ISSN 2047-3338

Fraud Detection in Credit Cards Using Machine Learning

Tehreem Zahid

Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan
tzahid113@gmail.com

***Abstract*— The rapid growth of electronic transactions in the modern financial landscape has led to an increased prevalence of fraudulent activities, particularly in the realm of credit and debit cards. This research paper explores the application of machine learning algorithms for the detection and prevention of fraud in card transactions. By leveraging the power of artificial intelligence and data analytics, financial institutions can significantly enhance their capabilities to identify and mitigate fraudulent activities, thereby safeguarding the interests of both consumers and businesses.**

***Index Terms*— Machine Learning, Data Science, Credit Card Fraud Detection and Algorithms**

I. INTRODUCTION

WITH the widespread adoption of digital payment systems, the risk of fraudulent transactions has become a critical concern for financial institutions. Traditional methods of fraud detection are no longer sufficient, necessitating the integration of advanced technologies such as machine learning. Machine learning is a field of artificial intelligence that focuses on the development of algorithms and statistical models that enable computers to learn and improve their performance on a specific task without being explicitly programmed. In essence, it allows machines to learn from data, identify patterns, and make predictions or decisions based on that learning. Machine learning plays an important role in agriculture, healthcare, industry, and e-commerce [14].

Classification of machine learning: Machine learning is classified into two categories first one is supervised machine learning which includes classification and regression, the algorithms used in it include Linear Regression, Polynomial Regression, Random Forest, Decision trees, Logistic Regression, Support Vector Machine, and Regression trees. The second one is Unsupervised Learning which includes clustering, dimensionally reduction, and association rule learning their algorithms are K-Means, Clustering, KNN (K-Nearest Neighbor), Hierarchical Clustering, Deep Learning, Distribution models and the third one is reinforcement learning which is further categorized in Q-learning Markov Decision Processes called decision making.

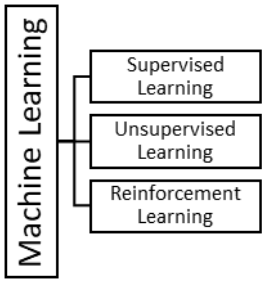

Fig. 1: Classification of Machine Learning

Advantages of machine learning: Machine learning helps in Automation, Prediction, Forecasting, and Scalability. Machine learning algorithms can detect unusual patterns in user behavior, helping to identify and prevent fraudulent activities such as unauthorized transactions or account takeovers. This enhances the security of e-commerce platforms. Machine learning algorithms provide security to e-commerce businesses. In security, there are three main things i.e., confidentiality, integrity, and availability [11]. By using machines learning algorithms, we can achieve these three things. To predict credit card fraud detection, we used machine learning algorithms and methods. The top ten ML algorithms are incorporated for the detection of credit card fraud.

A) Types of Credit Card Frauds

There are numerous methods used in credit card scams [14]. Many of its forms are discussed in detail in this paper:

i) Application fraud: It frequently happens in conjunction with identity theft. It happens when criminals apply for a new credit card on your behalf.

ii) Magnetic/Electronic Card Fraud: Fraudulent use of magnetic or electronic cards, including imprints. This signifies that the data stored on the card's magnetic strip is accessed by someone.

iii) CNP Fraud: CNP stands for Card Not Present. Offenders will commit CNP fraud against you if they find the

termination date of your card and the account number associated with it. This scam can be committed by cell phone, email, or the internet.

iv) Intercept Fraud: This type of fraud is also known as mail non-receipt credit card scams. In this case, you were waiting for a fresh card or an alternative, which an attacker may intercept.

v) Suspected Identity Fraud: To obtain a credit card, a suspect may provide a provisional address and a phony name.

vi) Doctored Credit Card Fraud: Credit card fraud involves removing the metallic band using a powerful magnet. This is done by the perpetrators themselves. Attempt to change the information on the card to that of legal cards.

vii) Fake Card Making Fraud: Making false cards is difficult. Thieves can create false cards by including a chip, magnetic band, and holograms.

Counterfeiters may use phony numbers and names to complete transactions with this sort of credit card.

viii) Acquisition of Account: The acquisition of accounts is one of the most common types of credit card fraud.

A thief can somehow gain access to all your data and documents. Normally, this is done online.

ix) Stolen or Lost Credit Card Fraud: Stolen or Lost Credit Card Fraud: Your card will be removed from your possession, either through robbery or because you dropped it. Offenders who obtain it will utilize it to conduct transactions.

x) Card ID Fraud: Card ID Fraud occurs when an offender obtains your card information and uses it to gain control of an existing card account or start a new one. All of this is being done under your name (Maniraj, 2019).

In this paper, various Machine Learning algorithms are implemented to verify which algorithm of Machine Learning provides the most efficient outcome and detects fraudulent activities they are as follows:

i. Logistic Regression (LR)
ii. Support Vector Machine (SVM)
iii. Decision Trees (DT)
iv. Random Forest (RF
v. KNN Model
vi. Gradient Boosting Model
vii. Bagging and Boosting

II. LITERATURE REVIEW

With improvements in digital technologies in e-commerce, the way of purchasing items also changed. Users use credit cards for online payments. A credit card is a small piece of plastic card that consists of all information related to a person which is issued by a bank or financial service providers. The illegal use of another person's credit card to get money or property either physically or digitally is known as credit card fraud. To predict credit card fraud, we used machine learning and its algorithms. For this research, we have reviewed different books, research papers, articles, and journals from Google Scholar and Research Gate. IEEE Xplore, Springer.com, and different sites. We used the strategy of extracting articles by using the keywords" Fraud Detection in Cards using Machine Learning".

(Johnathan, 2023) used the three supervised machine learning models as Logistic Regression, Random Forest, Decision Trees, and credit card fraud dataset to predict fraudulent transactions. Logistic regression is a statistical method used for the relationship between a categorical dependent variable and one or more independent variables. It is a type of regression analysis that is commonly used for classification tasks. Random Forest is also used for classification. A decision tree is a hierarchical tree-like structure used for decision-making in machine learning and data mining. It concluded that the Random Forest model gives the best results with about 94% accuracy in predicting credit card fraud using features age, class, and amount in the American dataset. It concluded that persons who are above 60 years of age suffer more from these fraudulent activities or online transactions occurring between the hours of 22:00 GMT and 4:00 GMT.

(Trivedi, 2020) It proposed a Feedback mechanism and used Unsupervised, Supervised machine learning algorithms to detect credit card fraud. Random Forest, Naïve Bayes, Logistic Regression, SVM, KNN, and Decision Trees are used for the prediction using the European dataset. After comparing these models, it concluded that Random Forest provided the best results for prediction of about 95.9889%.

(Islam, 2024) This research paper proposed a rule-based model whose main objective was to detect and minimize the fraudulent activities by the attackers. It was compared with the existing machine learning models including Random Forest, Decision Trees, KNN, and Logistic regression using two datasets called Pay Sim and Bank Sim datasets. The results showed that by using the Bank Sim dataset the accuracy was about 96% of Random Forest, Decision Tree 97%, and KNN 93%, and by using the Pay Sim dataset the accuracy was 98% of Random Forest, Decision Tree 83%. It concluded that the Random Forest algorithm gives the best result for prediction.

(Algorithms, 2019) It proposed a novel-based approach to detect fraud in financial transactions using the dataset. This paper reviews the past transactions of users. It performed the SMOTE (Synthetic Minority Over-Sampling Technique) operation on the dataset which did not provide good results. It concluded that Logistic regression, decision tree, and random forest are the algorithms that gave better results. This research paper used the two machine learning models to detect fraud in financial transactions using the dataset. These models include the Local Outlier Factor and Isolation Forest Algorithm. The results showed an accuracy of about 98% and a precision was about 33%.

(ALARFAJ, 2022) This research paper presented the comparative analysis of machine learning and deep learning algorithms to detect and prevent fraudulent activities in financial transactions using the European dataset. It used models including the Extreme Learning Model, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression and Deep learning models including the Baseline

Model, Convolutional Neural Network (CNN), Long short-term memory (LSTM), and Residual neural network (RNN). After performing the experiment and comparison, the results showed that deep learning models give an accuracy of about 89.72% which was much better than old and existing machine learning algorithms.

(Khalid, 2023) It proposed a novel ensemble approach that also combined the existing machine learning algorithms including Decision Tree, Random Forest, Support Vector Machine Logistic Regression, and deep learning model Convolutional Neural Network (CNN) which gives the best result by using the European datasets. Research shows that deep learning models provide better, and more efficient results as compared to existing machine learning models. It performed a comparative analysis between machine learning and deep learning models. It used the China banks dataset, European dataset, and Brazilian dataset. The results showed that CNN has a high accuracy was about 99.72%.

(Tiwari, 2021) It performed the comparative analysis of machine learning models (Logistic Regression, Random Forest, and Support Vector Machine with deep learning models including Artificial neural network, Convolution neural network, and recurrent neural network to detect Credit card fraud, but results showed that the Random Forest gives the best results. Used different datasets including German dataset, Private banks, and Pag Seguro (Brazilian Online Payment Service). It concluded that KNN and SVC give better results on small datasets but are not implemented or do not provide good accuracy on large datasets.

(Itoo, 2021) performed the comparative analysis of machine learning models including Logistic Regression, Random Forest, Naïve Bayes, and KNN using the European dataset from Kaggle. The results showed that Logistic regression gives a high accuracy of 94% as compared to KNN 75% and Naïve Bayes 91%.

(Khalid, 2024) It proposed the novel ensemble approach and various machine learning models used the machine learning models, to predict credit card fraud detection using the European dataset. The results showed that logistic regression gives better results as compared to the proposed approach. Comparing all the results it can be concluded that machine learning models' accuracy can be improved by the ensemble machine approach.

III. METHODOLOGY

One of the biggest concerns for financial institutions and consumers alike is credit card theft. Financial losses and harm to financial institutions' reputations can result from fraudulent transactions. The identification of fraudulent transactions has made substantial use of machine learning techniques. In this project, we use logistic regression, Support vector machine, Decision Tree, Random Forest Algorithm, KNN, and Gradient Boosting Technique to classify transactions as either legitimate or fraudulent based on their features. The data used in this project is a CSV file containing credit card transaction data and is obtained from Kaggle. The dataset contains 31 columns and 284,807 rows. The "Class" column is the target variable, which indicates whether the transaction is not fraud (Class = 0) or Fraud (Class = 1). Before training the model, we first separate the fraud data and non-fraudulent data. Since the data is imbalanced, with significantly more legitimate transactions than fraudulent transactions, we under-sample the legitimate transactions to balance the classes. The data was separated into training and testing sets using the train_test_split() function.

Table I: Dataset Information

Features	Null/Not Null	Datatypes
Time	Not Null	Float64
V1 TO V28	Not Null	Float64
Amount	Not Null	Float64
Class 0= Not Fraud,1= Fraud	Not Null	int64

In this research paper, we used the language python language and its libraries pandas, NumPy, and Scikit-learn for visualization. The preprocessing of the data set is done by removing the unwanted columns that are present in the data set. The column named ‘Time’ is dropped or removed here as it’s not needed.

Fig. 2 indicates the dataset's correlation matrix. This matrix clarifies that the class of attributes is independent of the amount of the transaction. It is also evident from the matrix that the transaction type is dependent on the attributes added by the PCA.

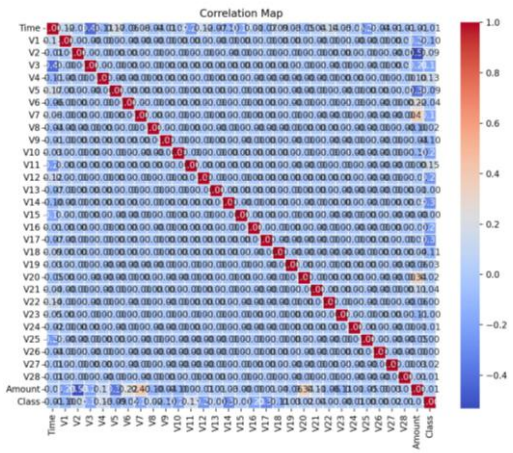

Fig 2: Correlation Matrix for Credit Card Dataset

IV. ALGORITHMS AND IMPLEMENTATION

1) Logistic regression is a widely used classification algorithm that models the probability of an event occurring based on input features. The logistic regression model is trained on the training data using the Logistic Regression () function from Scikit-learn. The trained model is then used to predict the target variable for the testing data. A logistic regression () algorithm is implemented. Firstly, train the dataset in this model and then evaluate the remaining dataset

with the help of the prediction method for the remaining data. This approach gives an accuracy of 94.86%.

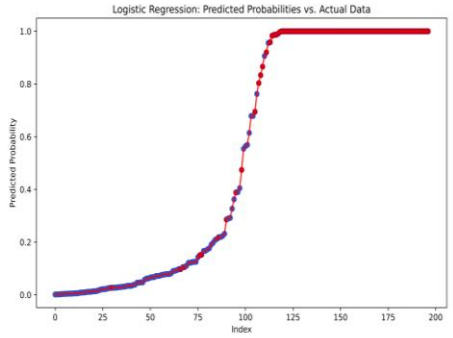

Fig. 3: Logistic Regression

2) Decision Tree is a machine learning algorithm used for both classification and regression tasks. It models decisions based on a tree-like structure where each internal node represents a "decision" based on the value of a feature, each branch represents the outcome of that decision, and each leaf node represents the final decision or prediction. This approach gives an accuracy of 88%.
3) Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It's particularly effective in high-dimensional spaces and is well-suited for both linear and non-linear classification problems. This approach gives an accuracy of 90%.
4) Random Forest is based on Supervised machine learning. It can be used for both regression and classification problems. It is a collection of multiple random decision trees which is called a forest. This approach gives an accuracy of 92%.
5) Gradient Boosting is a powerful ensemble learning technique used for both regression and classification tasks in machine learning. It minimizes the overall prediction error. This technique gives an accuracy of 93%.
6) K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for both classification and regression tasks. This technique gives an accuracy of 68%.
7) Bagging is also called Bootstrap Aggregating technique which is used to improve the accuracy of a model. It divides the original dataset into small subsets for more training of the model and getting more accuracy. It has two types including Bagging meta-estimator and Random Forest. Bagging is commonly applied to decision tree-based models like Random Forest, you can also use it with other base models, including logistic regression. This paper used the meta-estimator n =10 with logistic regression. By using the bagging technique, we can improve the accuracy of the model.

The pseudocode for this algorithm is as follows:

```
from sklearn.ensemble import Bagging Classifier
fromsklearn.linear_modelimport LogisticRegression

# Define the base logistic regression model
base_model = LogisticRegression()

# Define the bagging meta-estimator with logistic regression as the base model

bagging_model = BaggingClassifier(base_model, n_estimators=50, random_state=42)

# Train the bagging model

bagging_model.fit(X_train, Y_train)

# Predict on the testing set

y_pred_bagging = bagging_model.predict(X_test)

# Evaluate the model
accuracy=accuracy_score(Y_test,y_pred_bagging)
print('Accuracy:', accuracy)
```

Table II: Experimental Results for ML Algorithms

Machine Learning Algorithms	Prediction Accuracy
Bagging (Ensemble Approach)	95%
Logistic Regression	94%
Gradient Boosting	93%
Random Forest	92%
Decision Tree	88%
Support Vector Machine	90%
KNN	68%

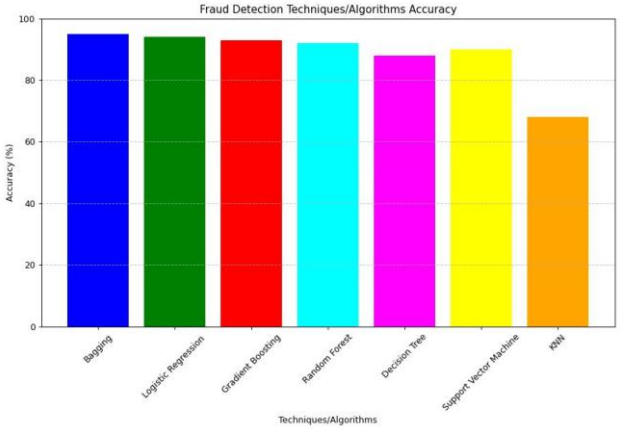

Fig. 4: Comparison Between Different Algorithms

V. CONCLUSION

Credit card fraud is a serious issue for both consumers and financial organizations. Fraudulent transactions can result in financial losses and harm the reputation of financial institutions. Machine learning techniques have been used greatly to detect fraudulent transactions. In this paper, we use logistic regression, Support vector machine, Decision Tree, Random Forest Algorithm, KNN, and Gradient Boosting Technique to classify transactions as either legitimate or fraudulent based on their features. Their accuracy showed that Logistic Regression with high accuracy 94% performed well as compared to the others. But there is a technique called Bagging through which we improved our accuracy by 94 to 95%. The paper concludes that the Logistic Regression, Bagging, and Gradient Boosting perform better as compared to Decision Tree and other algorithms. The limitations and future recommendations include that the Logistic Regression Algorithm and Gradient Boosting, Bagging can also be tested on different datasets and in different domains for credit card fraud detection for high accuracy.

REFERENCES

[1] Trivedi, N.K., Simaiya, S., Lilhore, U.K. and Sharma, S.K., 2020. An efficient credit card fraud detection model based on machine learning methods. International Journal of Advanced Science and Technology, 29(5), pp.3414-3424.

[2] Islam, S., Haque, M.M. and Karim, A.N.M.R., 2024. A rule-based machine learning model for financial fraud detection. International Journal of Electrical & Computer Engineering (2088-8708), 14(1).

[3] Dornadula, V.N. and Geetha, S., 2019. Credit card fraud detection using machine learning algorithms. Procedia computer science, 165, pp.631-641.

[4] Maniraj, S.P., Saini, A., Ahmed, S. and Sarkar, S., 2019. Credit card fraud detection using machine learning and data science. International Journal of Engineering Research, 8(9), pp.110-115.

[5] Alarfaj, F.K., Malik, I., Khan, H.U., Almusallam, N., Ramzan, M. and Ahmed, M., 2022. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access, 10, pp.39700-39715.

[6] Khalid, A.R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J. and Adejoh, J., 2024. Enhancing credit card fraud detection: an ensemble machine learning approach. Big Data and Cognitive Computing, 8(1), p.6.

[7] Tiwari, P., Mehta, S., Sakhuja, N., Kumar, J. and Singh, A.K., 2021. Credit card fraud detection using machine learning: a study. arXiv preprint arXiv:2108.10005.

[8] Itoo, F., Meenakshi and Singh, S., 2021. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13(4), pp.1503-1511.

[9] Afriyie, J.K., Tawiah, K., Pels, W.A., Addai-Henne, S., Dwamena, H.A., Owiredu, E.O., Ayeh, S.A. and Eshun, J., 2023. A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decision Analytics Journal, 6, p.100163.