An Investigative Approach to Payment Card Fraud Detection using Machine Learning Techniques

MSc Research Project

MSc Fintech

Nwabuogoh Anne Alu

Student ID: x22115871

School of Computing

National College of Ireland

Supervisor: Brian Byrne

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nwabuogoh Anne Alu
Student ID:	x22115871
Programme	MSc Fintech
Year	2023
Module	MSc Research Project
Supervisor	Brian Byrne
Submission Due Date	14/08/2023
Project Title	An Investigative Approach to Payment Card Fraud Detection using Machine Learning Techniques
Word Count	6020
Page Count	26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

An Investigative Approach to Payment Card Fraud Detection using Machine Learning Techniques

Nwabuogoh Anne Alu

x22115871

Abstract

In today's rapidly evolving digital landscape, the threat of payment card fraud has escalated, imposing substantial financial burdens on individuals and businesses. This study is motivated by the imperative to counteract this menace, aiming to investigate the effectiveness of diverse supervised and deep learning techniques, including Extreme Gradient Boosting (XGBoost), Logistic Regression, LightGBM, Long Short-Term Memory Recurrent neural network (LSTM-RNN), Random Forest and Multilayer Perceptron. To handle the problem of class imbalance, the hybrid approach was employed, SMOTE for oversampling and Edited Nearest Neighbors (ENN) for under-sampling. Notably, the model evaluation results highlight the prowess of boosting classifiers, especially LightGBM and XGBoost, in detecting credit card fraud, both techniques had an F1-Score of 0.63 and a PR-AUC score of 80% and 81.6% respectively.

***Keywords—LSTM, deep learning, LightGBM, MLP, Payment Card Fraud, Random Forrest, ANN*, SMOTE**

1.0 Introduction

Payment card fraud has become a pervasive and persistent threat in today's digital era, posing significant challenges to financial institutions, merchants, and consumers. As technology continues to advance, fraudsters are constantly adapting their tactics, exploiting vulnerabilities in payment systems, and compromising individuals' card data for fraudulent activities. This research project’s objective is to investigate the application of machine learning techniques in combating payment card fraud, recognizing its evolving nature and the need for innovative solutions.

Fraud and the Evolution of Payment Card Fraud

Fraud, in general, entails intentional deception or misrepresentation for personal gain, often involving illegal activities like identity theft, forgery, and financial scams. Payment card fraud specifically targets credit and debit cards, with fraudsters seeking unauthorized access to cardholder data for illicit purposes. The evolution of payment card fraud can be traced to the widespread adoption of electronic payment systems and increasing reliance on digital transactions.

With rapid technological advancement, fraudsters exploit new avenues in payment card systems. Traditional methods, like stolen or counterfeit cards, have evolved into more sophisticated techniques such as data breaches, skimming devices, and online phishing scams. The expanding e-commerce, mobile payments, and contactless transactions landscape widens the threat horizon, demanding continuous efforts to outpace fraudsters.

Rise of Payment Card Fraud

Statistics underscore the alarming rise in payment card fraud, underscoring the urgency to address this issue. This upward trend is expected to persist as fraudsters refine their approaches and exploit emerging technologies. High-profile data breaches, like the 2013 Target breach compromising over forty million users' payment card information and the 2017 Equifax breach exposing personal data of around 143 million people, highlight vulnerabilities in the payment card ecosystem.

Methods Employed by Fraudsters

Fraudsters employ diverse methods to access individuals' card data, enabling fraudulent transactions. Understanding these methods is crucial for developing effective countermeasures. Common tactics encompass Data Breaches, Skimming, Phishing, Social Engineering, Malware, and Ransomware.

Earlier Methods of Fraud Detection

Before the advent of advanced machine learning algorithms, the use of traditional methods for payment card fraud detection was the norm, some of these methods include:

Rule-Based Systems: Specific rules and criteria detected potentially fraudulent transactions, often relying on pre-established patterns. For instance, exceeding a predetermined transaction amount triggered a manual review alert.

Manual Review: To find potential fraud, human analysts would manually evaluate transactions that had been flagged or that had suspicious patterns. Determining whether a transaction was legitimate entailed looking at the specifics of the transaction, customer behavior, and other pertinent data. Manual review was efficient but time-consuming and had a limited potential to scale. These earlier methods of payment card fraud detection relied on predefined rules, manual intervention, and limited data analysis capabilities. Although they offered some protection, they were less successful at spotting intricate and evolving fraud patterns.

Machine Learning in Combatting Payment Card Fraud

The advancement of machine learning techniques has been proven as a powerful tool in the detection and prevention of payment card fraud. By leveraging large datasets and automated learning algorithms, machine learning models can analyze patterns, identify anomalies, and make accurate predictions regarding fraudulent activities. The application of machine learning in combatting payment card fraud offers several advantages, including real-time detection, scalability, and adaptability to evolving fraud patterns.

Machine Learning Techniques in Payment Card Fraud Detection

There exist quite a few ML techniques that have proven effective in combatting payment card fraud. These include:

Supervised learning algorithms, such as SVM, XGBoost, k-Nearest Neighbors, and random forests, learn from labeled data to classify correctly payment transactions to be either legitimate or fraudulent.

Unsupervised learning techniques, such as K-means, isolation forest, PCA and apriori algorithms, identify anomalies in patterns in customer transaction data. These approaches are particularly useful for detecting previously unseen fraud patterns and new threats.

Deep learning techniques, such as convolutional neural networks (CNN) and LSTM-RNN, excel at extracting complex features from transaction data. They can capture intricate patterns and relationships, enhancing fraud detection accuracy.

Payment card fraud continues to be a pressing concern, evolving alongside technological advancements. The escalating rise in fraud cases necessitates innovative approaches to combat this issue effectively. Machine learning techniques offer promising solutions by leveraging advanced algorithms and large datasets in detecting, analysing, and preventing payment card fraud..

1.1 Research Question and Objectives

The question this research seeks to answer is:

"*How well do different machine learning models perform in detecting payment card fraud*".

The research objectives include:

- Assess and contrast the efficiency of various models, then choose the most effective one.
- Utilize a sampling strategy to address the imbalance within the classes.
- Analyze and choose appropriate essential performance measures.

Section 2 provides a comprehensive analysis of the existing research on credit card fraud detection and section 3 outlines the research methodology employed to address the research question, presenting an in-depth overview of the various phases involved. Section 4 discusses the proposed system process flow, while the implementation of distinct models is deliberated in Section 5. Following this, Section 6 showcases and assesses the outcomes and discoveries from the experiments, accompanied by an extensive discourse on the results. Lastly, in Section 7, the core insights of this research are synthesized, and suggestions for future endeavors are put forth.

2.0 Related Work

2.1 Supervised Machine and Deep Learning Approach to Fraud Detection

(Singhai, et al., 2023) highlights the importance of detecting fraudulent card transactions using machine learning methods. Various algorithms have been utilized for credit card fraud detection and among these, the KNN algorithm stands out as a non-parametric classification approach that considers the mainstream class of its k-nearest neighbors. While KNN improves detection rates and reduces false alarms, selecting the optimal value of k remains a challenge. Additionally, the authors suggest the use of real-time datasets and diverse data to address data imbalance and heterogeneity in credit card fraud detection. Privacy concerns, however, pose ongoing challenges. The effectiveness of KNN in identifying credit card fraud is highlighted, surpassing rule-based systems and achieving high accuracy and combining KNN with other machine learning models can further improve fraud detection.

(Vejalla, et al., 2023) used RF, SVM, DT, LR and Naïve Bayes classifier in their research, the data used for the research was two days' worth of transactions from European credit card holders consisting of almost three hundred thousand records. The sample size for the training and test were varied and the algorithms and the results indicate that Random Forest algorithm performed better than the rest of the algorithms in all the samples.

(Alarfaj, et al., 2022) suggests that that while deep learning has shown promising results in various domains, only a few studies have explored its application in credit card fraud detection and acknowledge the challenge of class imbalance in credit card fraud datasets and how it is yet an unsolved problem. Their research was focused on supervised and unsupervised learning algorithms, using a combination of machine learning and deep learning algorithms to detect fraudulent transactions. The researchers started out with feature selection techniques to prioritize the features in order of importance, then applied the deep learning model, CNN, to further extract relevant variables.

The research paper by (Ugarković & Oreški, 2022) discusses the challenges of dealing with imbalanced datasets in machine learning algorithms. The paper proposes a hybrid approach that combines supervised and unsupervised machine learning techniques to handle the imbalanced data. Specifically, cluster analysis, as an unsupervised machine learning algorithm, is applied to the most significant variables identified through sensitivity analysis on predictive models developed using the decision tree, a supervised machine learning algorithm. The results of the study indicate that the hybrid approach of decision tree and cluster analysis shows promise as an effective tool for working with imbalanced data. The combined approach provided quality results and produced interpretable models for dealing with class-imbalanced datasets.

In their recent study, (Gupta, et al., 2023) synergistically employed a blend of machine learning algorithms and diverse data balancing strategies. The utilized algorithms encompassed Decision Tree (DT), Artificial Neural Network (ANN), XGBoost, and Logistic Regression. Concurrently, data balancing techniques like Random Oversampling, Random Under-sampling, and SMOTE were harnessed. The findings of the research underscored the potential pitfalls of directly applying machine learning algorithms to imbalanced datasets, as it can significantly skew outcomes in favor of the majority class. However, the introduction of data balancing techniques markedly rectified this bias, rendering the results more equitable and representative.

(Forough & Momtazi, 2022) proposed a novel solution for credit card fraud detection. They introduced a deep learning model that considers transaction sequences and addresses the issue of imbalanced datasets using a new under-sampling technique called Seq-US. The study explores the use of Long Short-Term Memory (LSTM) as a sequential classification model, incorporating a sliding window for contextual patterns. Additionally, Conditional Random Field (CRF) is employed to leverage transaction prediction dependencies. The Seq-US method demonstrated superior performance compared to other sampling approaches, enhancing model accuracy while reducing training time.

2.2 Class Imbalance

(Cicak & Avci, 2023) highlights the significance of addressing imbalanced data in machine learning, specifically in predictive maintenance tasks. The main objective of their research was to enhance classification success by employing resampling methods to tackle class imbalance. Various oversampling and under-sampling techniques were applied to create balanced datasets from synthetic and real-world datasets, and the performance of different classifiers is compared, with the Jrip algorithm showing the best results for imbalanced datasets. The authors discuss the challenges of applying oversampling methods on large datasets, suggesting that a hybrid approach to resampling could prove to be more effective.

On the other hand, (Tran & Dang, 2021), employs two resampling techniques, SMOTE and ADASYN, on credit card data to create balanced datasets. The research provides valuable insights into the effectiveness of the ML models for fraud detection after resampling the dataset. The use of different classification evaluation indices allows for a thorough assessment of the models. The comparisons between algorithms based on SMOTE and ADASYN, as well as between ML algorithms based on classification measures, contribute to building an efficient fraud detection system.

(Yakshit, et al., 2022) emphasize the importance of preprocessing imbalanced data to improve classification accuracy and reduce bias, highlighting the significance of using appropriate oversampling methods for better model performance in imbalanced dataset analysis. The research explores a range of methods, including oversampling techniques (ADASYN and SMOTE) combined with classification algorithms (SVM and CNN with SVM). The experimentation is performed on Google Colab, and the performance of different machine learning techniques, namely SVM and (CNN with SVM), is evaluated. The results indicate that the combination of SVM and CNN outperforms SVM alone, and SMOTE yields better results than ADASYN in terms of performance metrics such as precision, recall, and F1 score.

In their research, (Prajapati, et al., 2021) employed a combination of deep learning and machine learning algorithms, including Artificial Neural Network (ANN), XGBoost, and Random Forest, for credit card fraud detection. To address the imbalanced data issue, they applied three resampling techniques: SMOTE Oversampling, SMOTE Under-sampling, and SMOTE-Tomek (hybrid). The authors emphasized the importance of accuracy, F1-score, recall, and precision as evaluation metrics, particularly when dealing with imbalanced data. However, they highlighted that precision and recall hold greater significance in this context. The study's results demonstrated that Random Forest slightly outperformed ANN and XGBoost, showing slightly higher precision and recall. Furthermore, the research revealed that the SMOTE-Tomek hybrid technique proved to be the most effective in managing the imbalanced data, providing valuable insights into the selection of appropriate resampling methods for credit card fraud detection tasks.

Assaghir et al. (2019) conducted a study on credit card fraud detection and highlighted the impact of unbalanced datasets on inaccurate results and significant financial losses. To tackle this issue, they balanced the dataset and utilized various machine learning algorithms. The results revealed that C5.0, LR, Decision Tree, SVM, and ANN exhibited the highest levels of Sensitivity, Accuracy, and AUCPR. The authors cautioned that handling extreme class imbalance using certain techniques might lead to unfavorable outcomes and a high number of false positives. To avoid misleading assessments, they wisely employed three metrics to evaluate the model's performance. Additionally, the study emphasized the importance of dataset size, suggesting that larger datasets could potentially yield better results.

3.0 Research Methodology

This section of the study outlines the research techniques employed, following the principles of the Knowledge Discovery in Database (KDD) approach. The section covers essential aspects such as data selection, research procedure, and techniques used throughout this study.

3.1 Data Selection

Payment card data consists of transaction records customers of financial institutions, these data are of a sensitive nature, such that organizations in possession of this data safeguard it against data leakage and are bound by various data privacy laws to protect their customers data. The availability of public datasets is limited and those available have either been transformed using Principal Component Analysis (PCA) or have been generated using a transaction data simulator.

The choice of dataset for this study was sourced from Kaggle[1], the dataset contains simulated transaction data of credit card holders. The simulated data contains transactions from January 1st, 2019, to December 31st, 2020. The initial dataset was composed of two separate files that were merged to enhance the depth of data analysis, the final merged dataset comprises 1,852,394 credit card transactions conducted by 1,000 customers, encompassing 23 distinct transaction attributes This dataset was selected because the records still contained the data in their original form and had not been transformed using PCA. The table below shows the variables within the dataset.

S/N	Variable Name	Description
1	amt	The amount of the transaction
2	city-pop	The total population of the customer’s city

[1] https://www.kaggle.com/datasets/kartik2112/fraud-detection

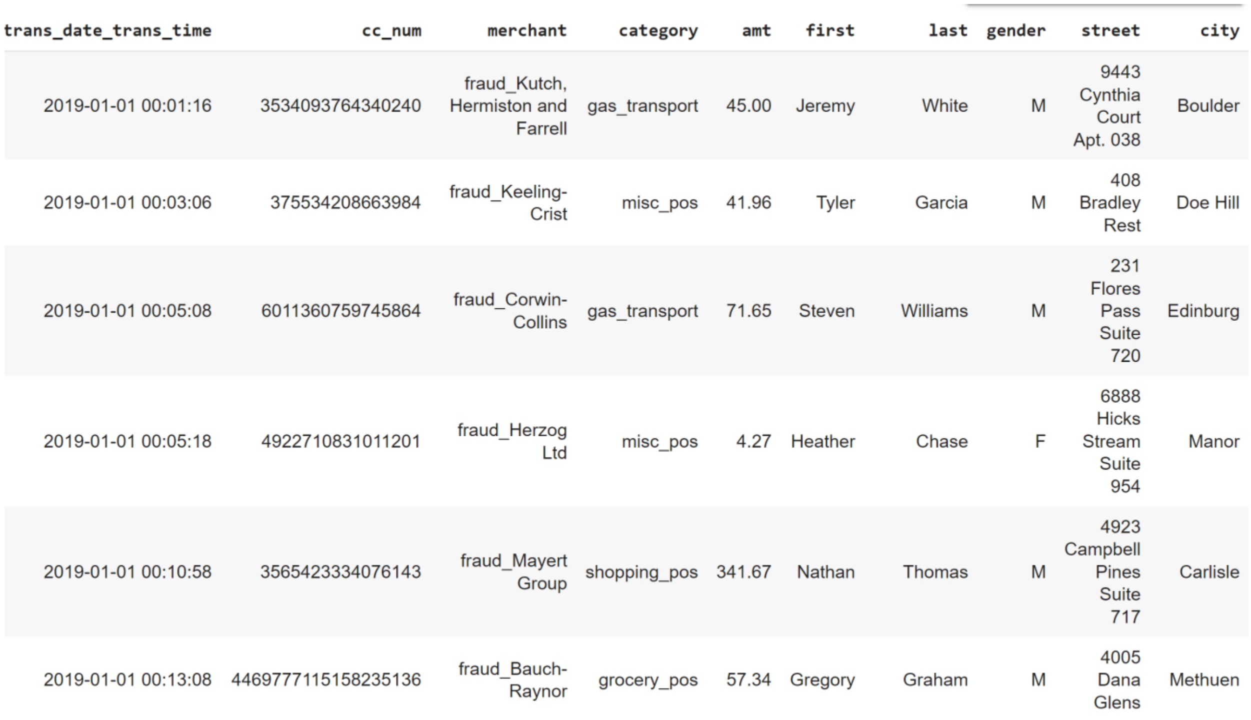
3	unix-time	Unix time of the transaction
4	trans-day-trans-time	Transaction date and time
5	dob	The customer’s date of birth
6	first	The customer’s first name
7	last	The customer’s last name
8	gender	the customer’s gender
9	merchant	The recipient of the transaction
10	merch-lat	The latitude of the merchant
11	merch-long	The longitude of the merchant
12	Street/city/state	The state, street and city of the customer
13	trans-num	The transaction reference number
14	cc-num	The customer’s card number (PAN)
15	job	The customer’s job type
16	category	The category of the transaction
17	zip	The customer’s zip code
18	lat & long	The customer’s longitude and latitude
10	IsFraud	Boolean representing if the transaction was fraudulent or genuine

Table 1: Data Description

3.2 Data Preprocessing

This step consists of several tasks that need to be performed on the data before the machine learning models can be trained on it. The original dataset was quite large and required more computational resources than what was provided on Google Colab, therefore, a stratified sample of the data was saved to a csv file, which was 40% of the original dataset. This was done using R in RStudio. The programming language used for this research was Python and a cloud based integrated development environment (IDE) was used for the coding aspect of the study.

The final dataset was read into Google Colab using the pandas package `read_csv` method and the different formats of the variables within the dataset were explored. On examination, the data types for the variables consisted of numerical, nominal, and datetime.



trans_date_trans_time	cc_num	merchant	category	amt	first	last	gender	street	city
2019-01-01 00:01:16	3534093764340240	fraud_Kutch, Hermiston and Farrell	gas_transport	45.00	Jeremy	White	M	9443 Cynthia Court Apt. 038	Boulder
2019-01-01 00:03:06	375534208663984	fraud_Keeling-Crist	misc_pos	41.96	Tyler	Garcia	M	408 Bradley Rest	Doe Hill
2019-01-01 00:05:08	6011360759745864	fraud_Corwin-Collins	gas_transport	71.65	Steven	Williams	M	231 Flores Pass Suite 720	Edinburg
2019-01-01 00:05:18	4922710831011201	fraud_Herzog Ltd	misc_pos	4.27	Heather	Chase	F	6888 Hicks Stream Suite 954	Manor
2019-01-01 00:10:58	3565423334076143	fraud_Mayert Group	shopping_pos	341.67	Nathan	Thomas	M	4923 Campbell Pines Suite 717	Carlisle
2019-01-01 00:13:08	4469777115158235136	fraud_Bauch-Raynor	grocery_pos	57.34	Gregory	Graham	M	4005 Dana Glens	Methuen

Figure 1: Snapshot of Dataset

The dataset was then checked for duplicates and missing data using the seaborn python package, there were no duplicates or missing data that needed to be catered for. The variable 'X' which was created by R as a serial number field was dropped and the initial stage and the variable '`y_sample`', the target variable, was renamed to isFraud. Features which had a high number of unique values (high cardinality) were also dropped during the experiment, as they were found to have little effect on model performance.

3.3 Data Transformation

The "gender" variable was transformed into numerical values, and the time-based variables like '`trans_date_trans_time`', '`dob`', and '`unix_time`' were converted to datetime values, while the features '`zip`' and '`cc_num`' were converted to string values to make them suitable for training the machine learning algorithms. The target variable was already in numerical form and did not require categorical encoding, but the shopping category variable ‘category’ was transformed into categorical column. Feature scaling was also considered for a few of the variables, but had no significant impact of the results, therefore it was reversed.

3.3.1 Feature Engineering

Some features lacked substantial information when considered in their original form, as a result feature engineering needed to be performed on them. The ‘age’ variable was derived from the ‘trans_date_trans_time’ and ‘dob’ features. Additional features ‘hour’, ‘week_day’ and ‘month’ were derived from the ‘trans_date_trans_time’ feature, also, the features ‘cust_merc_lat_dist’ and ‘cust_merc_long_dist’ were gotten from the features “lat’, ‘long’, ‘merch_lat’, and ‘merch_long’. The original features were then dropped after the creation of the new features, as they had become redundant.

3.3.2 Class Imbalance

The class distribution result indicates a severe class imbalance in the payment card transaction dataset, with approximately 99.4% of transactions being genuine and only 0.58% being frauds.

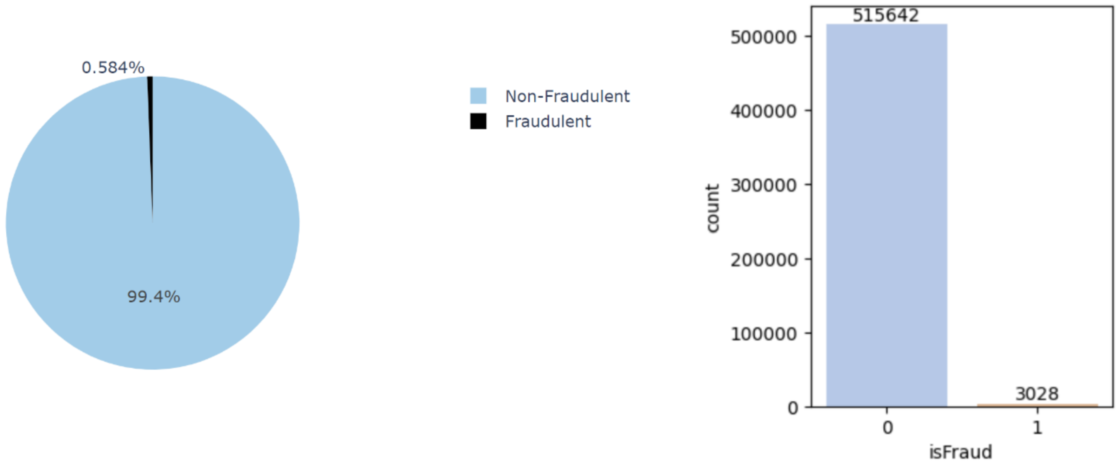


Figure 6: Class Distribution

This imbalance poses challenges for machine learning algorithms, as they may become biased toward the majority class and struggle to effectively detect the minority class. To address this issue, techniques like data resampling, cost-sensitive learning, and using appropriate evaluation metrics are essential for building a robust fraud detection model. The class imbalance problem will be handled in section 3.5. to mitigate the impact it might have on the model training and results.

3.4 Data Exploration

As part of the EDA, we examined the relationship between the target and the independent variables. This exploration was to identify the relevant features and provide insight into hidden patterns in the data.

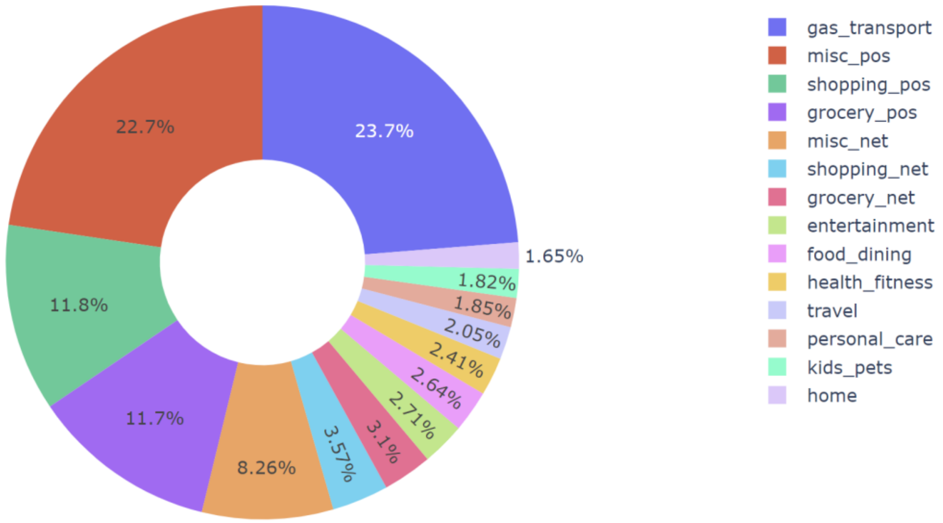


Figure 2: distribution of fraudulent transactions by shopping category

Figure 2. above shows the distribution of fraudulent transactions by shopping category, most of the transactions (10.2%) were from customers purchasing gas for their vehicles, and only 3.2% was spent on the home category.

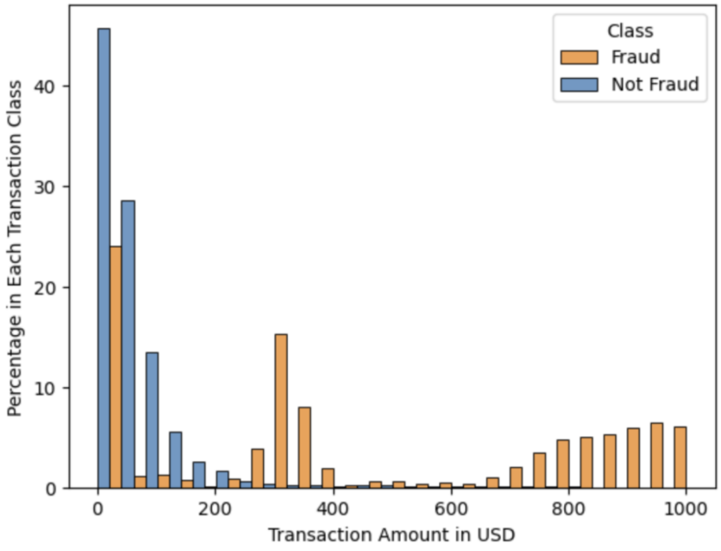


Figure 3: distribution of fraudulent transactions by amount.

The relationship between the target feature and the transaction amount was established in figure 3, showing notable peaks in fraudulent transactions of amounts $200 or lower.

Figure 4 shows the relationship between the transaction fraud and the day of the week the transactions were carried out, from the image, it can be seen that most frauds were committed during the weekend days.

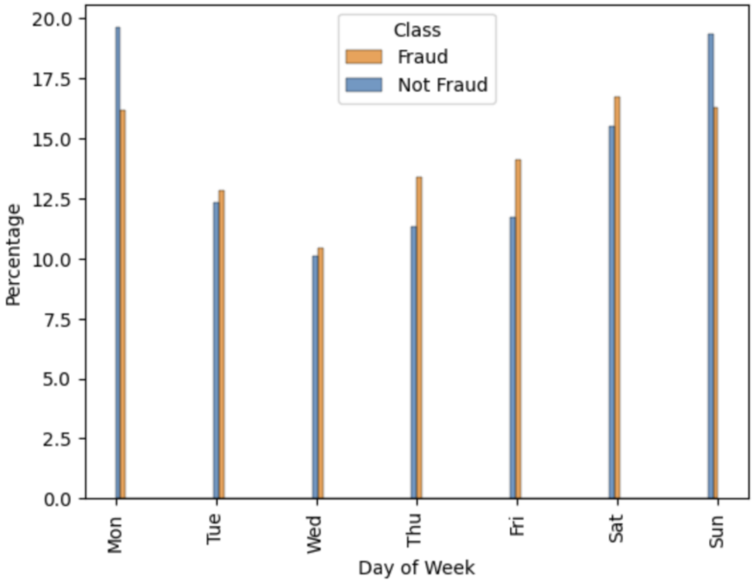


Figure 4: Distribution of fraud by day of week.

Also, the month which the transactions were carried out were examined for fraudulent patterns and figure 5 depicts the relationship, most transactions during the Christmas season appear to be legitimate, while transactions done in late spring to early summer had more frauds.

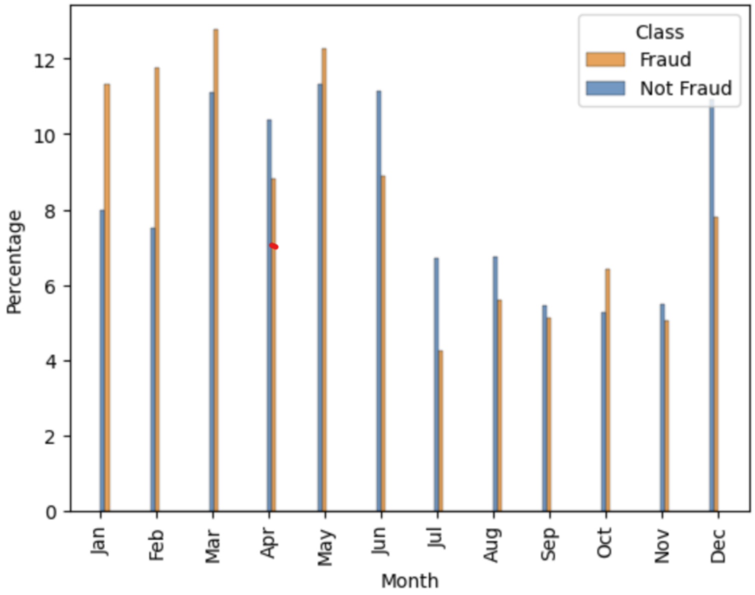


Figure 5: distribution of fraud by month

3.5 Handling Class Imbalance

Imbalanced data poses challenges in evaluating machine learning models, as using accuracy as the sole evaluation metric can be misleading. In the case of imbalanced classes, the model may achieve high accuracy by correctly classifying the majority class but perform poorly on the minority class, which is of greater importance in payment card fraud detection. Misclassifying the minority class can have severe consequences, making it crucial to focus on metrics like precision, recall, precision-recall curve, or F1 score that prioritize correctly categorizing the minority class and minimize false negatives (Cicak & Avci, 2023). According to a study by (Han, et al., 2022) dealing with imbalanced data creates a problem where most machine learning methods tend to overlook the minority class, resulting in inadequate learning of classification boundaries. To overcome this issue, the SMOTE-ENN technique developed by (Batista, et al., 2004) combines over- and under-sampling approaches, utilizing SMOTE and ENN, to establish dataset equilibrium. For this research, the balancing technique SMOTE-ENN was used, it is a hybrid method utilized in addressing imbalanced datasets within machine learning. It encompasses two steps: first, it introduces synthetic samples to the underrepresented class (via SMOTE), and second, it refines the dataset by eliminating instances through a k-nearest neighbor classifier (ENN). This combined approach is particularly valuable in scenarios like credit card fraud detection, where class imbalances are prevalent. The goal of SMOTE-ENN is to create a more equitable dataset, enhancing the model's capability to identify the minority class while mitigating the influence of noisy data.

3.6 Data Modelling

At the core of machine learning are the various algorithms that can be used to train the data and produce results that provide insights into the data and help with data analysis. These models are used for solving different kinds of problems in different domains. In this research, we employ six machine learning algorithms to analyze the payment card data.

3.6.1 Logistic Regression

Logistic regression is a popular supervised learning algorithm used in machine learning. It predicts dependent variables based on independent features and is particularly effective for categorical outcomes. By establishing probability factors, logistic regression efficiently categorizes new datasets, making it valuable for solving classification problems (Raju, et al., 2023). This estimation aids decision-making processes. Notably, the outcome of this model is a probability, always ranging between 0 and 1, irrespective of the context. Through a logit transformation, these probabilities are computed by dividing the probability of success by the probability of failure, resulting in a single numerical value. This concept is commonly referred to as the log odds, which is essentially the natural logarithm of odds (Prakash, et al., 2023). This logarithmic transformation enhances the predictive capabilities of the model, making it a valuable tool for assessing the likelihood of credit card fraud.

3.6.2 Random Forest

The inherent limitations of decision trees, such as vulnerability to overfitting and sensitivity to specific data patterns, can be effectively addressed through ensemble methods. These methods combine predictions from multiple trees, yielding more accurate results compared to a single tree. Among these, the random forest model stands out as a potent ensemble technique rooted in decision trees and bagging, which involves training multiple decision trees on distinct subsets of data obtained through bootstrapping, followed by aggregation. Each tree operates on randomized data, evaluating case proximity, and this diversity lends uniqueness to each tree in the forest, while maintaining consistent distribution.

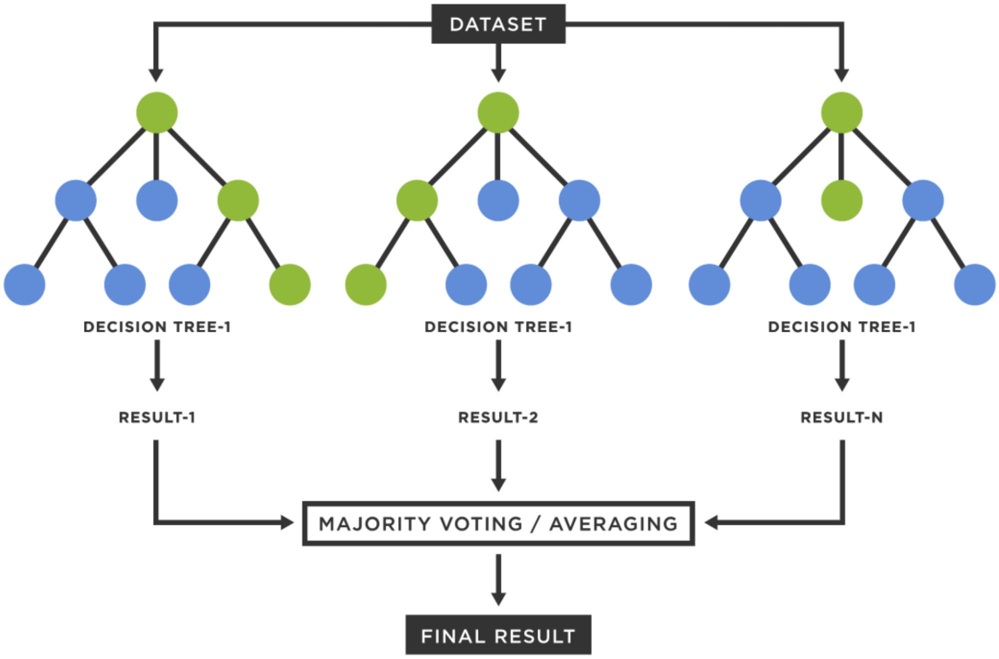


Figure 8: Random Forest Classifier

Random forests offer notable benefits such as exhibiting robust generalization by amalgamating decisions from diverse trees and mitigating overfitting issues associated with complex decision trees. Their independent construction makes them computationally efficient and resistant to outliers. In contemporary applications like fraud detection, where accuracy is paramount, random forests are widely favored due to their ease of use and high-performance outcomes (Chang, et al., 2022)..

3.6.3 Light Gradient Boosting Machine

LightGBM has become increasingly popular in various data analysis fields, from tackling health related issues (Fang, et al., 2023) to environmental and weather analysis (Fan, et al., 2019). It is an innovative iteration of Gradient Boosting Decision Tree introduced by Microsoft in 2017 and was designed to address the complexities of learning decision trees when working with extensive datasets and high feature dimensions.

Revamping the conventional GBDT approach, LightGBM introduces novel strategies to address the computational burden posed by large datasets and numerous features. Unlike traditional GBDT models that evaluate information gain across all potential split points for each feature, LightGBM employs advanced techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to intelligently reduce the computational load. This results in enhanced efficiency and improved performance, making LightGBM a compelling choice for scenarios involving extensive data and high-dimensional features (Zhang, et al., 2022).

3.6.4 Extreme Gradient Boost Algorithm

XGBoost is another boosting algorithm that stands out as a highly efficient and scalable rendition of gradient boosted decision trees. This technique constructs additive models through a sequential process, continually minimizing overall error by introducing models based on the previous step's errors. This culminates in an ensemble of base learners that collectively outperform individual classifiers, thanks to their gradual refinement, shallow tree depths, and uniform contributions. To fortify resilience against noise and overfitting, the gradient boosting method was enriched with a stochastic sampling scheme. XGBoost goes a step further with a well-regulated model, enhancing control over overfitting tendencies (Hajek, et al., 2022).

3.6.5 Multi-Layer Perceptron

The multilayer perceptron (MLP) is an artificial neural network structure utilized in supervised learning. It constructs a model f(.): Rm –> Ro, where 'i' represents input dimension and 'o' signifies output dimension. Comprising an input layer for signal reception, a hidden layer for computational processing, and an output layer for predictions, the MLP operates in a stepwise manner. Remarkably, even an MLP with a single hidden layer can effectively approximate continuous functions, highlighting its flexibility. User inputs undergo multiplication with weights, coupled with bias addition—a shared aspect among hidden units. This outcome then enters an activation function within each hidden layer, yielding a sequential outcome. Consequently, the MLP excels in mastering non-linear models (Tekkali & Vijaya, 2021).

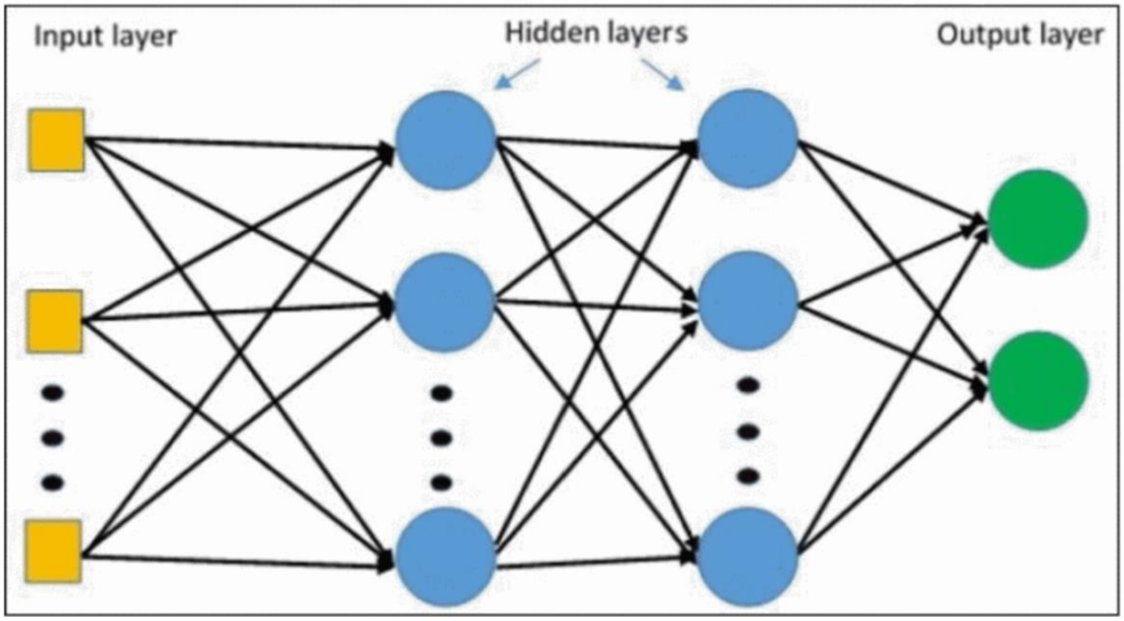


Figure 9: Multilayer Perceptron-ANN

The MLP's training is grounded in input-output dimensions, facilitating the modeling of correlations and refinement of model weights and biases for error minimization. This process hinges on the backpropagation algorithm, which orchestrates adjustments in weights and biases to enhance the model's accuracy and performance.

3.6.6 Long-Short-Term-Memory-Recurrent Neural Network

LSTM, an enhanced version of RNNs, is pivotal in deep learning-based intrusion detection models. It excels in prediction due to its memory-based efficiency and connection between input parameters and output predictions. LSTM is adept at learning long-term dependencies in sequence predictions, particularly valuable for extended patterns. This design is crucial for effectively capturing and predicting intricate temporal patterns, making it ideal for tasks like credit card fraud prediction.

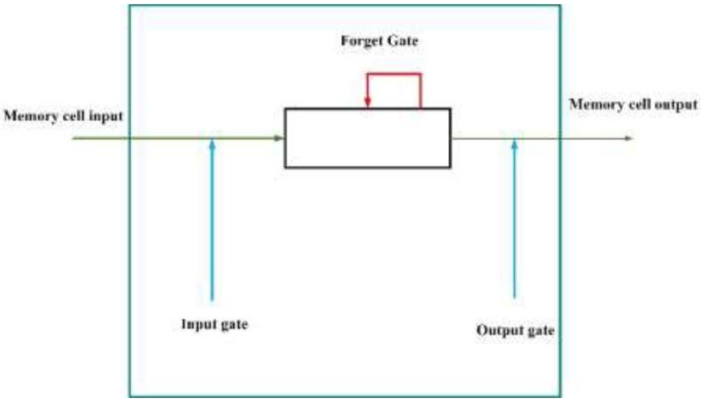


Figure 9: LSTM

Long Short-Term Memory (LSTM) addresses the shortcomings of ordinary Recurrent Neural Networks (RNNs), which suffer from gradient vanishing or explosion issues. LSTM introduces a memory cell that preserves its state over time. Filtering mechanisms regulate data flow to and from the memory cell. An input gate controls signal adjustment to the cell state, while a base station governs the influence of the cell state on hidden layer neurons. A forget gate aids in recalling prior states. The importance of each component lies in reducing uncertainty. Key determinants of LSTM's output quality include hidden layer neuron sum, training algorithm, inner transfer functions, and pass rate (Femila, et al., 2022).

4.0 Design Specification

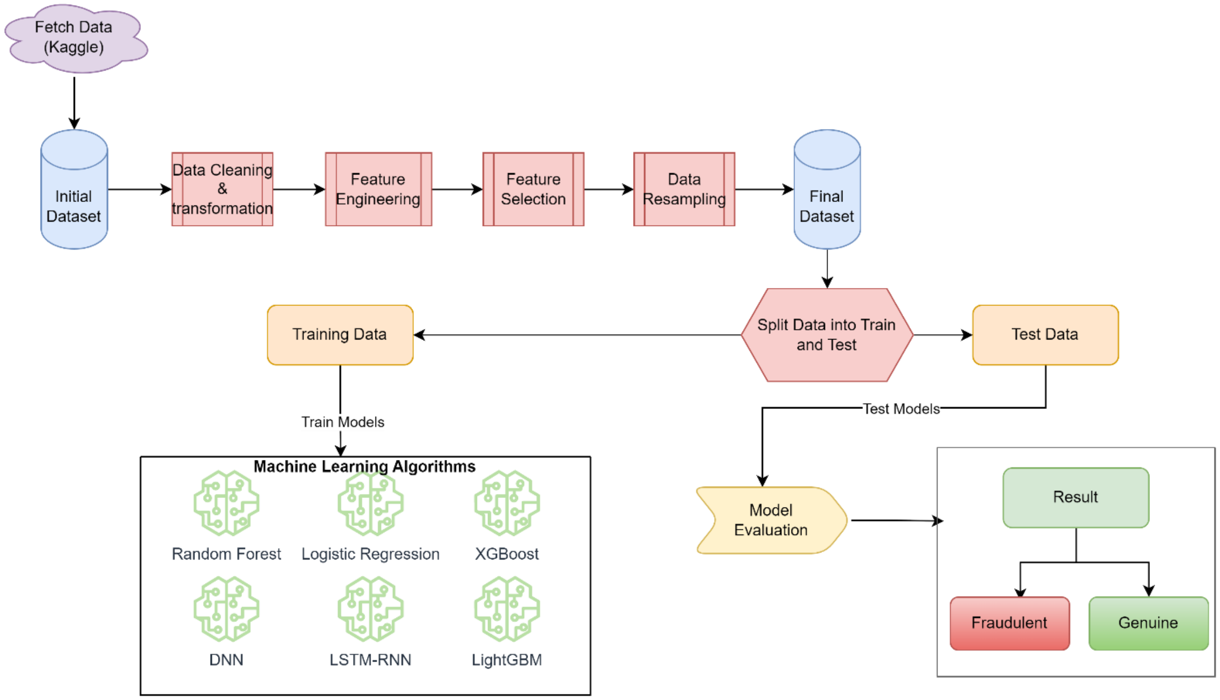


Figure 10: System Process Flow

For this research, Python was selected as the most suitable programming language due to its powerful capabilities in data science and machine learning, as well as its user-friendly syntax. An online IDE (Google Colab[2]) was utilized to execute the Python code, facilitating the research process.

The system architecture, as depicted in Figure 10, outlines the sequential steps undertaken in this study. Initially, the dataset was obtained from the Kaggle repository in CSV format. However, due to its large size (over 7 million records) and limited processing resources, only a 40% sample of the data was used. This reduced dataset was saved to an Excel file after ensuring it was stratified for balanced representation.

Subsequently, the data was imported into a pandas dataframe for further analysis and processing. This approach allowed for efficient manipulation and exploration of the dataset, enabling various machine learning algorithms to be applied for payment card fraud prediction. Following the implementation of classifiers, their performance is assessed using various evaluation metrics derived from the classification reports. These metrics allow for a comprehensive analysis of the models' predictive capabilities and their ability to distinguish between fraud and non-fraud instances. The results of the models, along with the research findings, are effectively communicated through visualizations. These visual representations help in providing clear

[2] https://colab.research.google.com/#

insights, making it easier to interpret and communicate the outcomes of the credit card fraud prediction study to stakeholders and decision-makers.

5.0 Implementation

This section focuses on the final implementation of the machine learning algorithm that were used in this research, which were discussed in section 3.4. The dataset used to train the models was sourced from Kaggle and contained 1,296,675 million records, of with a stratified sample of the data was selected to train the models, the final dataset used consisted of approximately 500,000 records. To address the class imbalance problem, the dataset was subjected to SMOTE-ENN discussed in section 3.3.2. to help the performance of the models.

The `scikit-learn` package was installed for the class balancing libraries, also for the models, `lightgbm` needed to be explicitly installed in Google Colab in order to import it into the code using Python pip package installer. The dataset was split before the model training could commence and each of the models was trained on the training data and subsequently tested on the test data. Several model evaluation tools were imported from the `sklearn.metrics` such as `confusion_matrix, ConfusionMatrixDisplay, roc_auc_score.` These tools were used in assessing the performance of the six classifiers after their training.

To enhance the predictive efficacy of the classification models, hyperparameter optimization was executed through Randomized Search. This involved identifying optimal hyperparameter values that maximize model performance using random combinations from the predefined search space. The Randomized Search was applied to some of the models utilizing the `sklearn.model_selection` library's RandomizedSearchCV() function. This approach was favored over the grid search technique, which tends to be computationally demanding for extensive datasets, also, the hyperparameter optimization was conducted using 3-fold cross-validation.

S/N	Model Name	Hyperparameters
1	Logistic Regression	`'solver': 'liblinear', 'penalty': 'l2', 'max_iter': 20, 'class_weight': None, 'C': 0.1`
2	Random Forest	Default Parameters
3	LightGBM	`'n_estimators': 500, 'max_samples': 1.0, 'max_features': 1.0, learning_rate: 0.2, subsample: 0.5, num_leaves: 4272, colsample_by_tree: 1`
4	XGBoost	`n_estimators=500, max_depth=6, learning_rate=0.3, subsample=0.75, min_child_weight=1, colsample_bytree=0.5, gamma=0.2`
5	MLP	Default Parameters

6	LSTM	Default Parameters

Table 2: Hyperparameter for Classifiers

6.0 Evaluation

The objective of this study was to compare the effectiveness of different machine learning algorithms in predicting fraudulent payment card transactions. To achieve this, we trained and assessed six classifiers using a synthetic dataset. While we balanced the training data using a modified SMOTE technique, the test dataset was deliberately kept highly imbalanced to mimic real-world transaction scenarios.

To evaluate the models' performance, we employed metrics beyond accuracy, as the imbalanced dataset makes accuracy less informative and for all the classifiers trained, the accuracy score was above 902%. Accuracy tends to favor the majority class, which doesn't provide a comprehensive assessment. Instead, we utilized a range of metrics including Precision, Recall, F1-Score, PR-AUC (Precision-Recall Area Under Curve), Geometric Mean, ands MCC (Matthews Correlation Coefficient) to evaluate their performance. The outcomes of these evaluations are presented in table 3 below.

Model Names	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	Geometric Mean (%)	MCC	PR-AUC (%)
Logistic Regression	92.6	75.74	5.66	10.54	84.16	0.19	19.50
Random Forest	99.4	82.5	51	63	91	0.65	75
LightGBM	99.4	84.2	50	63	92	0.65	80
XGBoost	99.36	84.3	50.1	63	92	0.65	81.6
MLP	95.2	82.1	9	16	88.7	0.27	25
LSTM	98.9	6.9	7.1	7	6.4	0.53	2.6

Table 3: Model Evaluation Results

6.1 Logistic Regression Evaluation

The logistic regression model demonstrated a recall of 0.76, indicating that it correctly identified about 76% of actual fraudulent transactions. However, the precision was notably low at 0.06, implying that only 6% of the predicted fraud cases were accurate. The F1 score, which balances the trade-off between precision and recall, was 0.11, indicating room for improvement in achieving a better balance between these metrics. The Matthews Correlation Coefficient (MCC) of 0.19 suggested a moderate overall performance of the model, while the Geometric Mean of 0.84 indicated a reasonable balance between sensitivity and specificity. The Precision-Recall Area Under Curve (PR AUC) score was 0.195, indicating that the model's ability to distinguish between the two classes was grossly suboptimal. Figure 11 below depicts a graphical representation of the model’s performance, consisting of (i) PR-AUC (ii) Confusion Matrix and (iii) feature importance chart.

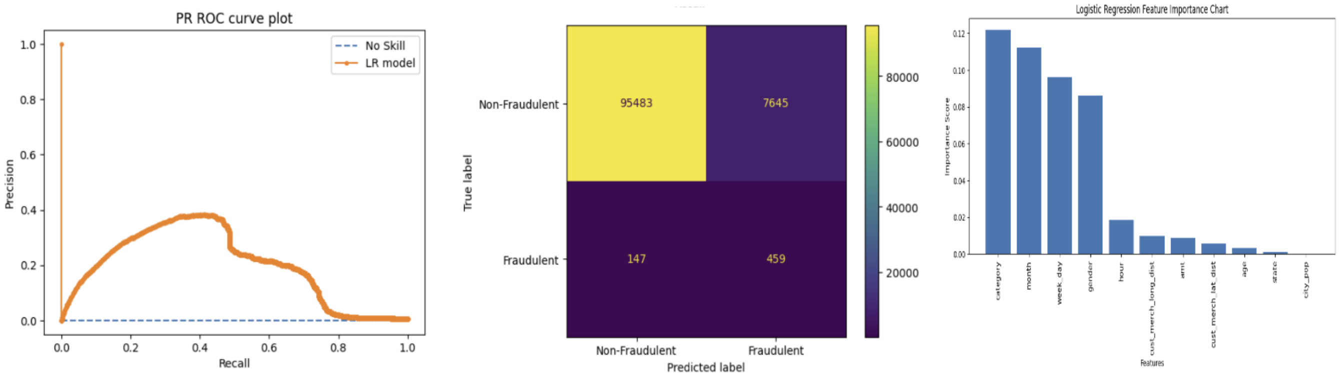


Figure 11: Logistic Regression Model Evaluation

In summary, while the logistic regression model showed a decent ability to identify fraudulent transactions based on recall, there is significant potential for improvement in achieving a higher precision and a better balance between various performance metrics.

6.2 Random Forest Evaluation

The Random Forest (RF) model exhibited a relatively high recall of 0.83, indicating its capability to correctly identify a substantial portion of actual fraudulent transactions. The precision of 0.51 suggests that over half of the predicted fraud cases were accurate, contributing to a better balance between precision and recall, as reflected in the F1 score of 0.63. The Matthews Correlation Coefficient (MCC) of 0.65 signifies a robust overall performance of the RF model. The Geometric Mean of 0.91 showcases a commendable equilibrium between sensitivity and specificity. The Precision-Recall Area Under Curve (PR AUC) score of 0.748 indicates that the model’s ability to distinguish between the two classes is relatively effective. Figure 12 shows a graphical representation of the model’s performance, consisting of (i) PR-AUC (ii) Confusion Matrix and (iii) feature importance chart.

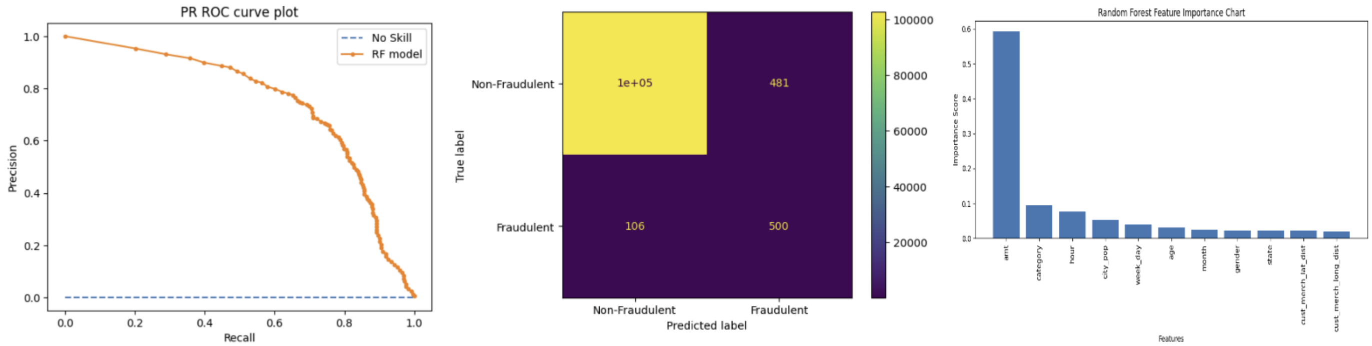


Figure 12: Random Forest Model Evaluation

The model demonstrates a balanced performance, with solid recall and precision values, a high F1 score, and favorable MCC and Geometric Mean scores. The model's ability to differentiate between classes is further supported by its respectable PR AUC score.

6.3 Light-GBM Model Evaluation

The LightGBM model demonstrates a recall of 0.84, slightly higher than that of the random forest model and indicates its ability to identify a substantial proportion of actual fraudulent transactions. The precision of 0.50 suggests that approximately half of the predicted fraud cases were accurate (like predictions based on a coin toss), resulting in a balanced trade-off between precision and recall, as evidenced by the F1 score of 0.63. The Matthews Correlation Coefficient (MCC) of 0.65, similar to that of the RF model, reflects the overall performance of the LightGBM model. while the high Geometric Mean of 0.92 showcases a balanced combination of sensitivity (correctly predicted positives) and specificity (correctly predicted negatives). The Precision-Recall Area Under Curve (PR AUC) score of 0.80 highlights the model's effectiveness in distinguishing between the two classes. According to the feature importance chart, the ‘amt’ variable seemed to be the most significant in predicting fraudulent transactions. Figure 13 shows some visualizations for Light-GBM model results.

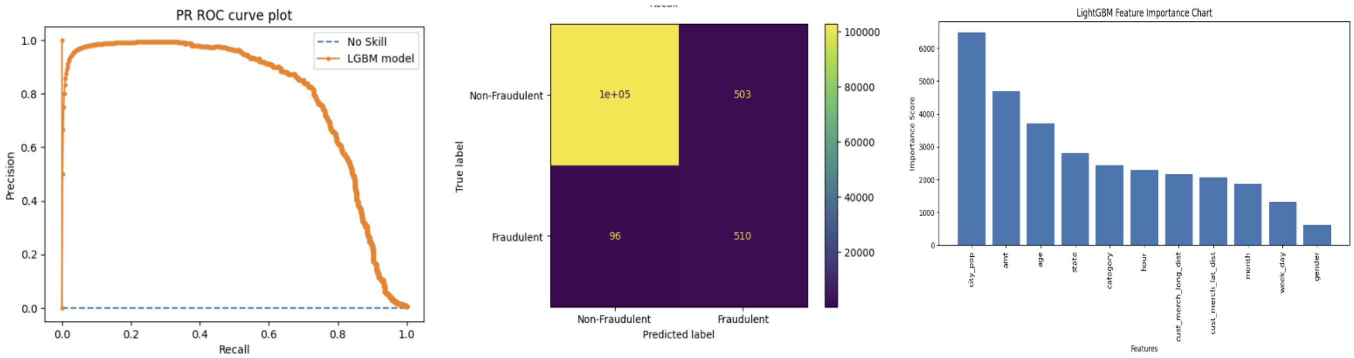


Figure 13: Light-GBM Model Evaluation

The model demonstrates balanced performance, with notable recall and precision values, a satisfactory F1 score, and favorable MCC and Geometric Mean scores. Its ability to differentiate between classes is further supported by the impressive PR AUC score.

6.4 XGBoost Model Evaluation

In comparison to the LightGBM model, the XGBoost model demonstrates a slightly higher Recall of 0.84, indicating its ability to capture more true positive instances. The Precision of 0.50 suggests a similar trade-off between true positive and false positive predictions. The F1 Score of 0.63 reflects a comparable balance between precision and recall performance. The Matthews Correlation Coefficient (MCC) of 0.65 is consistent with the LightGBM model, implying similar overall performance. The Geometric Mean of 0.92 still showcases a balanced combination of sensitivity and specificity, underlining its robustness.

Moreover, the PR AUC of 0.82 is slightly higher than that of the LightGBM model, indicating the XGBoost model's improved ability to rank positive instances. This suggests that the

XGBoost model performs on par with or slightly better than the LightGBM model across various evaluation metrics.

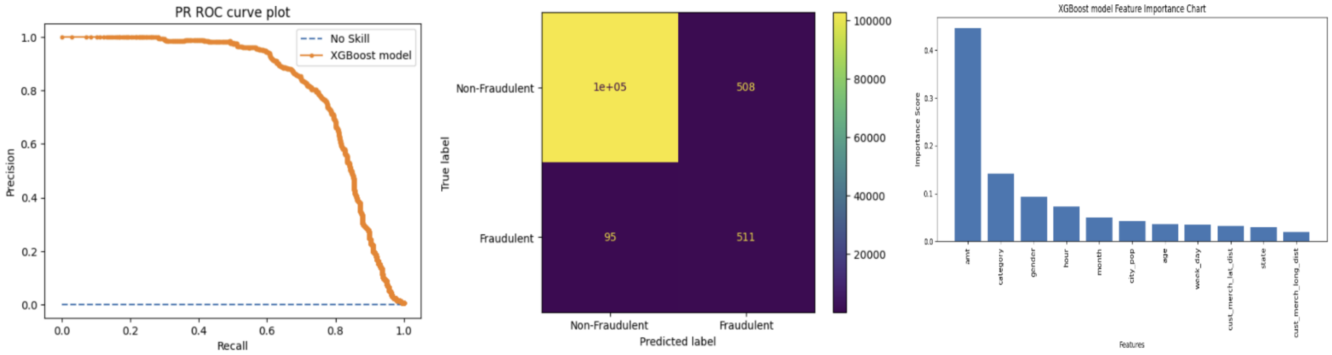


Figure 14: XGBoost Model Evaluation

Figure 14 depicts the PR-AUC curve, confusion matrix and feature importance chart for the model. Both LightGBM and XGBoost models performed better than the other models used, with Random Forest coming in behind them.

6.5 Multilayer Perceptron Model Evaluation

The deep learning classifiers did poorly compared to the supervised learning algorithms, the first deep learning model trained was the Multilayer Perceptron (MLP) model demonstrates a recall of approximately 0.82, signifying its ability to correctly identify around 82% of actual positive cases. However, its precision is notably low at about 0.09, indicating that only about 9% of the predicted positive cases are actual positives. This results in an F1 score of around 0.17, reflecting a relatively low harmonization of precision and recall.

The Matthews Correlation Coefficient (MCC) is approximately 0.27, which suggests a moderate level of agreement between predicted and actual classifications. The geometric mean is around 0.89, indicating a moderate equilibrium between sensitivity and specificity. The PR AUC score stands at about 0.25, indicating a relatively low capacity to strike a balance between precision and recall in the precision-recall trade-off, this can be seen in the PR-AUC curve in figure 15.

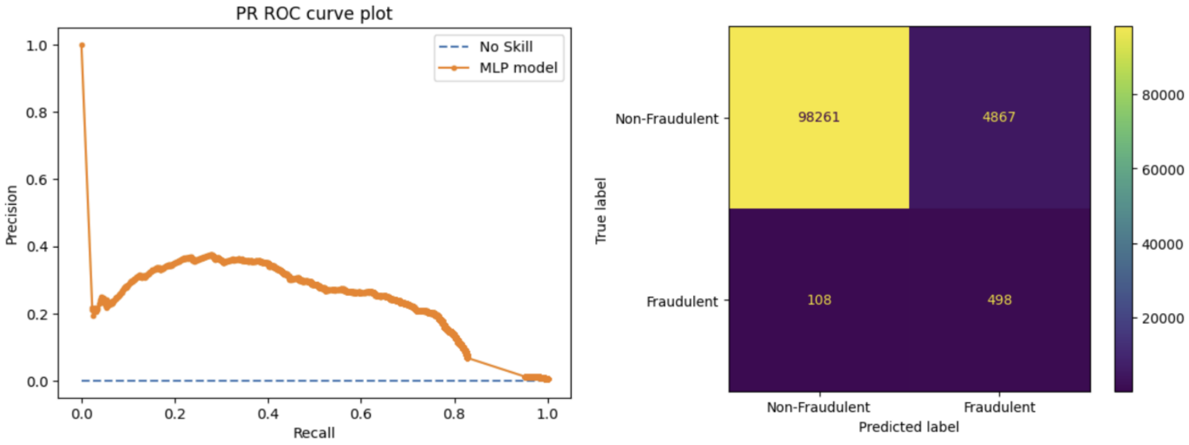


Figure 15: MLP Model Evaluation

The model exhibits a somewhat imbalanced performance, with notable recall and less remarkable precision values, and a relatively lower MCC and Geometric Mean scores. Its capacity to differentiate between classes is not as strong, which is reflected in the PR AUC score.

6.6 Multilayer Perceptron Model Evaluation

The LSTM model's performance, as indicated by the performance metrics, was relatively low, the recall, precision, and F1-score values are all around 0.07, suggesting that the model struggles to accurately identify positive instances. The Matthews Correlation Coefficient (MCC) of 0.065 further emphasizes this low level of agreement between the LSTM's predictions and the actual outcomes. Additionally, the Geometric Mean is approximately 0.53, which indicates that the balance between sensitivity and specificity is compromised in the LSTM model's predictions. The PR AUC score of 0.026 highlights that the model's ability to rank positive instances higher than negative ones is limited. Figure 16 shows the visual representation of the model’s evaluation results.

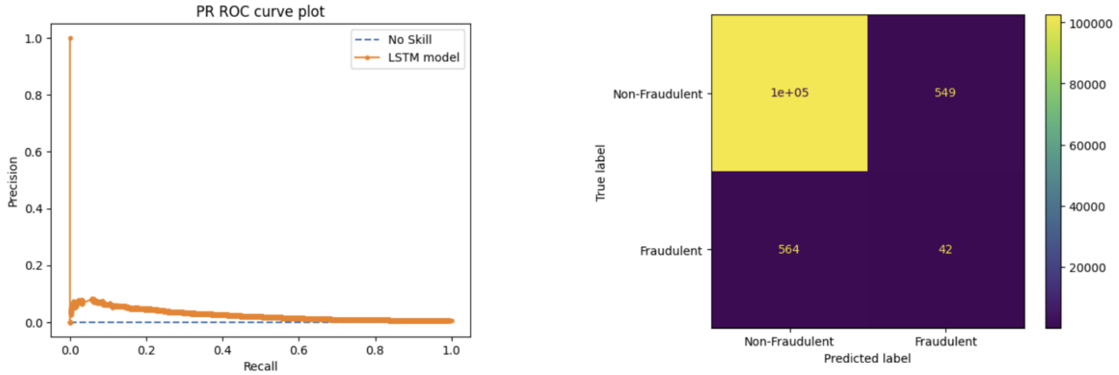


Figure 16: LSTM Model Evaluation

Overall, these results suggest that the LSTM model may not be performing as well as the other classifiers, such as LightGBM and XGBoost, in predicting payment card fraud based on the provided dataset.

6.7 Discussion

Analyzing the results presented in the performance metrics of various classifiers provides valuable insights into their effectiveness for credit card fraud detection. Among the evaluated classifiers, boosting methods, such as LightGBM, XGBoost, and Random Forest, outperformed other techniques, showcasing their potential to effectively identify fraudulent transactions. These models demonstrated notable recall rates, ranging from 0.821 to 0.844, indicating their capacity to correctly classify a substantial portion of actual fraudulent cases. This is a critical aspect for fraud detection, as overlooking fraudulent transactions can have significant financial implications.

While recall is a vital metric for capturing fraudulent transactions, it's also important to consider other factors, such as precision. Precision reflects the accuracy of positive predictions, and in this context, Random Forest exhibited the highest precision at around 0.51. This suggests that when Random Forest identified a transaction as fraudulent, it was more likely to be accurate compared to other models.

The F1 Score, which harmonizes precision and recall, provides an overall measure of a model's prediction accuracy. It's noteworthy that ball three boosting algorithms achieved F1 Scores of 0.63, indicating a good balance between precision and recall, and highlighting their potential for real-world application. Matthews Correlation Coefficient (MCC) offers an all-encompassing assessment of a classifier's performance, considering true positives, true negatives, false positives, and false negatives. Among the classifiers, LightGBM and XGBoost demonstrated MCC values around 0.648 and 0.65, respectively, indicating a well-rounded evaluation of their abilities.

The Geometric Mean (G-Mean) further showcases the equilibrium between sensitivity and specificity. Boosting models, particularly LightGBM and XGBoost, displayed higher G-Mean scores, signifying their capacity to maintain a balance between correctly identifying fraudulent transactions while minimizing false positives. The Precision-Recall Area Under the Curve (PR AUC) metric emphasizes a model's ability to differentiate between the positive and negative classes in highly imbalanced datasets. Here, XGBoost stood out with a PR AUC of 0.82, highlighting its capability to make precise predictions, even in cases where true fraudulent instances are scarce.

However, it's important to note that while boosting models excelled across multiple metrics, the LSTM model's performance remained lower. Its recall rate was relatively poor, and both its precision and F1 Score were significantly lower than those of other models, indicating challenges in effectively identifying fraudulent transactions. The ensemble-based boosting classifiers, particularly LightGBM and XGBoost, emerged as robust performers for credit card fraud detection in this study. These models showcased commendable recall rates, well-balanced precision and recall trade-offs, and a comprehensive evaluation of their predictive performance. The choice of the optimal classifier would depend on the specific trade-offs preferred in a real-world application scenario.

One major limitation to this research was the availability of real-world payment card data, the data used was generated by a simulator and therefore, may potentially limit the model's ability to generalize across different timeframes or geographical locations.

7.0 Conclusion

In conclusion, this study aimed to address the question of how effectively various machine learning models can detect payment card fraud. To achieve this, a series of research objectives were pursued, including the assessment and comparison of model efficiency, the implementation of a sampling strategy to counter class imbalance, the selection of pertinent performance metrics, and the identification of key predictors for credit card fraud prediction. Using feature engineering, new features were created, and SMOTE-ENN resampling technique was employed to handle the class imbalance.

Based on the model evaluation results, it is evident that boosting classifiers, especially LightGBM and XGBoost, exhibited remarkable performance in detecting credit card fraud. These models displayed impressive precision, with LightGBM achieving a precision of 0.503 and XGBoost achieving a precision of 0.51. Furthermore, their recall values were notably high, with LightGBM attaining a recall of 0.842 and XGBoost achieving a recall of 0.843. The F1-scores for these models were also strong, with LightGBM and XGBoost achieving an F1-score of 0.630. Additionally, their Matthews Correlation Coefficients (MCC) were substantial, with both models achieving 0.65. Furthermore, their area under the precision-recall curve (AUC-PR) scores were noteworthy, with LightGBM achieving an AUC-PR of 0.80 and XGBoost achieving an AUC-PR of 0.816. Collectively, these metrics underscore the effectiveness of LightGBM and XGBoost in achieving a harmonious balance between identifying legitimate and fraudulent transactions.

For future works, advanced sampling techniques such as Adaptive Synthetic Sampling (ADASYN) or Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC), could be explored to further address class imbalance, enhancing model performance on underrepresented instances, also refining feature engineering approaches is essential to extract more discriminative patterns from complex transactional data, potentially boosting predictive accuracy, especially for the deep learning models like RNN and ANN. Validating model performance on external datasets ensures generalizability, critical for real-world applications and a deeper dive into optimizing hyperparameters is vital to fine-tune models for optimal results. These avenues collectively aim to refine and elevate the effectiveness of credit card fraud detection techniques, contributing to enhanced security in financial transactions.

References

Alarfaj, F. K. et al., 2022. Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms. *IEEE Access,* Volume 10, pp. 39700-39715.

Assaghir, Z. et al., 2019. An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. *IEEE Access,* Volume 7, pp. 93010-93022.

Batista, G. E. A. P. A., Prati, R. C. & Monard, M. C., 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter,* 6(1), pp. 20-29.

Chang, V. et al., 2022. Digital payment fraud detection methods in digital ages and Industry 4.0. *Computers and Electrical Engineering,* 100(107734).

Cicak, S. & Avci, U., 2023. *Handling Imbalanced Data in Predictive Maintenance: A Resampling-Based Approach.* s.l., s.n., pp. 1-6.

Fang, M. et al., 2023. A hybrid machine learning approach for hypertension risk prediction. *Neural Comput & Applications,* 35(20), p. 14487–14497.

Fan, J. et al., 2019. Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural water management,* 225(105758).

Femila, J. R. et al., 2022. Autonomous credit card fraud detection using machine learning approach. *Computers and Electrical Engineering,* 102(108132).

Forough, J. & Momtazi, S., 2022. Sequential credit card fraud detection: A joint deep neural network and probabilistic graphical model approach. *Expert Systems,* 39(1), pp. 1-13..

Gupta, P. et al., 2023. *Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques.* s.l., s.n., pp. 2575-2584.

Hajek, P., Abedin, M. Z. & Sivarajah, U., 2022. Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework. *Information Systems Frontiers.*

Han, L. et al., 2022. An explainable XGBoost model improved by SMOTE-ENN technique for maize lodging detection based on multi-source unmanned aerial vehicle images. *Computers and Electronics in Agriculture,* 194(106804).

Prajapati, D., A. Mehta, T. J., Jhaveri, K. & Kelkar, V., 2021. *Credit Card Fraud Detection Using Machine Learning.* Mumbai, s.n., pp. 1-6.

Prakash, B. S., Amudha, V. & Meenakshisundaram, N., 2023. *Efficient Human Action Recognition using Novel Logistic Regression Compared over Linear Regression with Improved Accuracy.* Chennai, s.n., pp. 1-6.

Raju, G. C., Amudha, V. & Sajiv, G., 2023. *Comparison of Linear Regression and Logistic Regression Algorithms for Ground Water Level Detection with Improved Accuracy.* Chennai, s.n., pp. 1-6.

Singhai, A., Aanjankumar, S. & Poonkuntran, S., 2023. *A Novel Methodology for Credit Card Fraud Detection using KNN Dependent Machine Learning Methodology.* Salem, s.n., pp. 878-884.

Tekkali, C. G. & Vijaya, J., 2021. *A Survey: Methodologies used for Fraud Detection in Digital Transactions.* Coimbatore, s.n., pp. 1758-1765.

Tran, T. C. & Dang, T. K., 2021. *Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection.* Seoul, s.n., pp. 1-7.

Ugarković, A. & Oreški, D., 2022. *Supervised and Unsupervised Machine Learning Approaches on Class Imbalanced Data.* Osijek, s.n., pp. 159-162.

Vejalla, I., Battula, S. P., Kalluri, K. & Kalluri, H. K., 2023. *Credit Card Fraud Detection Using Machine Learning Techniques.* s.l., s.n., pp. 1-4.

Yakshit, et al., 2022. *Analyzing various Machine Learning Algorithms with SMOTE and ADASYN for Image Classification having Imbalanced Data.* Bhopal, s.n., pp. 1-7.

Zhang, W. et al., 2022. Efficient time-variant reliability analysis of Bazimen landslide in the Three Gorges Reservoir Area using XGBoost and LightGBM algorithms. *Gondwana Research.*