iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55567

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 11 Issue VIII Aug 2023- Available at www.ijraset.com

Comparative Study of Machine Learning Algorithms for Credit Card Fraud Detection

D. Pooja Sri[1], G. Praveen Babu[2]
Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad, University College Of Engineering, Science & Technology Hyderabad

Abstract: ***In the present world, advancement in technologies like e-commerce and financial technology (FinTech) has led to a surge in the daily volume of online card transactions and there are a lot of issues with credit cards. Identifying CC scams is currently among the most typical problems worldwide. An individual's credit card information may be fraudulently obtained and used by criminals for fraudulent purposes. Due to this significant increase in credit card fraud that impacts banks, merchants, and card issuers, it is essential to develop mechanisms that ensure the security and integrity of credit card transactions. This research examines divergent approaches to detecting credit card fraud using machine learning (ML) Models. The algorithms used are the Random Forest algorithm, Decision Trees, and the AdaBoost algorithm. The outcomes of these algorithms are based on accuracy, precision, recall, and F1-score and AUC-ROC score. Different models are compared and the algorithm that has good evaluation metrics is considered the best algorithm that is used to detect fraud.***
Keywords: Credit card fraud, Machine Learning, Random Forest, Accuracy, Confusion Matrix

I. INTRODUCTION

E-payment systems are crucial in today's culture of intense financial competition. The majority of financial institutions have improved the public's access to business services through Internet banking in the twenty-first century. The process of purchasing goods has been greatly simplified. Financial institutions frequently give their customers cards that make it simple for them to make purchases when they don't have cash on hand. Customers also benefit from credit cards because they are shielded from products that might be damaged, misplaced, or even stolen. Customers must verify the transaction with the merchant before making any purchases with their credit card. Credit card fraud occurs when a thief makes purchases using another person's credit card information without that person's permission. As a result of business migration to the Internet and the electronic financial transactions that take place in fostering the cashless economy, accurate fraud detection has become essential in protecting such transactions. There have been billion-dollar losses from credit card theft as a result of the widespread use of credit cards and insufficient security measures. To reduce their losses, all credit card issuing organizations must now implement reliable fraud detection systems. Because financial institutions are frequently reluctant to disclose such information, it is challenging to estimate losses with any degree of precision. As per the Nilson report [6], $408 billion in worldwide losses are projected over the next ten years, which is based on the global network payment card transactions projected worldwide.

II. LITERATURE SURVEY

Numerous researchers have suggested various ML algorithms to efficiently identify and minimize fraud. The primary goal of this research is to evaluate various machine learning models and select the most effective one for fraud detection. Online and offline fraud, card theft, data phishing, application fraud, and telecommunication fraud are some of the various types of fraud mentioned in the study paper [1]. As fraudsters become more sophisticated, their ability to commit fraud without compromising users' personal information or breaking the law increases daily.

As mentioned in the paper [2] in addition to the conventional techniques, hybrid techniques that combine majority voting and Adaboost techniques can be used to identify credit card fraud. To test the robustness of the detecting algorithms, the authors used the real-time publicly available credit card dataset from the financial institution and added noise to the sample data.

CCFD is always challenging, and models are unable to yield higher accuracy [3] as per the “Review of Machine Learning Approach on Credit Card Fraud Detection”, so a hybrid solution using the Artificial Neural Network (ANN) was proposed. It focuses on the federated learning framework for assuring the data privacy and security challenge.By combining the ANN with a federated learning approach, data availability and data confidentiality can be guaranteed to credit card holders in order to prevent spoofing fraud, which is difficult to detect when comparing real and fake information about the card details that are spoofed by cyber attackers.

Since the card-related information would be the same as that from a well-known, reliable source, it is very challenging to distinguish the fraudulent transaction from the original transaction.

III. PROPOSED METHODOLOGY

In the digital age, it is crucial for financial organizations to identify credit card fraud. The most cutting-edge and promising method for helping businesses stop the fraudulent activities that cause ever-increasing losses each year is currently machine learning. The customer data is analysed using a variety of machine intelligence algorithms using the proposed methodology. The analysis of the data set and the user's current dataset is supported by the classification process of algorithms, which then optimizes the performance of the model based on training and testing. Accuracy, precision, recall, and F1 score are taken into account when evaluating the performance. The model with the highest predicted performance in detecting fraud within a particular set of transactions is the most effective one. For the purpose of this study, credit card fraud is detected using the ML models below.

1) Random Forest
2) Decision trees
3) AdaBoost algorithm

In order to determine the best algorithm based on the evaluation metrics, the research compared and analyzed these machine learning techniques (Random Forest, Decision Tree, and AdaBoost). The most accurate model is then chosen, and it is used to categorize honest and dishonest transactions. A visual representation of the system's components, including its structure like input, process, and output, is called a system architecture. It primarily focuses on the system's step-by-step construction, including the key interactions and a high-level description of the problem solution.

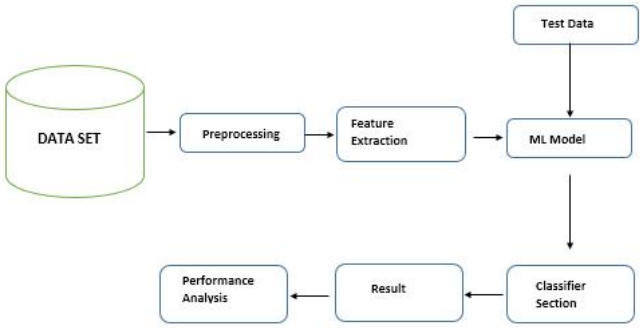


Fig. 1 System Architecture

The architecture diagram in Fig. 1 above provides information on the elements that make it possible for this research's comparative study. Data is first gathered from the source and sent for preprocessing, where it is cleaned up and processed by having redundant data, empty spaces, and null values removed. The following step involves extracting the features from the data that are crucial to building a model for additional classification. After data preparation, it will be divided into training and testing datasets, which are provided to ML models for fraud prediction.

When splitting the data, various strategies could be used, with the training dataset being used to train the model using the known statistics and the testing dataset being used to test the trained model's ability to predict fraud.

The results are then assessed using the metrics after a classifier is chosen to distinguish between fraudulent and legitimate transactions in the following step. The performance analysis is then completed to compare which algorithm performs better based on metrics such as accuracy score.

IV. IMPLEMENTATION

A. Data Acquisition

We have used real-time credit card transaction data from European cardholders to evaluate and contrast the ML models. In 2 days, there were 284,807 transactions, 492 of which were marked as fraudulent. Users conduct credit card transactions for online purchases, and these transactions need to be secure. To protect the confidentiality of user transaction data using the PCA (Principal Component Analysis) algorithm, people have converted transaction data into numerical format on Kaggle [4].

The Random Forest, Decision Forest, and AdaBoost algorithms are trained using the file "creditcard.csv," and after validation, it is tested to determine whether the dataset contains legitimate or fraudulent transaction transactions. The dataset is split into two categories as a result: preparation data and experiment data[5], where preparation data is given to the model for training, and experiment data is used to test the model.

B. Data Preprocessing

Data cleaning, which involves making the raw data ready (i.e., clean and formatted) so that the machine learning model can use it to make predictions, is the most important stage in the preprocessing of data. In addition to eliminating null values and blank spaces, it also entails separating the dataset into train and test sets, formatting the data as needed, and performing feature extraction for analysis.

PCA (principal components analysis) is one of the feature extraction methods used in this study. By converting the transactional data into a numerical format, PCA helped to reduce the data's dimensionality. The dependence of variables on one another is then determined by plotting a correlation matrix, as seen in Fig. 2.

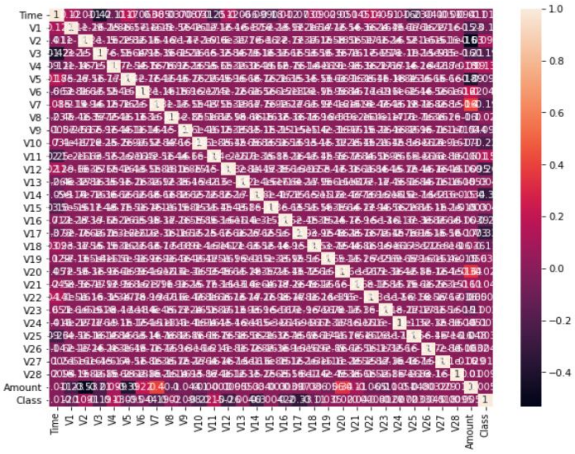


Fig. 2 Correlation Matrix

C. Splitting into the Train Set and Test Set

Only 0.172% of all card transactions in the Kaggle credit card dataset used in this study were fraudulent, which indicates a severe imbalance in the dataset. Figure 3 displays the dataset's visualization. Credit card issuers are unable to scrutinize each transaction to spot fraud because of the enormous volume of credit card transactions. In order to conduct the analysis, a sample of the dataset must be taken. To handle an unbalanced dataset and produce a balanced dataset, various sampling techniques from the synthetic minority oversampling technique (SMOTE) and adaptive synthetic (ADASYN) can be used [7].

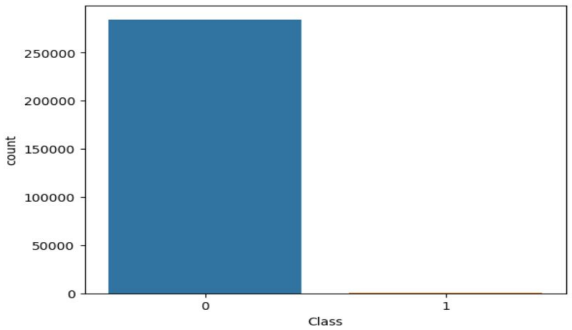


Fig. 3 Visualization of the dataset

The majority of financial transactions, including money transfers and payments made with a debit or credit card, are legitimate. Due to the fact that only a small percentage of occurrences are fraudulent, the dataset is unbalanced. The dataset's sampling strategy, the variables selected, and the detecting technique(s) employed have a significant impact on how well card transaction fraud is detected [8].

In this research, the dataset is balanced using the under-sampling technique by reducing the number of transactions from the majority class, which has typical transactions. After under-sampling, the dataset is visualized in Fig. 4.

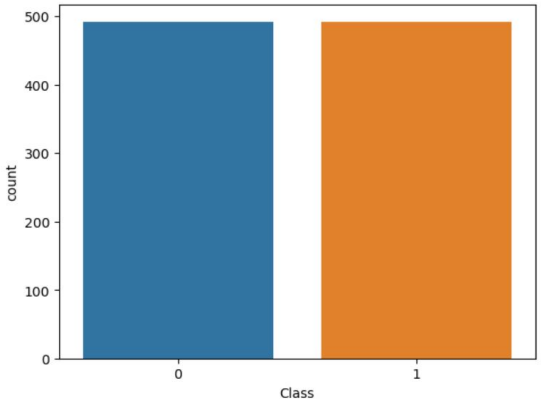


Fig. 4 Visualization of the dataset after under-sampling

D. Machine Learning Models

1) Random Forest Algorithm

A variety of classifier algorithms are combined internally by the random forest algorithm to create an accurate classifier model. A classification model will be trained using a decision tree. It uses various algorithms or decision trees at random, creating a forest tree as a result.

The predictive accuracy of the input dataset is increased by averaging the results from multiple subsets of the input dataset using the supervised approach classifier known as random forest. It builds decision trees on different datasets, taking into account their average for the classification process and majority vote for regression. When applied to datasets with continuous variables, random forest performs remarkably well.

The randomization of bootstrap sampling of a decision and selection of attributes cannot guarantee that all of them have the same stability in decision making, according to the research paper [9]. While all basic classifiers have the same weight, the random forest has a relatively high weight compared to others. In order to categorize the variables, Random Forest ranks their importance. It then uses multiple iterations of the same model to attempt to predict the outcome.

2) Decision Tree

A supervised ML model used for decision-making is a decision tree classifier. Both classification and regression problems can be solved using it. It divides a data set into smaller subsets and has a hierarchy tree-like structure. Every decision tree consists of a set of nodes and branches, where internal nodes stand in for the decision's features and external or leaf nodes for its results. Branches are the decision-making criteria or rules that resemble human thought. It will analyze the effects of each branch and visually outline the results of the complex decision-making problem. Decidual and categorical data are both handled by decision trees. A decision tree's process moves forward from the root node to predict the input dataset's class. This algorithm tracks the branch and moves to the following node by comparing the attributes of the actual dataset with those of the root attribute. The algorithm verifies the attribute value with the other sub-nodes twice before moving. A precise, condensed decision tree is what the Decision Tree model aims to create. For the purpose of detecting credit card fraud, the decision tree has two stages. The decision tree is first built using the provided training data, and then decision rules are used to classify incoming credit card transactions.

3) AdaBoost Algorithm

Adaptive Boosting, also known as the AdaBoost algorithm, is a boosting technique that is applied as an ensemble method in machine learning. By combining several straightforward models, the boosting method in machine learning aims to produce highly accurate models. The AdaBoost algorithm's working methodology is shown in Fig. 6.5, where the ensemble method takes into account all of N models' predecessors and combines the previous model with the weights of its successors. When teaching weak learners, it works best. As a result, AdaBoost is used to solve the classification problem by combining several weak learners into a single strong learner. ML Models are combined with the AdaBoost method [10] to enhance the classification problem and have a positive effect.

E. Evaluation Metrics

The effectiveness and performance of the statistical or machine learning model are quantified using evaluation metrics. These metrics provide information on how well the model is performing and help in comparing different models or algorithms. Here are a few metrics that were employed in this study to assess how well the models performed.

1) Accuracy
2) Confusion Matrix
3) Precision
4) Recall
5) F1-Score
6) AUC-ROC

V. EXPERIMENTAL RESULTS

The three machine learning methods are compared in this section by taking into account all the performance evaluation metrics (such as accuracy, precision, recall, F1 Score, Roc-AUC score, and confusion matrix), and the results are presented.

Fig. 5 to Fig 7, show how a confusion matrix for three different models functions based on the occurrences of False Positive and False Negative values, improving the model's accuracy.

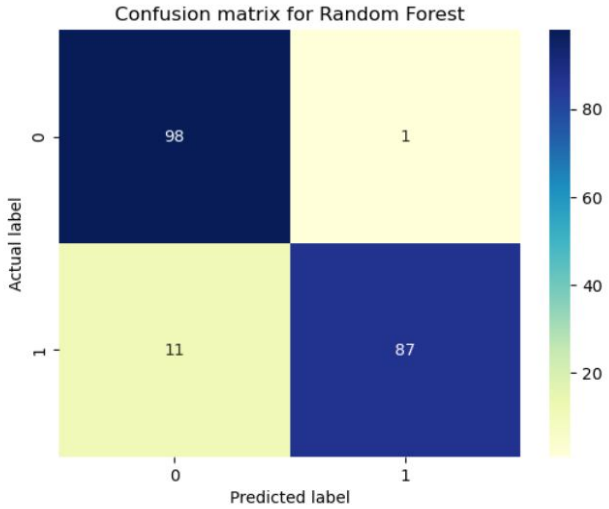


Fig. 5 Confusion Matrix for Random Forest

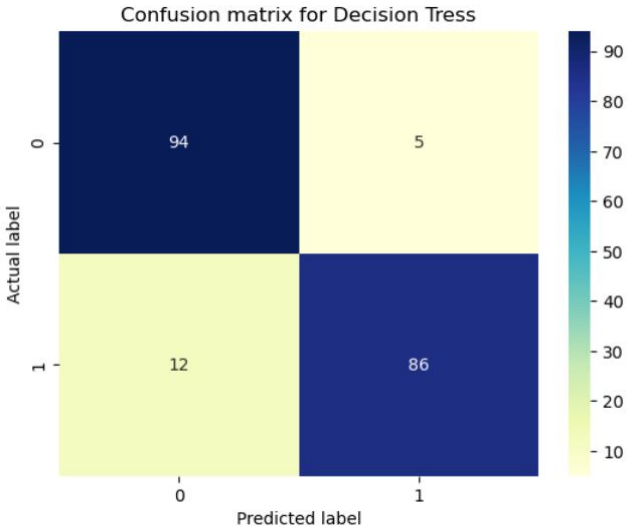


Fig. 6 Confusion Matrix for Decision Tree

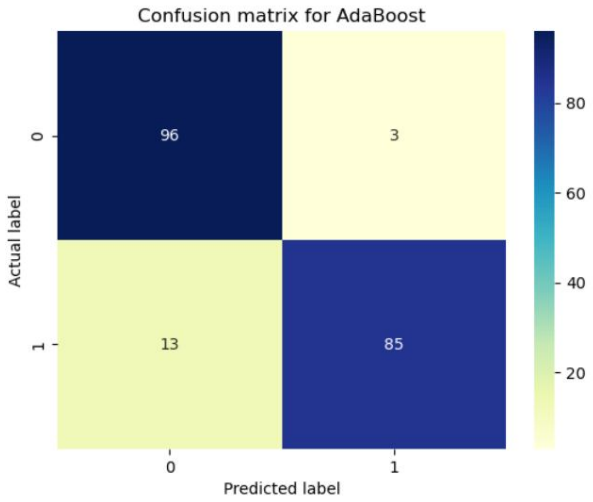


Fig. 7 Confusion Matrix for AdaBoost

The comparative outcomes of the machine intelligence algorithms, which are based on the values of evaluation metrics like Precision, Recall, F1 Score, and AUC-ROC Score, are shown in Figs. 8 and 9.

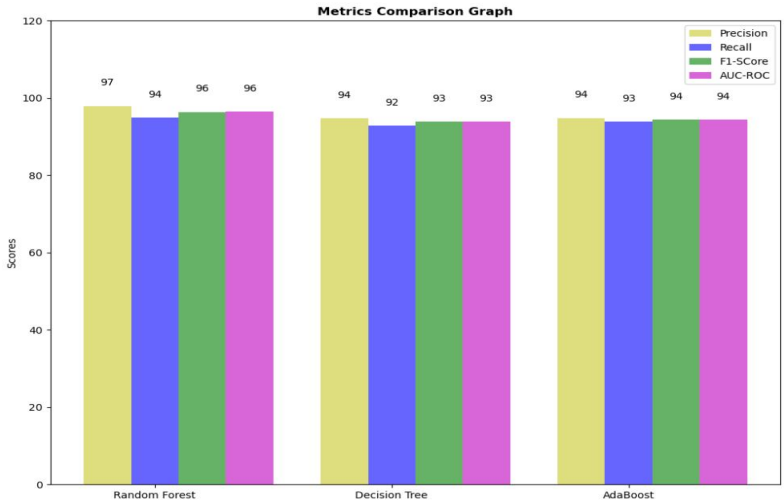


Fig.8 Metrics Comparison Graph

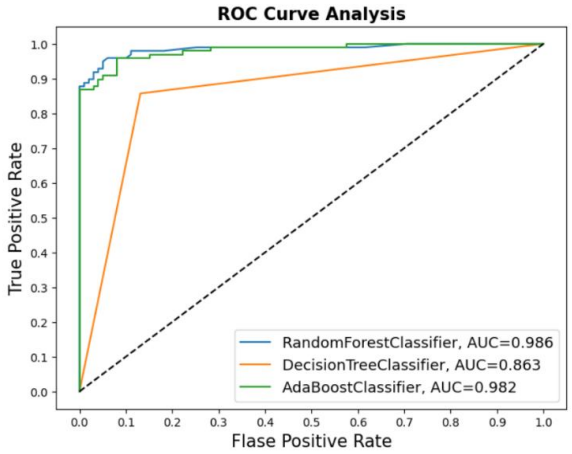


Fig. 9 ROC Curve Analysis

TABLE I Comparison Statistics of ML Models

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Random Forest	96.4%	97.8%	94.8%	96.3%	96.4%
Decision Tree	93.9%	94.7%	92.8%	93.8%	93.9%
Adaptive Boost	94.4%	94.8%	93.8%	94.3%	94.4%

The Random Forest algorithm outperforms the AdaBoost and Decision Tree algorithms in terms of making predictions of CC fraud, according to the statistics comparison of ML models shown in Table 1.

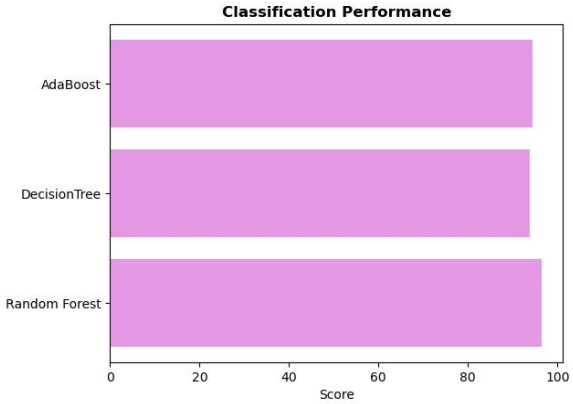


Fig. 10 Classification Performance

The accuracy of Random Forest is 96%, which is best compared to the accuracy of the AdaBoost model (94%), and the accuracy of the Decision tree (93%), as shown in the classification Performance Fig. 10.

VI. CONCLUSIONS

In terms of performance values for the evaluation metrics, Random Forest, AdaBoost, and Decision Tree are the next best accuracy models, according to the comparative analysis of the various machine learning algorithms used in this research. Statistics of the evaluation metrics are presented along with a comparison accuracy graph to help evaluate the models and choose the best model for the CCFD system. In conclusion, it has been demonstrated that data science intelligence techniques based on machine learning models are effective at identifying credit card fraud and saving millions of dollars by lowering cybersecurity attacks in the financial and banking sectors.

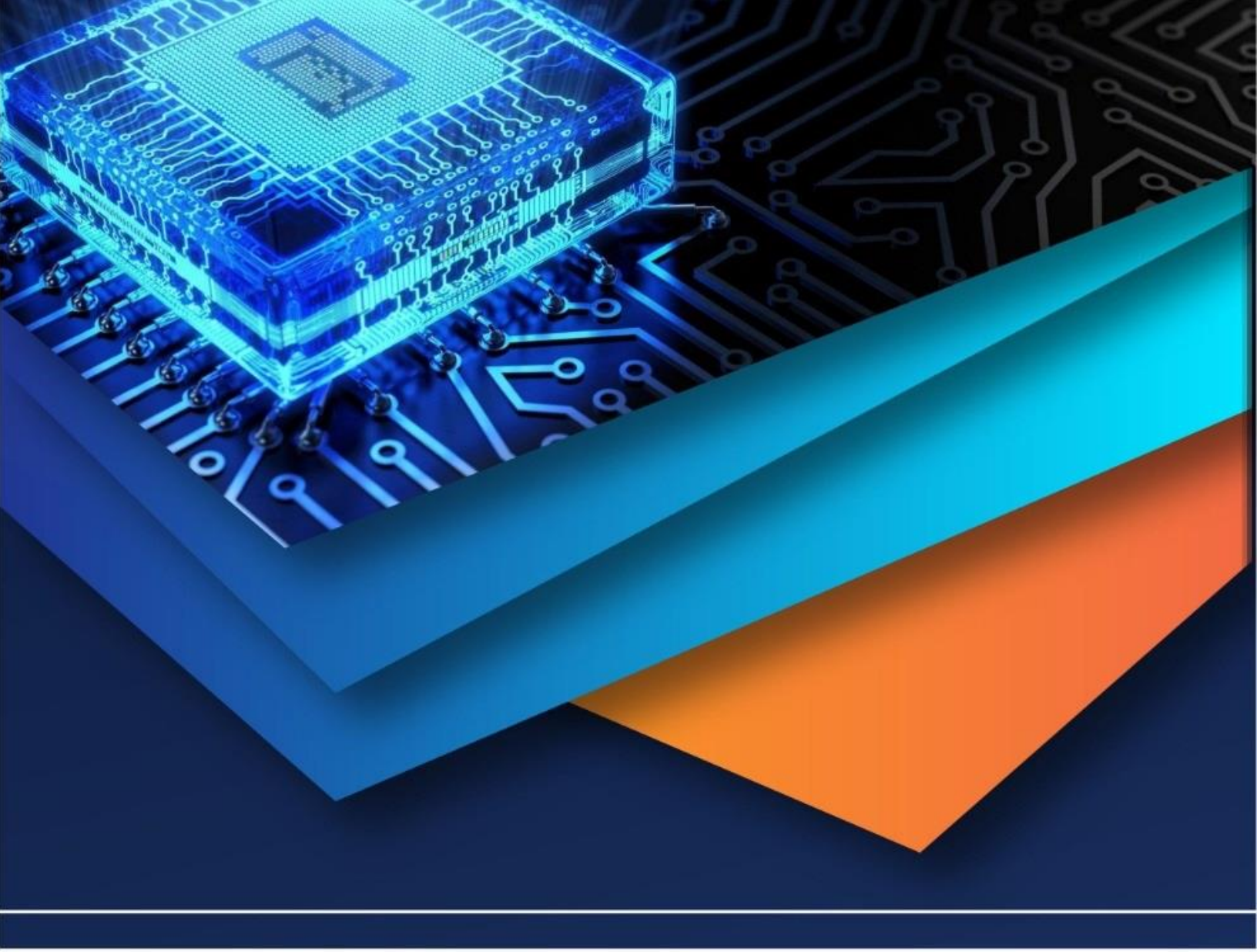
VII. FUTURE ENHANCEMENTS

As using real-time datasets makes model training efficient, avoids the issue of data imbalance, and enables tests with more accurate results, future work may involve extending the suggested methodology to real-time large datasets, such as those handled securely by banking and financial institutions. Various resampling techniques and algorithms may also be used to better predict fraudulent transactions.

REFERENCES

[1] D. Tanouz, R. R. Subramanian, D. Eswar, G. V. P. Reddy, A. R. Kumar, and C. V. N. M. Praneeth, "Credit Card Fraud Detection Using Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 967-972, doi: 10.1109/ICICCS51141.2021.9432308.
[2] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," in IEEE Access, vol. 6, pp. 14277-14284, 2018, doi: 10.1109/ACCESS.2018.2806420.
[3] Bin Sulaiman, R., Schetinin, V. & Sant, P. Review of Machine Learning Approach on Credit Card Fraud Detection. Hum-Cent Intell Syst 2, 55–68 (2022). https://doi.org/10.1007/s44230-022-00004-0
[4] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
[5] A. Priya, A. S. Narayanan, S. Madhu Bala and B. D. Patel, "Optimal Algorithm for Credit Card Fraud Detection," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2022, pp. 1091-1098, doi: 10.1109/ICAIS53314.2022.9742922

[6] https://nilsonreport.com/
[7] T. C. Tran and T. K. Dang, "Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection," 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, Korea (South), 2021, pp. 1-7, doi: 10.1109/IMCOM51814.2021.9377352.
[8] S. K. S, K. K. Shah, K. Kumar, K. K. Patel, and A. R. Sah, "Credit Card Fraud Detection Using Machine Learning Model," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-7, doi: 10.1109/MysuruCon55714.2022.9972647.
[9] Jemima Jebaseeli, T., Venkatesan, R., Ramalakshmi, K. (2021). Fraud Detection for Credit Card Transactions Using Random Forest Algorithm. In: Peter, J., Fernandes, S., Alavi, A. (eds) Intelligence in Big Data Technologies—Beyond the Hype. Advances in Intelligent Systems and Computing, vol 1167. Springer, Singapore. https://doi.org/10.1007/978-981-15-5285-4_18
[10] E. Ileberi, Y. Sun and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," in IEEE Access, vol. 9, pp. 165286-165294, 2021, doi: 10.1109/ACCESS.2021.3134330



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY