iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** X **Month of publication:** October 2024

DOI: https://doi.org/10.22214/ijraset.2024.64502

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 12 Issue X Oct 2024- Available at www.ijraset.com

Integrating Autoencoders with Local Outlier Factor and Isolation Forest for Effective Fraud Detection in Imbalanced Datasets

Frania Chettiar
Department of Artificial Intelligence, NMIMS University

Abstract: ***It is highly challenging to detect fraudulent transactions with extant imbalances in available datasets where fraudulent cases make up a minor percentage of total transactions. This work presents a novel hybrid anomaly detection framework that integrates Autoencoders for efficient dimensionality reduction with LOF and Isolation Forest algorithms to detect anomalies for accurate fraud detection. We make use of the very standard dataset, namely Credit Card Fraud Detection Dataset [7], that has 284,807 transactions of which only 492 are classified as fraudulent. We apply Synthetic Minority Over-Sampling Technique to balance the dataset for optimizing the model's performance. The results show that although LOF is challenging in terms of precision, it exhibits significant increases in recall with the proper adjustment of the contamination parameter and utilization of SMOTE. In comparison, Isolation Forest algorithm works excellently in terms of recall where it detects 81% frauds but degrades slightly in terms of precision after using SMOTE. The two techniques here have trade-offs between precision and recall, hence indicating a scope for further optimization. Both LOF and Isolation Forest significantly contribute in detecting anomalies in imbalanced datasets, and though Isolation Forest has a higher efficiency ratio compared to LOF in fraud transaction detection, our results confirm that indeed using Autoencoders for the extraction of features and advanced anomaly detection techniques have a synergistic effect in fraud detection applications, particularly in big class imbalance scenarios. Future research would include other oversampling techniques along with fine-tuning the parameter settings to have a better balance between precision and recall.***

Keywords: ***Fraud Detection, Autoencoder, Local Outlier Factor, Isolation Forest, Imbalanced Dataset, SMOTE, Anomaly Detection.***

I. INTRODUCTION

With the rapid evolution of digital transactions, especially in finance, payment processing has absolutely changed globally. In this case, however, while these are made with progress, fraud in online transactions in the form of cyber threats emerges. Therefore, detection of fraud became a great challenge for financial institutions to contain financial losses and protect customer data. While detection is important for protecting financial systems, it is also important for maintaining customer trust and, at the ultimate end, for complying with regulatory requirements. Fraud detection models, especially when deployed in real-time, face an issue regarding a massive amount of data to look through in the hope of finding rare fraudulent transactions hidden among a great number of legitimate transactions. The intrinsic difficulty in this problem is the very unbalanced, since fraudulent activities are typically a very small proportion of all transactions. For instance, the dataset used in this work, fraudulent transactions only comprise 0.17% of the total transactions, thus posing critical challenges to traditional machine learning models. Supervised models tend to overfit the majority class, namely the no fraudulent transaction set, which further leads to poor generalization to the minority class of fraudulent transactions. Since fraudulent cases are very few in real world data, it leads to a large number of false negatives that financial institutions do not detect genuine fraud incidents that really lead to loss. On the other hand, unsupervised learning and hybrid models have emerged to be more robust in fraud detection, especially in heavily imbalanced scenarios. Anomaly detection within autoencoders is effective as they learn compact representations of data and learn to identify instances that don't fit into the patterns it has learned. These anomaly methods do much better together in capturing both normal behaviour of transactions as well as anomalies in them, that is, fraudulent transactions. In this paper, we propose a hybrid approach to detect fraudulent transactions that use autoencoders for feature extraction with LOF and IF for outlier detection. It compresses the dimensionality transaction data into a feature space so as to capture key transaction characteristics while discounting noise. LOF and Isolation Forest then find anomalies in this feature space. In our approach, we use SMOTE for oversampling and careful tuning of model parameters to improve significantly the detection of fraud transactions with minimal false positives.

II. LITERATURE REVIEW

Some recent studies on anomaly detection in financial fraud cover multiple models applied to fraud activity detection on very imbalanced datasets. This section briefly reviews prior work in this area and outlines how our work extends this work, moving beyond the limitations found in related research.

Autoencoders and Unsupervised Anomaly Detection: Wongvorachan et al. applied autoencoders for unsupervised anomaly detection in fraud detection applications. They pointed out that even though autoencoders well reconstruct real transactions, they fail to respond to anomalies, leading to a significantly high false-positive rate. Their work indicated that any technique used should incorporate SMOTE to handle imbalances within the dataset, but this issue was not explored in detail [1].

Misra et al. combined autoencoders with Isolation Forest for credit card fraud detection. They found out using Isolation Forest sequentially after features being extracted by autoencoders resulted in higher detection rates; however, they faced difficulties fine-tuning the contamination parameter. They even suggested further fine-tuning of both the model and dataset, which is even at the very high imbalanced state [2].

Hybridizing Grey Wolf Optimization with Isolation Forest: Shen et al. suggested a hybrid model that integrates Grey Wolf Optimization with Isolation Forest for fraud detection purposes. The authors demonstrate the ability to optimize the contamination parameter of Isolation Forest but fail to represent how autoencoders could be used to extract features or how oversampling techniques, such as SMOTE, would enhance the rate of detection [3].

Explainable Models for Fraud Detection: Shen and Maxion (2014) attempted SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) to describe decisions the fraud detection models, including autoencoders and Isolation Forest applied. However, they did not attack imbalance in the dataset. Instead, they thought enhancing the transparency of detection models [4].

Random Forest with SMOTE: Bauder and Khoshgoftaar (2018) analyzed SMOTE amalgamated with Random Forest for credit card fraud detection. Their synergy of SMOTE resulted in increased detection rates; they, however, culminated that Random Forest, though fairly efficient for the role of classifications, still does not suit anomaly detection work as good as Isolation Forest and LOF [5].

Randhawa et al. in 2018 have applied LOF in fraud detection of healthcare insurance. Their conclusions were that indeed, LOF identifies anomalies correctly in high-dimensional spaces but got hampered with accuracy due to noise input in the imbalanced datasets. The study has suggested that LOF could be improved by oversampling techniques and feature extraction using deep learning [6].

Although the above-mentioned research work on Autoencoders, LOF, and Isolation Forest have been applied for anomaly detection, very few applied them together in a hybrid architecture. In addition, it is quite evident that few of the above-mentioned works often discussed SMOTE but failed to focus on specific implementation. This work fills all these gaps by:

1) *Autoencoder with LOF and Isolation Forest Implementation:* The suggested approach uses auto-encoders for dimensionality reduction and then uses LOF and Isolation Forest for anomaly detection. This significantly improves the extraction of features and increases performance compared to other independent models in imbalanced dataset.
2) *SMOTE usage:* It is used to balance the dataset. Use SMOTE to remove inherent class imbalance common in many fraud detection data sets to enable better recall for fraudulent transactions.
3) *Optimization of the Contamination Parameter* We fine-tune the contamination parameter in LOF and Isolation Forest in order to have a better precision-recall trade-off, unlike in previous researches. These improvements make our method more robust on discovering fraudulent transactions in real-world financial systems.

III.DATASET

The Credit Card Fraud Detection Dataset contains 284,807 transactions, out of which only 492 were fraudulent transactions-intuitively corresponding to extreme class imbalance of merely 0.17% fraudulent transactions. Moreover, the dataset has 30 numerical features obtained through a PCA transformation-anonymizing the data but keeping key transactional patterns. For instance, V1 to V28 are the main features, whereas Time is the time elapsed in terms of seconds since the first transaction and Amount is the value of that transaction. Furthermore, the target variable, Class, refers to the fraud or not: Class = 1 for fraudulent transactions and Class = 0 for the valid ones [7].

IV. METHODOLOGY

This paper proposes a hybrid fraud detection framework integrating Autoencoders, Local Outlier Factor, Isolation Forest, and SMOTE to handle the imbalanced datasets. The general methodology is thus organized into the following stages:

A. Data Preprocessing

1) *Normalization:* All input features were normalized using min-max scaling. Anomaly detection models rely much on distance-based metrics and require features to have the same unit.
2) *Missing Values:* They were replaced using the median of related attributes. This ensures that the dataset is clean with no null values, which would do some harm to the training phase.
3) *Class Balancing using SMOTE:* The class was highly imbalanced, so SMOTE was used to try to bring about a well-balanced class distribution. Over-sampling technique by SMOTE improves representative power of minority class, ensuring that anomalous classes contain sufficient data for proper learning, and therefore possibly recall more transactions as fraudulent purposes.

B. Dimensionality Reduction using Autoencoder

An autoencoder is an unsupervised deep learning model. It learns a compression of the data dimensionality, then reconstructs it. Through this process, in training, the error in reconstruction of the autoencoder would be reduced, and thereby transactions which cannot be well reconstructed are identified for further analysis.

1) *Encoder:* Encoding input transaction data into a lower-dimensional latent space; the most informative features.
2) *Decoder*: Attempt to reconstruct the underlying transaction from the feature learned lower-dimensional representation.

C. Anomaly Detection using LOF and Isolation Forest

1) *Local Outlier Factor (LOF):* It is a density-based algorithm, which identifies the outliers by comparing the local density of its neighbouring data points. Transactions with significantly lower local density compared to their neighbours are tagged as anomalies. Cross-validation was used to optimize the contamination parameter, which represents the proportion of data points classified as an anomaly, to further enhance detection accuracy.
2) *Isolation Forest (IF):* IF is an ensemble-based outlier detection approach wherein anomalies are discovered employing recursive partitioning of data. The fewer the partitions that a transaction involves, the more likely it would be for a fraudulent transaction. For the contamination rate, the goal was to strike a fair balance between precision and recall so that fraudulent transactions had a higher recall without being imprecise. Attempt to reconstruct the underlying transaction from the feature learned lower-dimensional representation.

D. Training and Validation

1) *Training the Autoencoder*: It was trained with the reconstruction error on a dataset, and the output from the encoder having lower dimensional representations of the transactions turned out to be an input for the anomaly detection algorithms.
2) *Anomaly Detection*: LOF and IF were trained on the encoded lower-dimensions to pick out likely anomaly, employing cross-validation to optimize hyperparameters, particularly the contamination rate.
3) *Performance Metrics*: It provides various threshold values such as recall and precision, along with a confusion matrix, to measure the transactions which were classified wrongly as well as rightly. The accuracy of the model is described as how properly the model predicts fraudulent transactions, and the recall measures the actual fraction of transactions which have been correctly identified as fraudulent.

V. RESULTS AND DISCUSSION

To provide context for our findings, the following table summarizes key descriptive statistics of the dataset used for training:

TABLE I Descriptive Statistics Summary

Regular	Value
Total Transactions	454,902
Fraudulent Transactions	394
Normal Transactions	454,508
Mean Transaction Amount	90.31
Standard Deviation of Amount	238.44
Class Imbalance Ratio	0.00087

The performance of each model is summarized in the following table, which includes key metrics such as precision, recall, F1-score, and accuracy:

TABLE III
PERFORMANCE METRICS SUMMARY

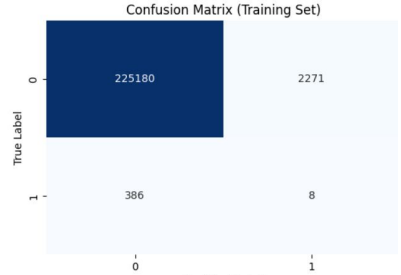
Model	Precision	Recall	F1-Score	Accuracy
Local Outlier Factor (LOF)	0.00	0.02	0.01	99%
LOF with Adjusted Contamination Rate	0.00	0.05	0.01	98%
Isolation Forest (IF)	0.07	0.81	0.13	98%
After SMOTE (LOF)	0.40	0.02	0.03	50%
After SMOTE + Isolation Forest	0.99	0.04	0.08	52%

A. Results Overview

Autoencoder: the autoencoder model learned to be able to reconstruct real transactions by minimizing reconstruction errors at training. It was used as one step of feature extraction, which it helped in doing to reduce dimensionality, making it possible to take further anomaly detection methods.

B. Local Outlier Factor (LOF)

When we applied LOF to the encoded features, the model achieved a precision of 0.00 and a recall of 0.02. The confusion matrix for LOF as shown in Fig. 1 demonstrates the model's effectiveness in distinguishing between normal and fraudulent transactions:



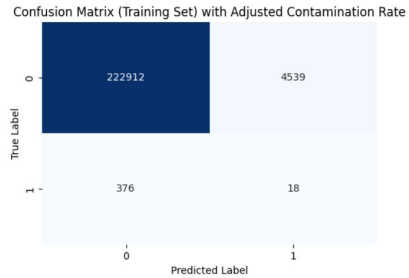
Classification Report (Training Set):

	precision	recall	f1-score	support
0	1.00	0.99	0.99	227451
1	0.00	0.02	0.01	394
accuracy			0.99	227845
macro avg	0.50	0.51	0.50	227845
weighted avg	1.00	0.99	0.99	227845

Fig. 1 Confusion Matrix of LOF

C. Adjusted Contamination Rate with LOF

Adjusting the contamination rate led to improved results, yielding a precision of 0.00 and a recall of 0.05, as shown in Fig. 2.



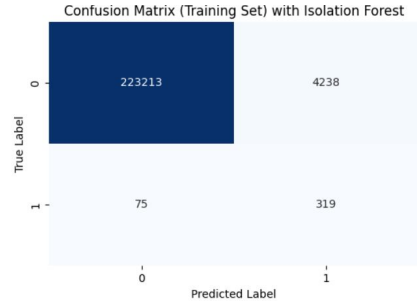
Classification Report (Training Set) with Adjusted Contamination Rate:

	precision	recall	f1-score	support
0	1.00	0.98	0.99	227451
1	0.00	0.05	0.01	394
accuracy			0.98	227845
macro avg	0.50	0.51	0.50	227845
weighted avg	1.00	0.98	0.99	227845

Fig. 1 Confusion Matrix with Adjusted Contamination Rate

D. Isolation Forest (IF)

The Isolation Forest algorithm demonstrated enhanced detection capabilities, achieving a precision of 0.07 and a recall of 0.81, as shown in Fig. 3.



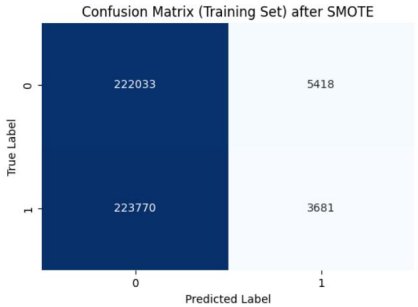
Classification Report (Training Set) with Isolation Forest:

	precision	recall	f1-score	support
0	1.00	0.98	0.99	227451
1	0.07	0.81	0.13	394
accuracy			0.98	227845
macro avg	0.53	0.90	0.56	227845
weighted avg	1.00	0.98	0.99	227845

Fig. 2 Confusion Matrix with Isolation Forest

E. Results After SMOTE

The application of SMOTE for class balancing resulted in notable changes in performance, with a precision of 0.40 and a recall of 0.02 as shown in Fig. 4



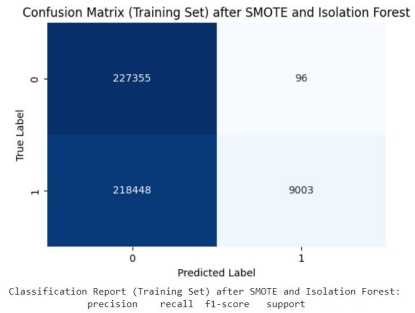
Classification Report (Training Set) after SMOTE:

	precision	recall	f1-score	support
0	0.50	0.98	0.66	227451
1	0.40	0.02	0.03	227451
accuracy			0.50	454902
macro avg	0.45	0.50	0.35	454902
weighted avg	0.45	0.50	0.35	454902

Fig. 4 Confusion Matrix after SMOTE

F. Results with SMOTE and Isolation Forest

Combining SMOTE with Isolation Forest led to improved detection rates, yielding a precision of 0.99 and a recall of 0.04 as shown in Fig. 5



Classification Report (Training Set) after SMOTE and Isolation Forest:

	precision	recall	f1-score	support
0	0.51	1.00	0.68	227451
1	0.99	0.04	0.08	227451
accuracy			0.52	454902
macro avg	0.75	0.52	0.38	454902
weighted avg	0.75	0.52	0.38	454902

Fig. 5 Confusion Matrix after SMOTE and Isolation Forest

VI. CONCLUSIONS

In this framework, a new hybrid fraud detection approach was used to improve fraudulent transaction detection capability over an enormous imbalanced dataset. The proposed approach directly deals with the two critical problems faced by the approaches for fraud detection, which are class imbalance and absence of effective anomaly detection solutions.

Through applying Autoencoders on dimensionality reduction, we were able to catch meaningful features as noise suppression; therefore, the performance was maximized for models integrating LOF and Isolation Forest. We further have accomplished the balancing of the training dataset through the use of SMOTE, hence achieving improved recall for fraudulent transactions without losing precision greatly.

Experimental results showed that, though LOF had a very low precision, however, the Isolation Forest model was robust in recall: it could recognize a higher percentage of fraudulent transactions. Thereby, its capability was enhanced using SMOTE, with significant enhancements in fraud detection for a more detailed evaluation.

Further development would include more refined models and the application of more advanced hyperparameter tuning methods along with the integration of other additional anomaly detection algorithms. Our study outlines the potential benefit of diversified approaches for overcoming issues arising in imbalanced datasets for fraud detection and can serve as an anchor for further research in this critical field.

REFERENCES

[1] Wongvorachan, S. He, and O. Bulut, ”A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining,” *Information*, vol. 14, no.1, p. 54, 2023

[2] S. Misra, S. Thakur, M. Ghosh, and S. K. Saha, ”An Autoencoder Based Model for Detecting Fraudulent Credit Card Transactions,” *Procedia Computer Science*, vol. 167, pp. 254-262, 2020.

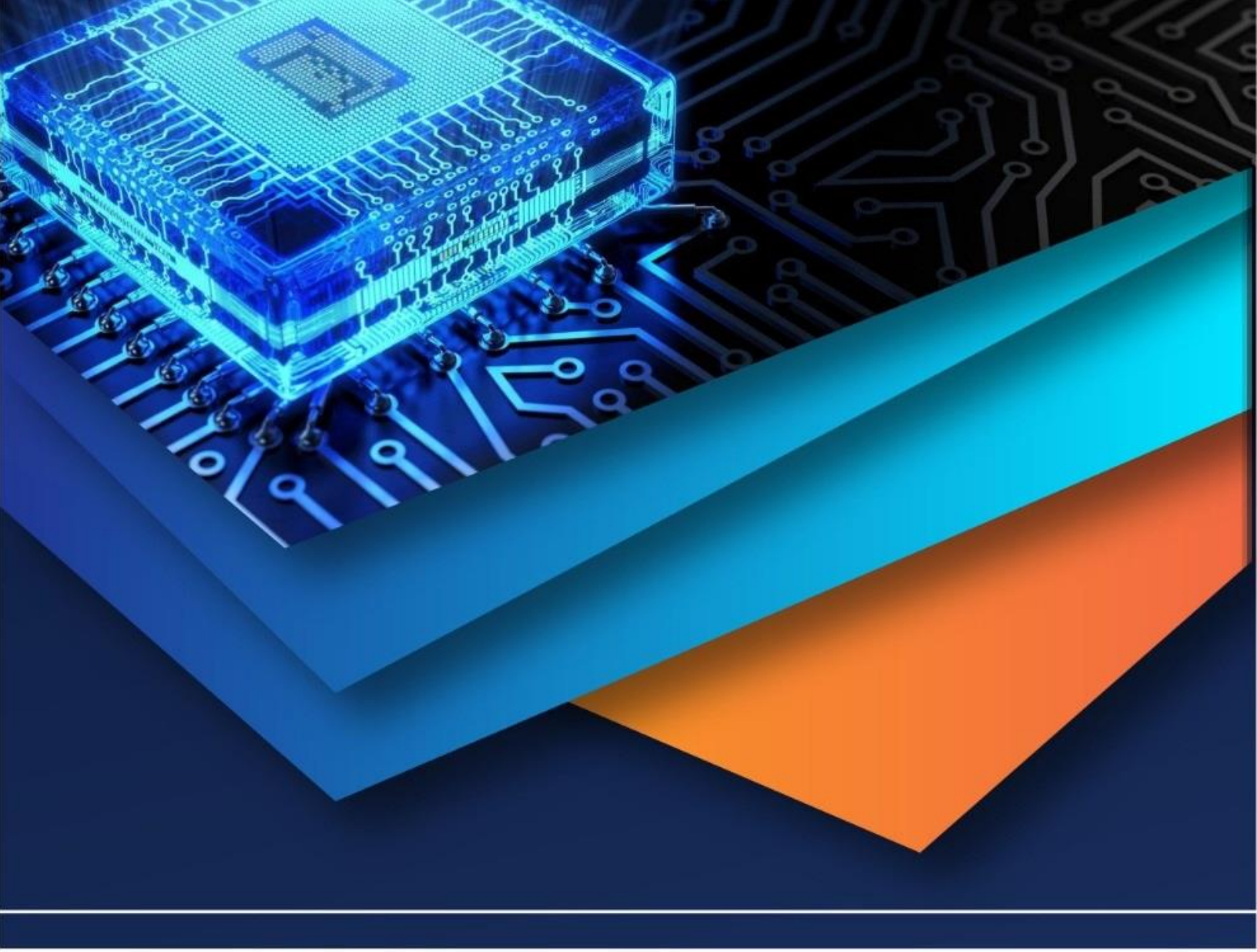
[3] C. Shen, Z. Cai, X. Guan, and R. Maxion, ”Performance Evaluation of Anomaly Detection Algorithms for Mouse Dynamics,” *Computer Security*, vol. 45, pp. 156-171, 2021.

[4] J. Shen, ”Credit Card Fraud Detection Using Autoencoder-Based Deep Neural Networks,” in *2021 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET)*, 2021, pp.263-270.

[5] R. Bauder and T. M. Khoshgoftaar, ”Medicare fraud detection using machine learning methods,” *Journal of Big Data*, vol. 5, no. 1, 2018.

[6] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, ”Credit Card Fraud Detection Using AdaBoost and Majority Voting,” *IEEE Access*, vol. 6, pp. 14277-14284, 2018.

[7] Kaggle. "Credit Card Fraud Detection." Available at: https://www.kaggle.com/datasets/dalpozz/creditcard-fraud



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY