[[1]](#footnote-1)

Credit Card Fraud Detection

Samarth Jain  
*Data Science and Artificial Intelligence  
Indian Institute of Technology Roorkee*

*Abstract*— In this evolving world credit card fraud poses a significant threat to financial institutions and consumers alike, necessitating fraud detection systems. In this project, we explore the efficiency of various machine learning algorithms in detecting credit card fraud , namely, XGBoost, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, AdaBoost, Gradient Boosting, and Artificial Neural Network (ANN) algorithms. Our study utilizes a comprehensive dataset containing labeled instances of fraudulent and legitimate credit card transactions. We have employed the F1 score, precision, and recall as evaluation metrics to assess the effectiveness of each model in identifying fraudulent transactions. Through extensive experimentation and analysis, we identify the strengths and weaknesses of each machine learning technique in the context of credit card fraud detection.

*Index Terms*— XGBoost, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, AdaBoost, Gradient Boosting, Artificial Neural Network (ANN), F1-score, Precision, Recall.

# INTRODUCTION

I

n an increasingly evolving world, where cashless transactions have become the new normal, the constant threat of credit card fraud have posed significant challenges to financial institutions and consumers alike. Detecting and thwarting credit card fraud is not just a matter of financial prudence but also a critical aspect of safeguarding individuals' and businesses' financial well-being.

This study focusses into the domain of credit card fraud detection, using a number of machine learning techniques to achieve this goal.

The existing fraud detection mechanisms often struggle to keep pace with the evolving new tactics of fraudsters, resulting in increasing financial losses and compromising security. Thus, there exists an urgent need to find advanced machine learning algorithms that can look into the accuracy and efficacy of fraud detection system.

Firstly, by using a number of machine learning algorithms—including XGBoost, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, AdaBoost, Gradient Boosting, and Artificial Neural Network (ANN)—to rigorous evaluation, we try to find the best approaches for identifying fraudulent transactions. Secondly, by assessing the performance of these models using metrics such as F1 score, precision, and recall, we find the insights into the strengths and limitations of each algorithm in real-world scenarios.

# Procedure

## Data Exploration and Preparation

For our study we utilize credit card fraud detection dataset , which is available on Kaggle. We conduct thorough examination of the dataset's dimensions, attribute types, and summary statistics, getting a good understanding of its composition. Additionally, we perform checks for missing values within the dataset to ensure data integrity.

## Data Pre-processing

Recognizing the importance of addressing class imbalance, we increase the minority dataset by duplicating the values. Furthermore, to standardize the dataset for model training, numerical features such as Time and Amount are normalized to a uniform scale relative to other features.

## Model Building

Our study uses the implementation of a different machine learning models, the details of which are as follows. Artificial Neural Network (ANN) architecture is a 2 layer neural network imlemeted using TensorFlow and keras, and sigmoid activation function. The batch normalization and dropout layers are included to prevent overfitting. XGBoost model consisting of array of decision trees, is trained to minimize the gradient of a chosen objective function. Comprising multiple decision trees trained on random subsets of the data with replacement, the Random Forest model aggregates individual tree predictions to reduce variance and improve generalization. The KNN classifies each data point based on the majority class of its k nearest neighbors in feature space. The SVM maximizes the margin between support vectors and fine-tuning of kernel parameters and regularization techniques for increasing classification accuracy. The logistic regression is used to estimate probabilities using a logistic function by using regularization techniques such as L1 and L2 regularization to prevent overfitting. AdaBoost is used train a series of weak learners, with each subsequent learner focusing on instances which were misclassified by the previous one. Gradient boosting is sued to construct an ensemble of weak learners, improving on errors of previous models using gradient descent optimization.

*.*

## Model Comparison

We compare the performances of different machine learning models for credit card fraud detection using metrics like F1 score, precision, and recall to find the analysis in training and testing datasets.

# Results

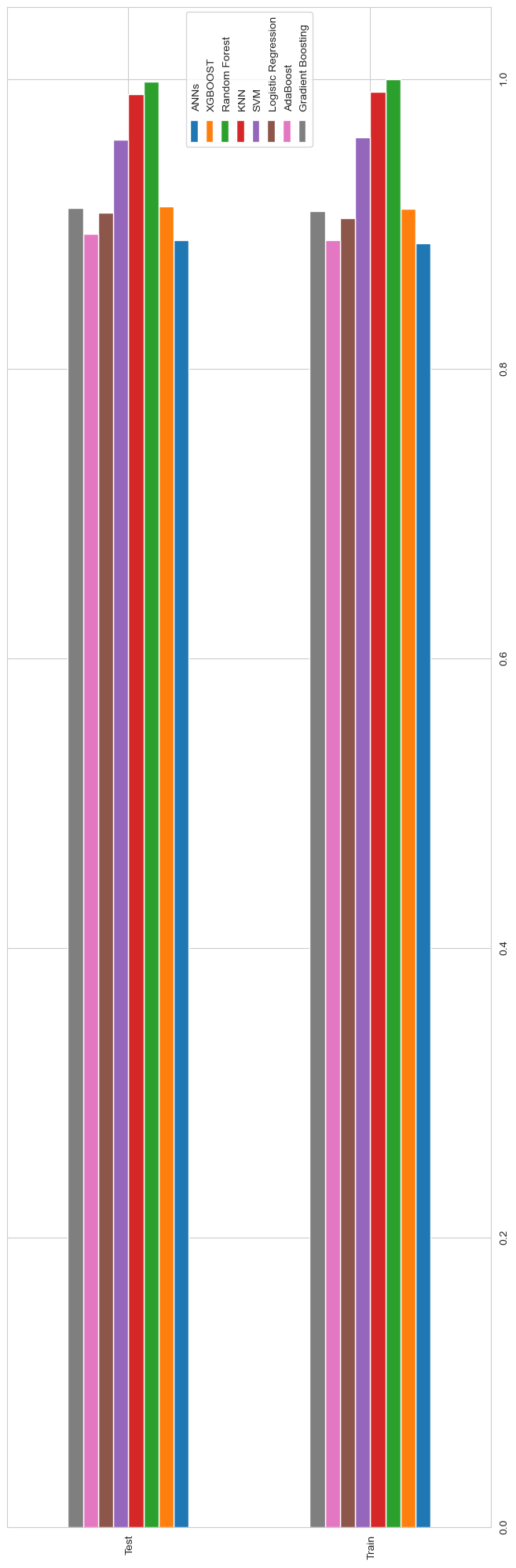
Our analysis reveals performance profiles along a number of machine learning algorithms. Among the models evaluated, Random Forest provides the best with F1 scores of 0.9999 on the training set and 0.9984 on the testing set. Following which XGBoost also finds the F1 scores of 0.9103 and 0.9123 on the training and testing sets,.

Moreover, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) exhibit F1 scores above 0.95 on both training and testing datasets. Logistic Regression, AdaBoost, and Gradient Boosting models also have F1 scores ranging from 0.8887 to 0.9078 on the testing set. While the ANN model returns the F1 score of 0.8864 and 0.8885.

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Train** | **Test** |
| **ANN** | **0.8864** | **0.8885** |
| **XGBoost** | **0.9103** | **0.9123** |
| **Random Forest** | **0.9999** | **0.9984** |
| **KNN** | **0.9912** | **0.9896** |
| **SVM** | **0.9597** | **0.958** |
| **Logistic Regression** | **0.9041** | **0.9078** |
| **AdaBoost** | **0.8887** | **0.8931** |
| **Gradient Boosting** | **0.9086** | **0.9111** |

# Conclusion

The paper finds the accuracy of different algorithms for credit card fraud detection. Random Forest shows the best accuracy among all, followed by XGBoost, KNN, and SVM. This study will indeed help the consumer and banking system to develop fraud prevention machanisms.



1. [↑](#footnote-ref-1)