


Analysis of Ensemble Machine Learning Models for Fraud Detection

1st Siddharth Chourasia 


Dept. of Computer Science and Engineering
KIET Group of Institutions

Ghaziabad, India
sidofficial1005@gmail.com

2nd Satyam Kesharwani 


Dept. of Computer Science and Engineering
KIET Group of Institutions

Ghaziabad, India

3rd Shanu Sharma 

Dept. of Computer Science and Engineering
KIET Group of Institutions

Ghaziabad, India

4th Swati Sharma 

Dept. of Computer Science and Engineering
KIET Group of Institutions

Ghaziabad, India
swati.cse@kiet.edu

5th Bharti Chugh 

Dept. of Computer Science and Engineering
KIET Group of Institutions

Ghaziabad, India

bharti.cse@kiet.edu

Abstract— Recent developments in electronic payment and e-commerce have led to a rise in digital fraud cases, including credit card fraud. Within the financial services industry, identifying credit card fraud is still a difficult task that can have significant effects on consumers' and financial institutions' reputations. Therefore, it is crucial to implement systems capable of detecting credit card fraud. This research utilizes the European Cardholder Data Collection to assess the efficacy of ensemble machine learning techniques in identifying credit card fraud. The European cardholder data set is subjected to a rigorous assessment and evaluation of series of machine learning models, such as Gradient Boosting, Random Forest and CAT Boost, to detect fraudulent transactions. It has been noted that the dataset is unbalanced, which may indicate that the models' performance is not very ideal. The study suggests using data sampling strategies to achieve a balanced distribution of data across various algorithms that yield optimal outcomes. The main goal is to identify the model that performs best for machine learning classification problems by comparing the performance of different models. The results of this study will support further initiatives to improve credit card transaction security and lower financial losses brought on by fraud.

Keywords — Credit card, Random Forest, Gradient boost, XGBoost, CatBoost.

I. INTRODUCTION

Credit card fraud continues to pose a substantial threat to the financial well-being of individuals and institutions alike. With the exponential growth and advancement of technology, the world has seen in the recent years, the growth of electronic payments and online transactions have also increased at large. Fraudsters have adapted to exploit vulnerabilities in these systems, resulting in staggering financial losses. In 2020, global losses from card fraud exceeded \$27 billion, as reported by renowned Nilson Report, a figure that underscores the magnitude of this issue [1]. To address this pressing concern, the financial industry and data science community have actively sought innovative solutions to detect and prevent credit card fraud. In this regard, machine learning has emerged as a

potential tool for detecting fraudulent activities by leveraging patterns and anomalies present in transaction data. Series of machine learning models, such as Gradient Boosting, Random Forest and Bagging, have been widely used to enhance the accuracy and robustness of credit card fraud detection systems [2]. These models have shown promise in various fraud detection applications, leveraging the collective intelligence of multiple base learners to improve classification accuracy.

The European Cardholder Data set, also known as the "Credit Card Fraud Detection" data set, has become a benchmark for evaluating the performance of fraud detection algorithms. This data set, which contains anonymized credit card transactions, simulates real-world financial data, enabling researchers to rigorously test the effectiveness of fraud detection methods [3]. The data set has a large class imbalance, reflecting the distribution of credit cards in real life, with the maximum transactions being legitimate with only a small percentage categorized as fraudulent, out of the nearly 1.4 million cases of identity theft in 2020, there were 393,207 incidents related to credit card fraud. These scams now trail only government document and benefit fraud in terms of frequency among identity theft cases reported this year. 2020 saw 365,597 instances of fraud committed with brand-new credit card accounts. Between 2019 and 2020, identity theft complaints surged by 113%, and reports of credit card identity theft saw a notable increased of 44.6%. The global economy lost \$24.26 billion to payment card fraud in the previous year. 38.6% of losses resulting from credit card scams were recorded in the United States in 2018[4].

While ensemble machine learning models have proven effective in handling class imbalances and detecting credit card fraud, the rapid evolution of deep learning techniques, offers new avenues for exploration. The primary goal of this research paper is to undertake comprehensive analysis of the performance of ensemble machine learning models and deep

learning models on the Benchmark European Cardholder Data Set. The outcomes of this research aim to provide insights into the strengths and weakness of ensemble machine learning models for credit card detection, offering a guidance to practitioners and researchers in selecting the most suitable model for their specific needs.

II. LITERATURE REVIEW

This paper [8] discusses the challenges faced during data pre-processing in credit card fraud detection. It addresses issues like strident data, missing values, and redundancy in initial data. Data pre-processing, a crucial step in the Knowledge Discovery in Databases (KDD) process, helps overcome data complexity and improves analysis conditions. The paper focuses on sorting transaction attributes, classifying values, and handling initial data through cleansing, transformation, integration, and reduction. The paper also addresses issues like scarcity of real-time data on account of data sensitivity, and privacy concerns. The authors use synthetic minoring oversampling and logistic regression models to increase fraudulent data.

This study [4] uses publicly available data, high-class imbalance data, and alterations in the form of fraud to identify fraud using credit cards. In order to reduce false negative rates, deep learning methods are applied and the European card benchmark data set is utilized. The transaction data set's best features are ranked using feature selection methods, and a deep learning model with extra layers is suggested for feature extraction and classification. Several CNN layer designs are used to analyse the CNN model's performance. An evaluation that compares the suggested CNN with baseline model to ML algorithms reveals that the proposed method demonstrate superior performance than the current methods. The three performance assessment metrics of recall, accuracy, and precision are used to gauge how accurate the classifiers are. Moreover, our study uses variety of machine learning techniques— Support Vector Machine, Random Forest, KNN, Logistic Regression, Decision Tree, XGBoost, Extreme Learning Method—to solve the imbalance of data sets in the data set.

Credit card fraud poses a significant issue for businesses, and a study aims to classify fraudulent activity using various features. The study [7] found that the XGBoost Classifier achieved the highest performance among the four classification algorithms, with an overall accuracy of over 99%. However, due to imbalanced data, the F1 score is used for comparison. The researchers aim to improve classifier performance using three data balancing approaches, with Random Over Sampling yielded the optimal results. The XGBoost Classifier achieved around 99.9% accuracy, with an F1 score of 0.856, precision score of 0.913, recall of 0.805, and accuracy score of 0.99.

This research group has found that feature extraction and data sampling techniques significantly impact credit card fraud detection. The study found that RUS data sampling and CAE feature extraction techniques yielded the best results, while Random Forest performed well for F1 score metric and Cat

Boost was the best for AUC metric. The researchers suggest further exploration of data sampling and feature extraction algorithms. The findings support previous studies on this topic. [9]

A group at the Jadavpur University studied an auto-encoder based model designed for detecting fraudulent credit card transaction. A two-stage model is proposed for identifying fraudulent credit card transactions. In initial stage, an auto-encoder to transform transaction attributes into a lower-dimensional feature vector. Subsequently, in the second stage, a classifier utilizes this feature vector as input.. The model outperforms systems that rely solely on a classifier or other auto-encoder-based systems in terms of F1-score [10].

S. Prakoonwit [11] reported that the number of cybercrime victims, including credit card fraud, is increasing. Various methods, such as B-SMOTE and K-CGAN, are used for detection and prevention. The K-CGAN method shows promising results in generating data indistinguishable from the original. Credit card fraud is a notable issue due to widespread usage of credit cards, even more after Covid-19. Businesses face challenges in obtaining real-world data for research. An extensive strategy is needed to combat credit card theft. There were 4 credit card data sets in the analysis. The authors argue that more research is needed to examine the efficacy of techniques across various data sets and applications.

This study [12] introduces credit card fraud detection system based on machine learning that outperforms the current systems. The evolutionary algorithm is used for feature selection. The internet is experiencing exponential growth, and the author discussed how machine learning techniques are utilized to identify credit card fraud. Results from a variety of algorithms, including the Genetic Algorithm-Artificial Neural Network and Genetic Algorithm-Random Forest, are shown, demonstrating increased accuracy over previous methods. The application of a GA-based feature selection method and the utilization of a data collection containing credit card transactions are also mentioned in this paper. When compared to alternative approaches, the experimental findings show excellent performance and great accuracy. The goal to validate the framework with more data sets is mentioned in the text's conclusion. 42 publications were assessed by the researchers. A few of the conclusions assert that they support earlier research on this subject: Other machine learning techniques including the GA feature selection approach far better than the cutting-edge approaches; the GA-RF achieved 2.28% higher accuracy than LR. The authors propose that trials with various feature vectors were used to validate the suggested technique. In comparison to utilizing the genetic approach, the results indicated a considerable loss in performance when employing a random feature vector.

The proliferation of credit card usage has resulted in a rise in fraud, prompting the adoption of machine learning algorithms for detecting and preventing fraud. A comparative analysis of detection techniques and data confidentiality has been conducted, leading to proposal of hybrid solution that integrates neural networks in a federated learning framework as an

effective and accurate method while maintaining privacy. In [13], it was noted that most financial institutions now also offer internet banking services to the public. The effectiveness of hybrid approaches and ensemble classifiers has been identified in credit card fraud detection, particularly in addressing challenges arising from data imbalance and heterogeneity. The proposed hybrid approach using federated learning can improve fraud detection while preserving data privacy. Numerous machine learning algorithms like random forest, ANN, SVM, and KNN have been used for fraud detection, with accuracy rates ranging from 95% to 99.96%. Leveraging federated learning and blockchain technology has the potential to improve both data security and scalability. Imbalanced data can be handled through techniques like oversampling and under sampling. The proposed research model combines ANN with federated learning to achieve higher accuracy and privacy. A collaborative effort between banks and financial institutions can lead to creation of an effective fraud detection system. However, there are limitations to the proposed method in real life deployment.

However, the proposed method is effective in preserving privacy and using real-time data sets, but the deployment in real-life scenarios is constrained due to the rules and regulations of banks and financial institutions. Adapting the method will be challenging as each institution has its own limitations and relies on internal resources. Additionally, there is a risk of hackers decoding patterns learned by the trained model. Further work is needed to instil confidence in banks and financial institutes regarding the adoption of this technology.

A. Motivation

To attain the greatest possible success against previous studies with the least number of resources used has been an essential aspect that has been an important motivating factor for this study. Every day that passes, there is an increase in credit card industry cybercrime. In order to strengthen the systems that are attempting to identify these types of frauds, it is now essential to recognize the delicate nature of the situation and identify cutting-edge patterns and defences. By using machine learning techniques, we can harness the power of computers and make excellent use of them. They work well at detecting cybercrime, especially when using real-world data sets. These methods can assist detect unauthorised transactions quickly and efficiently, which will lessen the annual loss that financial institutions suffer from cybercrime. [14].

B. Key Contributions

- To understand and analyse the effectiveness of different ensemble models
- Using various balanced and unbalanced environments.
- To encode and scale to normalize and fit the data for the classification
- Using different models for analysis.

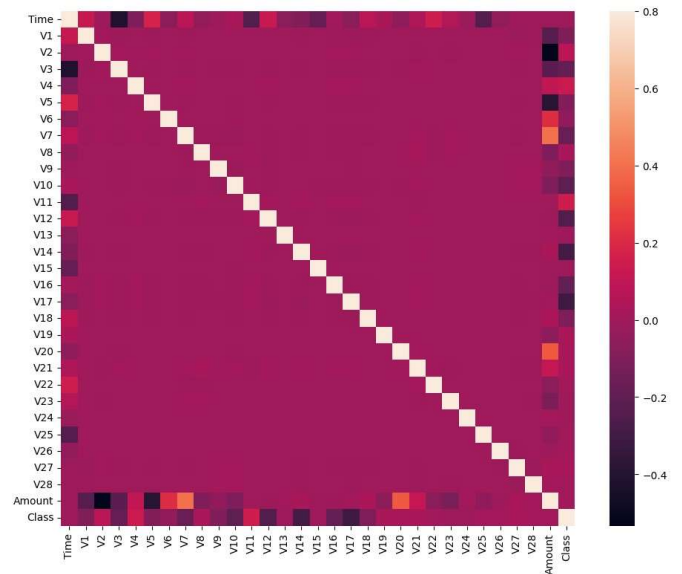


Fig. 1. Correlation Matrix for Unbalanced Dataset

III. PRELIMINARIES

A. Data Acquisition

For this proposed study, the European Cardholder Dataset [15] was used. The dataset is large with over 200K data instances. The dataset used in this study was collected from credit card transactions carried out by European cardholders in the year 2023, with sensitive information removed to ensure privacy and compliance with ethical guidelines.

B. Exploratory Data Analysis

After acquiring the dataset, EDA has been performed on this dataset to understand the different aspects of the dataset. The author has performed exploratory data analysis and found out the following. Figure 1 is the correlation matrix for unbalanced dataset.

TABLE I
EDA

Feature Name	Value
No. of Instances	284193
No. of features	31
No of Classes	2
No of Missing Values	0
No of Numeric Features	31
No of categorical features	0
Majority Class %age	99.83301
Minority Class %age	0.16699

C. Data Preprocessing

Before using the dataset, the author proposes to perform preprocessing and maintain and secure the integrity of the dataset. The dataset has some NA values which need to be handled accordingly. The data set has 31 columns on which

PCA has been applied and so forth. Those NA values have been replaced with using mean, which generally indicates a better handling, due to the fact that our data is highly imbalanced with the fraud percentage in the data set to be only the bare minimum of 0.16999% of the whole data.

The dataset also included 1081 duplicate values that have been removed for better performance and accuracy.

1) *Data Encoding*: In this study, the major challenge was to reshape the data for classification of a transaction as fraud or non-fraud. This could not be achieved directly by applying ML models to the dataset as it first needs to be encoded for a binary classification problem. For this, label encoders have been used to encode the data for our proposed solution.

2) *Data Normalization*: For this dataset, the author has used Robust Scaling method to normalize the data on interquartile ranges rather than using MinMax Scaler which is not robust to the outliers in the dataset.

D. Supervised Learning Models

Finally, after making the dataset ready for the ensemble models, pre-processing and balancing the data and reviewing several research papers, now the authors began with the testing of different ensemble models that have been proposed to use in this research.

These models were tested on both balanced and unbalanced data set and numerous performance metrics were used to test the performance of these models under different circumstances.

1) *Decision Tree Classifier*: Decision trees are versatile supervised machine learning algorithms commonly used for classification and regression tasks. They make predictions by recursively partitioning the data set based on feature attributes. Information Gain (IG) Formula (for classification):

$$IG(D, A) = H(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} \cdot H(D_v) \quad (1)$$

where H is the entropy [16].

2) *XGBoost (Extreme Gradient Boosting)*: XGBoost construct a series of decision trees sequentially, where each tree is designed to rectify the errors of its predecessors. Utilizing gradient descent optimization to minimize a differentiable loss function, XGBoost enhances model performance [17]. Objective Function:

$$\text{Objective}(XGBoost) = L(\theta) + \Omega(f) \quad (2)$$

3) *Gradient Boosting Algorithm*: Gradient Boosting is a versatile ensemble learning technique extensively employed for both classification and regression problems. It combines several weak learners, mainly decision trees, to formulate a robust predictive model. The gradient boosting process minimizes a loss function that quantifies the difference between predicted and actual values [18]. Boosting Formula:

$$F_m(x) = F_{m-1}(x) + \arg \min_h \sum_i L(y_i, F_{m-1}(x_i) + h(x_i)) \quad (3)$$

4) *CatBoost Algorithm*: CatBoost automates the encoding of categorical features and streamlines the learning process, making it an accessible choice for practitioners. In classification tasks, typically employing a logarithmic loss function (cross-entropy) to quantify the difference between predicted and actual class labels. [19]. Logarithmic Loss (Cross-Entropy) Formula:

$$L(y, p) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (4)$$

4) *Random Forest Classifier*: Random Forest, developed by Leo Breiman [20] is an ensemble learning technique that elevating classification accuracy through the fusion of numerous decision trees.. It introduces randomness by creating bootstrapped samples from the data set and using random feature selection when building each tree. The ultimate prediction is established by employing a voting mechanism in the context of classification tasks or through an averaging approach in regression tasks., with each tree's prediction contributes to the overall decision.

E. Resampling

In this dataset, the target attribute has a majority class of non-fraud cases so, the dataset is resampled using SMOTE over sampling technique. Detecting credit card fraud poses a classification challenge where the aim is to achieve classification for a given instance in the data set whether it is a fraud or not. A crucial first step in developing a successful fraud detection algorithm is balancing the data set. With a relatively low percentage of fraudulent transactions and the large majority of legal (non-fraudulent) transactions, the data set usually shows a severe class imbalance. Studies in [5] and other researchers in the field of fraud detection emphasize how crucial it is to solve class imbalance in order to achieve robust model performance.

To rectify this imbalance, oversampling the fraudulent transactions (minority class) or under-sampling the legal transactions (majority class) is necessary for balancing the data set. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE), as introduced in the paper [6], are commonly employed to generate artificial instances of the minority class, contributing to a more equitable distribution of data.

According to the research [7], the random oversampling method produces 0.99 precision and 0.99 accuracy scores when implemented on XGBoost, making it the most appropriate methodology for handling imbalanced data. However, in our proposed work, we have balanced our data set using the SMOTE over sampling strategy, which is recommended by the majority of researchers worldwide.

F. Performance Metrics

For the proposed work, the three standard performance metrics are used - Precision, Recall and F1 Score.

1) *Precision*: The percentage of accurate positive predictions out of all the positive predictions made by the model..

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (5)$$

2) *Recall*: The percentage of true positive predictions among all actual positive data instances is termed as recall.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (6)$$

3) *F1 Score*: The F1 score is a balanced evaluation metric that is calculated as the harmonic mean of recall and precision.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

These metrics are frequently utilized to evaluate the effectiveness of classification models. They provide valuable insights into the model's ability to precisely identify positive instances (precision), its capability to capture all positive instances (recall), and a harmonious balance between the two (F1 score).

IV. PROPOSED ARCHITECTURE

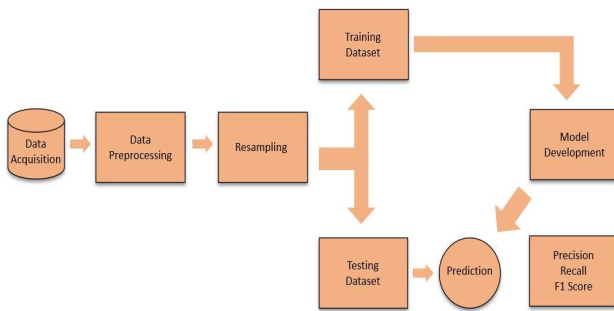


Fig. 2. Proposed Architecture

In figure 2, firstly the data has been acquired, then on this data EDA was performed and the data was pre-processed for further work where this dataset was resampled using balancing

techniques and then the dataset is partitioned into training and test data and the models were trained and the results were produced.

V. RESULTS AND DISCUSSIONS

Table II shows the performance evaluation of different models when an unbalanced dataset was used to train the models. The table shows a comparative analysis of how different models have performed when the data set was imbalanced.

TABLE II COMPARISON OF MODELS ON UNBALANCED DATA			
Model Used	Precision	Recall	f1 score
Decision Tree	81.05	83.69	82.35
XGBoost	98.75	84.95	91.33
Gradient Boost	88.76	73.15	80.20
CatBoost	97.61	84.51	90.61
Random Forest	97.47	83.70	90.06

Table III shows the comparative analysis of different models when the dataset was balanced using SMOTE oversampling. It gives an idea of how each model has performed when it was trained with a balanced dataset.

TABLE III COMPARISON OF MODELS ON BALANCED DATA			
Model Used	Precision	Recall	f1 score
Decision Tree	80.65	73.53	76.92
XGBoost	94.19	83.51	88.52
Gradient Boost	94.25	74.55	83.25
CatBoost	75.47	86.96	80.81
Random Forest	96.05	78.49	86.39

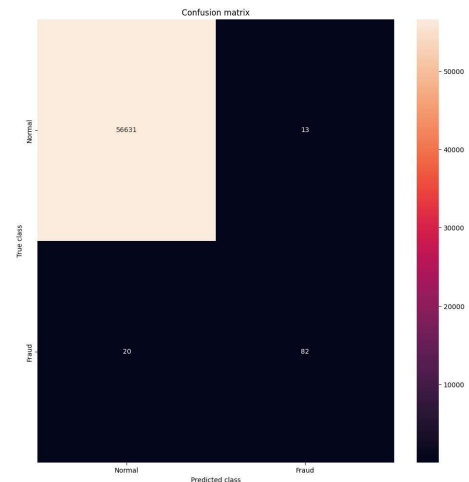


Fig. 3. Confusion Matrix for XGBoost

VI. CONCLUSION & FUTURE SCOPE

After careful evaluation and analysis of various models in Table II and Table III, it can be concluded that XG Boost Algorithm gives better and more accurate results under both balanced (88.52%) and unbalanced (91.33%) environments. Though, each model showed a significant decrease in the f1 score when the dataset was balanced, Gradient Boosting Algorithm performed better and showed a significant increase from 80.20% to 83.25% when the dataset was balanced.

The current area of study is field that still needs to be explored deeply under different criteria and much work is needed in the field of deep learning and using explainability to better understand the relation between different attributes. Also, it is proposed to use the novel GPT Architecture for detecting the frauds in the industry which will showcase the power of this advanced model beyond its general capacity and known use in natural language processing. Additional deep learning models could also prove to be a potential asset to this fraud detection system, which still remains unexplored.

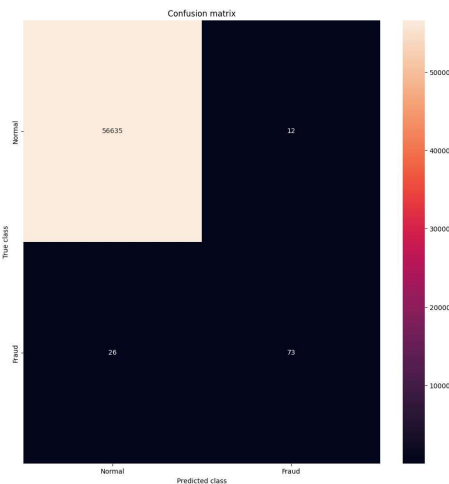


Fig. 4. Confusion Matrix for Gradient Boosting

REFERENCES

- [1] Nilson Report. "Card Fraud Losses Reach \$27.36 Billion." (2020).
- [2] Abreu, Pedro, and Sergio Guerreiro. "A novel approach for credit' card fraud detection using bayesian networks." *Expert Systems with Applications* 37.6 (2010): 3696-3702.
- [3] Dal Pozzolo, Andrea, et al. "Calibrating probability with undersampling for unbalanced classification." 2015 IEEE Symposium Series on Computational Intelligence. IEEE, 2015.
- [4] Alarfaj, F.K., Malik, I., Khan, H.U., Almusallam, N., Ramzan, M. and Ahmed, M., 2022. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE Access*, 10, pp.39700-39715.
- [5] Ribeiro, A., Santos, M. F., & Oliveira, S. (2019). Resampling strategies for imbalanced datasets: A systematic review. *Expert Systems with Applications*, 129, 67-82.

- [6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
- [7] Gupta, P., Varshney, A., Khan, M.R., Ahmed, R., Shuaib, M. and Alam, S., 2023. Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques. *Procedia Computer Science*, 218, pp.2575-2584.
- [8] Bathala, S. B., & Nagendra, M. (2019). Data pre-processing for credit card data set using synthetic minority over-sampling techniques.
- [9] Salekshahrezaee, Z., Leevy, J. L., & Khoshgoftaar, T. M. (2023). The effect of feature extraction and data sampling on credit card fraud detection. *Journal of Big Data*, 10(1).
- [10] Misra, S., Thakur, S., Ghosh, M., & Saha, S. K. (2020). An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction. *Procedia Computer Science*, 167, 254-262.
- [11] Strelcenia, E., & Prakoonwit, S. (2023). Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation. *AI*, 4(1), 172-198.
- [12] Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, 9(1).
- [13] Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of Machine Learning Approach on Credit Card Fraud Detection. *Human-Centric Intelligent Systems*, 2(1-2), 55-68.
- [14] Dantas, R.M., Firdaus, R., Jaleel, F., Mata, P.N., Mata, M.N. and Li, G., 2022. Systemic acquired critique of credit card deception exposure through machine learning. *Journal of Open Innovation: Technology, Market, and Complexity*, 8(4), p.192.
- [15] Kaggle, "Credit Card Fraud Detection," 2018. [Online]. Available: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>
- [16] Rokach, L. and Maimon, O., 2005. Decision trees. *Data mining and knowledge discovery handbook*, pp.165-192.
- [17] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- [18] He, Z., Lin, D., Lau, T. and Wu, M., 2019. Gradient boosting machine: a survey. *arXiv preprint arXiv:1908.06951*.
- [19] CatBoost documentation. "Loss Functions for Classification." Retrieved from <https://catboost.ai/en/docs/concepts/loss-functions-classification>
- [20] Breiman, L., 2001. Random forests. *Machine learning*, 45, pp.5-32.