iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 Issue: IV Month of publication: April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60531

Machine Learning Approaches to Credit Card Fraud Detection

Deepali P. Kadam[1], Rushabh S. Chiparikar[2], Mahesh A. Kamble[3], Medha H. Attarde[4]
Department of Information Technology, *N.Y.S.S. Datta Meghe College of Engineering*, Airoli, Navi Mumbai, India

Abstract: Credit card fraud continues to be a major problem in today's digital world, and it needs to be detected and prevented. With the rise of online shopping, the risk of fraud goes up, so we need to find better ways to stop it. Machine learning algorithms are a promising way to spot fraudulent transactions in large amounts of credit card data. But there are challenges to algorithm accuracy, such as not having enough data, having too few examples of fraud, and fraudsters using new tricks all the time. In this study we have looked how well three common machine learning algorithms—Random Forest, Decision Tree, and Logistic Regression—work at spotting credit card fraud using real-world data from transactions. To balance the uneven distribution of fraudulent and genuine transactions, we use SMOTE. We compare algorithms using precision and recall, which measure their ability to identify fraud accurately. Machine learning models analyse historical transactions, recognizing patterns that distinguish normal and fraudulent behaviour. This approach improves fraud detection for credit card transactions, ensuring financial security in online commerce.

Index Terms: Credit Card Fraud Detection, Synthetic Minority Over-sampling Technique (SMOTE), Machine learning algorithms, Random Forest, Decision Tree, Logistic Regression.

I. INTRODUCTION

Credit card fraud (CCF) is a type of identity theft in which someone other than the owner makes an unlawful transaction using a credit card or account details [1]. A credit card that has been stolen, lost, or counterfeited might result in fraud. Card-not-present fraud, or the use of your credit card number in e-commerce transactions has also become increasingly common as a result of the increase in online shopping [1]. Increased fraud, such as CCF, has resulted from the expansion of e-banking and several online payment environments, resulting in annual losses of billions of dollars. The total value of transactions using cards issued in the Single Euro Payments Area (SEPA) amounted to €5.40 trillion in 2021 [2]. In 2021, €1.53 billion (0.028%) of card transactions were determined to be fraudulent, representing a decline of 7% from the previous year. Despite a 12.3% increase in total card payment values, the amount of fraudulent transactions decreased by 11.2%. This resulted in a decrease in the proportion of fraudulent transactions from 0.026% in 2020 to 0.021% in 2021. This reduction in fraud share to its lowest level since data collection began in 2008 indicates the effectiveness of measures aimed at detecting and preventing fraud [2]. Some online payments only require basic card information, but this information can be stolen unnoticed through phishing attacks. Fraudsters misuse stolen card details for fraudulent purchases, often leaving victims unaware of the breach. Despite efforts to protect card details, factors like phishing sites or lost/stolen cards can lead to compromised information. Detecting fraud involves analysing consumer spending patterns using data analysis and machine learning. This helps identify suspicious transactions and reduce the risk of fraudulent purchases.

II. LITERATURE REVIEW

In the study "Credit Card Fraud Detection Using the State of the art Machine Learning and Deep Learning Algorithms," by Fawaz Khaled Alarfaj and Iqra Malik (2022), the authors confronted the very important issue nowadays—credit card fraud, more experienced through increasing transactions carried out online [1]. The numerical states highlight the extremely large financial losses for the cardholder and issuing banks that face the fraudulent activities and the supporting challenges of class imbalance, evolving fraud tactics, high false alarm rates, and data accessibility. However, the existing machine learning methods produce partial answers, as their effectiveness is impeded by the low accuracy. The authors propose solving this problem through the utilization of state-of-the-art deep learning algorithms. They compare the performance of these algorithms on the benchmark dataset of European cards first, with the aim to compare performatively against machine learning methods, then using convolutional neural networks.

Since they significantly improved accuracy at detection this way by various experiments and training, while tracking their improvements and presenting these in better terms, we further discuss ways of balancing data and ways that reduced the false negative rates—re-evidence of how robust their methods are on this real-world credit card fraud problem.

In the paper "Detecting Credit Card Fraud Using Selected Machine Learning Algorithms," edited by Maja Puh and Ljiljana Brkic (2019), they put down an issue in the background of credit card fraud detection in the paper [3]. They also point at increased interest in using machine learning algorithms for fraud detection but also at challenges coming with regularities. The authors in this paper have completed their analysis with fraudulent cases and compared the models, which had empirical and real credit card transactions. They limit the imbalance observed in class sizes and take care of the ever-changing fraud patterns, respectively, with respect to incremental learning. Measure techniques based on precision, recall metrics denote fraud detection effectiveness.

The paper "Credit card Fraud Detection using Machine Learning" authored by Tanouz and Subramanian (2021), in their paper have aptly stated about the growing problem pertaining to credit card fraudulent practices against the background of the increased popularity of their usage [4]. They have put forth developing a fraud detection algorithm which would be in a position to identify exactly those who when caught would be indulging in fraudulent activity. The introduced research deals with a heavily imbalanced dataset and tries to deal with it with the help of diversely ranged state-of-the-art classification machine learning-based algorithms: Logistic Regression, Random Forest, and Naïve Bayes. Performance evaluation is done by calculating accuracy, precision, recall, F1 score, confusion matrix, and ROC-AUC scores for those algorithms using the respective metrics.

In the paper "Enhanced credit card fraud detection based on attention mechanism and LSTM deep model" by Ibtissam Benchaji and Samira Douzi (2021), the authors address the escalating issue of credit card fraud, particularly in online transactions [5]. They aim to develop a novel system for credit card fraud detection by leveraging sequential modelling of data, incorporating attention mechanism and LSTM deep recurrent neural networks. The proposed model considers the sequential nature of transactional data to accurately identify important transactions predictive of fraudulent activities. By combining the UMAP method for feature selection, LSTM networks for incorporating transaction sequences, and attention mechanism to enhance performance, the authors achieve strong results in terms of efficiency and effectiveness. This research aims to significantly improve fraud detection systems for financial institutions, thereby reducing substantial losses incurred due to fraudulent activities.

In the paper "Credit Card Fraud Detection Based on Machine and Deep Learning" by Hassan Najadat and Ola Altiti (2020), it is answered that the surging rate of credit card fraud has been associated with the general skyrocketing trends the world over in the use of credit cards for daily transactions [6]. The research seeks to the present improvements to the detection of such fraudulent activities related to the use of credit cards. They modelled their approach using an IEEE-CIS Fraud Detection dataset from Kaggle to separate a genuine transaction from a fraud transaction in a priori. In their proposed approach, the authors design a Bidirectional LSTM-Max Pooling-BiGRUMax Pooling model with defining Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) architectures in the model. For comparing, other than these six machine learning classification algorithms are used. The result specifies that their model is superior to any of the above, which gives the score of the model at 91.37%, and thus clearly, it is showing the efficacy in the work of credit card fraud detection.

The authors in the paper "Credit Card Fraud Detection Based on Machine Learning Methods," Dejan Varmedja and Mirjana Karanovic (2019), tried that the problem of credit card should be solved with the help of using different kinds of machine learning algorithms [7]. They have provided a Credit Card Fraud Detection dataset and used the SMOTE technique in case of data overfitting problems from the data happening in the dataset. First, they do the feature selection, followed by splitting the data into training and test data. This study has compared the performance of Logistic Regression, Random Forest, Naive Bayes, and Multilayer Perceptron. Based on the findings, such high accuracy is presented by each algorithm in the detection of credit card fraud; thus, solidly suggesting that the model developed may be solid also in the detection of other irregularities.

In their research paper "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison," by Samidha Khatri and Aishwarya Arora (2020), tried to make an attempt to explain the problem of the growing rate of credit card fraud among credit card holders [8]. Compare time-proven algorithms for supervised learning, classifying a real transaction versus a fraud. The above research therefore gives some insights through the concept used in the machine learning on effectual methods of fraud detections for both the financial institutions and the card users in this way reducing the existing risks in today's economic scenarios environment of delinquencies on the credit cards.

In the paper "Random Forest for Credit Card Fraud Detection" by Shiyang Xuan and Guanjun Liu (2018). Here, the authors promise that they will "decrease the great risk of incurring financial losses from credit card fraud" nowadays, such losses and risks are considered to be one of the most serious [9].

They propose to perform feature extraction of normal and fraudulent behaviour at two levels: at first by using historical transaction data and then training features using two types of Random Forests with different base classifiers. The authors carried out the comparative analysis for two random forests and assessed their detection performance for credit card frauds using data from an e-commerce company in China. This is in a bid to come up with an effective solution for the detection of frauds, hence timeously recognition of fraud transactions and minimization of damage done on the entire account.

III. METHODOLOGY

Our research initiative follows a systematic methodology to enhance fraud detection accuracy. We start by putting together a big collection of records of credit card transactions. We make sure that this data includes a wide range of transaction types and examples of fraud. After that, we clean up the data by taking care of problems like missing values, outliers, and the fact that one class is much bigger than the other. We use Random Forest, Decision Tree, and Logistic Regression algorithms to choose features and train models using the data after it has been cleaned up. We use the Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes because one class is much larger than the other. After the models have been trained, we use metrics like accuracy, precision, recall, and F1 to evaluate how well each algorithm works. We combine the most successful algorithm with a user-friendly interface, enabling real-time fraud detection with actionable insights for stakeholders. By doing so, we reduce credit card fraud risks and protect financial transactions efficiently.

- In the first stage, we collect a large dataset of credit card transactions. This dataset is crucial for training and evaluating our fraud detection algorithms. We gather transaction data from banks, online stores, and payment companies. Each transaction record includes important information like the amount, date, cardholder details, and whether it was fraudulent or not. We make sure to include a variety of transactions to train our algorithms effectively. To ensure data protection, strict measures are in place during data collection, protecting sensitive information. This carefully compiled dataset is the foundation for our research, allowing us to create and validate fraud detection models that identify and reduce fraudulent credit card transactions with precision.

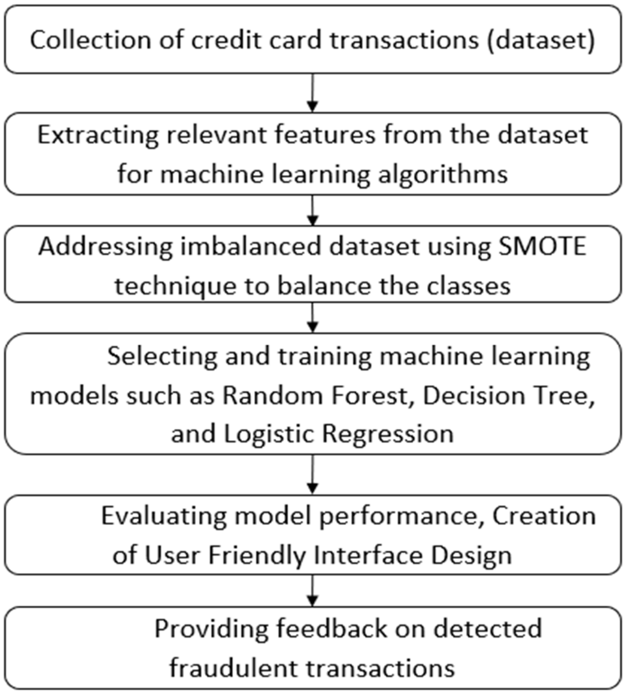

Collection of credit card transactions (dataset)
Extracting relevant features from the dataset for machine learning algorithms
Addressing imbalanced dataset using SMOTE technique to balance the classes
Selecting and training machine learning models such as Random Forest, Decision Tree, and Logistic Regression
Evaluating model performance, Creation of User Friendly Interface Design
Providing feedback on detected fraudulent transactions

Fig. 1. Proposed Methodology

1) Following the next step, we pinpoint and extract the most important pieces of information from the credit card transaction data that will help train our machine learning models. Feature extraction means going through the data to find important factors or qualities that help tell the difference between fake and real transactions. These features include transaction amount, time of transaction, merchant category, cardholder location, and transaction frequency, among others. By carefully choosing which features to use, we try to focus on the ones that will help us predict best while getting rid of any that are not needed or add extra work.

2) To address the challenge of uneven distribution between categories (fraudulent and legitimate transactions) in our credit card transaction data, we use the Synthetic Minority Oversampling Technique (SMOTE). SMOTE creates new instances of the underrepresented category (fraudulent transactions) by interpolating existing data points, balancing the representation of both categories. By producing synthetic samples, SMOTE reduces the bias towards the dominant category and allows our machine learning models to train on a more balanced dataset. To avoid favouring the typical group, this method lessens their disproportionate influence on the models. By applying SMOTE, we strengthen our fraud detection system, enabling it to find fraudulent transactions more accurately. As a result, our system becomes more dependable and effective, offering actionable insights for fraud identification.

3) Following the next step, we concentrate on selecting and training machine learning algorithms that excel at detecting credit card fraud. We select Random Forest, Decision Tree, and Logistic Regression due to their widespread use and proven accuracy in classification tasks, including fraud detection. After countering class imbalance using SMOTE, the preprocessed and balanced dataset is utilized to train each model. By meticulously training and cross-validating the models, we determine the optimal parameters for each algorithm, guaranteeing high performance. Our chosen models are thoroughly tested to assess their ability to correctly identify fraudulent transactions. We measure accuracy, precision, recall, and other important metrics. This evaluation stage is essential, as it determines the effectiveness of our models before we deploy and continue assessing our fraud detection system.

4) Following the next step, we evaluate the model performance, we thoroughly test our trained machine learning models for detecting credit card fraud. Using established measurements like accuracy, precision, recall, and F1-score, we gauge the models' effectiveness and reliability. Furthermore, we focus on creating a user-friendly interface that makes interacting with our fraud detection system simple. The interface is continually enhanced based on user feedback to ensure the best possible user experience and accessibility for everyone involved.

5) In final the stage, we prioritize delivering helpful feedback on identified fraudulent transactions to key parties involved. Utilizing our machine learning advancements, we've created a structured method to convey our findings clearly. The feedback we provide aims to give practical insights to aid in fraud prevention and management. Furthermore, we prioritize transparency and clarity in our communication to enable informed decisions and proactive steps against potential future fraudulent activities.

6) In summary, our project encompasses a systematic approach to credit card fraud detection, beginning with the meticulous collection of transaction data from various sources. Leveraging feature extraction techniques and addressing class imbalance using SMOTE, we train machine learning models such as Random Forest, Decision Tree, and Logistic Regression. Rigorous evaluation of model performance and continuous interface refinement ensure user-friendly interaction and reliable fraud detection. The provision of actionable feedback on identified fraudulent transactions enhances transparency and aids in fraud prevention and management.

IV. RESULTS

We measure the model's fraud detection performance using accuracy, precision, recall, and the confusion matrix. The confusion matrix provides a detailed view of the model's classifications. For fraud detection, a two-class problem, the matrix is 2x2. The rows represent the actual class (legitimate or fraudulent), while the columns represent the predicted class. The first row shows the count of correctly classified positives (fraudulent) and false positive (legitimate labelled as fraudulent). The second row shows the count of false negative (fraudulent labelled as legitimate) and correctly classified negatives (legitimate).

Predicted Result	Actual Result	
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Fig. 2. Confusion matrix

To evaluate the performance of fraud detection models, it's crucial to first understand the details provided by the confusion matrix. This matrix breaks down the model's classification results, providing insights into its accuracy, precision, recall, and F1-score.

1) *Accuracy:* It represents the overall correctness of the model's predictions and is calculated as the ratio of the number of correctly classified instances (both true positives and true negatives) to the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

2) *Precision:* Precision measures the accuracy of the model's positive predictions, calculated as the ratio of true positives to the sum of true positives and false positives.

$$Precision = \frac{TP}{TP + FP}$$

3) *True Positive Rate (TPR) or Sensitivity or Recall:* TPR measures the proportion of actual positive cases (fraudulent transactions) that are correctly identified as positive by the model.

$$Recall = \frac{TP}{TP + FN}$$

4) *F1-Score:* The F1-score provides a balanced assessment of a classification model's precision and recall, crucial for evaluating its performance, particularly in scenarios with imbalanced datasets.

$$F1\ \ score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Through this validation process, we assessed using a confusion matrix and calculated metrics such as accuracy which gives an overall measure of rightfulness, while precision focuses on predicting positive results correctly. Recall assesses a model's ability to spot actual true positives. The F1-Score strikes a balance between precision and recall, making it valuable for evaluating performance, especially when working with unbalanced datasets.

In our endeavour to enhance credit card fraud detection through machine learning, we employed a diverse array of algorithms, including Decision Tree, Logistic Regression, and Random Forest. We chose these algorithms because each of them has special skills in handling transaction data and spotting fraudulent patterns. Decision Tree was useful for dividing the data into groups and finding the most important features for spotting fraud. Logistic Regression was useful for making a model that could tell how likely it was that fraud would happen based on different transaction details. Finally, Random Forest, renowned for its ensemble learning approach, was harnessed to collectively improve predictive accuracy while mitigating overfitting and capturing the complexities inherent in credit card transactions. By leveraging a combination of these algorithms, we aimed to conduct a comprehensive analysis of fraudulent behaviour, thereby enhancing the efficacy of our credit card fraud detection system.

Among the selected algorithms, Random Forest emerged as the standout performer, achieving the highest accuracy at 99.99%. The remarkable success achieved with the Random Forest algorithm in our credit card fraud detection model can be attributed to its ensemble learning approach. Random Forest excels in managing complex, non-linear relationships inherent in the dataset. By aggregating multiple decision trees, Random Forest collectively enhances predictive accuracy while mitigating overfitting and capturing the intricacies present in the transaction data. The robust performance of Random Forest underscores its suitability for our specific task of fraud detection, highlighting its ability to effectively navigate the diverse and dynamic nature of credit card transactions. The strategic inclusion of Random Forest alongside other algorithms like decision tree, logistic regression, reflects our commitment to a comprehensive and nuanced analysis of fraud detection. Its superior performance underscores its effectiveness in handling the complexities of the dataset, reinforcing its pivotal role in our methodology for credit card fraud detection using machine learning.

The focus shifts towards developing a user-friendly interface, ensuring accessibility and ease of interaction for individuals navigating the system. A well-designed interface not only enhances user engagement but also encourages transparent and genuine responses, thereby improving the quality and reliability of the transactional data collected. Prioritizing a user-centric approach underscores the significance of cultivating a supportive environment that empowers users to interact confidently with the tool.

TABLE I
ACCURACY OF DIFFERENT ML METHODS

Methods	Accuracy (%)
Logistic Regression	94.50
Decision Tree	99.82
Random Forest	99.99

Feedback generation plays a pivotal role in the tool's functionality, offering users insightful analyses derived from their transactional patterns. Personalized and helpful feedback helps users not only understand how to spot fraud but also makes them want to take action to prevent it. By giving users information that is specific to them, the system makes them feel like they have more control and encourages them to take steps to protect their money.

V. LIMITATIONS

Utilizing machine learning for credit card fraud detection has its drawbacks. Firstly, accuracy relies on well-balanced training data. Imbalances or insufficient diversity can result in biased detections. Complex algorithms, like random forests can be highly accurate but difficult to explain, hindering decision-making. Also, these Models might not recognize novel fraud patterns not present in the training data. Ethical and regulatory concerns, such as Data privacy and compliance issues arise with machine learning use. Lastly, Machine learning is only a part of fraud prevention and necessitates continuous oversight, adjustments, and collaboration with human experts. To overcome these limitations, a comprehensive fraud prevention approach is crucial.

VI. CONCLUSION

Machine learning techniques, such as random forest, decision tree, and logistic regression, have proven effective in detecting credit card fraud. Our research indicates significant advancements in detection accuracy, particularly when using oversampling techniques like SMOTE. Despite these advancements, challenges like data quality, algorithm complexity, and ethical concerns persist. Ongoing research and stakeholder collaboration are crucial for improving fraud detection methods and strengthening financial transaction security. A comprehensive fraud prevention strategy requires a combination of technology, regulatory enforcement, and human oversight to ensure its effectiveness.

REFERENCES

[1] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," in IEEE Access, vol. 10, pp. 39700-39715, 2022, doi: 10.1109/ACCESS.2022.3166891.

[2] European Central Bank, “Report on card fraud in 2020 and 2021,” 2023; ISBN 978-92-899-6096-0, ISSN 2315-0033, doi:10.2866/531174 QB-BI-23-001-EN-N.

[3] M. Puh and L. Brkić, "Detecting Credit Card Fraud Using Selected Machine Learning Algorithms," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 1250-1255, doi: 10.23919/MIPRO.2019.8757212.

[4] 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.

[5] Benchaji, I., Douzi, S., El Ouahidi, B. et al. Enhanced credit card fraud detection based on attention mechanism and LSTM deep model. J Big Data 8, 151 (2021). https://doi.org/10.1186/s40537-021-00541-8.

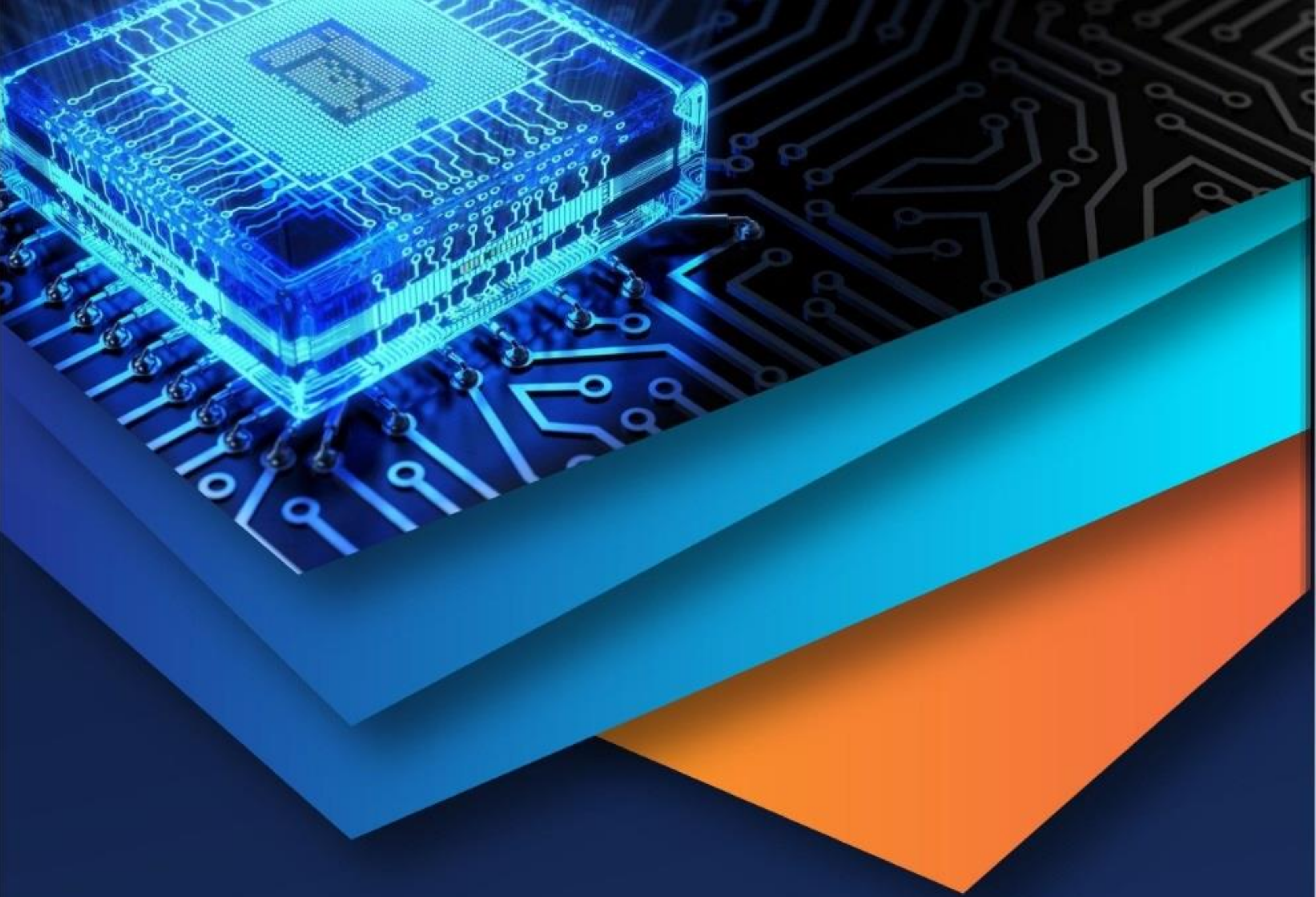
[6] H. Najadat, O. Altiti, A. A. Aqouleh and M. Younes, "Credit Card Fraud Detection Based on Machine and Deep Learning," 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 204-208, doi: 10.1109/ICICS49469.2020.239524.

[7] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2019, pp. 1-5, doi: 10.1109/INFOTEH.2019.8717766.

[8] S. Khatri, A. Arora and A. P. Agrawal, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 680-683, doi: 10.1109/Confluence47617.2020.9057851.

[9] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, "Random forest for credit card fraud detection," 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, China, 2018, pp. 1-6, doi: 10.1109/ICNSC.2018.8361343.

[10] R. Sailusha, V. Gnaneswar, R. Ramesh and G. R. Rao, "Credit Card Fraud Detection Using Machine Learning," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1264-1270, doi: 10.1109/ICICCS48265.2020.9121114



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

IMPACT FACTOR:
7.429