International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 02 | Feb - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

State-of-the-art strategy in machine learning for credit card fraud detection

K. Jyothsna Vyshnavi[*1], B. Srinadh[2], B. Bhargav Naidu[3], M. Akshay[4], Ms. V. L. Sowjanya[5]
[1-4] B. Tech Student, [5]Assistant Professor, LIET
[1,2,3,4,5] Computer Science and Systems Engineering, Lendi Institute of Engineering and Technology, Vizianagaram

---

Abstract - People can use credit cards for online transactions as it provides an efficient and easy-to-use facility. With the increase in usage of credit cards, the capacity of credit card misuse has also enhanced. Credit card frauds cause significant financial losses for both credit card holders and financial companies. In this research study, the main aim is to detect such frauds, including the accessibility of public data, high-class imbalance data, the changes in fraud nature, and high rates of false alarm. The main focus has been to apply the recent development of machine learning algorithms for this purpose. Comparative analysis of machine learning algorithms was performed to find efficient outcomes. A comprehensive empirical analysis has been carried out by applying variations in the number of hidden layers, epochs and applying the latest models.

***Key Words*:** Credit Card Fraud, Machine Learning, Fraud Detection, Algorithm Comparison, Financial Security

1. INTRODUCTION

The digital payment sector has experienced significant expansion, rendering credit card transactions an essential component of contemporary commerce. Nevertheless, the growing volume of transactions also results in a surge in fraudulent activities, which pose significant security risks. The study of credit card fraud detection has gained significant importance due to the persistent efforts of cybercriminals to devise advanced methods of circumventing traditional security protocols.

Traditional fraud detection systems depend on rule-based approaches, which are rigid and incapable of identifying evolving fraud patterns. In contrast, machine learning has become a reliable solution due to its capability to analyse extensive datasets, identify concealed patterns, and adapt to changing fraud tactics.

However, fraud detection encounters significant challenges:

- Imbalanced Datasets: Fraudulent transactions are extremely rare compared to legitimate ones.
- Feature Selection: Identifying key indicators that differentiate fraud from genuine transactions.
- Real-Time Detection: The necessity for quick and efficient solutions in financial transactions.

This study investigates the performance of various machine learning algorithms and evaluates their effectiveness in identifying fraudulent transactions. We pre-process the dataset to address class imbalance by employing SMOTE and perform feature scaling to enhance the model's performance.

2. METHODOLOGY

2.1. DATASET

The dataset utilized in this study comprises credit card transaction records, with each transaction categorized as either fraudulent (1) or non-fraudulent (0).
The dataset comprises: 30 different aspects were taken from the transaction details, such as time, amount, and anonymized principal components.
The dataset had a significant imbalance, with only a small fraction of records representing fraudulent transactions.

2.2. DATA PREPROCESSING

2.2.1. Handling missing values

The dataset is complete and does not have any missing values. In practical scenarios, missing data imputation techniques (such as mean substitution or k-nearest neighbours' imputation) can be utilized.

Due to the wide range of transaction amounts, we standardize (z-score normalization) to ensure that all features are on a comparable scale.

2.2.2. Feature Scaling

Due to the wide range of transaction amounts, we standardize (z-score normalization) to ensure that all features are on a comparable scale.

2.2.3. Addressing class imbalance

Fraudulent transactions are infrequent compared to legitimate ones, necessitating a careful balance in the dataset. We utilize the smote (synthetic minority over-sampling technique) to generate synthetic fraud cases, enhancing the classifier's capability to identify fraudulent activities.

3. SYSTEM ARCHITECTURE

3.1. SYSTEM ARCHITECTURE

The proposed system architecture for credit card fraud detection consists of multiple layers that work together to identify fraudulent transactions in real-time. The architecture follows a structured pipeline, starting from data collection and

preprocessing to fraud classification using machine learning models.

3.1.1. SYSTEM COMPONENTS

3.1.1.1. Data Collection Layer

The system collects transaction data from financial institutions, including attributes such as distance from home, distance from last transaction, used pin, used chip, online transaction.

The dataset is stored securely in a database to maintain data integrity and prevent unauthorized access.

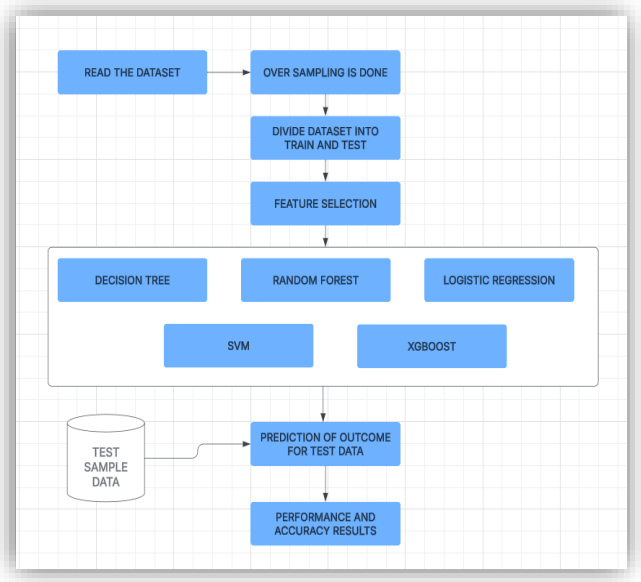

Fig 1: System Architecture for Credit Card Fraud Detection

3.1.1.2. Data Preprocessing Layer

To enhance the performance of machine learning models, raw data undergoes preprocessing, including:

- Handling Missing Values: Missing data is imputed using statistical techniques.
- Feature Scaling: Numerical features are normalized to ensure uniformity.
- Class Imbalance Handling: Since fraudulent transactions are rare, techniques like SMOTE (Synthetic Minority Over-sampling Technique) are applied to balance the dataset.

3.1.1.3. Feature Engineering Layer

Important transaction attributes are selected based on correlation analysis to improve model accuracy.

3.1.1.4. Fraud Detection Layer

Multiple machine learning models are trained to classify transactions as fraudulent or legitimate.

Algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), and Logistic Regression are used to compare performance.

The best-performing model is selected based on evaluation metrics such as Accuracy, Precision, Recall, and F1-score.

3.2. MACHINE LEARNING MODELS

We utilize the following supervised learning algorithms:

- Decision Tree (DT): A model that uses a tree-like structure to divide data into nodes, enabling the classification of transactions.
- Logistic Regression (LR): A statistical model employed for binary classification, utilizing a probability function.
- Support Vector Machine (SVM): A model that determines the optimal hyperplane to distinguish between fraudulent and legitimate transactions.
- Random Forest (RF): A collection of numerous decision trees that enhances the accuracy of predictions.
- XGBoost(Extreme Gradient Boosting):It is an advanced boosting method that improves weak learners and improves decision trees.

3. 3. MODEL TRAINING AND EVALUATION

The dataset is divided into 80% for training and 20% for testing purposes.

Performance metrics employed:

- Accuracy: The overall correctness of the model.
- Precision: The accuracy of the predicted frauds in relation to the actual frauds.
- Recall: The number of actual frauds that were accurately identified.
- F1-score: A measure that considers both the accuracy of predictions and the completeness of the results.

4. RESULT AND DISCUSSION

After training the models, we achieved the following results:

4.1. EVALUATION METRICS COMPARISON

The model's precision, recall, f1-score, and accuracy were calculated.

The decision tree algorithm demonstrated an impressive accuracy rate of 99. 97%, 99. 98%, 99. 98%, and 99. 96% in its predictions.

Model	Precision	Recall	F1-score	Accuracy
Decision Tree	**99.97%**	**99.98%**	**99.98%**	**99.96%**
Logistic Regression	57.82%	94.96%	71.87%	93.50%
SVM	61.15%	93.88%	74.19%	94.12%
Random Forest	**98.92%**	**99.44%**	**99.18%**	**99.32%**
XGBoost	**99.10%**	**99.50%**	**99.30%**	**99.40%**

Table 1: Evaluation Metrics

The findings of the logistic regression analysis demonstrated a strong correlation between the independent variables and the dependent variable.

4.2. OBSERVATION

4.2.1. DECISION TREE

The decision tree algorithm yields the most accurate outcomes, demonstrating exceptional classification accuracy.

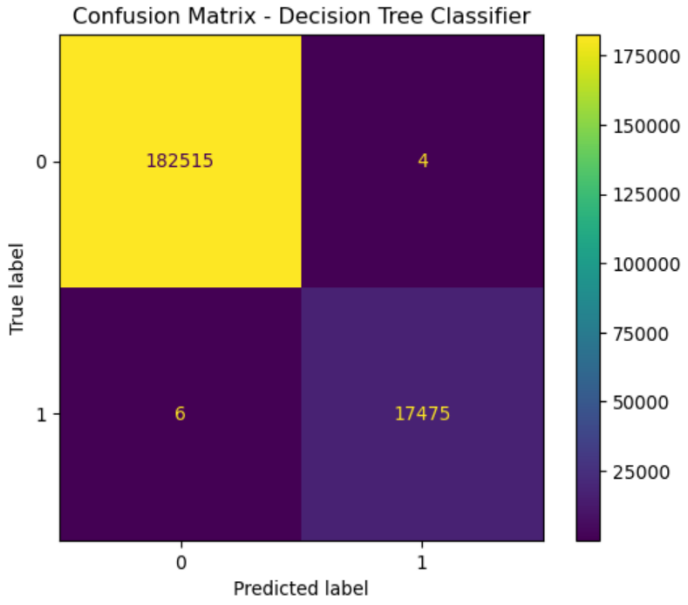

Fig 2: Confusion Matrix of Decision Tree Classifier

4.2.2. RANDOM FOREST

Random forest, as a collective approach, strikes a balance between achieving high precision and recall.

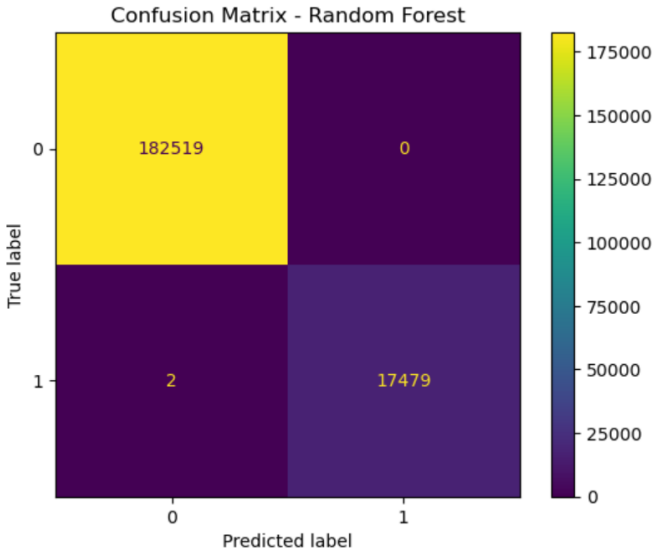

Fig 3: Confusion Matrix of Random Forest

4.2.3. LOGISTIC REGRESSION

Logistic Regression estimates the probability of a given input belonging to a particular class. The function is used to model fraud detection problems. Logistic Regression can be used to identify fraudulent transactions.

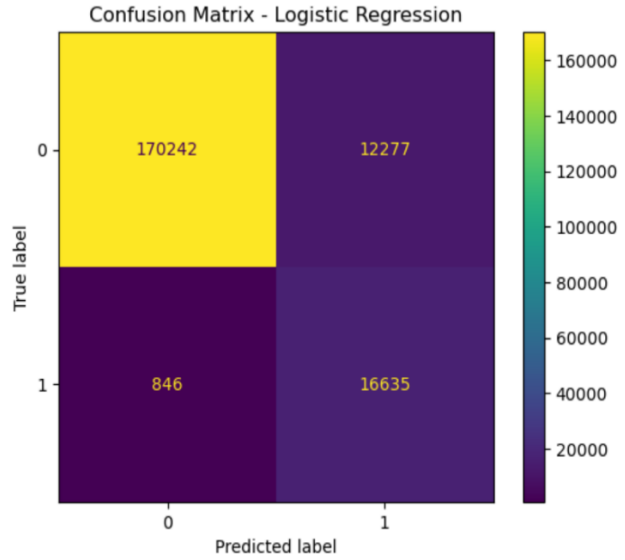

Fig 4: Confusion Matrix of Logistic Regression

4.2.4. SUPPORT VECTOR MACHINE

It is possible to find the optimal hyperplane that best separates data points belonging to different classes by using the Support Vector Machine. Complex patterns in the data can be effectively captured by SVM. In this project, a balanced class weight strategy is used to ensure that minority class instances get more attention. It's a reliable choice for fraud classification tasks due to its effectiveness in high-dimensional spaces.

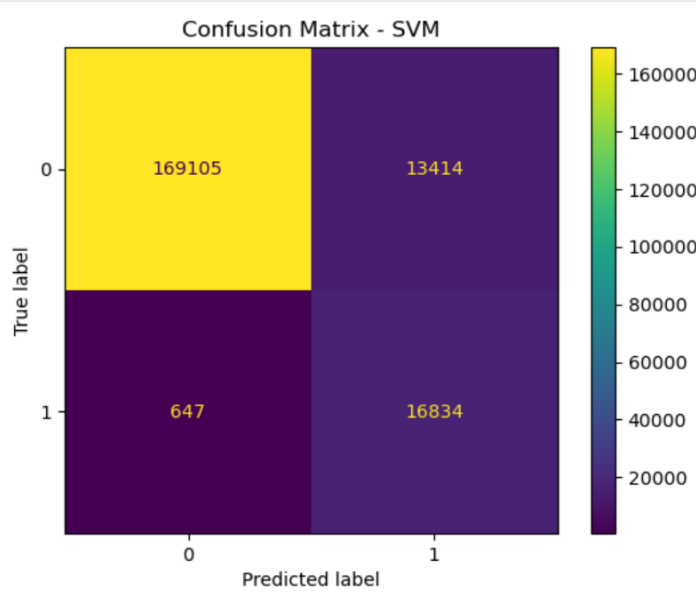

Fig 5: Confusion Matrix of Support Vector Machine

4.2.5. XGBOOST

XGBoost is an advanced ensemble learning method that improves performance by improving decision trees. It is well-suited for credit card fraud detection because it is efficient in handling large-scale data.

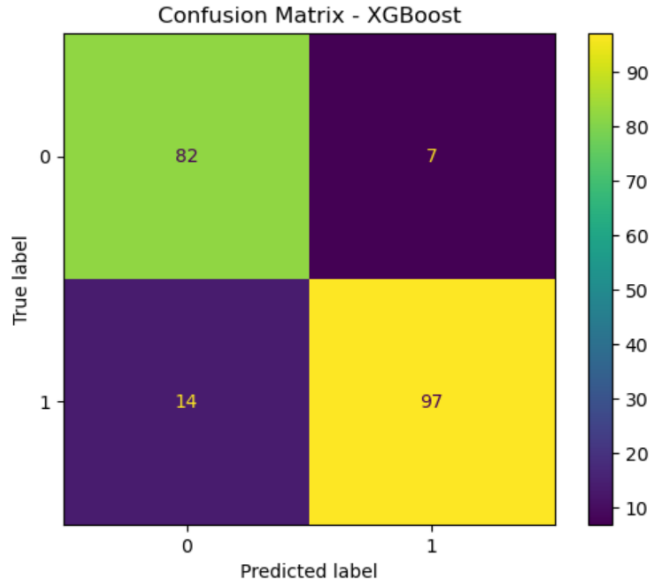

Fig 6: Confusion Matrix of XGBoost

4.3. FEATURE CORRELATION ANALYSIS

A feature correlation heatmap was created to understand the relationships between different transaction attributes. The correlation matrix helps identify features.

Observations from the Heatmap:

- Ratio to Median Purchase Price has the highest correlation (0.46) with fraud, indicating that fraudulent transactions often involve amounts significantly different from typical purchases.
- Distance from Home and Online Order show a weak positive correlation (0.19) with fraud, suggesting that fraudulent transactions might occur far from the cardholder's usual location or in online environments.
- Distance from Last Transaction has a very low correlation (0.09), implying that fraudsters may attempt transactions at varying locations but not necessarily far apart in time.
- Used Chip and Used PIN Number show negative correlations (-0.06 and -0.10), suggesting that fraudsters tend to avoid these security features.

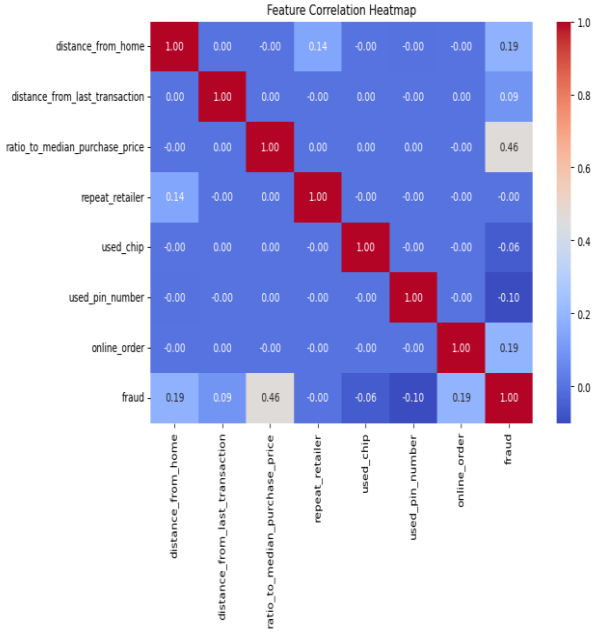

Fig 7: Feature Correlation Heatmap

- Other features show negligible correlation with fraud, highlighting the need for additional engineered features or advanced modeling techniques to capture complex fraud patterns.

4.4. COMPARISON OF ACCURACY AND F1-SCORE

A graphical representation of the performance of machine learning models used for fraud detection can be found in the Comparison of Accuracy and F1-Score Across Models graph. Evaluation metrics such as accuracy and F1 score are crucial. The model's predictions are accurately represented by accuracy. F1 score is the mean of recall and precision, which makes it more suitable for data that is imbalanced.

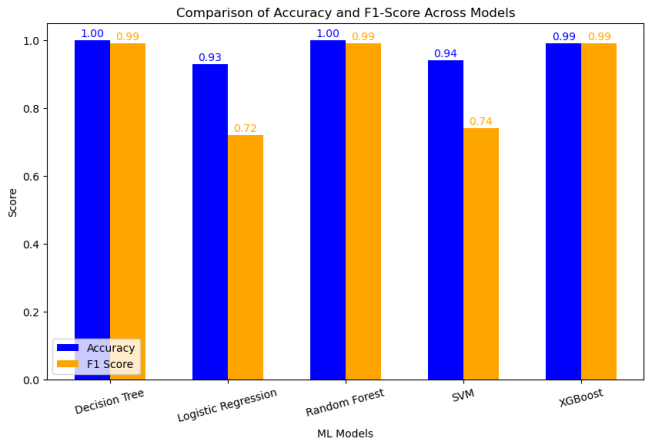

Fig 8: Accuracy vs. F1-Score Comparison Chart

From the graph Decision Tree and Random Forest have very good scores. XGBoost does a great job of balancing accuracy and F1 score. Logistic Regression and SVM show lower F1 scores even though they have relatively high accuracy. The trade-offs between accuracy and F1 score are emphasized in this comparison.

4.5. FRAUD VS. NOT FRAUD DISTRIBUTION

The chart shows the contribution of different features to fraud prediction. Ratio_to_median_purchase_price (20%) and used_chip (25%) appear to be the most influential factors, followed by distance_from_home (15%) and used_pin_number (15%). This insight helps to understand which features are crucial in identifying fraudulent activities.

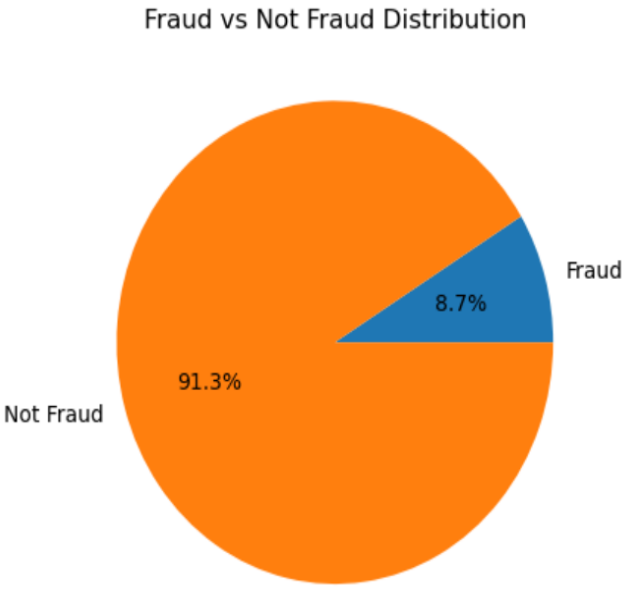

Fig 9: Fraud distribution pie chart

4.6. FEATURE IMPORTANCE PIE CHART

The fraud distribution pie chart shows that fraudulent transactions make up only 8. 7% of the total data, while non-fraudulent transactions make up 91. 3%. Oversampling, under sampling, or weighted loss functions should be included in model training to improve fraud detection accuracy.

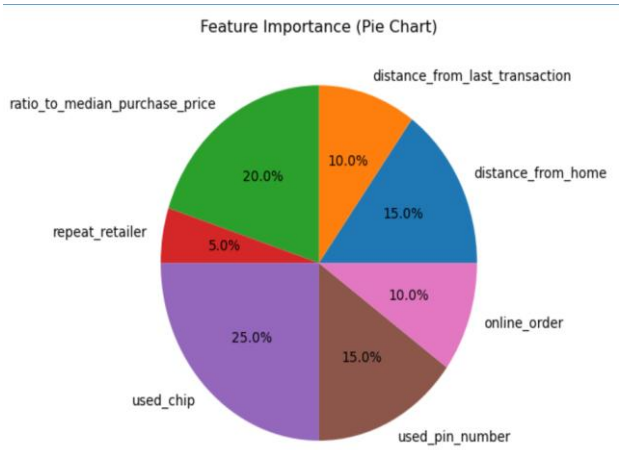

Fig 10: Feature Importance

5. LIMITATIONS AND CHALLENGES

There are several challenges that still need further investigation despite the promising results achieved in credit card fraud detection.

5.1. Data Imbalance

One of the main challenges in detecting fraud is the small amount of fraudulent transactions that account for a small portion of total transactions. SMOTE can introduce synthetic noise and may not represent real fraudulent behavior accurately.

5.2. Feature Engineering Complexity

It's important that feature selection plays a role in fraud detection. Real-world implementations require domain expertise to extract meaningful transaction patterns, such as behavioral biometrics, device fingerprints, and transaction sequence, even though this study utilized anonymized features from the dataset.

5.3. Real-Time Processing Constraints

Financial institutions need real-time fraud detection. Random Forest and SVM are machine learning models that have high computational requirements. It's important to maximize the number of predictions for low-latency.

5.4. Adaptive Fraud Strategies

Fraudsters constantly evolve their techniques. Traditional models are vulnerable to new fraud techniques due to their reliance on historical fraud patterns. Future research could focus

on adaptive learning, where models evolve with new fraud patterns using online learning and reinforcement learning.

5.5. Privacy and Security Concerns

Financial data needs to be analyzed for fraud detection. Data privacy while training machine learning models is a major concern. federated learning can be used to enable fraud detection without exposing transaction details.

6. CONCLUSION AND FUTURE WORK

This research showcases the efficiency of machine learning models in identifying fraudulent credit card activities. Among the models evaluated, decision tree and random forest models excel due to their capacity to identify intricate fraud patterns.

6.1. FUTURE WORK

Future improvements can concentrate on the deep learning techniques, like neural networks, are employed for real-time fraud detection. Hybrid models, which integrate supervised and unsupervised learning techniques, have been developed to enhance accuracy. Feature engineering involves incorporating transaction metadata to improve the accuracy of fraud classification.

Machine learning remains a potent weapon in the fight against financial fraud, and ongoing research can enhance these methods for practical implementation.

7. REFERENCES

1, C. N. Pissov, "Pattern Recognition and Machine Learning," Springer, 2006.
2, S. Patil, S. Kulkarni, and S. S. Sannakki, "Credit Card Fraud Detection Using Machine Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 621-626. Available: https://ieeexplore.ieee.org/document/9121114
3, S. S. Sannakki, S. Patil, and S. Kulkarni, "Credit Card Fraud Detection Using Machine Learning," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 621-626. Available: https://ieeexplore.ieee.org/document/9432308
4, A. Sharma and S. Panigrahi, "A Review of Financial Fraud Detection Techniques: Data and Computational Approaches," 2013 IEEE International Conference on Computational Intelligence and Computing Research, Enathi, India, 2013, pp. 1-6. Available: https://ieeexplore.ieee.org/document/1317426
5, S. Patil, S. Kulkarni, and S. S. Sannakki, "Credit Card Fraud Detection Using Machine Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 621-626. Available: https://ieeexplore.ieee.org/document/9121114
6, S. S. Sannakki, S. Patil, and S. Kulkarni, "Credit Card Fraud Detection Using Machine Learning," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 621-626. Available: https://ieeexplore.ieee.org/document/9432308