Fraud Transaction Detection Using Machine Learning

A report submitted to the

School of Technology, Woxsen University, Kamkole

for the successful completion

of

Minor Project - 1

in

B.Tech Program Name

by

Abhigna Ragala	21WU0102051
Abhijeeth Ragala	21WU0102050
Nihitha Vadlamuri	21WU0102049

Supervised by

Prof. Amogh Deshmukh

Designation

School of Technology

School of Technology

Woxsen University, Kamkole - 502345, Sangareddy District,

Telangana, India

July 2023

CERTIFICATE

This is to certify that the report entitled " **Fraud Transaction Detection Using Machine Learning**", submitted by **Abhigna Ragala (21WU0102051), Abhijeeth Ragala(21WU0102050), Nihitha Vadlamuri (21WU0102049)** to the School of Technology, Woxsen University, for the successful completion of the minor project - 1, is a record of bonafide research work carried out by them under my supervision and guidance. To the best of my knowledge, the work embodied in this Project have not been submitted to any other university or institute for the award of any other degree or diploma.

Dt: Supervisor Name

Place: Designation

DECLARATION

We hereby declare that the report entitled "**Fraud Transaction Detection Using Machine Learning**", submitted to the School of Technology, Woxsen University, in partial fulfilment of the requirements for the project is the original and independent work carried out by us, under the supervision of in the School of Technology, Woxsen University, Hyderabad. This Project has not formed the basis for the award of any Projects/Degree /Diploma /Fellowship /Associateship of this University or any other institution.

Name of the Students Signature

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who have supported us in the successful completion of this project on "Fraud Transaction Detection Using Machine Learning."

First and foremost, we are deeply grateful to our mentor, whose invaluable guidance, constructive feedback, and constant encouragement have been instrumental in shaping this project. Their expertise and insights have greatly enhanced our understanding and application of machine learning techniques in detecting fraudulent transactions.

We also extend our heartfelt thanks to our professors and mentors at Woxsen University for their continuous support and for providing us with the necessary resources and environment to carry out this research.

Special thanks to our peers and colleagues, whose collaboration and exchange of ideas have been pivotal in refining our project. Their critical assessments and suggestions have been immensely helpful.

We are also indebted to the various authors and researchers whose works we have referenced throughout this report. Their contributions to the field of machine learning and fraud detection have laid the groundwork for our project.

List of Abbreviations

1. **PCA**: Principal Component Analysis
2. **ROC Curve**: Receiver Operating Characteristic Curve
3. **AUC**: Area Under the Curve
4. **RUS**: Random Under Sampling
5. **SMOTE**: Synthetic Minority Over-sampling Technique
6. **ESOA**: Egret Swarm Optimization Algorithm.
7. **SVM**: Support Vector Machine
8. **XGBoost**: eXtreme Gradient Boosting
9. **CATBoost**: Categorical Boosting
10. **NB**: Naive Bayes
11. **LR**: Logistic Regression.
12. **TP**: True Positive
13. **TN**: True Negative
14. **FP**: False Positive
15. **FN**: False Negative
16. **Recall**: Recall (also known as Sensitivity or True Positive Rate)
17. **F1 Score**: F1 Score (harmonic mean of Precision and Recall).
18. **KNN**: K-Nearest Neighbors
19. **LDA**: Linear Discriminant Analysis
20. **RF**: Random Forest

Abstract

Accurate identification of fraudulent credit card transactions is crucial for safeguarding customers from unauthorized charges. Data Science, and specifically Machine Learning, play an indispensable role in addressing this challenge. This project demonstrates the application of machine learning to model a credit card fraud detection dataset. The task involves constructing a model based on historical fraudulent transactions to classify new transactions as fraudulent or legitimate. Our goal is to achieve a high fraud detection rate while minimizing false positives.

Credit card fraud detection is a classic example of a classification problem. In this endeavor, we explore data analysis, preprocessing, and the implementation of anomaly detection algorithms on PCA-transformed credit card transaction data. We utilize various machine learning models including Random Forest, XGBoost, Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and CATBoost. To address the issue of class imbalance, Random Under Sampling (RUS) is applied, balancing the dataset by reducing the number of legal transactions to match the number of fraudulent ones.

Our evaluation metrics include accuracy, precision, recall, F1 score, and ROC AUC score, ensuring a comprehensive assessment of model performance. The application of PCA helps in reducing dimensionality and computational complexity, enhancing the overall effectiveness of the models. This project underscores the potential of machine learning techniques in detecting and preventing credit card fraud, ultimately contributing to more secure financial transactions.

Contents

List of Figures

List of Tables

1. INTRODUCTION

In the ever-evolving landscape of financial transactions, the detection and prevention of fraudulent activities stand as paramount challenges for security and integrity. As traditional methods struggle to keep pace with the sophistication of modern fraud techniques. Specifically, our investigation centers on the comparative analysis of two powerful machine learning algorithms - Random Forest and Logistic Regression.

The significance of this study lies in its commitment to addressing the complexities of identifying fraudulent transactions within large datasets. Random Forest, renowned for its ensemble learning capabilities, and Logistic Regression, a time-tested and interpretable method, emerge as promising candidates for discerning patterns indicative of fraudulent behavior within transactional data.

The complexity of this task necessitates the formulation of meticulous cost-sensitive objective functions and loss functions, tailored to the imbalanced nature of transaction datasets where legitimate transactions vastly outnumber fraudulent ones. The integration of machine learning algorithms into this framework aims to map predicted values onto binary labels, distinguishing between genuine transactions (labeled as 0) and fraudulent transactions (labeled as 1).

By immersing ourselves in this detailed exploration, we aspire not only to contribute to the academic discourse surrounding fraud detection but also to provide practical insights that can empower financial institutions and businesses to fortify their defenses against the ever-evolving landscape of fraudulent transactions. This comprehensive study seeks to illuminate the intricate dance between deception and detection, ultimately fostering a more resilient and adaptive approach to safeguarding the integrity of financial transactions.

2. LITERATURE SURVEY

[1] Hajrek and team proposed a model[1] for fraud Detection in Mobile Payment Systems using an XGBoost-based Framework ,Dataset having Legal Transactions = 6.36M , Fraud Transactions = 8.2k ,3:1 Ratio (75% Training and 25% Testing), RUS (Random Under Sampling) Models used are XGBoost (eXtreme Gradient Boosting) Time taken = 207.0 secs and RUS + XGBoost Time taken = 2.4 secs . Highest cost savings can be achieved by combining random under-sampling and XGBoost methods

[2] Fraud detection in capital markets: A novel machine learning approach Dataset having AAER(Accounting and Auditing Enforcement Releases) benchmark dataset collected by the UCB's Center (University of Brekeley) , Dataset consists of 42 Different Variables. Models used are SMOTE(Synthetic Minority Oversampling Technique) and ESOA(Egret Swarm Optimization Algorithm) having Accuracy = 96.27%

[3] Feature-wise attention based boosting ensemble method for fraud detection, Dataset : Private Datasets
Feature Wise Attention Mechanism: Takes multiple features as input ,For each feature, the attention mechanism calculates a weight or importance score. The features are multiplied by their corresponding attention weights, and the results are summed to create a weighted representation of the input features.
Models used are AdaBoost, got an Accuracy of 92.05% and AM Boost got an Accuracy of 93.08%.

[4] Fraudulent Transaction Detection in FinTech using Machine Learning Algorithms, Datasets: Two Datasets
Task1: Number of Fraud Transactions = 2094, Number of legal Transactions = 92588
Task2: Number of Fraud Transactions = 2654, Number of legal Transactions = 97346
Both the Datasets containing 20 columns
Algorithms used are Random Forest, Decision Tree, KNN

[5] Credit Card Fraud Detection: An Improved Strategy for High Recall Using KNN, LDA, and Linear Regression

Algorithms used are KNN (K Nearest Neighbours) + (LR)Linear Regression + (LDA)Linear Discriminant Analysis

Based on Conditions pKNN, pLDA, pLR, pOR

Dataset:

Dataset1: Recall Score = 93.62%

Dataset2: Recall Score = 97.01%

Dataset3: Recall Score = 100%

Dataset4: Recall Score = 93.62%

[6] A Survey and a Credit Card Fraud Detection and Prevention Model using the Decision Tree Algorithm

Algorithms used are Decision Tree

Survey on Credit Card Fraud

[7] Autonomous credit card fraud detection using machine learning approach

Algorithms used are Naive Bayes : Accuracy = 69.53%

SVM(Support Vector Machine): Accuracy = 86.56%

ANN (Artificial Neural Network): Accuracy = 91.85%

LSTM-RNN(Long short-term memory-recurrent neural network) : Accuracy = 100%

[8] Ensemble Learning with Supervised Machine Learning Models to Predict Credit Card Fraud Transactions

Algorithms used are all Traditional Machine Learning Algorithms(SVM, KNN, Decision Tree, Random Forest)

Ensemble Learning (Boosting Algorithms)

Ensemble Learning performs better with 100.0% accuracy, 97.3% precision, 73.5% recall, and 83.7% f1-score against other ML classifiers.

3. PROPOSED METHODOLOGY

To develop the study, we have followed a procedure that helped for developing this research work. Overall, the process of our work is shown in Figure-1.

A. Dataset and Features Description

B. Dataset Preparation

C. Applying Machine Learning Techniques

D. Developing Classification Model

E. Model Performance Analysis

A. Dataset and Features Description:

Our Dataset consists of total 2,84,807 entries of transactions. In that there are legal transaction and fraud transactions. Legal transactions consist of 2,84,315 and fraud transactions consists of 492.There are total 31 colums. The dataset's description has been shown in the table below.

B. Dataset Preparation

We have focused on the dataset for preparing the training dataset. We labelled genuine transactions (labeled as 0) and fraudulent transactions (labeled as 1). Then we have transformed properly and we have made ready of this dataset. Finally, we have taken the decision for keeping 80% of train data and 20% is for testing.

C. *Applying Machine Learning Techniques:*

For the building model, we have used four classification algorithms named Logistic Regression, Random Forest, Decision Tree.

1) Logistic Regression: Logistic regression is a machine learning strategy that has taken advantage of statistics. The logistic function is the root of logistic regression so the main concept of this algorithm comes from this function. This function is additionally referred to as the sigmoid function. Among all algorithms of machine learning, Logistic regression is the most popular and comes after linear regression. They can be compared in diverse manners but their usage is different. The requirement of linear regression comes when to predict values and at the time of classification, logistic regression is used.

2) Random Forest: Random Forest is an ensemble classifier that uses decision tree algorithms in a randomized way. This algorithm is employed in regression as well as classification. It belongs to supervised machine learning. Leo Breiman was the developer of Random Forest which is considered the greatest classifier algorithm for a wide range of data. Any kind of pruning is not used here to grow the trees. This algorithm demonstrates randomness in two particular cases, to make a bootstrap dataset and to make decision trees from this dataset. This algorithm generates the result very fast and the accuracy of the prediction is very high. A wide range of input can be handled by this algorithm easily. This subsurface randomization scheme is combined with the bagging to prove each new tree by replacing the training data set.

3) SVM (support vector machine): Support Vector Machines (SVM) is a powerful machine learning algorithm rooted in both statistical and optimization principles. It seeks to find an optimal hyperplane that maximizes the margin between different classes in a dataset, effectively separating them. The kernel trick is a key mathematical concept, enabling SVM to implicitly map data into higher-dimensional spaces and handle non-linear relationships between features. Widely used for both classification and regression tasks, SVM is known for its ability to find decision boundaries that maximize the margin between support vectors, the closest data points from each class. Its versatility makes SVM effective in scenarios with complex data relationships, making it a popular choice in various applications, from image classification to text categorization.

4) Naïve Bayes (Bernoulli NB): Naïve Bayes, specifically the Bernoulli Naïve Bayes (Bernoulli NB) variant, is a probabilistic machine learning algorithm commonly used for binary classification tasks, such as spam detection or sentiment analysis. It is grounded in Bayes' theorem and assumes independence among features given the class label, making it

computationally efficient and particularly suitable for high-dimensional datasets. In the context of Bernoulli NB, which is tailored for binary features, it models the presence or absence of each feature and estimates the likelihood of a particular class based on these binary occurrences. Despite its "naïve" assumption of feature independence, Bernoulli NB often performs well in practice and is particularly effective when dealing with sparse and discrete data, making it a popular choice for text classification tasks.

5) XG Boost: XGBoost, short for Extreme Gradient Boosting, is a powerful and widely used machine learning algorithm known for its efficiency and effectiveness in various predictive modeling tasks. It belongs to the ensemble learning category and builds a strong predictive model by combining the outputs of multiple weak learners, typically decision trees. XGBoost excels in handling complex relationships within data, offering robustness against overfitting and providing impressive predictive accuracy. It incorporates regularization techniques, parallel computing, and a unique gradient boosting framework, making it highly scalable and suitable for large datasets. XGBoost has become a go-to algorithm in data science competitions and real-world applications due to its ability to deliver high performance across a range of tasks, including classification, regression, and ranking.

6) CAT Boost: CatBoost, short for Categorical Boosting, is a high-performance gradient boosting library designed for categorical feature support in machine learning. Similar to XGBoost, CatBoost is an ensemble learning algorithm that combines the predictions of multiple weak models, typically decision trees, to create a strong predictive model. What sets CatBoost apart is its ability to efficiently handle categorical features without the need for extensive preprocessing, making it particularly useful for datasets with a mix of categorical and numerical attributes. CatBoost employs a novel algorithm for gradient boosting, introducing an oblivious decision tree structure and utilizing an innovative method for handling categorical variables during the training process. This makes CatBoost both user-friendly and efficient, often leading to competitive performance with minimal hyperparameter tuning. It has gained popularity for its ease of use, impressive out-of-the-box performance, and suitability for various machine learning tasks, including classification and regression.

D. *Developing Classification Model:*

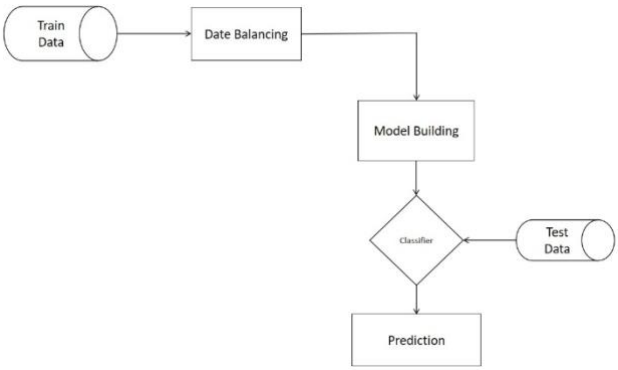

Fig – 1 Classification Model

Train Data: The initial dataset used to teach the model patterns and relationships within the data, divided into training and test sets for effective learning and evaluation.

Data Balancing: Ensuring a balanced distribution of classes in the training data, crucial for addressing imbalances and preventing bias in the model. Techniques include oversampling, undersampling, and methods like SMOTE.

Model Building: Involves selecting and configuring a suitable machine learning algorithm based on the problem type and data characteristics. Options range from XG boost,CAT boost , support vector machines, and neural networks to logistic regression, random forests, and Naive Bayes.

Classifier: The trained model becomes a classifier capable of making predictions on new, unseen data by learning patterns and relationships from the training dataset.

Test Data: A separate dataset not used during training, employed to evaluate the model's generalization performance and assess how well it performs on unfamiliar data.

Prediction: The process of using the trained model to make predictions on new data. The model takes input data and produces an output, such as a class label or numerical value, based on its learned patterns.

E. *Model Performance Analysis:*

$$TP - Rate = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1)$$

FP - Rate: This False Positive rate is always measured by calculating the total number of negative predicted numbers to the total number of negative numbers [17].

$$FP - Rate = \frac{False\ Positive}{FalsePositive + True\ Negative} \quad (2)$$

Precision: This value is determined from the total number of predicted positive values to all of the possible positive cases.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (3)$$

Recall: Recall is the value of the total number of predicted positive results to the total number of actual positives values.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

F- Measure: This value is used to present the overall statistics. This value is the weighted harmonic mean value of the recall and precision [18].

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

Fig – 2 Model Performance Analysis

4. RESULTS

Models	Accuracy	Precision	F1	Recall
Logistic Regression	93.59	90.21	89.73	89.25
Random Forest	96.63	100	94.9	90.3
XG Boost	95.95	97.7	93.4	89.48
CAT Boost	93.91	96.74	90.81	85.59
SVM	91.56	98.58	84.67	74.2
Naive Bayes(BernoulliNB)	96.29	98.79	93.65	89.01

Table – 1 Machine Learning models

Based on the F1 score Random Forest, XG Boost, Naïve Bayes performs well. So, we combined Those three models to base estimators.

Base estimators = [Random Forest, XG Boost, Naïve Bayes]

Models	Accuracy	Precision	F1	Recall
Logistic Regression	93.59	90.21	89.73	89.25
Random Forest	96.63	100	94.9	90.3
XG Boost	95.95	97.7	93.4	89.48
CAT Boost	93.91	96.74	90.81	85.59
SVM	91.56	98.58	84.67	74.2
Naive Bayes(BernoulliNB)	96.29	98.79	93.65	89.01
base_estimators	96.95	96.59	94.98	93.4

Table – 2 Machine Learning models with Base Estimators

We performed Base Estimators i.e., Bagging. So based on F1 scores Base Estimators have performed well (94.98%)

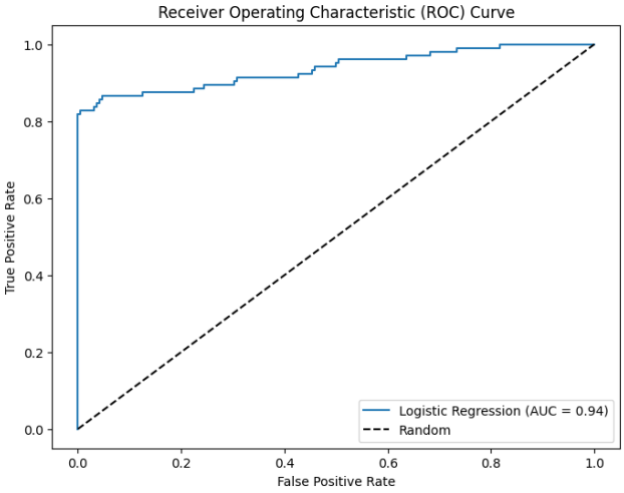

Fig -3 Logistic Regression ROC Curve

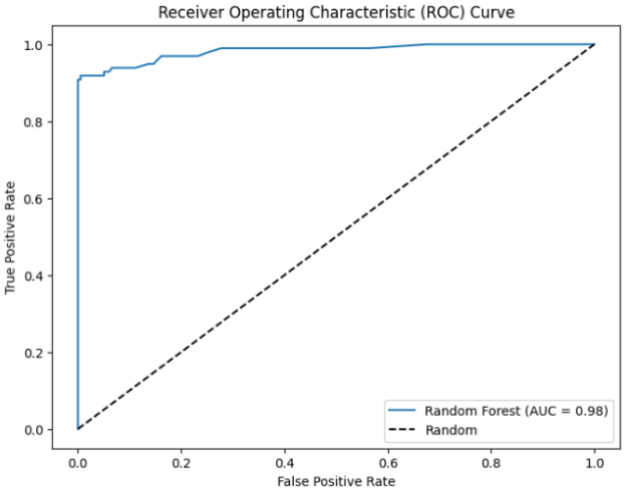

Fig - 4 Random Forest ROC Curve

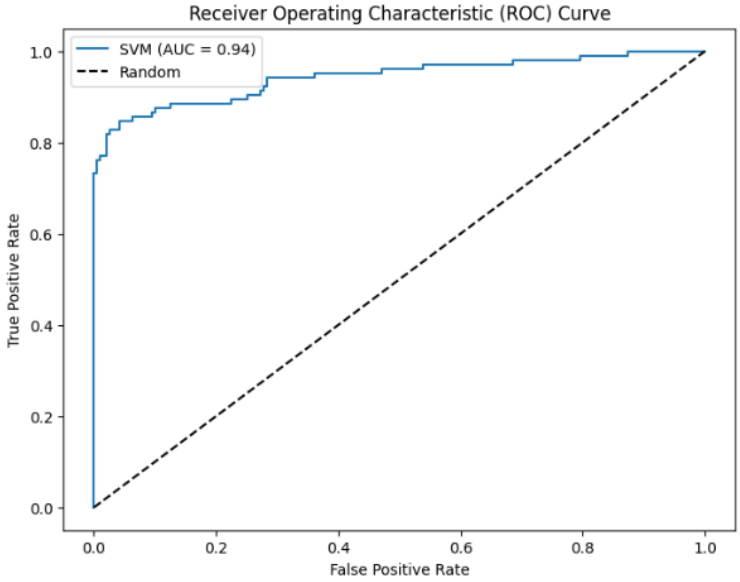

Fig – 5 SVM ROC Curve

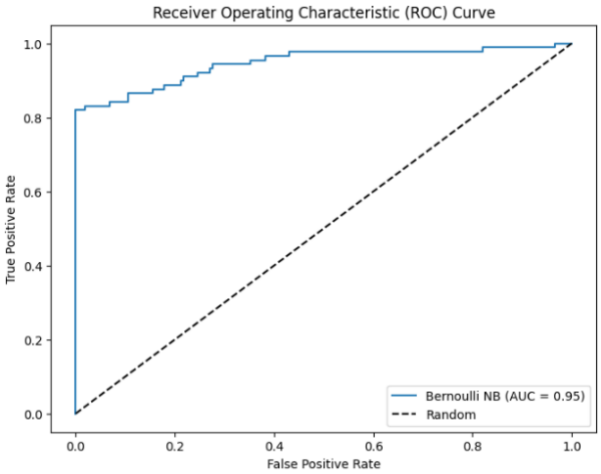

Fig - 6 Bernoulli NB ROC Curve

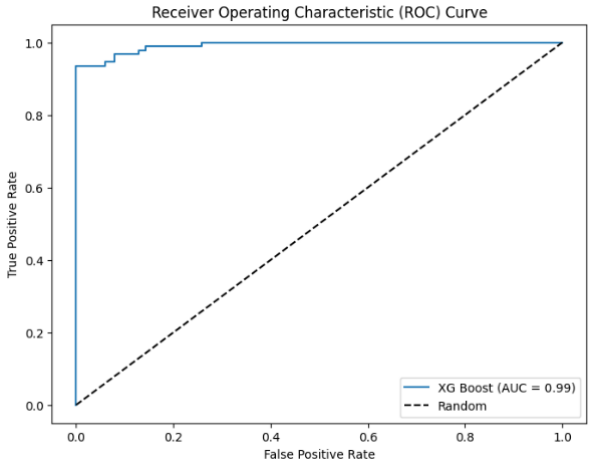

Fig – 7 XG Boost Curve

5. CONCLUSION

Our study on identifying fraudulent transactions has given us important insights, stressing the need for clever and flexible methods as fraud techniques evolve. By using advanced computer programs like Random Forest, Logistic Regression, XG Boost, CAT Boost, SVM, Naïve Bayes, and Base Estimators, we've successfully spotted tricky patterns linked to fraud. This research not only adds to the discussions in academics about detecting fraud but also gives practical ideas for banks and businesses. Our system finds a balance between traditional rule-based methods and the tricky patterns seen in fake transactions, dealing with the issue where most transactions are real.

In our exploration of algorithms, we discovered that Base Estimators work really well in handling the challenges of detecting fraud. Including Base Estimators in our smart system makes a big difference in creating a balanced and flexible approach. As we keep improving our methods, our research encourages a proactive approach, offering decision-makers the tools they need to strengthen defenses against the always-changing world of fraud in digital transactions. Our goal is to build a smarter system that's better at detecting fraud, making financial systems safer, transactions more secure, and ensuring people can trust their money is well-protected.

References

1. Hajek, P., Abedin, M.Z. and Sivarajah, U., 2023. Fraud detection in mobile payment systems using an XGBoost-based framework. Information Systems Frontiers, 25(5), pp.1985-2003.

2. Yi, Z., Cao, X., Pu, X., Wu, Y., Chen, Z., Khan, A.T., Francis, A. and Li, S., 2023. Fraud detection in capital markets: A novel machine learning approach. Expert Systems with Applications, p.120760.

3. Cao, R., Wang, J., Mao, M., Liu, G. and Jiang, C., 2023. Feature-wise attention based boosting ensemble method for fraud detection. Engineering Applications of Artificial Intelligence, 126, p.106975.

4. AbdulSattar, K. and Hammad, M., 2020, December. Fraudulent transaction detection in FinTech using machine learning algorithms. In 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT) (pp. 1-6). IEEE.

5. Chung, J. and Lee, K., 2023. Credit Card Fraud Detection: An Improved Strategy for High Recall Using KNN, LDA, and Linear Regression. Sensors, 23(18), p.7788.

6. Roseline, J.F., Naidu, G.B.S.R., Pandi, V.S., alias Rajasree, S.A. and Mageswari, N., 2022. Autonomous credit card fraud detection using machine learning approach. Computers and Electrical Engineering, 102, p.108132.

7. Baker, M.R., Mahmood, Z.N. and Shaker, E.H., 2022. Ensemble Learning with Supervised Machine Learning Models to Predict Credit Card Fraud Transactions. Revue d'Intelligence Artificielle, 36(4).