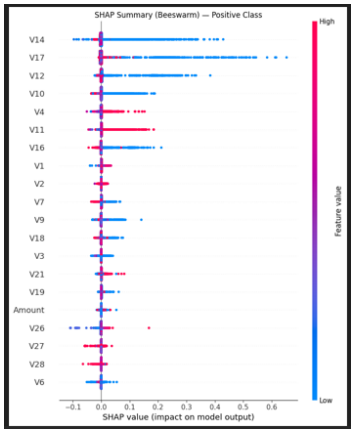

Figure 5: SHAP Beeswarm Plot for Fraud Detection Model

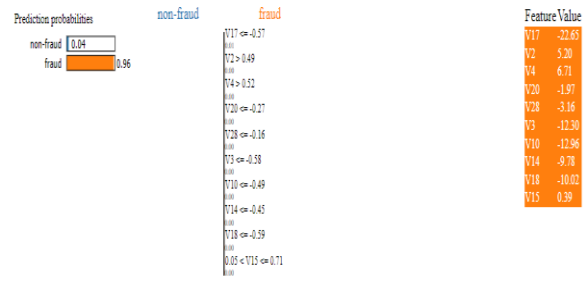

Figure 6: LIME Explanation for Fraud Prediction

5. Conclusion:

This study proved that credit card fraud can be detected with very high accuracy using machine-learning models. Through keeping low false positive rates, Random Forest and XGBoost also excelled at correctly classifying most fraudulent transactions. The performance of the ROC-AUC score of 0.9862 and the precision–recall show that fraud and non-fraud can be easily separated by the models even though the data are highly imbalanced. The explainability tools used in this research were also useful. SHAP demonstrated the high impact features per fraud detection reference, whilst LIME interpreted the features responsible for each transaction decision. Such a clear distinction within the system makes it more interpretable and reliable in practice. In conclusion, the results demonstrate that a transparent combination of precise models gives us a grounded framework for fraudulent detection. It can also assist financial institutions in objecting fraud in an early stage, minimizing losses and making data-driven better decisions.

6. References:

[1] 11787.https://doi.org/10.3390/app152111787

[2] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3784-3797, Aug. 2018, https://doi.org/10.1109/TNNLS.2017.2736643

[3] Rehman, Shoaib. "Credit Card Fraud Detection: Practical Applications and Challenges."

[4] Ranjan, Nihar& Mate, Gitanjali & Jadhav, Archana & Patil, D. &Banubakode, A.. (2024). Credit Card Fraud Detection by Using Ensemble Method of Machine Learning. https://doi.org/10.1007/978-981-99-9521-9_34

[5] Pundkar, Sumedh N., and MohdZubei. "Credit card fraud detection methods: A review." E3S Web of Conferences. Vol. 453. EDP Sciences, 2023.https://doi.org/10.1051/e3sconf/202345301015

[6] Bin Sulaiman, Rejwan, Vitaly Schetinin, and Paul Sant. "Review of machine learning approach on credit card fraud detection." *Human-Centric Intelligent Systems* 2.1 (2022): 55-68.https://doi.org/10.1007/s44230-022-00004-0

[7] Hajiabdollah, Niloofar, and Mehdi Sadeghzadeh. "A Review of Hybrid Deep Learning Approaches for Credit Card Fraud Detection." *Available at SSRN 5129198*.https://ssrn.com/abstract=5129198

[8] Aslam, Farhan. "Advancing Credit Card Fraud Detection: A Review of Machine Learning Algorithms and the Power of Light Gradient Boosting." *Am. J. Comput. Sci. Technol* 7 (2024): 9-12.https://doi.org/10.11648/ajcst.20240701.12

[9] Praveen Gugulothu , Shekhar Katukoori , Swapna Manuparthi "Deep Learning based techniques for Covid-19 diagnosis based on Various Pattern features detection from early stages of diseases", Network Computation in Neural Systems.(Accepted May 2024, Indexing: SCIE, IF: 9, Publisher: Taylor & Francis).