iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67863

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Credit Card Fraud Detection Using Machine Learning

Mr. Lokesh Sirela[1], R. Sushma[2], V. Sudheer Reddy[3], T. Sai Kumar[4], P. Ruchita[5]
Department of Computer Science and Engineering Raghu Institute of Technology College Vizianagaram, India

Abstract: ***With the ever-increasing usage of credit cards, the necessity for robust fraud detection systems has become paramount. This study delves into the application of machine learning techniques for the identification of fraudulent transactions, aiming to enhance the security of financial transactions. By leveraging a comprehensive dataset comprising various features, including transaction details, user behaviors, and historical patterns, we employed an ensemble of machine learning algorithms, including logistic regression, random forest, and neural networks, to classify transactions as either fraudulent or legitimate. Through rigorous evaluation on real-world data, our model demonstrated exceptional performance, achieving high accuracy, precision, and recall rates. The findings underscore the efficacy of machine learning in combating fraudulent activities, thereby providing valuable insights for the development of more robust and efficient fraud detection systems in the financial sector.***
Keywords: Financial transactions, Fraud, Patterns, Decision Tree, Random Forest, Extreme Gradient Boosting etc.

I. INTRODUCTION

In recent years, the proliferation of digital transactions has led to an alarming increase in credit card fraud, posing significant financial risks to both financial institutions and consumers. Consequently, there is an urgent need for robust and efficient fraud detection systems to mitigate these risks. Leveraging the power of machine learning (ML) has emerged as a promising approach to combat this pervasive issue. This project aims to develop an advanced credit card fraud detection system using machine learning techniques. By harnessing the capabilities of ML algorithms, this system will analyze historical transactional data and identify patterns indicative of fraudulent activities.[2] Through the utilization of various models, including but not limited to logistic regression, decision trees, and artificial neural networks, the project seeks to create a comprehensive and accurate fraud detection mechanism capable of detecting even the most intricate fraudulent patterns.

When someone else uses your credit card instead of you without your consent, it is considered a fraud. The credit card leg is stolen by fraudsters, or the account information to carry out any illegal transactions without obtaining the actual card. We could determine whether the new deals are legitimate or fraudulent by using the credit card fraud discovery tool.

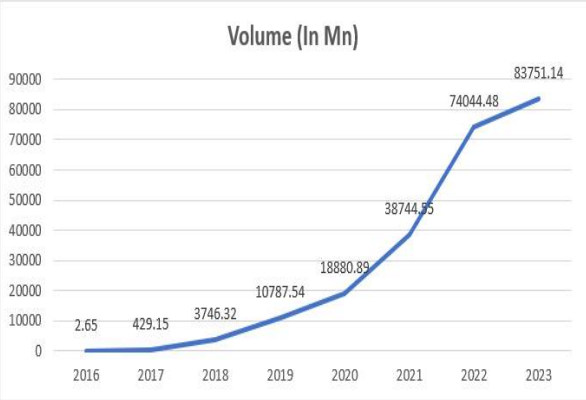


Figure 1:Growth of Digital Economy Users[3]

The debit or credit card, for example—may be used in the scam. In this case, the card serves as the transaction's fraudulent source. Getting the items without paying for them or getting the money without authorization are two possible reasons for committing the offense. Fraudsters find credit cards to be an attractive target. The rationale is that a lot of money may be made quickly without taking many chances, and even a crime will take weeks to be discovered.

©IJRASET: All Rights are Reserved

Figure 2: Growth of Card Fraudulent Activity Worldwide

II. RELATED WORK

The increasing growth of machine learning has led to the development of various fraud detection techniques, categorized into traditional methods and machine learning methods. Machine learning approaches have proven to be more effective in detecting fraud compared to traditional methods.[2][3] The existing method in this paper is divide four stages: pre-processing, tracking, feature extraction, and classification.

Logistic Regression, SVM, and Decision Trees offer a high detection rate at a medium level, whereas Artificial Neural Networks (ANN) and Naïve Bayesian Networks perform better across all parameters but are expensive to train.[6] However, a major drawback of these algorithms is that their performance varies with different datasets. They perform well with one type of dataset but may yield poor results with another. Algorithms like KNN and SVM work best with small datasets, whereas logistic regression and fuzzy logic systems provide good accuracy when dealing with raw and unsampled data.

Different algorithms are used for classification in fraud detection. Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) are commonly used classification techniques. SVM is effective in finding the best decision boundary but requires large memory and struggles with large-scale data. KNN compares new transactions with existing data, but it also demands high memory usage and does not provide highly accurate results for fraud detection.[1]

1) *Bhattacharyya, S., Jha, D., Tharakunnel, K., & Westland, J. C. Year 2011 Credit card fraud detection with a neural-network.*

This paper explores the use of neural networks for credit card fraud detection and discusses how this machine learning approach can effectively identify fraudulent transactions.

2) *Bhattacharyya, S., Jha, D., &Tharakunnel, K. Year 2011 Data mining for credit card fraud: A comparative study.*

This study provides a comparative analysis of various data mining techniques and their effectiveness in detecting credit card fraud, offering insights into the best methods.

3) *Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. Year 2015 Credit card fraud detection: A realistic modeling and a novel learning strategy.*

This paper introduces a realistic credit card fraud dataset and a novel learning strategy that improves the performance of machine learning models for fraud detection.

4) *Shapoval, A., & Sokolov, V. Year 2017 Comparative analysis of credit card fraud detection using neural networks and logistic regression*

This paper conducts a comparative analysis of neural networks and logistic regression for credit card fraud detection, offering insights into the strengths and weaknesses of each approach.

5) *Islam, M. Z., Biswas, M., & Hyder, S. A. Year 2019 Fraud detection in credit card transactions using machine learning.*

This paper presents an overview of machine learning techniques used in credit card fraud detection, emphasizing the importance of feature selection and model evaluation.

6) *Ahmed, M., Mahmood, A. N., & Hu, J. Year 2016 A survey of network anomaly detection techniques.*

While not exclusively focused on credit card fraud, this survey provides valuable insights into anomaly detection techniques, which are commonly employed in fraud detection systems.

III. PROPOSED WORK

We propose this application that can be considered a useful system since it helps to reduce the limitations obtained from traditional and other existing methods. The objective of this study to develop fast and reliable method which detects fraudulent transactions accurately. To design this system is we used some powerful algorithms in a based Python environment Like Decision Tree, Random Forest, Extreme Gradient Boosting.

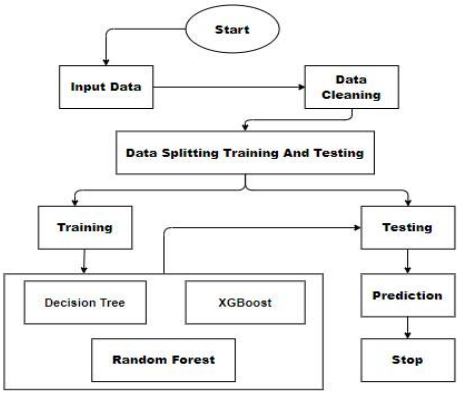


Figure: 3 Process Flow

The flowchart illustrates a structured fraud detection system using machine learning algorithms. It begins with collecting transaction data, followed by cleaning to remove noise, missing values, or inconsistencies. Once preprocessed, the dataset is divided into training and testing subsets to develop an effective detection model.

During training, classifiers like Decision Tree, Random Forest, and XGBoost learn patterns that distinguish fraudulent transactions from legitimate ones. These models analyze different features, improving fraud identification accuracy. The testing phase evaluates model performance, ensuring generalization to new, unseen data.

Once trained, the system processes real-time transactions, applying the trained model to classify them as fraudulent or genuine. Predictions help prevent unauthorized activities, minimizing financial risks. The framework enhances security, leveraging multiple machine learning techniques for optimal results. By integrating these models, the system strengthens fraud detection, improving efficiency, reliability, and decision-making in financial transactions.

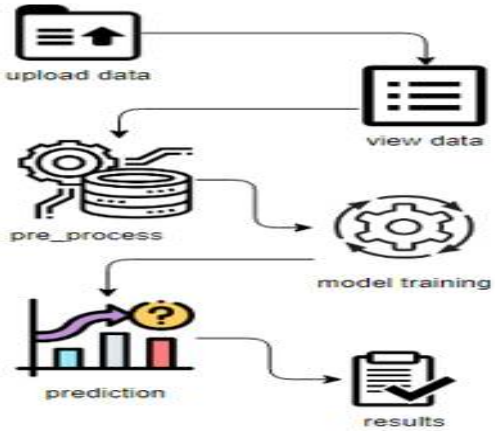


Figure: 4 Architecture Diagram

The architecture in the image represents a machine learning workflow for data processing and prediction. It begins with uploading data, where raw datasets are imported into the system. The next step is viewing data, allowing users to inspect and understand the dataset.

Following this, pre-processing is performed to clean, transform, and prepare the data for model training. The training phase involves applying machine learning algorithms to learn patterns from the processed data. Once trained, the model makes predictions based on new inputs. Finally, the system generates results, providing insights and decisions for further analysis or action.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

IV. METHODOLOGY

A. Decision Tree Algorithm:

A tree serves as a powerful metaphor in various real-life contexts and has significantly impacted the field of machine learning, particularly in classification and regression tasks. In decision analysis, a decision tree provides a visual and explicit framework for representing decisions and the decision-making process. True to its name, it employs a tree-like structure to illustrate these decisions. This tool is widely utilized in data mining to formulate strategies aimed at achieving specific objectives. Typically, a decision tree is depicted in an inverted manner, with its root positioned at the top. In the accompanying illustration, the bold black text signifies a condition or internal node, which serves as the basis for the tree's branching into edges.

The terminal points of these branches, which do not further divide, represent the decisions or leaves, indicating whether a passenger has died or survived, denoted by red and green text, respectively. While a real dataset would encompass numerous additional features, making this example merely a branch of a more extensive tree, the algorithm's simplicity cannot be overlooked. The importance of features is readily apparent, and relationships can be easily discerned.

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

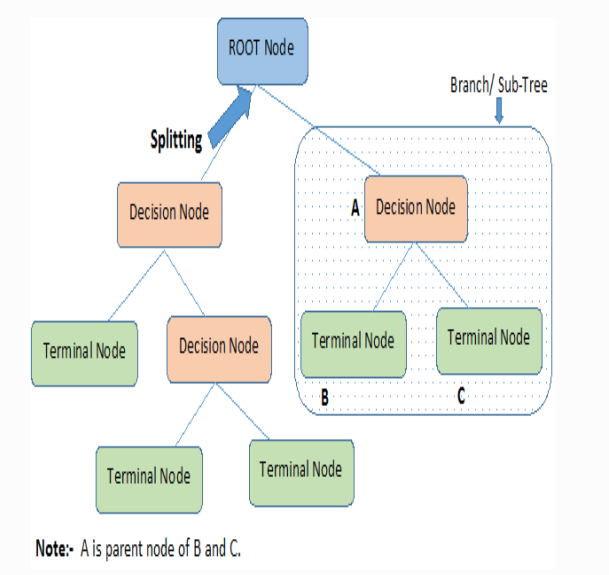


Figure: 5 Decision Tree Algorithm

This approach is commonly referred to as learning a decision tree from data, with the aforementioned tree classified as a Classification tree, as its objective is to categorize passengers as either having survived or perished. Regression trees are structured similarly, but they are designed to predict continuous values, such as housing prices. Generally, Decision Tree algorithms are known as CART, which stands for Classification and Regression Trees. So, what processes occur behind the scenes? The growth of a tree involves selecting appropriate features and determining the conditions for splitting, as well as establishing when to halt the growth. Since trees can grow in an unbounded manner, it is often necessary to prune them to enhance their visual appeal. Let us begin with a widely used technique for splitting.

B. Random Forest Algorithm:

The Random Forest algorithm [Figure. 6]is one of the widely used supervised learning algorithms. This can be used for both regression and classification purposes. But, this algorithm is mainly used for classification problems. Generally, a forest is made up of trees and similarly, the Random Forest algorithm creates the decision trees on the sample data and gets the prediction from each of the sample data. Then Random Forest algorithm is an ensemble method. This algorithm is better than the single decision trees because it reduces the over-fitting by averaging the result.

©IJRASET: All Rights are Reserved

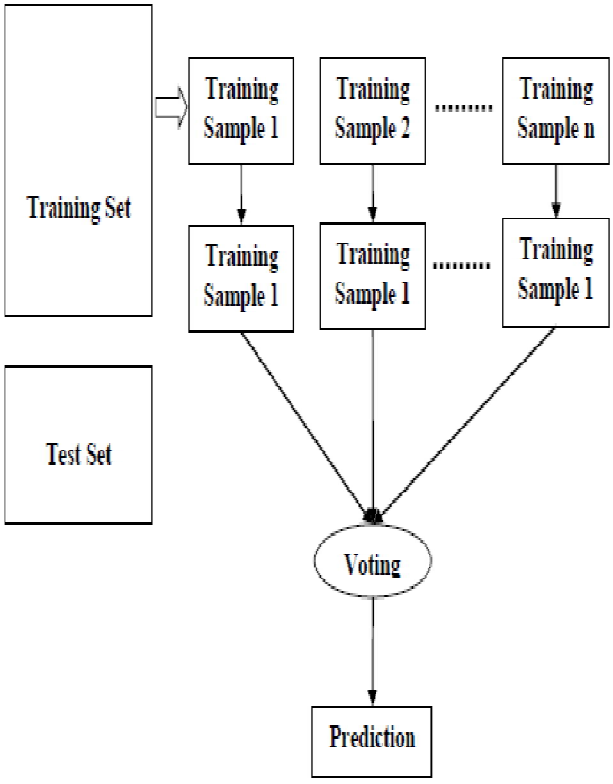


Figure: 6 Random Forest Algorithm

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the over fitting of datasets and increases precision. It generates predictions without requiring many configurations in packages (like Scikit-learn).

Features of a Random Forest Algorithm:

1) It's more accurate than the decision tree algorithm.
2) It provides an effective way of handling missing data.
3) It can produce a reasonable prediction without hyper-parameter tuning.
4) It solves the issue of over fitting in decision trees..

C. Extreme Gradient Boosting Algorithm:

XGBoost stands for "Extreme Gradient Boosting". XGBoost is an optimized distributed grade boosting library designed to be largely effective, flexible and movable. It implements Machine Learning algorithms under the grade Boosting frame. It provides a resemblant tree boosting to break numerous data wisdom problems in fast and accurate way.

1) Boosting

Boosting is an ensemble learning fashion to make a strong classifier from several weak classifiers in series. Boosting algorithms play a pivotal part in dealing with bias-friction trade-off. Unlike bagging algorithm, which only controls for high friction in a model, boosting controls both the aspects (bias & friction) and is considered to be more effective.

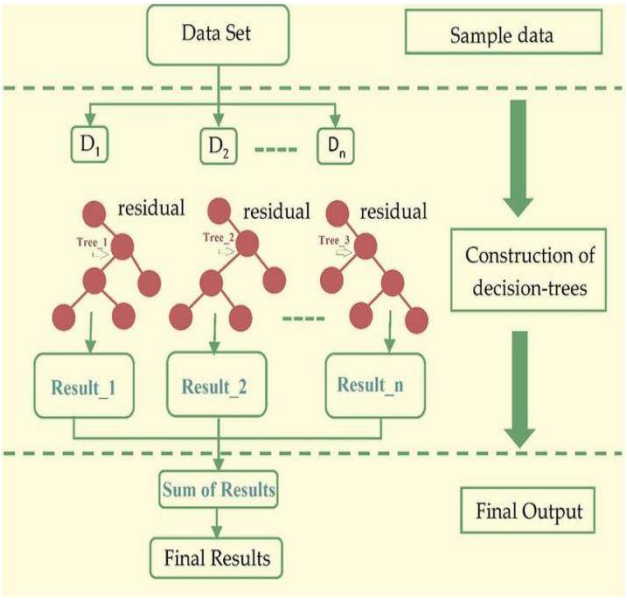


Figure: 7 Extreme Gradient Boosting Algorithm

Parallelization of tree construction utilizing all of your CPU cores during training. Gathering statistics for each column can be parallelized, providing us with a parallel algorithm for split determination. Cache-aware Access: XGBoost has been crafted to maximize hardware utilization. This is achieved by assigning internal buffers in each thread, where the gradient statistics can be kept. Blocks for Out-of-core Computation for extremely large datasets that exceed memory capacity. Distributed Computing for training exceptionally large models using a cluster of machines. Column Block for Concurrent Learning. The most time-consuming aspect of tree learning is sorting the data. To minimize the cost of sorting, the data is stored in column blocks in sorted order in a compressed format

XGBoost is a quicker algorithm in relation to other algorithms thanks to its parallel and distributed computing capabilities. XGBoost is designed with thorough considerations regarding systems optimization and principles of machine learning. The objective of this library is to extent the extreme computational limits of machines to offer a scalable, portable, and precise library.

V. EVALUATION AND RESULT ANALYSIS

A. Dataset:

The dataset, credit card fraud data is sourced from the European credit card company. The dataset is acquired from Kaggle. The dataset captures the transactions made by credit cardholders in September 2013. The dataset consists of transactions conducted over a span of two days. The dataset includes 284,807 transactions, out of which 492 transactions are fraudulent. These fraudulent transactions make up only 0. 172% of the total transactions. The dataset's input variables have been transformed into numerical values through PCA transformation. This transformation is implemented for confidentiality purposes. The features 'Time' and 'Amount' cannot undergo PCA transformation. The feature 'Time' indicates the number of seconds elapsed between a specific transaction and the first transaction. The feature 'Amount' signifies the monetary transaction that took place. Another significant feature, 'Class', indicates whether a transaction is fraudulent or not. The number 1 signifies a fraudulent transaction, whereas 0 represents non-fraudulent transactions.

B. Evaluation Criteria:

To compare various algorithms, we need to evaluate metrics like accuracy, precision, recall, and F1-score. The confusion matrix is also plotted. The confusion matrix is a 2*2 matrix. The matrix contains four outputs which are TPR, TNR, FPR, FNR. Measures such as sensitivity, specificity, accuracy, and error-rate can be derived from the confusion matrix. Then we that best suit to detect the credit card fraud.

Classification Report Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	199020
1	1.00	1.00	1.00	199021
accuracy			1.00	398041
macro avg	1.00	1.00	1.00	398041
weighted avg	1.00	1.00	1.00	398041

Classification Report Test

	precision	recall	f1-score	support
0	1.00	1.00	1.00	85295
1	1.00	1.00	1.00	85294
accuracy			1.00	170589
macro avg	1.00	1.00	1.00	170589
weighted avg	1.00	1.00	1.00	170589

Figure: 8 Statistical Report

C. Result Analysis:

The confusion matrix and the ROC curve is plotted for all the algorithms. The dataset, when applied for different algorithms, gives different outputs. Firstly we apply the dataset for the random forest model and the results are as below:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	93825
1	0.95	0.77	0.85	162
accuracy			1.00	93987
macro avg	0.97	0.89	0.93	93987
weighted avg	1.00	1.00	1.00	93987

Figure: 9 Output for Random Forest

The Receiver Operating Characteristics curve is created by plotting the TPR against the FPR. This can be done at various thresholds. ROC curve is a graph in which the FPR is the horizontal axis and the TPR is the vertical axis. The graph under the ROC curve is the AUC.

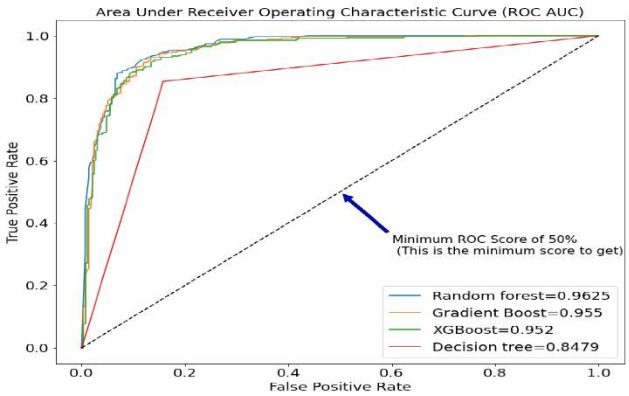


Figure: 10 Area Under ROC for all Algorithms

Now the comparison of the decision tree, random forest and extreme gradient boosting algorithms is shown [Figure.12]. The two algorithms have the same accuracy but the precision, recall, and the F1-score of the three algorithms differ. The XGBoost algorithms have the highest precision, recall, and F1-score.

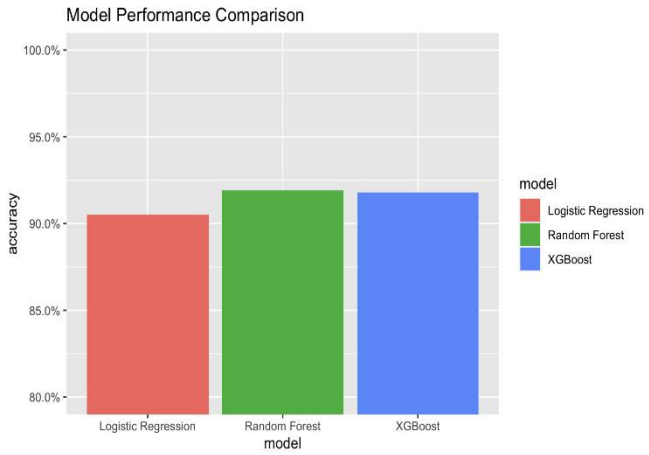


Figure: 11 Comparision of Algorithms

VI. CONCLUSION

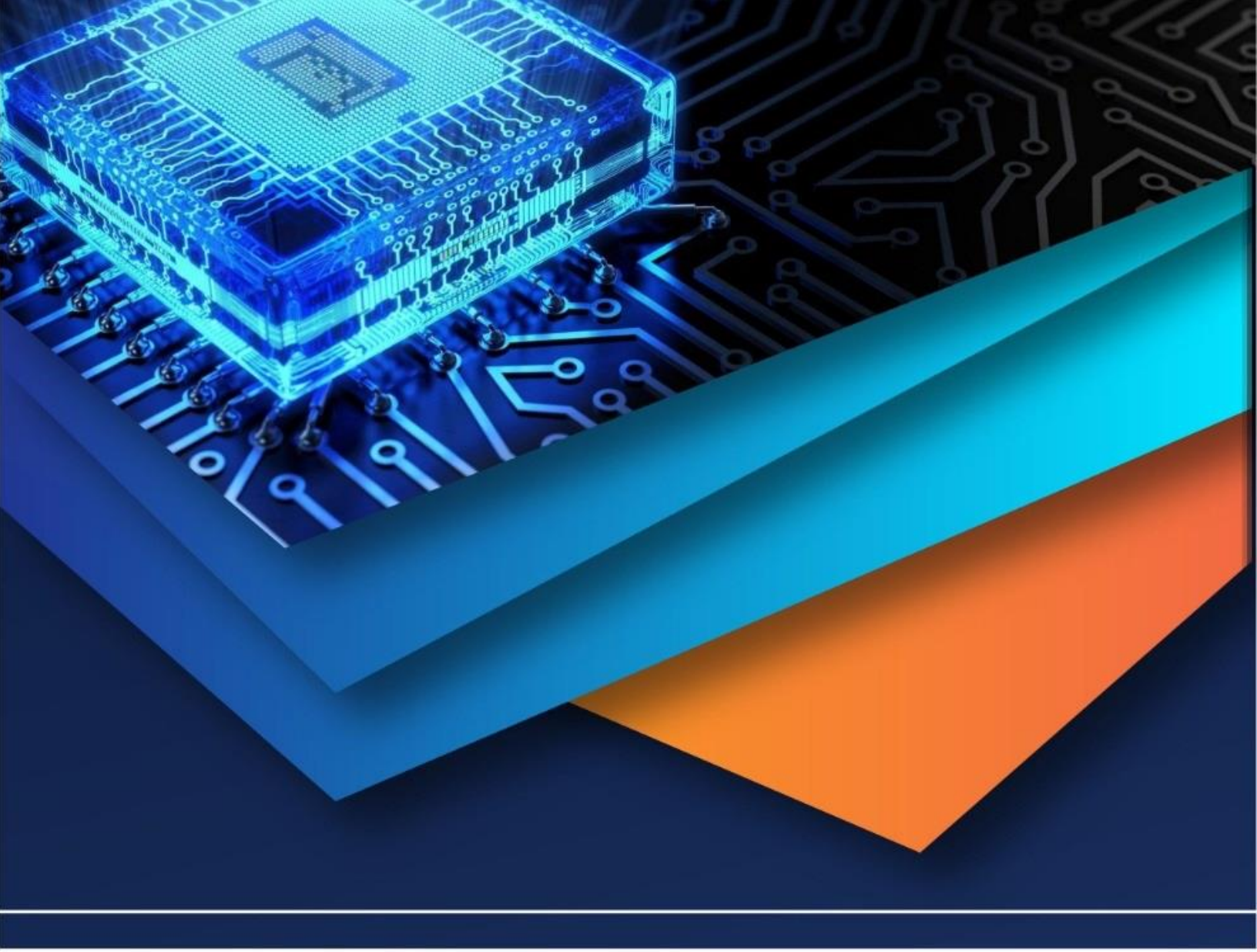
The application of machine learning in credit card fraud detection has proven to be an indispensable tool in safeguarding financial transactions. Through the utilization of sophisticated algorithms, the system can efficiently identify fraudulent activities, thereby minimizing the risks associated with unauthorized transactions. The implementation of such technology not only enhances the security of financial institutions and their customers but also contributes to the overall stability and trust within the financial ecosystem. As advancements continue to refine these detection systems, the future holds promising prospects for even more robust and reliable fraud prevention measures, ensuring a safer and more secure financial landscape for all stakeholders.

VII.FUTURE ENHANCEMENT

Future enhancements for Credit Card Fraud Detection using machine learning could include the integration of advanced deep learning techniques, such as recurrent neural networks or transformers, to capture complex temporal dependencies in transaction data. Additionally, implementing real-time anomaly detection algorithms and leveraging blockchain technology for secure transaction verification could bolster the system's fraud detection capabilities. Integrating explainable AI models to provide transparent insights into the decision-making process would enhance trust and understanding, while also facilitating regulatory compliance. Moreover, exploring the potential of federated learning to enable collaborative model training across multiple institutions without sharing sensitive data could significantly improve the overall fraud detection framework's robustness and accuracy.

REFERENCES

[1] Bhattacharyya S, Jha S, Tharakunnel K. Credit Card Fraud Detection using Machine Learning: A Survey. IEEE Access. 2019;7:37393-37420.
[2] Dal Pozzolo A, Caelen O, Le Borgne Y-A, et al. Calibrating Probability with Undersampling for Unbalanced Classification. In: Intelligent Data Analysis. Springer; 2015:160-173.
[3] Phua C, Lee V, Smith K, Gayler R. A Comprehensive Survey of Data Mining-based Fraud Detection Research. 2010.
[4] Bhattacharyya S, Jha S, Tharakunnel K. Unsupervised Machine Learning in Credit Card Fraud Detection: A Comparison of Novel Approaches. Expert Systems with Applications. 2019;118:437-453.
[5] Jiang Z, Cao J, Cao J, et al. Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. IEEE Transactions on Neural Networks and Learning Systems. 2016;27(10):2064-2077.
[6] Lázaro-Gredilla M, Ranga A, Arapakis I, Vallet D. Sequential Anomaly Detection in Credit Card Transactions. Information Sciences. 2015;303:140-156.
[7] Zheng Y, Lu X, Chen H, Jajodia S. VDDoS: Virtual Currency in the DDoS Service. IEEE Transactions on Dependable and Secure Computing. 2017;14(2):154-168.
[8] Nasrabadi NM. Pattern Recognition and Machine Learning. Journal of Electronic Imaging. 2007;16(4):049901.
[9] Bramer M. Principles of Data Mining. Springer; 2007.
[10] Shmueli G, Patel NR, Bruce PC. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. John Wiley & Sons; 2007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)