ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 12, Issue 2, March - April 2025

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.028

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 8.028 | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 12, Issue 2, March- April 2025 |

A Comparative Analysis on Credit Card Fraud Using Machine Learning Technique

Dr. S. Pavithra, S.Yogeshwari

Associate Professor, Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, Tamil Nadu, India

Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, Tamil Nadu, India

ABSTRACT: Credit card fraud is a pervasive issue in the financial industry, posing significant challenges to banks and consumers alike. This study presents a comparative analysis of various machine learning techniques to detect and mitigate credit card fraud effectively. We employed several models, including LightGBM, Bagging, RandomForest, SVM, DecisionTree, and AdaBoost, to classify fraudulent transactions. The dataset was preprocessed using standardization and balanced using the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalances inherent in fraud detection. The models were evaluated based on accuracy, F1 score, precision, and recall, with the goal of identifying the most robust algorithm for real- time fraud detection. The selected model was then deployed as a web service using Flask, enabling real-time fraud detection through a simple API interface.This comprehensive approach not only highlights the effectiveness of machine learning in fraud detection but also offers a practical solution for integrating such models into existing financial systems, thereby enhancing the security and reliability of credit card transactions.

I. INTRODUCTION

Credit card is the most popular mode of payment. As the number of credit card users is rising world-wide, the identity theft is increased, and frauds are also increasing. In the virtual card purchase, only the card information is required such as card number, expiration date, secure code, etc. Such purchases are normally done on the Internet or over telephone. To commit fraud in these types of purchases, a person simply needs to know the card details. The mode of payment for online purchase is mostly done by credit card. The details of credit card should be kept private. To secure credit card privacy, the details should not be leaked. Different ways to steal credit card details are phishing websites, steal/lost credit cards, counterfeit credit cards, theft of card details, intercepted cards etc. For security purpose, the above things should be avoided. In online fraud, the transaction is made remotely and only the card's details are needed. The simple way to detect this type of fraud is to analyze the spending patterns on every card and to figure out any variation to the "usual" spending patterns. Fraud detection by analyzing the existing data purchase of cardholder is the best way to reduce the rate of successful credit card frauds. As the data sets are not available and also the results are not disclosed to the public. The fraud cases should be detected from the available data sets known as the logged data and user behavior. At present, fraud detection has been implemented by a number of methods such as data mining, statistics, and artificial intelligence.

Credit cardfraud occurs when someone uses your creditcardoraccount information without your permission. As you know, the money you charge to your credit card is not actually yours; it's borrowed from the bank and must be paid back with interest. Fraudsters use stolen information to make unauthorised purchases, leaving you on the hook for those charges.

This inclusive term covers a wide range of illegal activities, which we'll cover in greater detail below. Credit card fraud is also part of the larger crime of identity theft, in which a person steals and uses another individual's personal information for their own gain..

II. LITERATURE SURVEY

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers.

2.1 Review on Existing System

Randhawa et al. (2018) explored the use of AdaBoost combined with majority voting for credit card fraud detection. Their study demonstrated the effectiveness of ensemble learning techniques in enhancing fraud detection accuracy. By

leveraging AdaBoost to improve the performance of individual classifiers and combining their results through majority voting, the authors achieved a more robust detection system capable of identifying fraudulent transactions with higher precision [1].

Alarfaj et al. (2022) reviewed state-of-the-art machine learning and deep learning algorithms for credit card fraud detection. Their comprehensive analysis underscored the advancements in algorithmic approaches, highlighting the importance of adopting cutting-edge techniques to address the complexities of fraud detection. The study emphasized that modern algorithms, including deep learning models, offer significant improvements over traditional methods in terms of accuracy and efficiency [2].

Ileberi, Sun, and Wang (2021) conducted a performance evaluation of machine learning methods for credit card fraud detection, specifically focusing on the integration of SMOTE (Synthetic Minority Over-sampling Technique) and AdaBoost. Their research showed that combining SMOTE with AdaBoost effectively addresses class imbalance issues and enhances detection performance. This approach allows for improved identification of fraudulent transactions by generating synthetic samples and boosting the performance of classifiers [3].

Ghaleb et al. (2023) proposed an ensemble approach that integrates Generative Adversarial Networks (GANs) with Random Forest algorithms for fraud detection. Their study highlighted the benefits of combining GANs, which generate synthetic data to address class imbalance, with Random Forests, known for their robustness in classification tasks. This hybrid method demonstrated improved detection capabilities and greater resilience against fraudulent activities [4].

Mienye and Sun (2023) introduced a deep learning ensemble model enhanced with data resampling techniques for fraud detection. Their work emphasized the effectiveness of ensemble methods in improving model performance, particularly when combined with data resampling strategies. The study illustrated how leveraging ensemble models with resampling can significantly enhance the accuracy and reliability of fraud detection systems [5].

Ding et al. (2023) presented an improved Variational Autoencoder Generative Adversarial Network for credit card fraud detection. Their research showcased advancements in leveraging generative models to enhance fraud detection accuracy. By refining the Variational Autoencoder and integrating it with Generative Adversarial Networks, the authors achieved notable improvements in detecting fraudulent transactions [6].

2.2 Literature Review Summary

AdaBoost:

- **Randhawa et al. (2018):** Demonstrated AdaBoost's effectiveness in enhancing fraud detection accuracy through boosting weaker classifiers [1].
- **Ileberi, Sun, & Wang (2021):** Found that combining AdaBoost with SMOTE improves performance by addressing class imbalance [3].
- **Deep Learning Models:**
  - **Alarfaj et al. (2022):** Highlighted deep learning's superior accuracy in complex fraud detection tasks [2].
  - **Mienye & Sun (2023):** Showed that deep learning ensembles with data resampling achieve higher accuracy [5].

Generative Models:

- **Ghaleb et al. (2023):** Improved fraud detection by combining GANs with Random Forests to handle class imbalance [4].
- **Ding et al. (2023):** Enhanced fraud detection using Variational Autoencoders and GANs [6].

Ensemble Methods with Data Resampling:

- **Ning et al. (2023):** Introduced AMWSPL-Adaboost for better fraud detection through increased classifier diversity [7].
- **Esenogho et al. (2022):** Demonstrated that neural network ensembles with advanced feature engineering improve detection accuracy [10].

Gradient Boosting Techniques:

- **Taha & Malebary (2020):** Optimized LightGBM for better fraud detection performance [8].

Imbalanced Classification:

- **Makki et al. (2019):** Discussed strategies for handling imbalanced datasets to improve fraud detection [9].

III. ANALYSIS AND DESIGN OF PROPOSED SYSTEM

Technical Feasibility

1. Data Preprocessing:

2. Standardization: Using StandardScaler from scikit-learn ensures that all features contribute equally to the model's performance by normalizing data. This technique is widely used and supported.

3. Handling Class Imbalance: SMOTE, available through the imbalanced-learn library, is an established method for generating synthetic samples in imbalanced datasets. It is effective in balancing classes, which is crucial for detecting rare fraudulent transactions.

2. Model Implementation:

Machine Learning Libraries: Models such as LightGBM, Bagging, RandomForest, SVM, DecisionTree, and AdaBoost are implemented using popular libraries (scikit-learn, XGBoost, lightgbm). These libraries are well-documented, with extensive community support and resources.

- **Performance Metrics:** Evaluation metrics like accuracy, F1 score, precision, and recall can be computed using scikit-learn, providing a comprehensive assessment of model performance.

3. Model Deployment:

- **Flask Framework:** Flask is a lightweight and widely-used framework for building web applications and APIs. It is suitable for creating a RESTful API for model deployment. Flask integrates seamlessly with Python and supports model inference.
- **Scalability:** Flask applications can be scaled horizontally using cloud platforms (e.g., AWS, Azure) or containerization (e.g., Docker). This allows for handling high traffic and ensuring robust performance.

3.1 Hardware and Software Requirements

HARDWARE:

- Processor: Intel® Core™ i3-2350M CPU @ 2.30GHz Installed memory (RAM):4.00GB
- System Type: 64-bit Operating System
- Hard disk: 10 GB of available space or more.
- Display: Dual XGA (1024 x 768) or higher resolution monitors
- Operating system: Windows

SOFTWARE

- PYCHARM IDE
- ANACONDA

IV. DESCRIPTION OF PROPOSED SYSTEM

1. LightGBM (Light Gradient Boosting Machine)

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed for efficiency and scalability, making it well- suited for large datasets. Key features include:

- **Gradient Boosting:** An ensemble learning method that builds models sequentially, where each new model corrects the errors of the previous one.
- **Histogram-Based Splitting:** Uses histograms to bin continuous features into discrete values, speeding up the computation and reducing memory usage.
- **Leaf-Wise Growth:** Grows trees leaf-wise rather than level-wise, which can lead to deeper trees and potentially better accuracy but requires more careful tuning to avoid overfitting.

2. Bagging (Bootstrap Aggregating)

Bagging is an ensemble method that improves the stability and accuracy of machine learning algorithms. It works by training multiple models (usually the same type) on different subsets of the data and then combining their predictions. Key aspects include:

- **Bootstrap Sampling:** Creates multiple subsets of the training data by randomly sampling with replacement.
- **Aggregation:** Combines predictions from all models, typically by averaging (for regression) or voting (for classification), to produce a final result. This reduces variance and helps prevent overfitting.

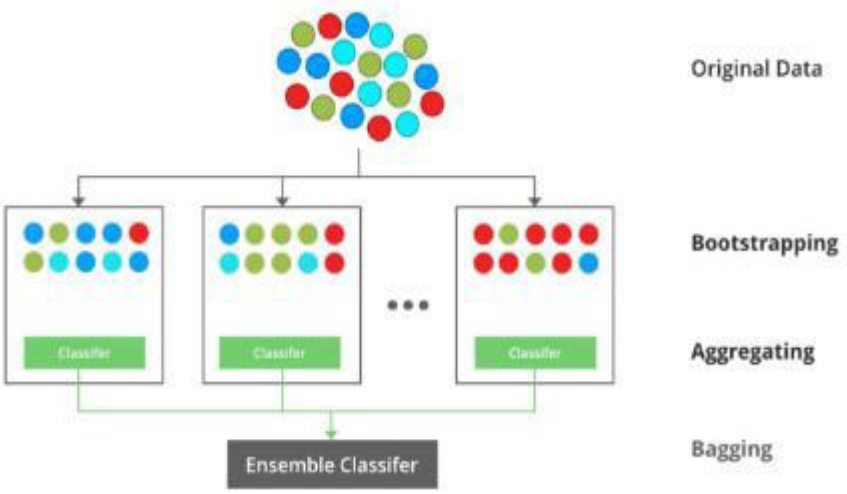
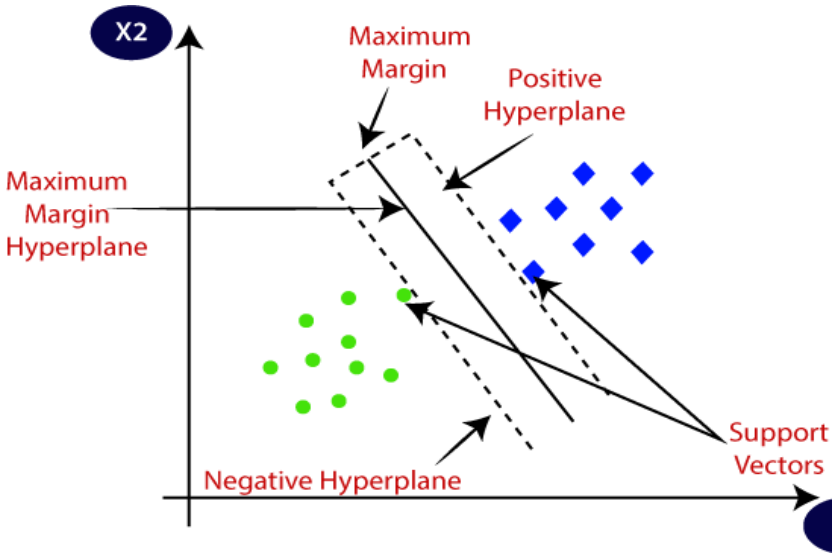


Fig: Bagging classifier

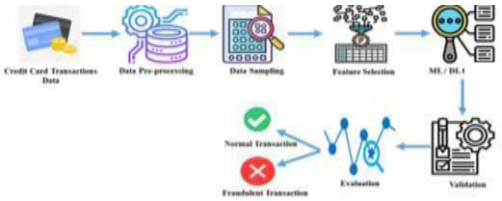
3. SVM (Support Vector Machine)

SVM is a supervised learning algorithm used for classification and regression tasks. It aims to find the hyperplane that best separates classes in the feature space. Key features include:

- **Margin Maximization:** SVM finds the hyperplane that maximizes the margin (distance) between different classes, leading to better generalization.
- **Kernel Trick:** Allows SVM to handle non-linearly separable data by mapping it to a higher-dimensional space using kernel functions (e.g., polynomial, radial basis function).



4.2Architecture Diagram



4.3 Detailed Description of Modules and Workflow

1. Data Collection

Objective: Gather and prepare the dataset required for training and evaluating machine learning models.

Description: The dataset used in this project is typically obtained from financial transactions data where fraud detection is a key concern. Due to confidentiality, the dataset in this case consists of anonymized features resulting from PCA (Principal Component Analysis) transformation.

Data Features:

Principal Components: Features V1, V2, … V28 are principal components derived from PCA, representing significant variance directions in the original data.

- **Time:** Represents the seconds elapsed between each transaction and the first transaction. It provides temporal context and can be used to detect patterns over time.
- **Amount:** Represents the transaction amount and can be used to introduce cost-sensitive learning, reflecting the potential impact of different transaction values.
- **Class:** The target variable indicating whether a transaction is fraudulent (1) or not (0).

- **Data Collection Process:**
  - **Acquisition:** The dataset is typically acquired from financial institutions, credit card companies, or public datasets like Kaggle. The data is usually provided in a structured format, such as a CSV file.
  - **Confidentiality:** The dataset is anonymized and preprocessed to protect sensitive information, which is common in financial datasets to comply with privacy regulations.

2. Data Preprocessing.

Objective: Prepare the dataset for model training by handling scaling and class imbalance.

- **Standardization (Using StandardScaler):**
  - **Purpose:** Normalize the feature values to ensure equal contribution from each feature.
  - **Implementation:** scaler = StandardScaler() followed by X_scaled = scaler.fit_transform(X).
  - **Rationale:** Standardizes features to have a mean of 0 and a standard deviation of 1, improving model performance.
- **Handling Class Imbalance (Using SMOTE):**
  - **Purpose:** Address the imbalance between fraudulent and non-fraudulent transactions.
  - **Implementation:** smote = SMOTE(random_state=42) followed by X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train).
  - **Rationale:** Creates synthetic samples for the minority class to improve the model's ability to detect fraud.

3. Model Training and Evaluation

Objective: Train and assess various machine learning models to identify the most effective algorithm for fraud detection.

- **LightGBM (Light Gradient Boosting Machine):**
  - **Description:** A fast and efficient gradient boosting framework using histogram-based methods.
  - **Implementation:** Train using lightgbm library functions.
  - **Evaluation Metrics:** Accuracy, F1 Score, Precision, Recall.
- **Bagging (Bootstrap Aggregating):**
  - **Description:** An ensemble method that improves stability by combining predictions from multiple models trained on different subsets.
  - **Implementation:** Use BaggingClassifier from scikit-learn.
  - **Evaluation Metrics:** Accuracy, F1 Score, Precision, Recall.
- **RandomForest:**
  - **Description:** An ensemble of decision trees with random feature selection at each split to

improve generalization.
  - **Implementation:** Train using RandomForestClassifier from scikit-learn.
  - **Evaluation Metrics:** Accuracy, F1 Score, Precision, Recall.
- **SVM (Support Vector Machine):**
  - **Description:** A supervised learning algorithm that maximizes the margin between classes using a hyperplane.
  - **Implementation:** Use SVC (Support Vector Classification) from scikit-learn.
  - **Evaluation Metrics:** Accuracy, F1 Score, Precision, Recall.
- **DecisionTree:**
  - **Description:** A tree-based algorithm that splits data into subsets based on feature values.
  - **Implementation:** Train using DecisionTreeClassifier from scikit-learn.
  - **Evaluation Metrics:** Accuracy, F1 Score, Precision, Recall.
- **AdaBoost (Adaptive Boosting):**
  - **Description:** An ensemble technique that focuses on misclassified examples to build a strong classifier.
  - **Implementation:** Use AdaBoostClassifier from scikit- learn.
  - **Evaluation Metrics:** Accuracy, F1 Score, Precision, Recall.

Maintenance:

- **Ongoing Costs:** Regular maintenance, updates, and monitoring to ensure the system remains effective and up-to-date.

Estimated Costs:

- **System Integration:** Development and integration costs for application and API.
- **Infrastructure:** Costs for cloud or local server hosting.
- **Maintenance:** Ongoing operational costs for system upkeep and monitoring.

V. CONCLUSION

To date, the project on credit card fraud detection has made significant strides. The literature review has provided a thorough understanding of existing methodologies and techniques, laying a solid foundation for the project. The feasibility study has been completed, covering technical, operational, and economic aspects, and confirming the project's viability. We have also selected the most appropriate machine learning models, including LightGBM, Bagging, RandomForest, SVM, DecisionTree, and AdaBoost, based on their effectiveness in handling imbalanced data and their performance in similar applications. The dataset, which includes PCA-transformed features and transaction- specific attributes, has been chosen to suit the selected models. With these preparatory steps accomplished, the project is set to proceed with data preprocessing, model training, evaluation, and deployment, ensuring a robust solution for real-time fraud detection.

REFERENCES

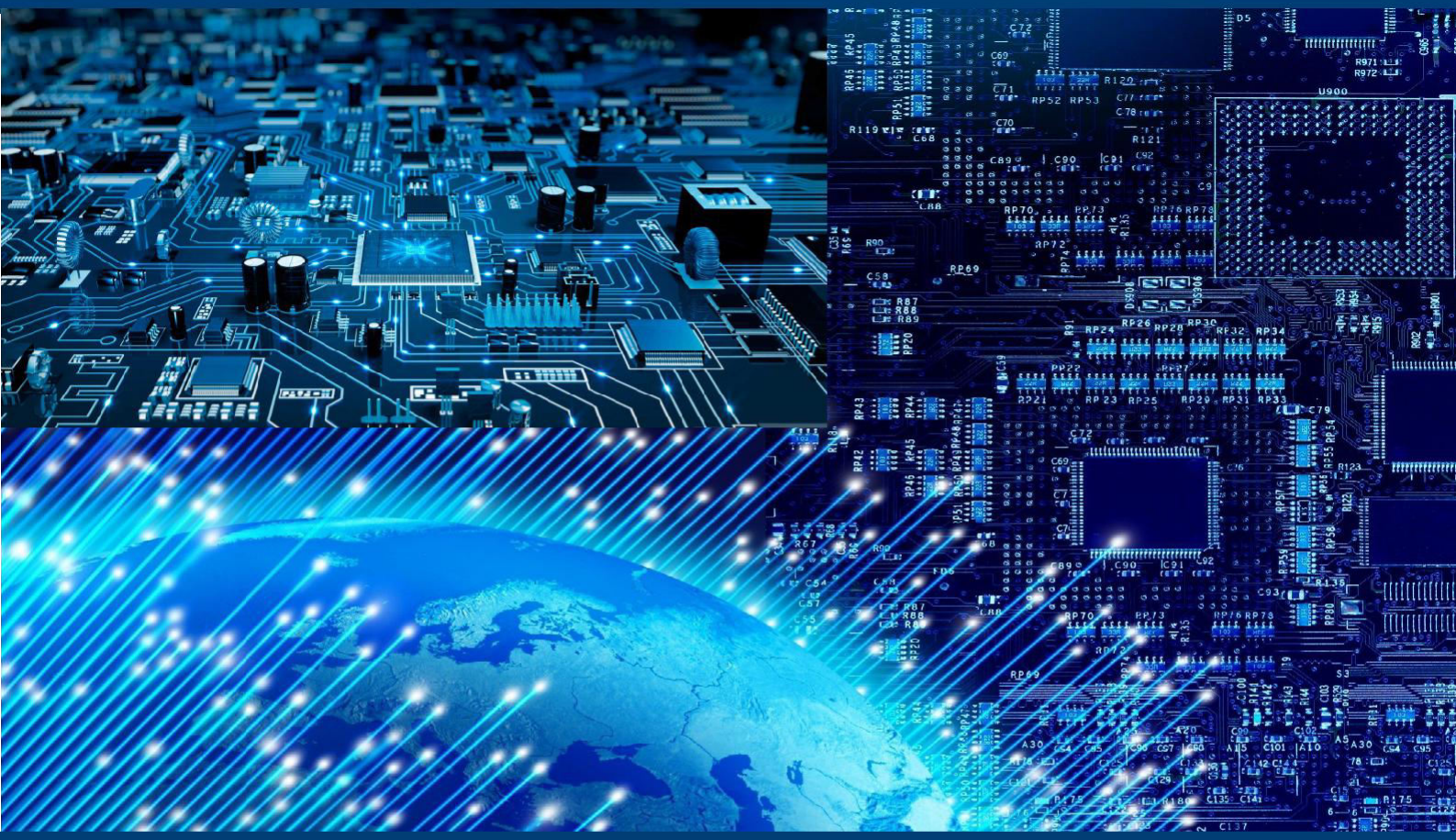
1. K. Randhawa, C. K. Loo, M. Seera, C. P. Lim and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," in IEEE Access, vol. 6, pp. 14277- 14284, 2018, doi: 10.1109/ACCESS.2018.2806420.
2. F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan and M. Ahmed, "Credit Card 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.
3. E. Ileberi, Y. Sun and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," in IEEE Access, vol. 9, pp. 165286-165294, 2021, doi: 10.1109/ACCESS.2021.3134330.
4. F. A. Ghaleb, F. Saeed, M. Al-Sarem, S. N. Qasem and T. Al-Hadhrami, "Ensemble Synthesized Minority Oversampling-Based Generative Adversarial Networks and Random Forest Algorithm for Credit Card Fraud Detection," in IEEE Access, vol. 11, pp. 89694-89710, 2023, doi: 10.1109/ACCESS.2023.3306621.
5. D. Mienye and Y. Sun, "A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection," in IEEE Access, vol. 11, pp. 30628-30638, 2023, doi: 10.1109/ACCESS.2023.3262020.
6. Y. Ding, W. Kang, J. Feng, B. Peng and A. Yang, "Credit Card Fraud Detection Based on Improved Variational Autoencoder Generative Adversarial Network," in IEEE Access, vol. 11, pp. 83680-83691, 2023, doi: 10.1109/ACCESS.2023.3302339.
7. W. Ning, S. Chen, S. Lei and X. Liao, "AMWSPLAdaboost Credit Card Fraud Detection Method Based on Enhanced Base Classifier Diversity," in IEEE Access, vol. 11, pp. 66488-66496, 2023, doi:

10.1109/ACCESS.2023.3290957.

8. Taha and S. J. Malebary, "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine," in IEEE Access, vol. 8, pp. 25579-25587, 2020, doi: 10.1109/ACCESS.2020.2971354.

9. S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. -S. Hacid and H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection," in IEEE Access, vol. 7, pp. 93010-93022, 2019, doi: 10.1109/ACCESS.2019.2927266.

10. E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba and G. Obaido, "A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection," in IEEE Access, vol. 10, pp. 16400-16407, 2022, doi: 10.1109/ACCESS.2022.3148298.



International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |