iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.77483

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 14 Issue III Mar 2026- Available at www.ijraset.com

Credit Card Fraud Detection

Ms. Aqsa Fatima[1], Zoya Farooqi[2], Mohd Adnan Sheikh[3], Haroon Umar[4]
[1]*Assistant Professor, Dept. of Computer Science & Engineering, Integral University, Lucknow*
[2, 3, 4]*Dept. of Computer Science & Engineering, Integral University, Lucknow*

Abstract: ***Credit card fraud detection by using machine learning techniques in imbalanced datasets has emerged as an important area of study in recent years, considering the increased usage of online transactions. Here, we discuss in detail the detection of credit card transactions by using machine learning techniques and avoiding imbalanced datasets in credit card transactions. Here, we used the Random Forest algorithm to detect credit card transactions and imbalanced datasets were handled by using the Logistic Regression algorithm. Synthesized Minority Over-sampling Technique was used to train the models. Here, we used the SMOTE method, and finally, evaluation of models was carried out by using accuracy, precision, recall, F1-score, and confusion matrix. Our result demonstrated that Synthesized Minority Over-sampling Technique was important in detecting credit card transactions. Also, it was evident that our Random Forest classifier was more accurate compared to our Logistic Regression classifier in detecting credit card transactions.***

Keyword: Credit Card Fraud Detection, Machine Learning, Random Forest, Logistic Regression, Class Imbalance, SMOTE, Performance Evaluation.

I. INTRODUCTION

The increased popularity of digital payment systems, online banking, and e-commerce sites has increased the usage of credit cards across the globe. Though this increased popularity has provided more comfort to users, it has also increased the instances of fraudulent financial activities to a great extent. The consequences of credit card fraud include serious financial, reputational, and operational risks for financial organizations. Thus, it becomes a critical need to develop a reliable and efficient system for detecting such frauds in modern financial systems.

Though the use of machine learning algorithms has several advantages, the detection of credit card fraud is a major problem due to the extreme class imbalance in the data set. In most of the data sets used to train the model, the fraction of the data set representing the fraud transactions is very low compared to the total number of transactions. This leads to a biased prediction model towards the majority class, i.e., the legitimate transactions, and thus the accuracy of the model is very high, though the fraud detection is poor.

To overcome the imbalance problem, resampling methods such as the Synthetic Minority Over-sampling Technique (SMOTE) are commonly employed. The SMOTE method creates artificial data samples for the minority class, thereby balancing the classes and helping the machine learning algorithms effectively learn the patterns of the fraudulent transactions. The SMOTE method helps improve the recall and F1-score values, which are critical in the fraud detection system.

In this research, the implementation of the Random Forest classifier as the primary classifier is done in consideration of its potential to deal with nonlinear relationships in data. The Logistic Regression classifier is also employed as a baseline classifier to compare its performance with that of the Random Forest classifier. The performance of these classifiers is tested under two different conditions: with the application of SMOTE and without the application of SMOTE. The performance of these classifiers is evaluated using multiple evaluation metrics.

From the experimental results, it is clear that handling the class imbalance problem helps to improve the fraud detection ability. In particular, the proposed approach using the Random Forest algorithm and SMOTE has shown improved recall and predictive accuracy compared to the Logistic Regression algorithm. This again proves the significance of handling the class imbalance problem and the use of ensemble methods in the detection of credit card frauds.

II. LITERATURE REVIEW

The growing use of electronic and digital payment systems has resulted in an increased number of financial transactions, which in turn has caused an increase in the number of credit card fraud cases. One of the biggest issues in credit card fraud detection is the highly imbalanced nature of the transaction data, where the number of fraudulent transactions is a very small fraction compared to the number of genuine transactions.

Machine learning has been shown to be effective in overcoming the challenges of credit card fraud detection. In this context, Alrasheedi [1] performed a comparative study of various machine learning algorithms for fraud detection and found that ensemble-based methods performed better than traditional linear models like Logistic Regression.

There have been a few studies that have focused on the effect of class imbalance on the performance of machine learning models. Sinap [2] compared different models on imbalanced credit card transaction data and concluded that while Logistic Regression is interpretable, it lacks the ability to capture intricate patterns of fraudulent activities. However, Random Forest performed better in recognizing these patterns because of its capacity to handle nonlinearities.

To counter the problem of imbalance in the data, resampling methods have been used extensively. Ahmed [3] presented an ensemble learning approach along with hybrid sampling methods, including the Synthetic Minority Over-sampling Technique (SMOTE). The findings showed that balancing the data improves the performance of models substantially, especially Random Forest, by decreasing false negatives and improving the F1-score.

A detailed insight into the latest trends in credit card fraud detection using credit card fraud detection systems was provided in a systematic review published in the Journal of Big Data [4]. The review emphasized the success of machine learning and deep learning algorithms, concluding that traditional ensemble learning algorithms perform outstandingly well if aided by the right imbalance handling strategies. The review also emphasized the need for the use of precision, recall, and F1-score in addition to accuracy for the evaluation of fraud detection systems.

Further validation was obtained from Albalawi and Dardouri [5], who compared the performance of traditional machine learning algorithms with deep learning algorithms for imbalanced datasets. The results confirmed that Random Forest performs better than Logistic Regression if data balancing strategies are used. While deep learning algorithms did produce encouraging outcomes, their intensive computation makes machine learning algorithms based on ensemble learning more appropriate for real-time fraud detection systems.

A broader insight into the issues associated with class imbalance problems was provided in a systematic review published in Computers journal [6]. The authors analyzed the implementation of various machine learning algorithms for detecting credit card fraud with original class imbalance problems and stated that most research uses a large number of resampling methods to solve class imbalance problems. The results also emphasized the importance of balancing evaluation and practical implementation.

In a similar study published in the Journal of Big Data, Breskuvienė and Dzemyda [7] analyzed the results of implementing various machine learning algorithms for detecting fraud with highly imbalanced classes and found that feature selection, imbalance optimization, and ensemble methods play a significant role in improving the results of detecting fraud, especially in cases where the number of minority class instances is extremely low.

The main aim of this research work is to comparatively evaluate the performance of Logistic Regression as well as Random Forest classifiers under different experimental conditions with and without the application of SMOTE. In addition to this, the feasibility of the proposed technique is demonstrated using an interactive environment using the Streamlit library.

III. METHODOLOGY

The research problem has been tackled following a systematized approach based on Machine Learning techniques to detect credit card fraud. The particular dataset we use is the Kaggle fraud train dataset, containing many data records that are imbalanced. During data preprocessing, redundant data is removed, categorical attributes are encoded, and all the data attributes are normalized. After that, the data is split into training and test data sets. The problem of class imbalance in the dataset is tackled through the Synthetic Minority Over-Sampling Technique (SMOTE), followed by predictions made through Logistic Regression (LR) and Random Forest (RF), along with assessment of their performances through accuracy, precision, recall, and F1 metrics. The best model is deployed in Streamlit to perform prediction on credit card fraud in an interactive manner.

A. Data Collection & Preprocessing

The system used the fraudTrain.csv dataset, which contains 1,296,675 records. Preprocessing involved removing personal identifiers, encoding categorical variables, like merchant and category, using Label Encoding, and normalizing transaction amounts with StandardScaler.

B. Handling Class Imbalance

SMOTE was applied to the training set to artificially create fraudulent transaction samples in order to prevent model bias because of the extreme imbalance (less than 1% fraud).

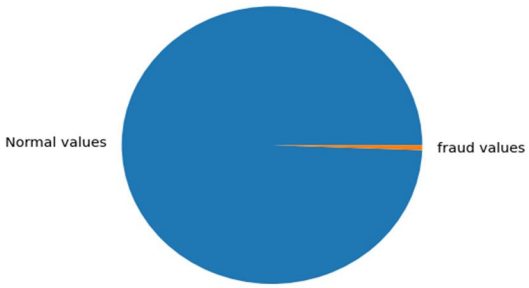


Figure 1. Showing Imbalance Data

C. Modeling Strategy

While comparing the models, the baseline model chosen for this comparative analysis was LR because it is simple, easy to interpret, and widely used for binary classification problems. Random Forest was chosen as the main model due to its ability to handle non-linear relationships in transaction-based data.

That included cleaning the data, encoding categorical features, and scaling numerical features. The cleaned dataset was then divided into a training set and a test set, which composed of 70% for training and 30% for testing, respectively. Both models were fitted on the training data and evaluated on the test data, considering the challenge of high class imbalance. To address this, SMOTE was applied only to the training data to prevent data leakage during evaluation.

Both models were in the same experimental environment and were evaluated using consistent metrics. Therefore, the proposed approach compares well with the ability of both Logistic Regression and Random Forest to detect fraud concerning credit card transactions.

IV. RESULTS & DISCUSSION

This section provides the experimental outcomes based on the application of Logistic Regression and Random Forest classification models on the credit card fraud detection dataset. Two different scenarios have been considered to test the performance of Logistic Regression and Random Forest classification models on the given dataset first without any balancing on the original imbalanced dataset and then on the dataset based on which the Synthetic Minority Over-Sampling Technique (SMOTE) algorithm was applied to check the performance of Logistic Regression and Random Forest classification models on imbalanced and class-balanced datasets.

A. Results without SMOTE

Firstly, a Random Forest classifier was trained on the original dataset, which was imbalanced in that fraudulent transactions comprised less than 1 percent of total transactions in the dataset.

Accuracy: ~99%

Precision: ~89%

Recall: ~61%

F1-score: ~72%

In the case of the baseline Random Forest model, even though the overall accuracy was very high, the results have to be viewed with caution due to the extreme imbalance of the dataset. The model seems to be overly accurate for the majority or genuine class, even though the recall value was low. This shows the number of false transactions being categorized as true, and such cases have to be avoided when it comes to real-world applications. This was again confirmed with the confusion matrix, which showed that while most correct transactions were correctly classified, a sizeable portion of fraud cases were still not being detected, showing the limitation or inherent deficiency in solely relying on the accuracy parameter with highly skewed data sets.

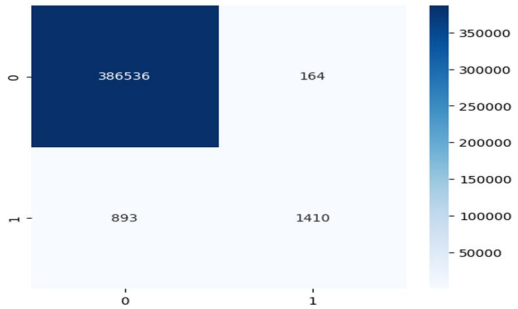


Figure 2. Confusion matrix without using smote

B. Results with SMOTE

To address the problem of class imbalance, SMOTE was used to synthetically perform oversampling for the fraud transaction class in the training dataset. The balanced dataset is then used to retrain the Random Forest model and measure its performance on the original test set.

Accuracy: ~99.4%

Precision: ~75%

Recall: ~75%

F1-score : ~75%

After SMOTE, the recall improved substantially from ~61% to ~74%. These improvements point out that this model became way better in picking up fraudulent transactions. The slight decrease in precision was due to the higher number of false positives; otherwise, its overall balance between precision and recall improved as reflected by the higher F1-score. Therefore, the SMOTE confusion matrix illustrates more fraud cases that were correctly detected and fewer missed frauds compared to the baseline model. These results confirm that SMOTE effectively mitigates class imbalance and enhances the minority-class learning capability of the Random Forest classifier without sacrificing overall accuracy.

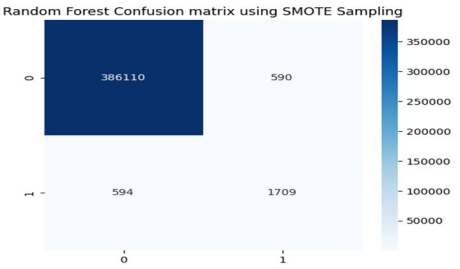


Figure 3. Confusion Matrix with smote technique

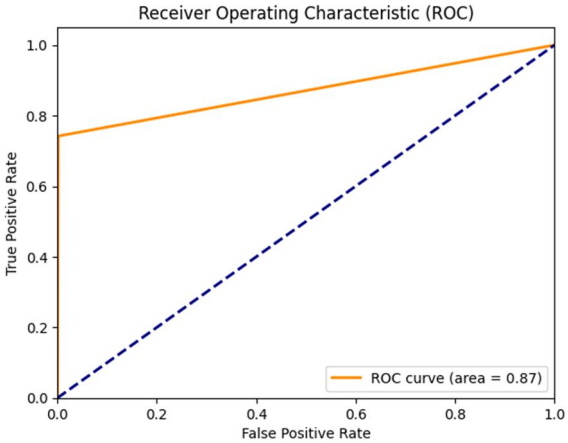


Figure 4. ROC Curve

The experimental evaluation confirms that class imbalance handling is critical for fraud detection systems. While both models performed reasonably well, we can observe that the **Random Forest model trained with SMOTE** emerged as the most reliable and accurate classifier for identifying fraudulent transactions. Its superior accuracy, improved recall, and robustness make it the preferred choice for deployment in real-time credit card fraud detection applications.

```
print(classification_report(y_test,pred_RF))
print(f'Random Forest Classfication report without sampling')
```

```
[43]
...
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    386700
           1       0.90      0.61      0.73      2303

    accuracy                           1.00    389003
   macro avg       0.95      0.81      0.86    389003
weighted avg       1.00      1.00      1.00    389003
```

Figure 5. Classification of performance metrices on imbalance data

```
print(classification_report(y_test,pred_RF1))
print(f'Random Forest Classfication report with smote sampling')
```

```
[58]
...
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    386700
           1       0.74      0.74      0.74      2303

    accuracy                           1.00    389003
   macro avg       0.87      0.87      0.87    389003
weighted avg       1.00      1.00      1.00    389003
```

Figure 6. Classification of performance metrices on balanced data

C. Comparison of Logistic Regression & Random Forest using SMOTE Technique

The results suggest that the efficacy of the proposed solution is evident from the fact that the accuracy of the proposed solution using Logistic Regression (LR) collaborated remarkably well and managed to attain a fairly accurate accuracy of approximately 94%. Therefore, it becomes evidently evident that the proposed solution using Logistic Regression, though fairly reasonable in terms of accuracy, managed to attain a decent accuracy. At the same time, the proposed solution using Random Forest (RF) managed to attain a significantly high accuracy of approximately 99.4%, which is a remarkably high accuracy. Hence, it is safe to conclude that Random Forest is a significantly better solution than that of the proposed solution using logistic regression.

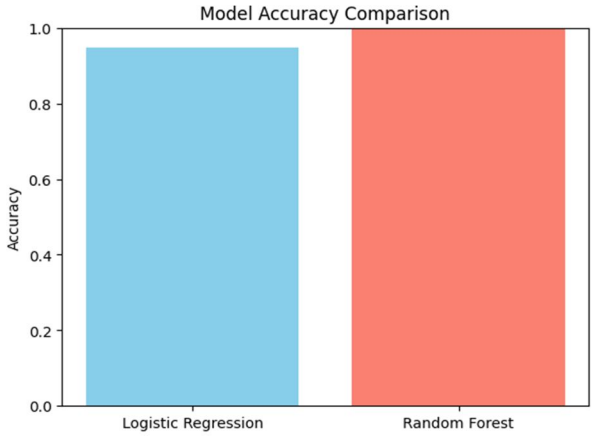


Figure 7. Comparison Plot Graph

V. FUTURE SCOPE

It has to be mentioned that there are very many possible ways in which the credit card fraud detection model developed in this study would improve in the future. For instance, training on a large real-world dataset from a variety of financial organizations would improve its robustness. A hybrid model including Random Forest and more advanced algorithms like XGBoost, LightGBM, and deep learning models would improve its ability to detect credit card fraud. Deep learning models like LSTM would help derive deeper insights into credit card user behavior. Advanced algorithms in managing classification imbalance would improve its ability to distinguish.

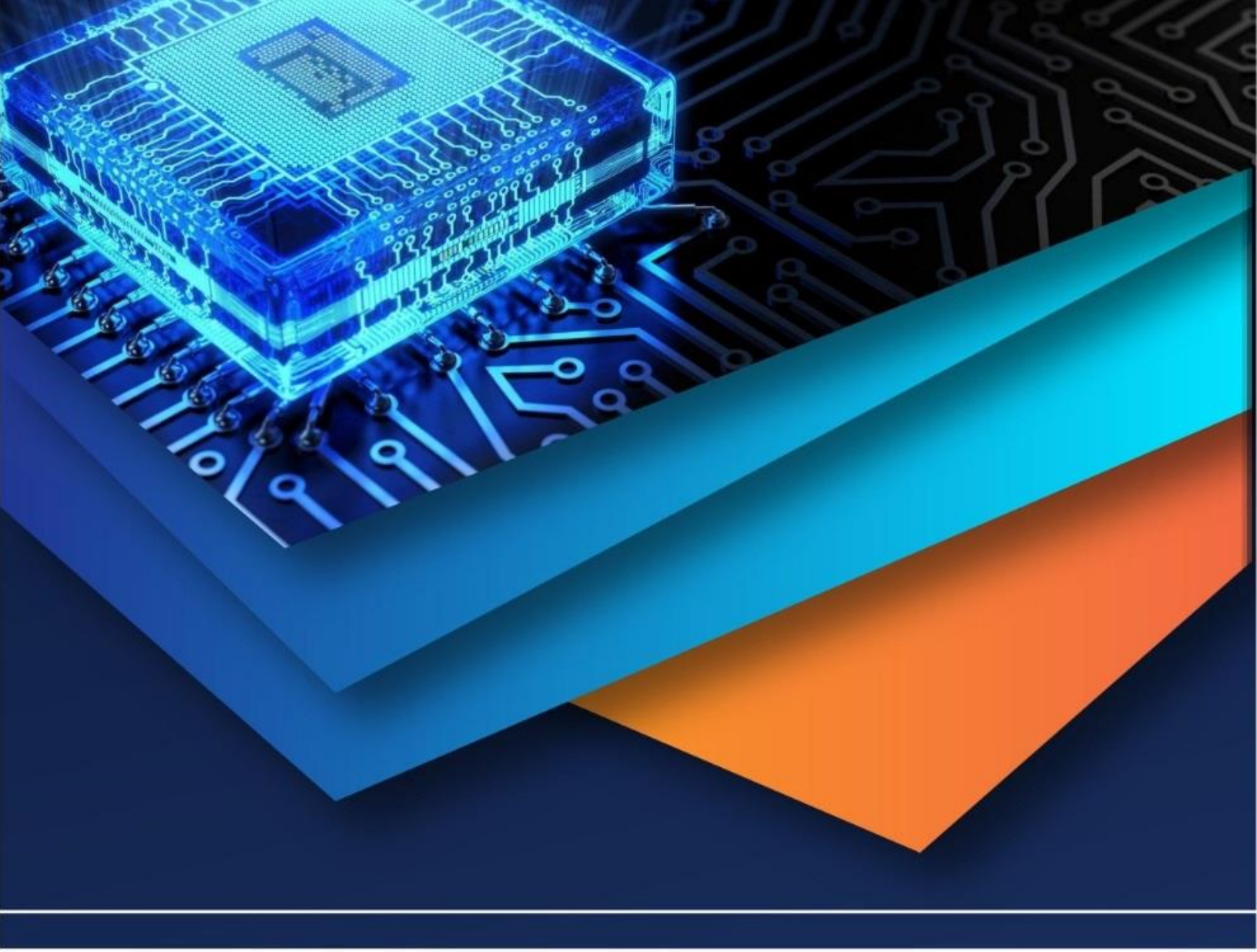
Furthermore, the platform may be evolved to facilitate a live streaming deployment environment that caters to an online and adaptive learning approach, which may result in immediate fraud detection. It may include an integration of explainable AI to enhance transparency and usability. Additionally, it may draw information from more than one source to enhance the accuracy of the predictions. It may provide several promising directions to enhance credit card fraud detection in realistic and evolving financial domains.

VI. CONCLUSION

In this paper, an approach for detecting Credit Card Frauds using Machine Learning techniques such as Logistic Regression and RandomForest has been explained. Most importantly, the approach has addressed the problem of dealing with an imbalanced data problem using SMOTE techniques. From the experimental outcome of the paper, the prediction bias for Legitimate Transactions can clearly be identified when the model is trained using an Imbalanced dataset. On the other hand, the prediction has been highly balanced for the minority class by using SMOTE techniques. Apart from that, the RandomForest approach has high efficacy over Logistic Regression by detecting the Credit Card Frauds more accurately by using the simplest form of decision-making mechanisms. The proposed system has room for exploration but is yet to be applicable for real-time scenarios. Nevertheless, the paper has provided an approach for contributing to the development of an efficient framework for dealing with Credit Card Fraud detection techniques.

REFERENCES

[1] Alrasheedi, M.A. Enhancing Fraud Detection in Credit Card Transactions: A Comparative Study of Machine Learning Models. *Comput Econ* (2025). https://doi.org/10.1007/s10614-025-11071-3
[2] V. Sinap, “Comparative analysis of machine learning techniques for credit card fraud detection: Dealing with imbalanced datasets”, TUJE, vol. 8, no. 2, pp. 196–208, 2024, doi: 10.31127/tuje.1386127.
[3] Khanda Hassan Ahmed, Stefan Axelsson, Yuhong Li, Ali Makki Sagheer, A credit card fraud detection approach based on ensemble machine learning classifier with hybrid data sampling, Machine Learning with Applications, Volume 20, 2025, 100675, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2025.100675.
[4] Hafez, I.Y., Hafez, A.Y., Saleh, A. *et al.* A systematic review of AI-enhanced techniques in credit card fraud detection. *J Big Data* **12**, 6 (2025). https://doi.org/10.1186/s40537-024-01048-8
[5] Albalawi Tahani , Dardouri Samia ,Enhancing credit card fraud detection using traditional and deep learning models with class imbalance mitigation,Frontiers in Artificial Intelligence,Volume 8 – 2025,2025,DOI=10.3389/frai.2025.1643292,ISSN=2624-8212
[6] Baisholan, N., Dietz, J. E., Gnatyuk, S., Turdalyuly, M., Matson, E. T., & Baisholanova, K. (2025). A Systematic Review of Machine Learning in Credit Card Fraud Detection Under Original Class Imbalance. Computers, 14(10), 437. https://doi.org/10.3390/computers14100437
[7] Breskuvienė, D., Dzemyda, G. Enhancing credit card fraud detection: highly imbalanced data case. *J Big Data* **11**, 182 (2024). https://doi.org/10.1186/s40537-024-01059-5



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY