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.2024.61169

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 IV Apr 2024- Available at www.ijraset.com

Machine Learning-based Anomaly Detection for Fraudulent Banking Activities

Mrs. T.Ganga Bhavani[1] , Marisetti Yamini Sai[2], B Geethika Poornima[3], Mohammed Jafar Sadiqe[4], Ramanapudi Kishan Sai[5]

[1]*Assistant Professor* , [2, 3, 4, 5]*B.tech Students Department of Information Technology, Pragati Engineering College, Surampalem , Andhra Pradesh, India*

Abstract: ***In this research, we look at how class weight-tuning hyperparameters can be used to balance the relative weights of fraudulent and authentic transactions. We utilize Bayesian optimization to optimize hyperparameters while taking into account real-world issues like imbalanced data. We suggest weight-tuning as a pre-process for unbalanced data, as well as CatBoost and XGBoost, to improve the performance of the LightGBM method by addressing these real-world difficulties. Finally, we employ deep learning to adjust the hyperparameters (particularly our suggested weight-tuning technique) in order to increase overall performance. We do experiments on real-world data to validate the recommended approaches. In addition to the usual ROC-AUC, we employ recall-precision measures to better address unbalanced datasets. CatBoost, LightGBM, and XGBoost are assessed independently using a 5-fold cross-validation methodology. Furthermore, the integrated algorithms' performance is evaluated using a majority voting ensemble learning approach. The findings indicate that LightGBM and XGBoost meet the optimum level condition of ROC-AUC = 0.95, precision 0.79, recall 0.80, F1 score 0.79, and MCC 0.79. Deep learning and Bayesian optimization are used to modify the hyperparameters, yielding ROC-AUC = 0.94, precision = 0.80, recall = 0.82, F1 score = 0.81, and MCC = 0.81. This represents a huge improvement over the cutting-edge techniques that we utilized for comparison.***

Keywords: Bayesian optimization, data Mining, deep learning, ensemble learning, hyper parameter, unbalanced data, machine learning.

I. INTRODUCTION

Financial transactions have expanded dramatically in recent years due to the development of financial institutions and the acceptance of web-based e-commerce. Online banking fraud has always existed, and detecting fraudulent transactions can be difficult [1], [2]. Credit card fraud patterns have evolved alongside credit card technology. Fraudsters go to great lengths to appear authentic, and credit card fraud is always evolving. Fraudsters try to give the appearance that Zhan Bu, the assistant editor who handled the manuscript's review and gave it approval for publishing, is a trustworthy source. They continuously stimulating these systems in an attempt to understand how fraud detection systems function, making fraud detection increasingly difficult. But technology can be utilized to combat fraud [4]. Fraud must be recognized as soon as possible in order to be avoided in the future [5]. Fraud is defined as dishonest or illegal misrepresentation with the purpose to get money or personal profit. Credit card fraud is the unlawful use of credit card information to make in-person or online purchases. Cardholders typically supply the card number, expiration date, and card verification number over the phone or online, making fraud in digital transactions possible [6]. To reduce losses caused by fraud, two tactics can be used: fraud detection and prevention. Fraud prevention is a proactive method that prevents fraudulent activity from occurring. Nonetheless, fraud detection is required whenever a fraudster attempts to complete a fraudulent transaction [7]. In the banking industry, categorizing data as valid or fraudulent is considered a binary classification task [8]. The volume of financial data and the size of the datasets that compose the transaction data make it either impossible or extremely time-consuming to manually evaluate and detect trends for fraudulent activities. Machine learning technologies are thus critical for predicting and detecting fraudulent behavior [9]. Dealing with large datasets and detecting fraud are made easier by machine learning techniques and processing capacity. Machine learning and deep learning algorithms can also be used to solve real-time problems [10]. In this paper, we offer a successful credit card fraud detection system that has been validated using publicly available datasets. It employs optimization algorithms such as logistic regression, XGBoost, LightGBM, and CatBoost, as well as combined majority voting techniques, deep learning, and hyperparameter settings.

©IJRASET: All Rights are Reserved

A perfect fraud detection system would identify more fraudulent situations with high precision—that is, all results should be correctly recognized, promoting client faith in the bank while safeguarding it from losses caused by incorrect detection. The key contributions of this study can be summarized as follows: • We use a weight-tuning hyperparameter to solve the unbalanced data problem as a pre-processing step, and we use Bayesian optimization to detect fraud. For even better results, we recommend combining LightGBM with CatBoost and XGBoost. We pick the XGBoost approach because it trains quickly on big datasets, provides a regularization term that prevents overfitting by evaluating the tree's complexity, and requires little work to tune the hyperparameters. We also employ the Catboost algorithm because, when compared to other machine learning algorithms, it produces good results and does not require hyperparameter tinkering to manage overfitting. • We describe an ensemble learning strategy based on majority voting that combines LightGBM, XGBoost, and CatBoost. We use real, unbalanced datasets to evaluate the efficacy of the integrated techniques in detecting fraud. In addition, we recommend employing deep learning to tweak and perfect the hyperparameters.We validate the effectiveness of the proposed methodologies through extensive tests on real-world data. To accommodate for unbalanced datasets, we apply recallprecision in addition to the widely used ROC-AUC. In addition, we employ the F1_score and MCC metrics to evaluate performance. The results reveal that the recommended strategies outperform the previously and currently employed approaches. We make the source codes 1 widely available for usage by other researchers and undertake evaluations using publicly available datasets. The reminder in this paper is organized as follows: Section II discusses the current state of the art. The suggested strategy for identifying credit card fraud is discussed in Section III. It comprises the dataset, pre-processing, feature extraction and selection, algorithms, framework, and evaluation metrics. Section IV covers the test evaluation results, while Section V presents the paper's conclusion.

II. LITERATURE SURVEY

Researchers have proposed numerous ways for preventing fraudulent transactions and detecting credit card fraud. This is an overview of current, relevant, and innovative works. Halvaiee and Akbari develop a novel model known as the AIS-based fraud detection model (AFDM). The Immune System Inspired Algorithm (AIRS) is used to improve the accuracy of fraud detection. Their research suggests that their proposed AFDM can reduce system reaction times by up to 40%, reduce costs by up to 85%, and improve accuracy by up to 25% when compared to simple algorithms [11]. By analyzing the periodic pattern of transaction time using the von Mises distribution, Bahnsen et al. developed a novel set of features and a transaction aggregation method. They also examine the impact of alternative feature sets on outcomes using a genuine credit card dataset, as well as a novel cost-based criterion for evaluating credit card fraud detection systems. More specifically, they build on the transaction aggregation method by creating extra offers based on recurring transaction patterns [12]. Randhawa et al. study the use of machine learning approaches to detect credit card fraud. To analyze the presented datasets, they first use common models such as Naive Bayes, stochastic forest and decision trees, neural networks, logistic regression, linear regression (LR), and support vector machines.

They also provide a hybrid technique that combines majority voting with AdaBoost. They also introduce noise into the data samples in order to determine resilience. Using publicly available statistics, they run tests to illustrate the effectiveness of majority voting in detecting credit card fraud [6]. Porwal and Mukund suggest a clustering-based method for identifying outliers in large datasets that are robust to shifting trends [13]. Their technique is founded on the assumptions that good user behavior persists over time and that good behavior data points have a consistent geographical signature across various groups. They provide an example of how changes in this data can be utilized to detect fraudulent activity. It has been established that as an evaluation criterion, the area under the precision-recall curve outperforms ROC [13]. The authors of [14] propose a group learning technique based on training set segmentation and grouping. Their proposed approach tries to address the significant imbalance in the dataset while also ensuring the integrity of the sample characteristics.

The key characteristic of their proposed framework is its capacity to train all base estimators at the same time, which improves its efficacy. To solve the issue of data imbalance, Itoo et al. use an oversampling technique that includes three distinct dataset ratios. The authors applied three machine learning methods: Naive Bayes, logistic regression, and K-nearest neighbor. The following metrics are used to evaluate algorithm performance: accuracy, sensitivity, precision, area under the curve, F1-score, and accuracy. The results show that the logistic regression-based model outperforms the other frequently used fraud detection strategies discussed in the article [15]. The authors of [16] describe a fraud detection system that blends the cost-sensitive learning paradigm with meta-learning ensemble techniques.

They undertake a variety of tests, and the cost-sensitive ensemble classifier outperforms ordinary ensemble classifiers in identifying unknown data while maintaining an acceptable AUC value. Altyeb et al. [17] offer an intelligent technique for detecting fraudulent

credit card transactions. The parameters of a LightGBM are optimized using the recommended Bayesian-based hyperparameter optimization technique.

They do research utilizing credit card transaction data made available to the public. These datasets include both real and fraudulent transactions. The evaluation results are reported as the area under the receiver operating characteristic curve (ROC-AUC), accuracy, precision, and F1-score. Xiong et al. present a learning-based approach to the fraud detection problem. They increase the performance of the suggested model by applying feature engineering approaches. The model is trained and verified on the IEEE-CIS fraud dataset. According to their findings, the model outperforms more traditional machine-learning approaches such as Bayes and SVM on the dataset under consideration [18].

III. SYSTEM ANALYSIS

A. *Existing System*

Machine learning-based credit card fraud detection. First, they assess the datasets using conventional models such as Naive Bayes, stochastic forest and decision trees, neural networks, logistic regression, linear regression (LR), and support vector machines. Additionally, they propose merging AdaBoost with majority voting to form a hybrid system. They also add noise to the data samples to test their robustness. They run studies on publicly available datasets to show how successfully majority voting detects credit card theft.

DISADVANTAGES OF THE EXISTING SYSTEM

1) Interpretability: Complex machine learning models, such as deep neural networks, can be difficult to understand. The ability to evaluate and articulate model conclusions is crucial in the banking business for building trust and maintaining regulatory compliance.
2) Overfitting and Underfitting: Models might underfit and fail to understand the complexity of the data, or they can overfit the training set, collecting noise instead of underlying patterns. To solve these challenges, thorough model validation and change are necessary.
3) Interpretability: Deep neural networks and other complicated machine learning models can be challenging to understand. To create confidence and ensure regulatory compliance, the banking sector must be able to understand and communicate model decisions.
4) Overfitting and Underfitting: Models may overfit the training set, amassing noise rather than underlying patterns, or they may underfit, failing to recognize the data's complexity. These issues require appropriate model changes and validation.
5) computing Resources: In cases where resources are restricted, models that require a lot of computing power may be unable to work.
6) Regulatory Compliance: The banking industry must strictly adhere to legal rules and compliance standards. When evaluating the model, it is critical to examine ethical and legal implications.
7) Adversarial Attacks: Scammers may attempt to influence the system by exploiting its weaknesses. The quest to create models that can withstand hostile attacks is currently ongoing.
8) Scalability: The fraud detection system must be able to scale as transaction volumes increase. One of the most critical tasks is to ensure that the system can handle a large number of transactions efficiently.

B. *Proposed System*

The data is adequately preprocessed before being divided into training and testing parts in the proposed fraud detection architecture. We next use Bayesian optimization on the training set to select the appropriate hyperparameters for performance enhancement. Following the cross-validation approach to analyzing the algorithms' performance in an imbalanced set, we use a range of evaluation metrics, including accuracy, precision, recall, the F1-score, the Matthews correlation coefficient (MCC), and AUC diagrams.

IV. SYSTEM DESIGN

A. *System Architecture*

Below diagram depicts the whole system architecture.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



Fig 1. Methodology followed for proposed model

V. SYSTEM IMPLEMENTATION

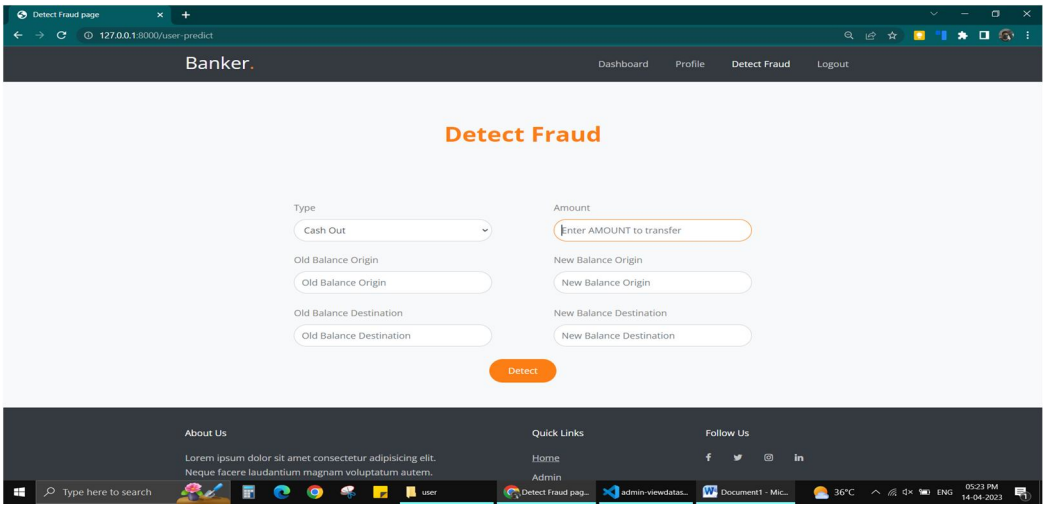
MODULES

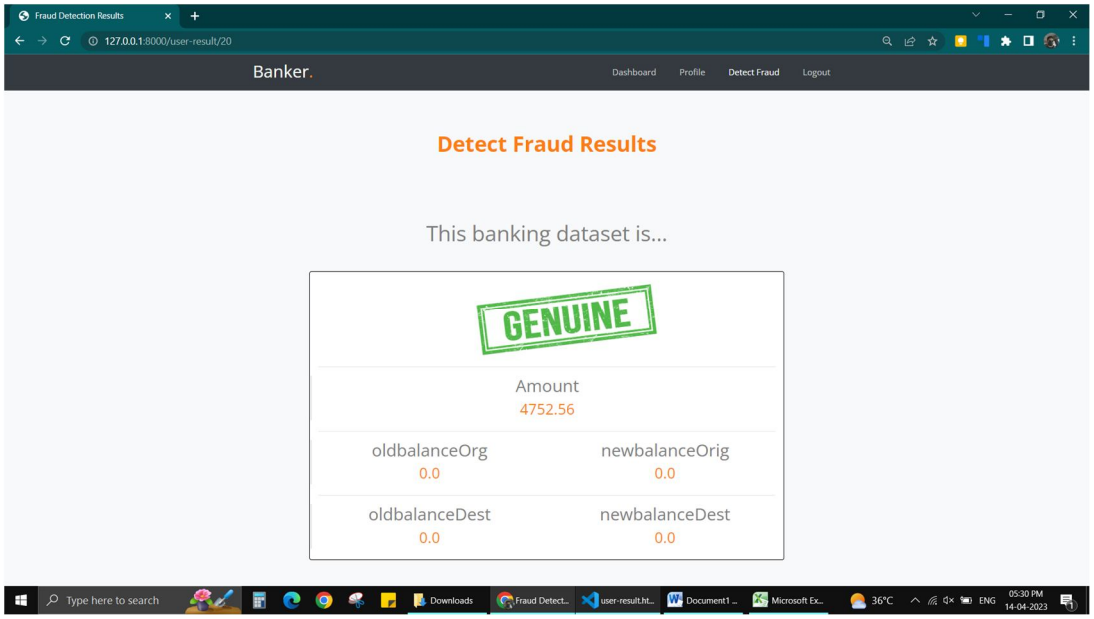
1) Gathering information and preprocessing: Assemble relevant datasets containing financial transactions, including both legitimate and fraudulent operations. Preprocessing includes resolving missing results, addressing outliers, and resampling to reduce the impact of class imbalances.
2) Selection and Feature Engineering: Determine which criteria are most significant for identifying fraud. This module evaluates the dataset and adds or alters existing features to improve the machine learning model's ability to detect patterns associated with fraudulent transactions.
3) Training Machine Learning Models: Use a variety of machine learning techniques, including support vector machines, decision trees, random forests, logistic regression, and gradient boosting. Train these algorithms to detect and identify using preprocessed data.
4) Real-time transaction verification: Using the previously learned machine learning model, develop a real-time transaction verification module. With the help of this module, transactions should be validated more quickly and efficiently, allowing for early fraud detection and prevention.
5) Model Assessment and Continuous Observation: The trained machine learning model's performance will be evaluated using metrics such as recall, accuracy, precision, and F1 score. Set up ongoing observation procedures to assess the model's performance over time. This allows for quick updates and adjustments in response to new fraud patterns.

VI. RESULTS AND DISCUSSION

To evaluate the efficacy of the proposed framework, we use boosting algorithms combined with a stratified 5-fold cross-validation method and a Bayesian optimization strategy. Following the extraction of the hyperparameters, we analyze each algorithm individually using the majority voting technique. We investigate algorithms with triple and double precision. Performance is evaluated based on the comparison results presented.

©IJRASET: All Rights are Reserved





VII. CONCLUSION AD FUTURE WORK

This paper recommends using machine learning approaches to detect fraud in financial applications. The publicly available UCI dataset is analyzed. The supplied dataset has a noticeable imbalance and a significant bias toward the majority of samples. This problem is solved using the synthetic minority over-sampling technique (SMOTE). XGBoost is the boosting strategy used to implement the KNN and Random Forest algorithms. The model performed well, with a 97.74 percent success rate. After reviewing the data, we discovered that clients aged 19 to 25 were more likely to be dishonest than other categories.

REFERENCES

[1] R. Rambola, P. Varshney and P. Vishwakarma, “Data Mining Techniques for Fraud Detection in Banking Sector,” 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-5, doi: 10.1109/CCAA.2018.8777535.

[2] N. Malini and M. Pushpa, “Analysis on credit card fraud identification techniques based on KNN and outlier detection,” 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2017, pp. 255-258, doi: 10.1109/AEEICB.2017.7972424.

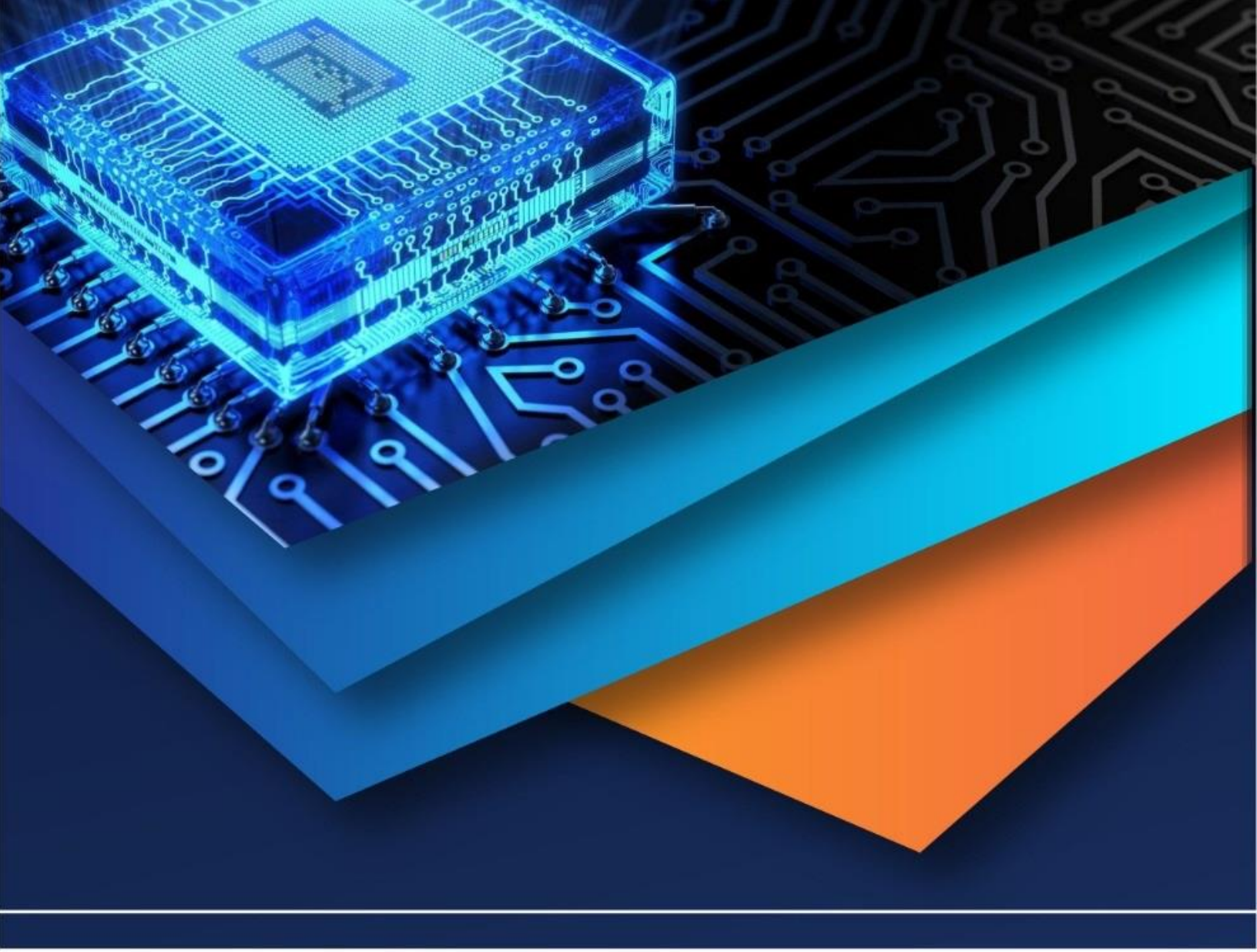
[3] Ishan Sohony, Rameshwar Pratap, and Ullas Nambiar. 2018. Ensemble learning for credit card fraud detection. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (CoDS-COMAD ’18). Association for Computing Machinery, New York, NY, USA, 289–294. DOI:https://doi.org/10.1145/3152494.3156815

[4] C. Wang, Y. Wang, Z. Ye, L. Yan, W. Cai, and S. Pan, “Credit Card Fraud Detection Based on Whale Algorithm Optimized BP Neural Network,” 2018 13th International Conference on Computer Science Education (ICCSE), Colombo, 2018, pp. 1-4, doi: 10.1109/ICCSE.2018.8468855

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[5] I. Benchaji, S. Douzi and B. ElOuahidi, "Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection," 2018 2nd Cyber Security in Networking Conference (CSNet), Paris, 2018, pp. 1-5, doi: 10.1109/CSNET.2018.8602972.
[6] John O. Awoyemi, Adebayo Olusola Adetunmbi, and Samuel Adebayo Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI), pages 1–9, 2017.
[7] Fabrizio Carcillo, Andrea Dal Pozzolo, Yann-A¨el Le Borgne, Olivier Caelen, Yannis Mazzer, and Gianluca Bontempi. Scarff: a scalable framework for streaming credit card fraud detection with spark. Information Fusion, 41:182–194, 2018.
[8] Galina Baader and Helmut Krcmar. Reducing false positives in fraud detection: Combining the red flag approach with process mining. International Journal of Accounting Information Systems, 2018.
[9] Ravisankar P, Ravi V, Raghava Rao G, and Bose, Detection of financial statement fraud and feature selection using data mining techniques, Elsevier, Decision Support Systems Volume 50, Issue 2, p491-500 (2011) SVM
[10] K. Seeja, and M. Zareapoor, "FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining," The Scientific World Journal, 2014, pp. 1-10. KNN, SVM [11] C. Tyagi, P. Parwekar, P. Singh, and K. Natla, "Analysis of Credit Card Fraud Detection Techniques," Solid State Technology, vol. 63, no. 6, 2020, pp. 18057-18069. Credit card faud
[11] C. Chee, J. Jaafar, I. Aziz, M. Hassan, and W. Yeoh, "Algorithms for frequent itemset mining: a literature review," Artificial Intelligence Review, vol. 52, 2019, pp. 2603–2621. Litrature review AI
[12] S. Kiran, J. Guru, R. Kumar, N. Kumar, D. Katariya, and M. Sharma, "Credit card fraud detection using Naïve Bayes model based and KNN classifier," International Jounral of Advance Research, Ideas and Innovations in Technology, vol. 4, 2018, pp. 44-47. KNN Naïve Byers
[13] Pumsirirat, A.; Yan, L. Credit Card Fraud Detection Using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. Available online: https://thesai.org/Downloads/Volume9No1/Paper_3- Credit_Card_Fraud_Detection_Using_Deep_ Learning.pdf (accessed on 23 February 2021). DL
[14] PwC's Global Economic Crime and Fraud Survey 2020. Available online: https://www.pwc.com/fraudsurvey (accessed on 30 November 2020). Fraud surver.
[15] Pourhabibi, T.; Ongb, K.L.; Kama, B.H.; Boo, Y.L. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decis. Support Syst. 2020, 133, 113303. Fraud detection.
[16] Lucas, Y.; Jurgovsky, J. Credit card fraud detection using machine learning: A survey. arXiv 2020, arXiv:2010.06479. Credit card fraud.
[17] Podgorelec, B.; Turkanovi´c, M.; Karakati˘c, S. A Machine Learning Based Method for Automated Blockchain Transaction Signing Including Personalized Anomaly Detection. Sensors 2020, 20, 147. Anomaly detection.
[18] Synthetic Financial Datasets for Fraud Detection. Available online: https://www.kaggle.com/ntnu-testimon/paysim1 (accessed on 30 November 2020). Fraud detection.
[19] Ma, T.; Qian, S.; Cao, J.; Xue, G.; Yu, J.; Zhu, Y.; Li, M. An Unsupervised Incremental Virtual Learning Method for Financial Fraud Detection. In Proceedings of the 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 3–7 November 2019; pp. 1–6. Financial fraud detection.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)