International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Research Article)

Check for updates

An examination of machine learning-based credit card fraud detection systems

Himanshu Sinha *

Kelley School of Business Indiana University, Bloomington, Naperville IL Unites State.

International Journal of Science and Research Archive, 2024, 12(02), 2282–2294

Publication history: Received on 07 July 2024; revised on 15 August 2024; accepted on 17 August 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.12.2.1456

Abstract

As a result of the e-commerce industry's explosive growth, credit cards are now frequently used for online purchases. In recent years, banks have faced a significant issue with credit card fraud (CCF) due to the difficulty in detecting fraudulent activity within the credit card system. Machine learning is the solution to the problem of CCFD during transactions. An analysis starts with a study of the Kaggle-provided CCFD dataset. There is a considerable disparity between the classifications in the dataset, which has 284,807 transactions total, of which only 492 are deemed fraudulent. The preprocessing methods are used to prepare the data, which includes the handling of the missing values, the detection of outliers, and the encoding of the categorical variables. Five different classification models are tested and evaluated employing different metrics like precision, F1-score, accuracy, and recall. These models are SVM, Random Forest (RF), Bagging, XGBoost, and DT. In terms of spotting fraudulent transactions, XGBoost is the model that has the highest accuracy rate of 99% among the others. To further strengthen the effectiveness and reliability of fraud detection systems, future research might investigate ensemble methodologies and integrate real-time data streams, guaranteeing thorough defence against financial crime.

Keywords: Credit Card Fraud Detection (CCFD); Security; Financial Fraud; Dataset; Machine Learning**.**

1. Introduction

The proliferation of online banking has coincided with an increase in the occurrence of suspicious financial transactions. Constantly, fraudulent individuals devise novel and impregnable methods to circumvent fraud detection systems. The act of deceit, which results in monetary benefit for the perpetrator, constitutes financial fraud. CCF stands as the most dominant type of financial deception[1]. The introduction of credit cards and the expansion of online shopping have brought about substantial conveniences for retailers and consumers alike[2]. Unfortunately, CCF has increased significantly since the start of the digital revolution. Globally, CCF poses a serious threat to financial organisations, as well as to consumers. It may include account hijacking, identity theft, and unauthorised transactions. Effective solutions are urgently needed for CCF, considering the financial consequences and the decline in confidence in electronic payment methods[3][4].

The disclosure of credit card information is imperative for safeguarding privacy. Phishing websites, loss or theft of credit cards, counterfeit credit cards, theft of card information, intercepted cards, and so forth, are all methods of obtaining credit card information. Avoid the aforementioned items for reasons of security. Online deception involves the remote execution of a transaction that requires only the card details. At the time of purchase, verification through a PIN, manual signature, or card imprint is not mandatory. Authentic cardholders are typically oblivious to the fact that their card information has been compromised or appropriated [5][6]. Any variations from the "usual" spending patterns may be readily identified by looking at the spending patterns linked to each card and detecting this kind of fraud. Eliminating successful credit card forgeries is most effectively accomplished through fraud detection through the analysis of cardholders' extant purchase data. The absence of access to the data sets and the non-disclosure of the results. Cases of

* Corresponding author: Himanshu Sinha

deception ought to be identified using the logged data and user behaviour that are readily accessible data sets[7][8]. Manual assessments are distinguished by their labour-intensive process, substantial financial investment, and vulnerability to human fallibility. As a result of advent of ML algorithms, the banking sector has begun to investigate more advanced and automated methods for detecting fraud.

Credit card fraud is detected using a variety of techniques, such as ML, DL, and statistical methods[9][10][11][12]. Credit card transaction abnormalities are found and examined using statistical methods including clustering, regression, and hypothesis testing. Machine learning techniques, on the other hand, use algorithms to analyse previous data and identify fraudulent activities in real-time. Deep learning techniques use neural networks to recognise complex patterns and characteristics in large, complicated datasets autonomously, leading to very precise fraud detection[13][14][15].

Data imbalance, caused by an unequal distribution of fraudulent and non-fraudulent transactions in the dataset, is a major problem with CCFD. Biassed model results and ineffective fraud detection skills might result from this imbalance. The use of machine learning approaches such as data balancing, oversampling, under-sampling, and the synthetic minority oversampling technique (SMOTE) to handle unbalanced credit card fraud data has been addressed in several studies[16][17]. Still, a thorough investigation of these methods' efficacy is absent.

CCFD relies heavily on supervised ML models and methods. In order to build a more powerful and precise fraud detection system, these methods combine several separate models. To improve overall prediction power and tackle issues like unbalanced datasets, ensemble techniques like bagging, boosting, and stacking are often used [18][19]. Ensemble methods enhance fraud detection effectiveness while decreasing the likelihood of false positives and negatives by capitalising on the strengths of many base models. Ensemble learning is a powerful and versatile method that may be used in the ongoing fight against credit card fraud[20]. In spite of these achievements, research on the computing efficiency of supervised ML models is essential.

1.1. The contribution of the study

Here are key research contributions from the examination of ML-based CCFD systems:

- **Comprehensive Data Preprocessing:** Implemented extensive data preprocessing techniques such as handling missing values, detecting outliers, and encoding categorical variables, along with feature selection, to ensure high-quality data for model training and improved model performance.
- **Addressing Class Imbalance:** Applied SMOTE to effectively address the significant class imbalance in the Kaggle Credit Card Fraud Detection dataset, ensuring balanced training data for the machine learning models.
- **Evaluation of Multiple Classification Models:** Conducted a comparative analysis of five classification models—XGBoost, SVM, RF, DT, and Bagging—evaluating their performance in detecting CCF based on metrics like accuracy, recall, precision, and F1-score.
- **Performance Metrics and Visualization:** Utilized confusion matrices and ROC curves to visualise and assess an efficacy of each model, providing a clear understanding of their strengths and weaknesses in fraud detection.
- **Superior Model Performance:** Found that XGBoost achieved the highest accuracy (99%) among the models tested, while Bagging attained the highest F1-score (95%), highlighting the effectiveness of different models in specific performance metrics for CCFD.

1.2. Structure of the study

For the sections that follow, this study is organised as follows: The literature pertinent to our investigation is examined in **Section 2**. the study's methodology; **Section 3** details the approaches used for conducting the research. In **Section 4**, discuss the intended results and evaluations of the study. **Section 5** contains the results and future intentions of our research investigation.

2. Literature Review

CCFD is an area that has seen a number of academic investigations. CCFD has been a subject of many research investigations, some of which are shown below.

In, Yuhes Raajha et al., (2023) concentrates on contemporary CCF practices and real-time fraud detection methods. Various ML algorithms, including FSVM, RF, LR, and SVM, were implemented to classify legitimate and fraudulent transactions on a dataset gathered from credit card users. CCFD scheme comparisons employing these classification models were executed with sensitivity, specificity, accuracy, and precision. Results from the comparison indicated that, among the algorithms tested, the FSVM performed best with a 98.61% accuracy rate[21].

In, Geng and Zhang, (2023) presents a CCFD network that is unsupervised and uses dual adversarial learning. Our technique places an emphasis on taking both the original and intermediate characteristics into account at the same time, which is different from the traditional approaches to anomaly identification. The results show that our method outperforms state-of-the-art fraud detection methods with an accuracy of 0.9224, an F1 score of 0.9208, and an MCC of 0.8456, as tested on the European cardholder dataset[22].

In, Singh et al., (2022) delves into the most recent developments and practical applications of credit card fraud prevention with machine learning. This document compares and contrasts four different ML algorithms based on how accurate they are. The Cat boost algorithm has been shown to have the highest success rate (99.87%) in identifying instances of credit card fraud. For the CCFD dataset, Kaggle was contacted[23].

In, Devika et al., (2022) identify cases of transaction fraud using a variety of ML-based methodologies and models. In this case, LR and similar algorithms are taken into account. The fraud will be identified by choosing the algorithm with the highest level of accuracy. With an accuracy of 0.99% or higher, the LR method would be able to detect fraud with ease. The administrator may easily submit a dataset to this web application for fraud detection after signing in and performing authentication[24].

In, Rai and Dwivedi, (2020) provides a way to detect fraudulent behaviour in credit card data employing NN-based unsupervised learning. The suggested technique achieves better results than the current methods of AE, LOF, IF, and K-Means clustering. Compared to the state-of-the-art approaches, the suggested NN-based fraud detection system has a success rate of 99.87%. While comparing AE, IF, LOF, and K Means, the equivalent accuracy percentages are 97%, 98%, 98%, and 99.75%[25].

In, Najadat et al., (2020) utilise the IEEE-CIS Fraud Detection dataset supplied by Kaggle to demonstrate a method for discerning between legitimate and fraudulent transactions. Our model is BiLSTM-BiGRU-MaxPooling, which is composed of a Bi-GRU and Bi-LSTM. Furthermore, six ML classifiers were implemented: LR, NB, Voting, Ada boosting, RF, and DT. In contrast to the outcomes produced by ML classifiers, our model demonstrated superior performance, attaining a score of 91.37%[26].

Table 1 Summary of the related work on Credit Card Fraud Detection with various approaches

References	Methodology	Dataset	Performance	Limitations & research gap
Yuhes Raajha *et al.*, [21]	Supervised classification using ML algorithms	Dataset collected from credit card users	When comparing FSVM with other methods, 98.61% accuracy in identifying valid and fraudulent transactions was obtained.	Lack of details on the specific characteristics and source of the dataset. Limited discussion on real-world implementation challenges.
Geng and Zhang, [22]	Unsupervised anomaly detection with dual adversarial learning	European cardholder dataset	Dual adversarial learning approach achieved an accuracy of 92.24%, surpassing existing fraud detection techniques.	Absence of explanation on the interpretability of the model's decisions. Lack of exploration on generalisation to other datasets.
Singh *et al.*, [23]	Supervised classification using ML algorithms	Kaggle dataset	Cat boost algorithm achieved the highest accuracy of 99.87% in detecting credit card fraud.	Insufficient comparison with state-of-the-art techniques. Lack of transparency regarding dataset characteristics and potential biases.
Devika *et al.*, [24]	Supervised classification using ML algorithms	Credit card data	Logistic Regression model attained 99% accuracy in fraud detection. A web application framework for fraud detection was proposed.	Limited discussion on the scalability and efficiency of the proposed web application. Evaluation restricted to a specific dataset without generalisation analysis.
Rai and Dwivedi, [25]	Unsupervised Learning Using Neural Network	credit card data	Proposed NN-based method achieved 99.87% accuracy, outperforming	Inadequate explanation of the computational complexity of the proposed NN architecture.

			existing unsupervised methods like AE, IF, LOF, and K-Means clustering.	Evaluation conducted on a specific dataset without validation on diverse datasets
Najadat *et al.*, [26]	Combination of BiLSTM, BiGRU, and ML classifiers	IEEE-CIS Fraud Detection dataset provided by Kaggle	BiLSTM-BiGRU model achieved 91.37% accuracy in predicting legitimate or fraud transactions.	Limited discussion on feature engineering and selection process. Evaluation is confined to a single dataset without generalisation analysis.

2.1. Research gaps

There are three main areas where the studies that were given have not addressed the research needs. First, the difficulties of scalability, computational efficiency, and the interpretability of model decisions—all of which are important in the real-world application of the suggested fraud detection methodologies—do not receive sufficient consideration in most research. Secondly, the results cannot be applied to many real-life situations since the datasets used were not openly discussed about their biases and peculiarities. Finally, while evaluating the relative efficacy and practical application of their procedures, several publications fail to compare them with state-of-the-art methodologies, which is a crucial step. Closing these knowledge gaps will help push CCFD forward and encourage the creation of better detection technologies.

3. Methodology

An analysis of CCFD systems based on ML was performed in this research endeavour. A dataset utilised for this purpose was Kaggle's CCFD, which consists of 31 numerical features. A major class imbalance exists in the dataset since out of 284,807 transactions, only 492 were found to be fraudulent. Several data preprocessing techniques were applied to get the data ready for the subsequent analysis. Encoding categorical variables, dealing with missing values, and labelling outliers were part of these activities. Besides, feature selection was performed to identify the characteristics that were the most important for the improvement of the model's performance. The challenge of class imbalance was solved by synthetic samples that were created through the SMOTE. Soon, the dataset was divided into two parts, 80% of which was used for training and 20% of which was used for testing. The performance of five classification models—XGBoost, SVM, RF, DT, and Bagging—was evaluated for their ability to detect CCF. To assess an efficacy of each model, confusion matrices were used to generate and visualise performance indicators like F1-score, precision, recall, and accuracy the overall process of CCFD system displays in Figure 1 data flow diagram.

3.1. Data Collection

The CCFD dataset from Kaggle was used in this study. This dataset contains thirty-one numerical characteristics. Out of the 284,807 transactions contained in the dataset, 492 were fraudulent while the other ones were authentic. Considering that just 0.173% of transactions were deemed fraudulent, it is clear that the sample is highly biased.

3.2. Data Preprocessing

"Data Pre-processing" means transforming raw data into a tidy dataset. Data collected from several sources is not always obtained in a processed manner, rendering it unsuitable for review. Preprocessing refers to the steps used to prepare the dataset for algorithm input. The following preprocessing methods are described below:

- **Handling missing values:** To address missing values, there are several options. Options include using a sophisticated imputation approach, filling missing values with zeros or the mean, or removing occurrences with missing data [27].
- **Outlier detection:** Inconsistent data input, incorrect observations, or very extreme data points are common causes of outliers, which are data points that differ substantially from the main dataset.
- **Encoding Categorical Variables:** Variables that may take on a small, defined range of values are called categorical variables.

3.3. Feature selection

Feature selection is an essential process in which the necessary elements are chosen from a given dataset in accordance with understanding. The dataset utilised in this study comprises numerous features, from which we selected only those that are essential for enhancing performance measurement.

3.4. SMOTE for Data balancing

When working with datasets that display substantial class imbalances, sampling methods assume a critical role in the domain of CCFD[28][29]. Prominent among the methods utilised to generate synthetic samples within the minority class via interpolation between pre-existing instances is the SMOTE[30]. To mitigate the potential for overfitting caused by the inclusion of redundant cases in the training set, the SMOTE is implemented. Equation (1) is utilised by SMOTE.

$$x_{syn} = x_i + x_{knn} - x_i)x\ t \qquad (1)$$

An arbitrary number between 0 and 1 is denoted by t, while xi represents the feature vector and xknn signifies the KNN[31].

3.5. Train-Test split

One of the most crucial aspects of a dataset for evaluating a model and understanding its traits is partitioning it into training and testing sets. There was an allocation of 80% of the data for training and 20% for testing in this report.

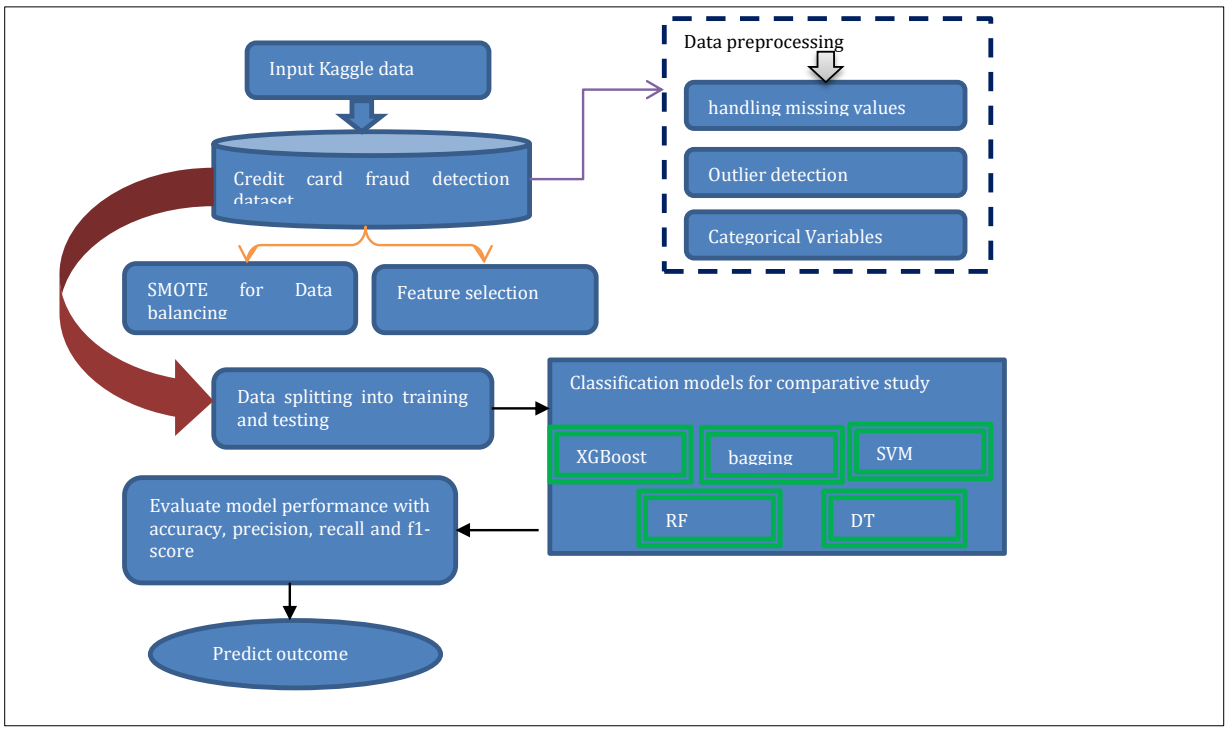

Figure 1 Data flow diagram for credit card fraud detection system

3.6. Classification Models

For this comparative study, compare various models like XGBoost, RF, DT, bagging, and support vector machines that present below:

3.6.1. XGBoost

Chen and Guestrin [32][33] suggested the XGBoost algorithm, which is based on the GBDT frame. The system's outstanding performance in Kaggle's machine-learning competitions has garnered significant attention. XGBoost, in contrast to the GBDT, prevents overfitting by incorporating a regularisation term into the objective function. This defines the objective function in Euq.2:

$$O = \sum_{i=1}^{n} \mathrm{L(yi, F(xi))} + \sum_{k=1}^{t} R(f_k) + C \qquad (2)$$

The regularisation term at the k time iteration is represented by $R(fk)$, while C is a constant term that can be omitted selectively. $R(fk)$ represents the regularisation term in euq.3.

$$R(f_k) = \alpha H + \frac{1}{2}\eta \sum_{j=1}^{H} \mathrm{w}_j^2 \qquad (3)$$

The complexity of leaves is denoted as α, the number of leaves is denoted as H, the penalty parameter is represented as η, and the output result of every leaf node is denoted as wj. The leaf node denotes the section of the tree that cannot be divided, and the leaves specifically show the expected categories based on the classification criteria. Also, XGBoost uses second-order Taylor series of the objective function rather than the first-order derivative used in GBDT. The objective function may be obtained by using the MSE as the loss function in euq.4.

$$O = \sum_{i=1}^{n}[p_i\, w_{q(x1)} + \frac{1}{2}(qiw^2_{q(xi)})] + H + \qquad \frac{1}{2}\eta \sum_{j=1}^{H} \mathrm{w}_j^2 \ldots\ldots\ldots \qquad (4)$$

gi representing the first derivative of the loss function and hi standing for the second derivative, and $q(xi)$ standing for a function that allocates data points to their respective leaves[34].

The total loss value is the result of adding together all the individual loss values. The total of the loss values of the leaf nodes in the DT may be used to compute the final loss value, because samples correlate to them.

3.6.2. Support vector machine (SVM)

The SVM method is a member of the classifier family of supervised hyperplane-based algorithms, which includes parametric and discriminative classifiers [35][36][37]. SVMs aren't built to handle multi-class classification jobs; they're only good for binary classification in hyperplane-based situations. It is advisable to locate a hyperplane and algorithm that yields the highest margin while requiring the fewest number of points possible.

3.6.3. Random forest (RF)

A supervised ML method called Random Forest may be used to resolve problems with regression and classification. During the training process, it constructs multiple decision trees and determines the outcome through a majority vote in order to enhance accuracy and generate more dependable forecasts. In order to improve precision, the aggregation and entropy criteria of Bootstrap are implemented[38]. The random forest is calculated with Gini search in equ.5 and 6.

$$IG(N_P, a) = Gini\ (N_P) - \sum_{i-1}^{c} \frac{|N_1|}{|N_P|}\ Gini\ (N_1) \qquad (5)$$

$$Gini(Np) = 1\ - \sum_{j-1}^{m} p_j^2 \qquad (6)$$

where Np denotes the amount of data present at node Np and |Ni| represents the amount of data at node Ni, $0 \le I \le c$. denotes the number of distinct data labels at node Np, and Pj represents the proportion of data with the jth label relative to the total number of data at node Np. The letter "j" denotes the label's number.

3.6.4. Decision tree (DT)

The DT method has extensive use in the domain of ML. The algorithm is applied to a given dataset in order to perform regression or classification analysis. The algorithm partitions the data into multiple subgroups in accordance with a series of inquiries. The procedure commences at the root node, also referred to as the primary node, wherein every sample is stored. Multi-split or binary operations are employed to partition each node into secondary nodes[39].

3.6.5. Bagging

Bagging is a technique whereby different classifiers are trained on different subsets of features and data to provide somewhat different classification hypotheses. After combining the classifiers, the ensemble is constructed. This approach enhances generalizability through the reduction of variance[40] [41].

3.6.6. Performance matrix

The visualisation of the confusion matrix and the computation of precise performance metrics are imperative when assessing the efficacy of a data mining classification algorithm. These will facilitate the evaluation of individual method performance and the comparison of the performance of various methods.

3.6.7. Confusion Matrix

Actual classes versus predicted classes are listed in a confusion matrix in tabular format. A value of each quadrant corresponds to the quantity of samples. It facilitates comprehension of the models' predicted True Positives, False

Positives, True Negatives, and False Negatives. Hence, it facilitates an evaluation of a model's classification performance. Displayed below is a confusion matrix.

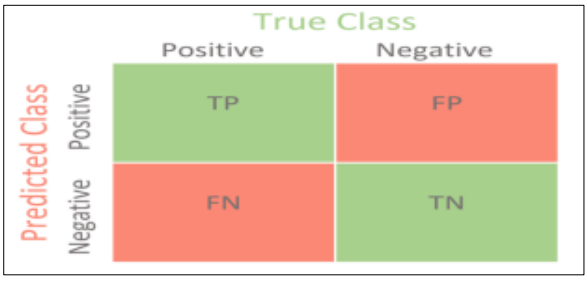

Figure 2 Representation of confusion matrics

- **TP -** True Positive: Denotes the positively labelled tuples for which the classifier provided an accurate label.
- **FP -** False Positive: Denotes the affirmative tuples that the classifier erroneously classified.
- **FN -** False Negative: is used to describe the negative tuples that the classifier mislabeled.
- **TN -** True Negative: is used to describe the negative tuples that the classifier properly labelled.
- Accuracy

As a famous measure, accuracy is not useful when judging the performance of a collection that is not fair. It is the total number of correct guesses added up over all previous predictions. It is computed using equation (7):

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (7)$$

- Precision

A proportion of correct predictions is referred to as precision or positive predictive accuracy (PPA), calculated as equ.8.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (8)$$

- Recall

As a percentage of all hits, recall shows how many of them were correctly recognised. It is computed using the equation (9):

$$\text{Recall} = \frac{TP}{\text{TP+FN}} \qquad (9)$$

- F1-score

The F1-score is a measurement that takes into account both sensitivity and precision for a single factor. A higher number means there is more unity. Its range is from 0 to 1. It is determined using equation (10):

$$\text{F1} = \frac{2*(\text{precision}*\text{recall})}{\text{precision+recall}} \qquad (10)$$

The ML models in this research were evaluated using these four metrics. The scores for each metric are given to help with decision-making and offer objective measurements for evaluating the models' performance. Additionally, to guarantee the models' performance level, we consistently checked the accuracy of using the CCFD approach.

4. Result analysis and discussion

A result of a comparative study across performance metrics like recall, f1-score, accuracy, and precision are provided in this section for the CCFD. A graphical representation of model performance like a ROC curve and confusion metrics are present in this section.

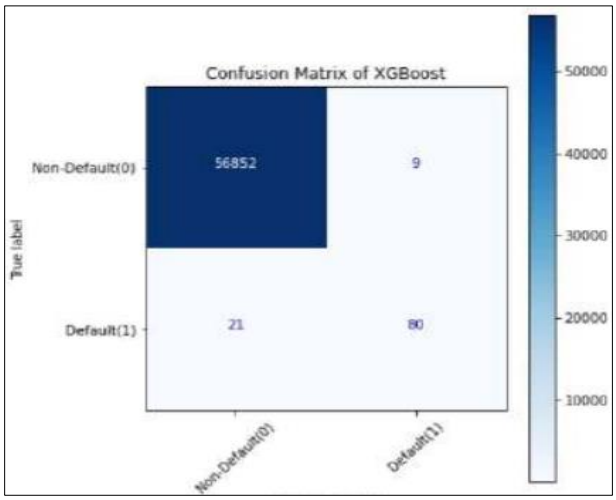

Figure 3 Confusion matrix for XGBoost model

A XGBoost model's confusion matrices displays in figure 3 reveal that the model correctly identified 56,852 non-default transactions but missed 80 default transactions, with 40,000 non-default transactions incorrectly classified as default.

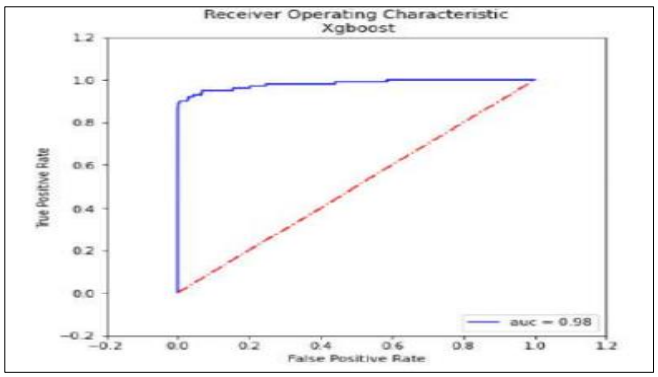

Figure 4 ROC curve of Xgboost model

Figure 4 illustrates the XGBoost model's ROC curve, which is 0.98. An AUC of 1 indicates perfect performance, so a value of 0.98 suggests the model is performing very well at classifying between the two classes.

4.1. Comparative Analysis

The results of comparing several credit card fraud detection technologies show that their efficacy varies across many performance metrics like recall, f1-score, accuracy, and precision. In This comparison, various models like SVM, RF, bagging, XGBoost, and DT. Table II shows the XGBoost model achieves high accuracy is 99% with 89% precision score.

Table 2 Comparison between different models performance

Models	Accuracy	Precision	Recall	F1-core
SVM[42]	94.9	95.9	95.1	95.1
RF[43]	98	82	90	85
DT[44]	88.16	95	81	87
Bagging[15]	94	95	95	95
XGBoost[45]	99	89	79	84

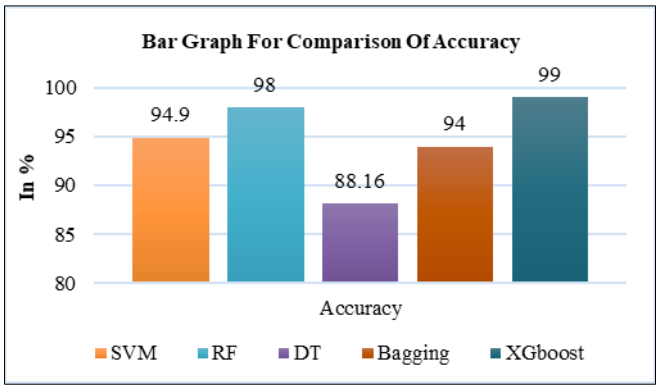

Figure 5 Bar Graph for comparison of accuracy

Figure 5 shows a Bar Graph for comparison of different models' accuracy. According to overall efficacy in accurately identifying cases, XGBoost obtains the highest rate at 99%. Following closely behind is Random Forest with an accuracy of 98%. Bagging and SVM both demonstrate high accuracy at 94% and 94.9%, respectively

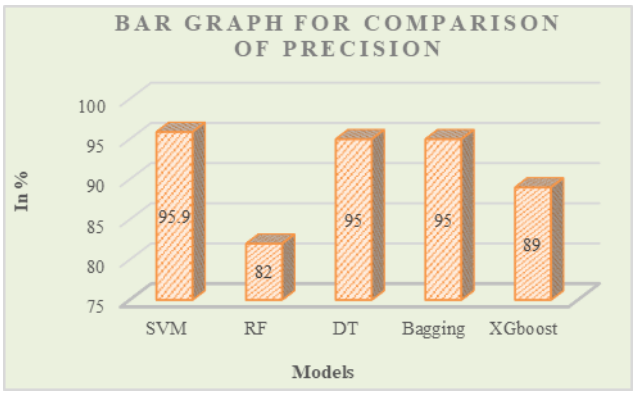

Figure 6 Comparison of models Precision

Bar graph for the comparison of models' Precision is shown in Figure 6. In this figure, Random Forest, decision tree, and Bagging achieve a higher precision score of 95%. XGBoost lags with a precision of 89%, suggesting a slightly higher rate of false positives.

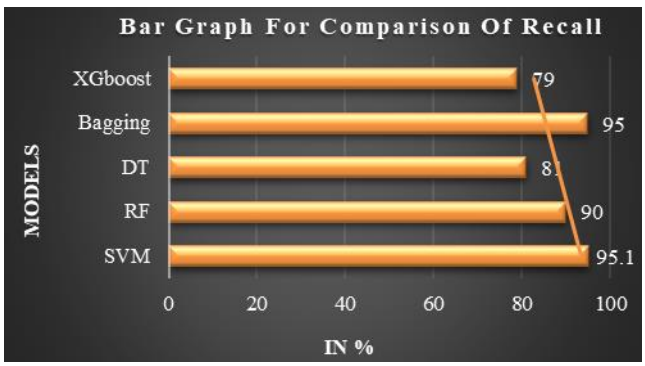

Figure 7 Bar Graph for comparison of Models Recall

The above Figure 7 shows a Bar Graph for comparison of Models Recall. Bagging and SVM show the highest recall rates at 95%. The Decision Tree follows with 81% recall. XGBoost and Random Forest have lower recall rates (79% and 90%).

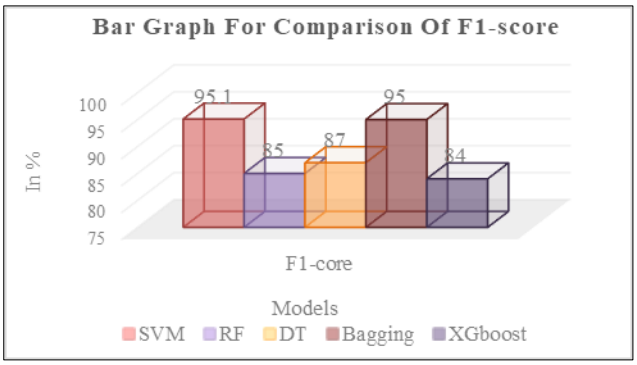

Figure 8 Bar Graph for comparison of f1-score

The following Figure 8 displays a Bar Graph for comparison of a f1-score in different models. In comparison, bagging achieves the highest F1-score of 95% while XGBoost and Decision Tree demonstrate F1-scores of 84% and 87%. In comparison, evaluate the model's performance XGBoost's highest performance and achieve 0.99% accuracy, 89% precision, 79% recall, and 84% f1-score.

A comparative research of various CCFD models, including SVM, Random Forest, Decision Tree, Bagging, and XGBoost, reveals significant differences in their performance metrics like precision, accuracy, recall, and F1-score. XGBoost outperforms other models according to overall accuracy, achieving a remarkable 99%, indicating its robust capability in identifying fraudulent transactions. However, it lags in precision (89%) and recall (79%) compared to models like Bagging and SVM, which both exhibit high recall rates of 95%. Bagging stands out with the highest F1 score of 95%, highlighting its balanced performance across all metrics. The ROC curve of XGBoost, with an AUC of 0.98, underscores its effectiveness in distinguishing among fraudulent and non-fraudulent transactions. This analysis underscores an importance of selecting the appropriate model based on specific performance criteria, as different models exhibit varying strengths and weaknesses in CCFD.

5. Conclusion

Credit card fraud is a major business concern. These frauds may cause major personal and company losses. As a consequence, companies are spending more money in developing innovative ideas and techniques for identifying and preventing fraud. A primary goal of this study was to investigate a range of ML algorithms that are designed to identify fraudulent transactions. The research utilised the CCFD dataset. Preprocessing entailed encoding categorical variables, managing absent values, and identifying outliers. On the basis of F1-score, accuracy, precision, recall, and SVM, RF, Bagging, XGBoost and DT performance metrics, five classification models were implemented and compared. XGBoost achieved an exceptional accuracy rate of 99%, establishing itself as the leading performer in the accurate detection of fraudulent cases. However, it is crucial to acknowledge certain constraints, including the possibility of overfitting stemming from the generation of synthetic samples and the dependence on a solitary dataset. Further investigation may be warranted into an integration of ensemble techniques and real-time data streams in order to enhance an efficacy of fraud detection systems. Finally, future studies may evaluate the model's capacity to scale to bigger datasets and keep up with increasing computing needs. To guarantee efficient processing as the dataset size increases, this strategy might make use of distributed computing or parallel processing.

References

[1] M. Azhan and S. Meraj, “Credit card fraud detection using machine learning and deep learning techniques,” in Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020, 2020. doi: 10.1109/ICISS49785.2020.9316002.

[2] L. Einav, P. Klenow, J. D. Levin, and R. Murciano-Goroff, “Customers and Retail Growth,” SSRN Electron. J., 2021, doi: 10.2139/ssrn.3983915.

[3] S. Arora and P. Khare, “Optimizing Software Pricing : AI-driven Strategies for Independent Software Vendors,” no. May, pp. 743–753, 2024.

[4] P. Khare and Sahil Arora, “Predicting Customer Churn in SaaS Products using Machine Learning,” Heal. Inf. Sci. Syst., 2020, doi: 10.1007/s13755-019-0095-z.

[5] K. Sravya, C. M. Kasthuri, K. R. Meghana, and A. S. Poornima, "Credit card fraud detection using machine learning algorithms - Study of customer behaviour," 11th Int. Conf. Adv. Comput. Control. Telecommun. Technol. ACT 2020, vol. 8, no. 2, pp. 143–150, 2020.

[6] Sahil Arora and Pranav Khare, "AI/ML-Enabled Optimization of Edge Infrastructure: Enhancing Performance and Security," Int. J. Adv. Res. Sci. Commun. Technol., vol. 6, no. 1, pp. 1046–1053, 2024, doi: 10.48175/568.

[7] S. Mathur., "Supervised Machine Learning-Based Classification and Prediction of Breast Cancer," Int. J. Intell. Syst. Appl. Eng., vol. 12(3), pp. 0–3, 2024.

[8] S. Daliri, "Using Harmony Search Algorithm in Neural Networks to Improve Fraud Detection in Banking System," Comput. Intell. Neurosci., 2020, doi: 10.1155/2020/6503459.

[9] S. A. Pranav Khare, "THE IMPACT OF MACHINE LEARNING AND AI ON ENHANCING RISK-BASED IDENTITY VERIFICATION PROCESSES," no. 05, pp. 1–6, 2019.

[10] P. Khare, "AI-Powered Fraud Prevention : A Comprehensive Analysis of Machine Learning Applications in Online Transactions," vol. 10, no. 9, pp. 518–525, 2023.

[11] P. Khare, "Enhancing Security with Voice : A Comprehensive Review of AI-Based Biometric Authentication Systems," vol. 10, no. 2, pp. 398–403, 2023.

[12] V. Rohilla, S. Chakraborty, and R. Kumar, "Deep learning based feature extraction and a bidirectional hybrid optimized model for location based advertising," Multimed. Tools Appl., 2022, doi: 10.1007/s11042-022-12457-3.

[13] S. Mathur and S. Gupta, "Classification and Detection of Automated Facial Mask to COVID-19 based on Deep CNN Model," in 2023 IEEE 7th Conference on Information and Communication Technology, CICT 2023, 2023. doi: 10.1109/CICT59886.2023.10455699.

[14] S. A. Pranav Khare*1, "THE IMPACT OF MACHINE LEARNING AND AI ON ENHANCING RISK-BASED IDENTITY VERIFICATION PROCESSES," Int. Res. J. Mod. Eng. Technol. Sci., vol. 06, no. 05, pp. 1–10, 2024.

[15] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, "Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach," Big Data Cogn. Comput., 2024, doi: 10.3390/bdcc8010006.

[16] P. Gupta, A. Varshney, M. R. Khan, R. Ahmed, M. Shuaib, and S. Alam, "Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques," Procedia Comput. Sci., vol. 218, pp. 2575–2584, 2022, doi: 10.1016/j.procs.2023.01.231.

[17] I. A. Mondal, M. E. Haque, A. M. Hassan, and S. Shatabda, "Handling Imbalanced Data for Credit Card Fraud Detection," in 24th International Conference on Computer and Information Technology, ICCIT 2021, 2021. doi: 10.1109/ICCIT54785.2021.9689866.

[18] M. S. Siraj, M. W. haqqani, and D. K. M. Quadry, "A novel credit card fraud detection using supervised machine learning model," Int. J. Multidiscip. Res. Growth Eval., 2024, doi: 10.54660/.ijmrge.2024.5.1.313-324.

[19] J. Thomas, "Optimizing Bio-energy Supply Chain to Achieve Alternative Energy Targets," J. Electr. Syst., vol. 20, no. 6, 2024.

[20] hanzeng wang, "Credit card fraud detection using supervised machine learning methods," 2022. doi: 10.1117/12.2628121.

[21] R. M. Yuhes Raajha, A. Kavin, D. Rajkumar, R. Reshma, R. Santhosh, and N. Mekala, "An Analytical Approach to Fraudulent Credit Card Transaction Detection using Various Machine Learning Algorithms," in Proceedings of the 2023 2nd International Conference on Electronics and Renewable Systems, ICEARS 2023, 2023. doi: 10.1109/ICEARS56392.2023.10085157.

[22] J. Geng and B. Zhang, "Credit Card Fraud Detection Using Adversarial Learning," in 2023 International Conference on Image Processing, Computer Vision and Machine Learning, ICICML 2023, 2023. doi: 10.1109/ICICML60161.2023.10424872.

[23] A. Singh, A. Singh, A. Aggarwal, and A. Chauhan, "Design and Implementation of Different Machine Learning Algorithms for Credit Card Fraud Detection," in International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2022, 2022. doi: 10.1109/ICECCME55909.2022.9988588.

[24] M. Devika, S. R. Kishan, L. S. Manohar, and N. Vijaya, "Credit Card Fraud Detection Using Logistic Regression," in 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2022, pp. 1–6. doi: 10.1109/ICATIECE56365.2022.10046976.

[25] A. K. Rai and R. K. Dwivedi, "Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme," in Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, 2020. doi: 10.1109/ICESC48915.2020.9155615.

[26] H. Najadat, O. Altiti, A. A. Aqouleh, and M. Younes, "Credit Card Fraud Detection Based on Machine and Deep Learning," in 2020 11th International Conference on Information and Communication Systems, ICICS 2020, 2020. doi: 10.1109/ICICS49469.2020.239524.

[27] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," J. Big Data, 2021, doi: 10.1186/s40537-021-00516-9.

[28] R. N. N.-S. Sini Sunny, jennifer Houg, Shibu Navaneeth, Sayed Aniqa, Afortude John Kofi, "Abstract P2073: Hyperbaric Oxygen Therapy Protects The Myocardium From Reductive Stress-induced Proteotoxic Remodeling," Circ. Res. logo, vol. 133, 2023, [Online]. Available: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=afRDlOoAAAAJ&citation_for_view=afRDlOoAAAAJ:9yKSN-GCB0IC

[29] A. Sayed, "Face mask detection model based on deep CNN technique using AWS," Int. J. Eng. Res. Appl., vol. 13, no. 5, pp. 12–19, 2023, doi: 10.9790/9622-13051219.

[30] "Prevalence of Anemia and its Determinants among the Rural Women of Khyber Pakhtunkhwa-Pakistan," Ann. Hum. Soc. Sci., 2023, doi: 10.35484/ahss.2023(4-iv)04.

[31] R. Alsmariy, G. Healy, and H. Abdelhafez, "Predicting cervical cancer using machine learning methods," Int. J. Adv. Comput. Sci. Appl., 2020, doi: 10.14569/IJACSA.2020.0110723.

[32] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. doi: 10.1145/2939672.2939785.

[33] R. K. Vinita Rohilla, Sudeshna Chakraborty, "Car Auomation Simulator Using Machine Learning," Proc. Int. Conf. Innov. Comput. Commun., 2020, [Online]. Available: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=zlcFgwEAAAAJ&citation_for_view=zlcFgwEAAAAJ:2osOgNQ5qMEC

[34] V. Rohilla, M. S. S. Kumar, S. Chakraborty, and M. S. Singh, "Data Clustering using Bisecting K-Means," in Proceedings - 2019 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2019, 2019. doi: 10.1109/ICCCIS48478.2019.8974537.

[35] J. Thomas, "The Effect and Challenges of the Internet of Things (IoT) on the Management of Supply Chains," vol. 8, no. 3, pp. 874–878, 2021.

[36] M. E. Mavroforakis and S. Theodoridis, "A geometric approach to support vector machine (SVM) classification," IEEE Trans. Neural Networks, 2006, doi: 10.1109/TNN.2006.873281.

[37] V. Rohilla, S. Chakraborty, and M. Kaur, "An Empirical Framework for Recommendation-based Location Services Using Deep Learning," Eng. Technol. Appl. Sci. Res., 2022, doi: 10.48084/etasr.5126.

[38] R. K. Vinita Rohilla, Sudeshna Chakraborty, "Random Forest with harmony search optimization for location based advertising," Int J Innov Technol Explor Eng, vol. 8, no. 9, pp. 1092–1097, 2019, [Online]. Available: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=zlcFgwEAAAAJ&citation_for_view=zlcFgwEAAAAJ:WF5omc3nYNoC

[39] R. K. Vinita Rohilla, Sudeshna Chakraborty, "Deep learning based feature extraction and a bidirectional hybrid optimized model for location based advertising," Multimed. Tools Appl., vol. 81, no. 11, pp. 16067–16095, [Online]. Available: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=zlcFgwEAAAAJ&citation_for_view=zlcFgwEAAAAJ:zYLM7Y9cAGgC

[40] S. Arora and S. R. Thota, "Using Artificial Intelligence with Big Data Analytics for Targeted Marketing Campaigns," no. June, 2024, doi: 10.48175/IJARSCT-18967.

[41] P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, "A Review of Ensemble Methods in Bioinformatics," Curr. Bioinform., 2010, doi: 10.2174/157489310794072508.

[42] Bharti Kudale, Swapnil Birajdar, Abhishek Hattekar and S. G. Sameer Kulkarni, "Credit Card Fraud Detection Using Machine Learning," Proc. - Int. Conf. Dev. eSystems Eng. DeSE, no. 01, pp. 168–172, 2023, doi: 10.1109/DeSE60595.2023.10469583.

[43] P. Sharma, S. Banerjee, D. Tiwari, and J. C. Patni, "Machine learning model for credit card fraud detection-A comparative analysis," Int. Arab J. Inf. Technol., 2021, doi: 10.34028/iajit/18/6/6.

[44] K. Kumain, "Analysis of Fraud Detection on Credit Cards using Data Mining Techniques," Turkish J. Comput. Math. Educ., 2020, doi: 10.17762/turcomat.v11i1.13590.

[45] Sanmati Marabad, "Credit Card Fraud Detection Using Machine Learning," Proc. - Int. Conf. Dev. eSystems Eng. DeSE, pp. 168–172, 2023, doi: 10.1109/DeSE60595.2023.10469583.