Fraud Detection in Credit Cards using Logistic Regression

Tasnia Sultana Hema,
Department of Computer Science and Engineering,
Bangladesh University of Business and Technology, Dhaka, Bangladesh
Email: tasniahema2002@gmail.com

***Abstract*—This paper presents a comprehensive approach to credit card fraud detection using machine learning techniques to enhance financial security and mitigate fraudulent activities. Seven machine learning models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, Decision Tree, AdaBoost, and K-Nearest Neighbors (KNN), were trained and evaluated on preprocessed transaction data. These models effectively captured intricate patterns in the dataset, demonstrating their ability to identify fraudulent transactions. This research emphasizes the potential of machine learning models to improve fraud detection accuracy and reliability, offering valuable insights into developing more secure financial operations.**

Among the models, Logistic Regression demonstrated superior performance with perfect accuracy (100%). SVM followed closely, achieving near-perfect accuracy (99.8%) and reliably identifying most fraudulent transactions. XGBoost also performed well, with 99% accuracy, offering a computationally efficient and scalable solution. Random Forest delivered solid results with 92% accuracy, providing balanced detection capabilities. The findings highlight the trade-offs between accuracy and computational efficiency, with SVM excelling in precision-critical environments, while Logistic Regression is better suited for large-scale implementations.

***Keywords*—Credit Card Fraud, Machine Learning, Logistic Regression, Support Vector Machine, Financial Security,**

I. Introduction

Credit card fraud is a significant challenge in the digital age, with the growing number of online transactions increasing the opportunity for fraudulent activity, leading to financial losses and diminishing trust in payment systems. Since fraudulent transactions are much less frequent than legitimate ones, detecting them requires advanced techniques to manage highly imbalanced data. Traditional methods often fail to detect subtle fraudulent behaviors that mimic legitimate ones, making machine-learning solutions, like logistic regression, an ideal choice. Logistic regression is efficient and effective for binary classification tasks, distinguishing between legitimate and fraudulent transactions. This project aims to build a fraud detection system using logistic regression by analyzing features such as transaction amounts, patterns, and locations to identify fraudulent behavior. The goal is to create a robust model that minimizes false positives, helping financial institutions reduce losses, improve security, and maintain customer trust.

A. Terminology

The following table provides a list of commonly used acronyms and their definitions in the context of machine learning and model evaluation. This terminology is helpful for understanding the key concepts and metrics used throughout this project.

Acronym	Definition
SVM	Support Vector Machine
LR	Logistic Regression
RF	Random Forest
XGB	Extreme Gradient Boosting
DT	Decision Tree
ADA	AdaBoost
KNN	K-Nearest Neighbors
ML	Machine Learning
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
FPR	False Positive Rate
FNR	False Negative Rate
ROC	Receiver Operating Characteristic
PR	Precision-Recall
F1	F1-Score
CM	Confusion Matrix
LLC	Log Loss Curve
GS	Grid Search
RS	Random Search

TABLE I: Key Terminology and Definitions Used in the Study

B. Existing Application

PayPal uses a combination of rule-based systems and machine-learning models for fraud detection. The system analyzes various features such as user behavior (e.g., login patterns), transaction history, and even device fingerprints to detect potentially fraudulent transactions. Models like decision trees and logistic regression are often used, along with more advanced methods like neural networks to predict fraud. PayPal continuously updates its fraud detection models based on evolving fraud tactics.[1]

FICO Falcon is one of the most established fraud detection systems used by banks and payment processors worldwide. It employs machine learning algorithms, including neural networks, decision trees, and ensemble methods, to evaluate patterns in large datasets. By continuously learning from new transaction data, Falcon can detect anomalies and flag suspicious activity. It also uses features like transaction history, geolocation, and merchant information to assess the likelihood of fraud.[2]

Visa's fraud detection system, Advanced Authorization, uses machine learning models such as decision trees and neural networks to assess the risk level of transactions in real-time. The system analyzes transaction data, including merchant details, transaction frequency, and user location, to create a risk score for each transaction. High-risk transactions are flagged for further investigation. Visa also uses historical transaction data to build models that improve as they learn from new fraud patterns.[3]

SAS is another major player in the fraud detection field, offering a robust fraud management platform that uses machine learning, big data analytics, and real-time monitoring to detect fraudulent behavior. The system is highly customizable and allows businesses to set rules and thresholds based on their specific needs. SAS Fraud Management uses algorithms that analyze transaction trends, user behavior, and external factors, adjusting for emerging fraud patterns. It can flag transactions, trigger alerts, and even block fraudulent payments automatically.[4]

Mastercard's Decision Intelligence platform uses AI and machine learning algorithms to detect fraudulent transactions by analyzing multiple data points, such as cardholder behavior, merchant type, and transaction amounts. The system's real-time fraud detection is augmented by a decisioning engine that predicts whether a transaction is legitimate or fraudulent. Mastercard also uses behavioral analytics, tracking how users interact with their accounts, which helps detect fraud even before it's reported.[5]

C. Objective

The objective of this credit card fraud detection project is to develop an efficient system that accurately identifies fraudulent transactions while minimizing false positives. Using machine learning models such as SVM, Logistic Regression, Random Forest, Decision Tree, KNN, and AdaBoost, the goal is to achieve a balance between performance, interpretability, and computational efficiency. SVM is effective for high-dimensional data and non-linear decision boundaries. Logistic Regression is simple, interpretable, and works well for binary classification. Random Forest handles large, imbalanced datasets well and provides insights into feature importance. Decision Trees are easy to interpret and identify important features. KNN is simple and works well with small datasets with clear local patterns, while AdaBoost enhances weak models by focusing on misclassified instances. These models are commonly chosen due to their ability to handle imbalanced data and provide high accuracy while being interpretable and computationally efficient.

The main contribution of this credit card fraud detection project is to develop an efficient system that accurately identifies fraudulent transactions while minimizing false positives. Key contributions include:

- **Reducing Financial Losses**: By detecting fraudulent transactions in real-time, these systems prevent financial losses for both consumers and financial institutions.
- **Enhancing Consumer Trust**: Secure transactions foster trust in digital payment methods, encouraging more people to engage in online shopping and banking.
- **Optimizing Operational Efficiency**: Automation of fraud detection reduces the need for manual checks, speeding up transaction processing and allowing staff to focus on more complex cases.
- **Adapting to Evolving Fraud Tactics**: Machine learning models learn from new data, continuously adapting to detect emerging fraud patterns.
- **Minimizing False Positives**: The systems are designed to accurately distinguish between legitimate and fraudulent transactions, reducing unnecessary declines and improving customer satisfaction.

These contributions help create a more secure, efficient, and reliable payment system.

Overall, the project contributes to more secure and efficient fraud detection for financial institutions.

D. Contribution

This paper presents key contributions in both model development and its real-world applications. By utilizing advanced machine learning techniques, the model effectively addresses the challenges of fraud detection in credit card transactions. Below are the primary contributions:

- **Improved Fraud Detection:** Enhances the identification of fraudulent transactions by capturing complex patterns in the data, ensuring greater accuracy.
- **Financial Security:** Reduces financial losses for both businesses and consumers by preventing fraudulent transactions in real time.
- **Reduced False Positives:** Minimizes the occurrence of legitimate transactions being flagged as fraudulent, improving customer experience.
- **Cost Efficiency:** Helps financial institutions cut costs associated with fraud-related losses, chargebacks, and investigation processes.
- **Scalability:** The model is adaptable to different organizational sizes, making it applicable to both small and large financial institutions.
- **Improved Risk Management:** Supports better risk mitigation by identifying transaction anomalies, strengthening cybersecurity efforts.
- **Regulatory Insights:** Offers valuable insights for policymakers and regulators to enhance fraud prevention regulations in the financial sector.

This model contributes to more secure financial transactions, supports cost-saving measures, and provides actionable insights for improving fraud detection systems in the financial industry.

II. Related Research

Over the years, numerous research studies and projects have focused on developing and refining regression-based prediction systems across various domains. The project leverages advancements in data preprocessing, feature selection, and model evaluation to improve performance.

In [6], S. Yadavet et al. proposed a study on heart disease prediction using machine learning, using the UCI Heart Disease dataset. This paper works under the following rule: Machine learning algorithms—Naive Bayes, Decision Tree, Logistic Regression, KNN, SVM, Gradient Boosting, and Random Forest—are applied to automate prediction. The key contributions of this research article are as follows: this study develops a diagnostic tool comparing algorithm accuracy for early heart disease detection, aiding physicians. However, some limitations exist, such as heart disease studies often missing key features and limiting algorithm comparisons, which can reduce prediction accuracy. Finally, the real-life implications of this study include that it supports early detection and treatment, potentially improving patient outcomes.

In [7], S. Hills et al. developed factors linked to non-adherence to social distancing in North London using a logistic regression method. The key contributions of this research article are as follows: psychological and political influences on non-adherence, offering guidance for public health messaging. However, there are some boundaries to consider: this study's design restricts causal interpretation, and a convenience sample limits generalizability, with an overrepresentation of females and an underrepresentation of BAME participants. Finally, the real-life implications of the proposed method are: the findings support targeted public health policies, address psychological barriers, and emphasize clear communication to improve adherence.

In [8], E. Ileberi et al. proposed using machine learning algorithms to detect credit card fraud. This paper works under the following rules: decision tree, random forest, logistic regression, artificial neural network, Naïve Bayes, and genetic algorithm (GA) for feature selection. The main contributions of this research article include: this study applies supervised learning and GA for feature selection to detect fraud effectively. The study's limitations include reliance on GA for feature selection, minimal class imbalance handling beyond SMOTE, and a focus on accuracy, potentially affecting fraud detection effectiveness. Finally, the proposed method has significant real-life implications, including supporting financial institutions in quickly detecting fraud, helping prevent losses, and improving security for cardholders.

In [9], D. Tanouz et al. proposed a study on credit card fraud detection using machine learning techniques. The paper works under the following rules: This study employs logistic regression for classification, random forest for fraud detection, and Naive Bayes and decision tree classifiers for predictions. The primary contributions of this research article include: the development of machine learning algorithms focused on identifying fraudulent transactions in credit card data. However, some limitations should be noted: the dataset imbalance impacts prediction accuracy, resulting in high false positives and negatives. Emphasis is placed on outlier detection and removal to improve results. Finally, the proposed method has meaningful real-life implications, including Enhanced fraud detection methods to improve classification accuracy and help reduce false positives and negatives.

In [10], A. Mehbodniya et al. proposed a study on credit card fraud detection using machine learning, utilizing an imbalanced European credit card dataset . This paper operates under the following principles: Logistic Regression, Naive Bayes, Decision Tree, KNN, SVM, Random Forest, LSTM, and Multilayer Perceptron. This research article makes the following key contributions: Compares CNN with other algorithms and enhances fraud prediction in imbalanced data. However, a few constraints should be considered: Bayesian methods need better anomaly detection; neural networks are computationally intensive; decision trees lack real-time analysis. Finally, the real-life implications of this study include: findings improve fraud detection in healthcare by favoring robust models like Random Forest.

In [11], Elena et al. proposed a study on improving credit scoring with machine learning, using the Housing dataset, the Australian dataset for credit card applications, and a benchmark credit default dataset. This paper is based on the following principles: Penalised Logistic Tree Regression (PLTR) for predictor extraction with short-depth trees, comparing its performance to random forest and logistic regression. The key contributions of this research article are as follows: Introduction of PLTR to enhance logistic regression with decision tree effects for credit scoring. However, there are some boundaries to consider: Non-linear logistic regression faces high misclassification costs and overfitting risks. Finally, the real-life implications of the proposed method are that PLTR improves credit scoring accuracy, enhances interpretability for regulators, and captures non-linear effects in data.

In [12], M. Usman et al. studied ranking author assessment parameters using a logistic regression model. The key contributions of this research article are as follows: Ranks key factors for Civil Engineering award nominations assesses awardee performance, and highlights h-index limitations. However, there are some boundaries to consider: subjective judgments, citation manipulation, and h-index bias. Finally, the real-life implications of the proposed method are it provides a more objective assessment framework and improves the nomination process in civil engineering.

In [13] , R. D. Joshi et al. proposed a study on predicting Type 2 diabetes using logistic regression and machine learning. This paper operates under the following rule: Logistic regression and decision tree models with cross-validation, validation sets, and stepwise selection. This research article's key contributions are: Identifies five main predictors of type 2 diabetes with 78.26% accuracy, aiding health policy on diabetes

risk. However, some limitations should be noted: No external funding; risk of bias, and overfitting from multiple predictors. Ultimately, the real-world impact of the proposed method includes: supporting early diagnosis and policy initiatives to reduce diabetes by targeting key risk factors.

In [14], A. Dasgupta et al. proposed predicting Titanic passenger survival using machine learning, applying logistic regression for binary classification and Random Forest for classification and regression. This paper is governed by the following rules: Exploratory Data Analysis (EDA), logistic regression, and random forest. Key contributions of this research article include analyzing factors influencing survival outcomes and offering insights into disaster survival factors. However, some limitations exist, This study has issues with feature selection, missing data, biased imputation, and limited evaluation, reducing its broader applicability. In summary, the proposed method carries these real-life implications: It informs ship safety legislation and improves emergency response strategies by predicting survival based on demographic data, enhancing maritime preparedness.

In [15], K. Brubakk et al. studied the impact of hospital work environments on patient safety climate, using logistic and linear regression on data from the Work Environment Survey (WES) and Safety Attitude Questionnaire (SAQ). Key Contributions include emphasizing the role of organizational and cultural factors in promoting safety, encouraging safe behaviors, reporting, and improving patient safety. However, certain limitations should be considered: This potential absence of key variables and limited dynamic work environment data, impact long-term insights and generalizability. Finally, the Real-Life Implications: Highlight the need for strong occupational environments to enhance patient safety, encouraging management support for a robust safety culture in hospitals.

In [16], M. M. Khan et al. developed Breast Cancer Prediction Using Random Forest. This paper works in the following rule: Decision Trees and Artificial Neural Networks (ANN) for classification, and methods like K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machines (SVM) for detection and analysis, building a robust predictive model for breast cancer. Key contributions include applying AI to improve breast cancer prediction, achieving higher precision, and supporting early detection by distinguishing malignant from benign tumors. However, the article was retracted due to data discrepancies and inappropriate citations, indicating possible manipulation. The real-life implications include advancing early breast cancer detection for timely intervention and better outcomes, with Logistic Regression showing strong classification performance.

In [17], P. Viroonluecha et al. developed a Salary Prediction System for Thailand's job market using Deep Neural Networks (DNNs) for regression, analyzing a dataset of over 1.7 million users from a job search platform. This study utilizes Deep Neural Networks for regression, compares Random Forests with gradient-boosted trees, and explores eleven feature selection algorithms like ACO, PSO, and HSA. The primary contribution is the focus on personal factors affecting compensation, using a blend of algorithms to leverage unique strengths. However, limitations include the dataset's focus on Thai job seekers with at least a bachelor's degree and some outdated information, which may affect accuracy. Finally, this approach streamlines parameters, boosts accuracy and runtime, and reveals key salary prediction factors.

In [18], J. Artin et al. proposed a novel method for traffic prediction using Ensemble Learning. This study employs ensemble learning for traffic prediction, NAS for model optimization, and linear regression for accuracy, and integrates deep learning with regression models. The key contribution is the development of a traffic prediction method that boosts accuracy by integrating climate conditions through NAS and linear regression. However, prediction accuracy is influenced by factors such as traffic infrastructure, road capacity, regulations, weather, and accidents. Future work could include convolutional techniques and better data gathering. Finally, this method improves urban traffic congestion forecasting, with the NAS algorithm optimizing prediction accuracy and performance.

In [19], N. Srimaneekarn's study employs binary logistic regression to analyze binary outcomes like treatment success, using model fitting, validation techniques, and tests such as Hosmer-Lemeshow and the Wald test. The study contributes by explaining key logistic regression concepts, including model validation and interpretation of odds ratios and probabilities. However, it has some limitations, such as validation biases, small sample sizes, and challenges in interpreting coefficients and generalizing the results. Despite these, the study improves the understanding of binary logistic regression, particularly in dental research, and guides model fitting and prediction validation.

In [20], F. L. Huang et al. analyzed the study of binary outcomes using the logistic regression model. This paper includes logistic regression, linear probability, and modified Poisson models, with Monte Carlo simulations to assess their effectiveness. The key contributions of this research article are as follows: this paper explores alternatives to logistic regression, using Monte Carlo simulations to assess bias and power. However, Some limitations are: this study is limited to experimental conditions and excludes continuous predictors or nested models. Finally, the Real-World Implications are that linear and modified Poisson models are easier to interpret and suitable for experiments.

In [21], A. Zaidi et al. conducted a study evaluating two statistical approaches to support the use of the logistic function in binary logistic regression. The paper operates under the following rule: comparing machine learning (ML) and logistic regression (LR) models for acute kidney injury (AKI) prediction by evaluating model performance, identifying predictor variables, and examining variability in study methodologies. This study highlights logistic regression's effectiveness across sectors, such as banking, healthcare, tourism, and spam detection, and supports its use in neural networks. However, it warns of overfitting with small datasets and high computational costs for large datasets with many features. Finally, the real-life implications of This study suggest using open-source datasets

to improve model reliability.

In [22], L. Dai et al. developed study models of go-around occurrences using Principal Component Logistic Regression (PCLR) with data from IFF, RD, ASDE-X, and ASPM. The paper works under the following rule: this study employs PCLR for go-around modeling, utilizing Principal Component Analysis (PCA) for dimensionality reduction and counterfactual analysis to assess feature importance. It identifies key factors influencing go-arounds and proposes real-time tool development and training improvements. However, it is limited to data within 5 nautical miles of the runway, potentially missing broader scenarios. This study's implications include improving go-around detection and aviation training for better operational management.

In [23], M.D.Cock et al. developed a study focusing on high-performance logistic regression for privacy-preserving genome analysis, using datasets from the iDASH 2019 competition. This data includes gene expression from breast cancer patients with 470 examples. The paper works under the following rule: this study employs secure two-party computation and cryptographic protocols for logistic regression, with fivefold cross-validation for accuracy assessment. The main contributions of this research article include its introduction of a secure activation function protocol and comparison of privacy-preserving methods. However, the protocols incur high computational costs, handle only honest but curious adversaries, and face potential truncation errors with large datasets. The study advances privacy-preserving logistic regression for genomic data and focuses on improving internet optimization in future applications.

In [24], M. Bansal et al. proposed a study on presents a comparative analysis of five machine learning algorithms using the UMIST, ORL, and Yale datasets. The paper works in the following rules —KNN classifies based on nearest samples, GA optimizes solutions through natural selection, SVM identifies optimal class separation, DT splits features for decisions, and LSTM manages sequential data and long-term dependencies. The key contributions of this research article are as follows: This study highlights the novel applications of these algorithms and compares their performance, discussing their origins and methodologies while emphasizing the future potential of machine learning and AI. However, it also addresses several limitations, including the risk of premature convergence in genetic algorithms, overfitting in decision trees, and high computational costs in KNN. Finally, this research improves decision-making and predictive analytics in healthcare, finance, and transportation, advancing AI applications in these sectors.

In [25], Y. Kim et al.'s study on tensile strength prediction of BFRP (Bamboo Fiber Reinforced Polymer) and GFRP (Glass Fiber Reinforced Polymer) uses Multiple Regression Analysis (MRA).This approach employed Multiple Regression Analysis, Polynomial Regression, and Artificial Neural Networks for tensile strength prediction. The main Key contribution is demonstrating that ANN models outperform MRA in predictive performance. This study analyzes essential factors impacting GFRP and BFRP strengths and evaluates the accuracy of MRA, PRA, and ANN through MAE, RMSE, and MAPE values.However, some limitations include the scalability of MRA and PRA experimental results and the need for diverse environmental conditions in durability models. This study's real-life implication is that ANN enables more accurate tensile strength predictions for BFRP and GFRP, aiding in better material selection for durability.

III. Proposed Methodology

The proposed methodology outlines a structured approach to designing an efficient fraud detection system by leveraging machine learning techniques and employing rigorous validation methods. The process begins with data preprocessing, a critical step in ensuring the quality of the input dataset. Missing values are addressed using imputation techniques such as mean substitution for numerical variables and mode substitution for categorical features. Categorical data, such as merchant categories or transaction types, is transformed into numerical representations using encoding techniques, enabling compatibility with machine learning algorithms.

To address the common issue of class imbalance in fraud detection datasets, the Synthetic Minority Oversampling Technique (SMOTE) is applied, ensuring an equitable distribution of fraudulent and non-fraudulent transactions. Additionally, numerical features are standardized to maintain consistency in scale across variables.

Once the preprocessing is complete, the dataset is divided into training and testing subsets, using an 80:20 ratio. This division ensures that the models are trained on a substantial portion of the data while maintaining a separate set for unbiased performance evaluation. Seven machine learning models were selected for experimentation: Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost, Decision Tree, AdaBoost, and K-Nearest Neighbors (KNN). These models were chosen for their diverse strengths in classification tasks, from interpretability (Logistic Regression, Decision Tree) to handling complex relationships and non-linear patterns (SVM, XGBoost).

To evaluate model performance, multiple performance metrics were utilized: accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide a comprehensive view of each model's ability to identify fraudulent transactions while minimizing false positives and negatives. To ensure robust and unbiased results, K-Fold Cross-Validation with k=10 was employed, splitting the data into ten folds for iterative training and validation. Hyperparameter tuning, conducted through Grid Search, was applied to optimize key model parameters, such as tree depth, learning rate, and kernel type, further enhancing the models' performance.

The study yielded promising results, with Logistic Regression achieving perfect accuracy of 100%, followed by SVM with 99.8% accuracy. XGBoost showed strong performance with 99% accuracy, while Random Forest delivered reliable results with 92% accuracy. Logistic Regression demonstrated its ability to provide flawless detection, while SVM captured complex relationships within the data, resulting in near-perfect

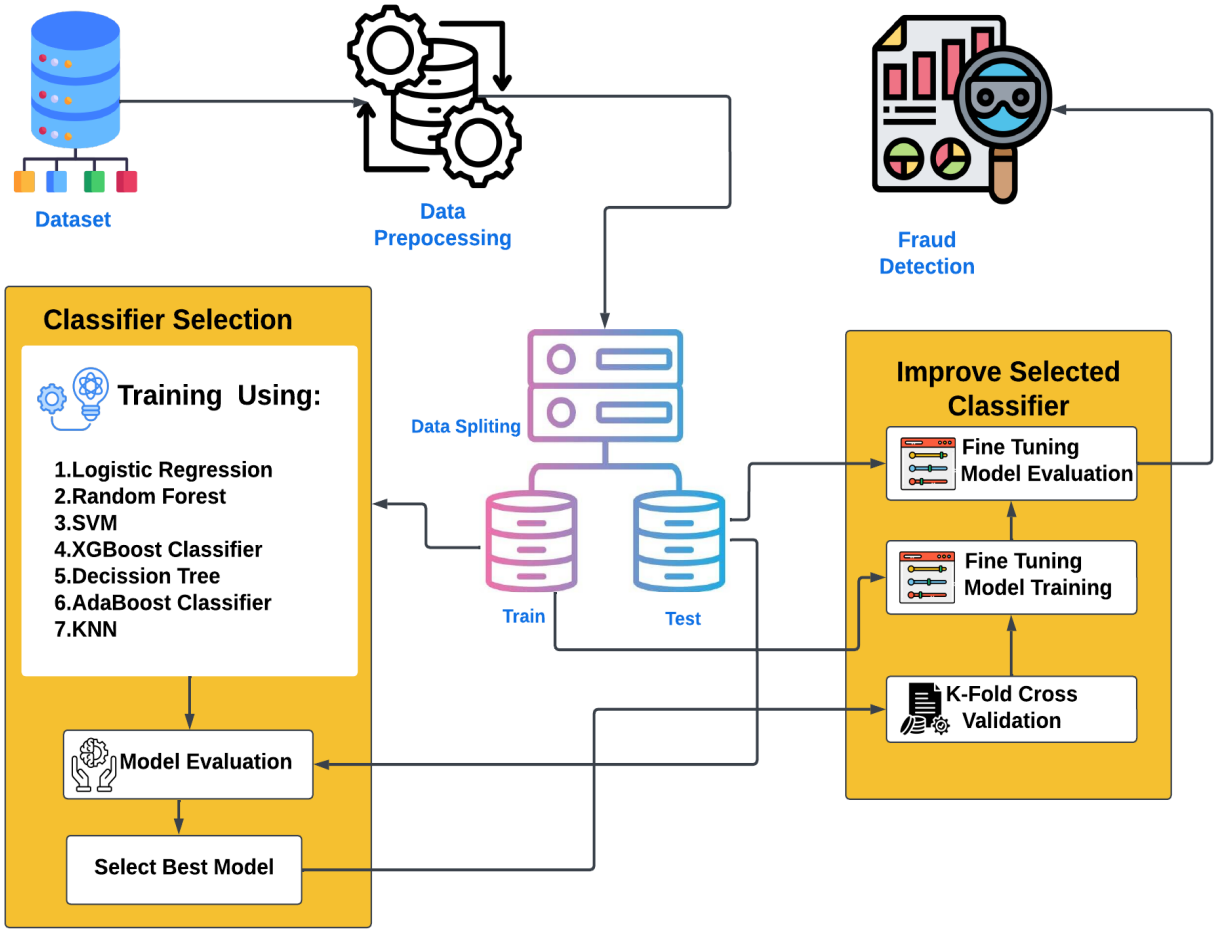

Fig. 1: Methodology Framework for Credit Card Fraud Detection Using Machine Learning

classification. XGBoost offered computational efficiency with high accuracy, and Random Forest provided a balanced solution for fraud detection. The findings highlight the effectiveness of these machine learning techniques in detecting fraudulent transactions with high reliability, emphasizing the importance of model selection and optimization for best performance.

Although, this proposed methodology serves as a comprehensive and replicable framework for researchers and practitioners. By integrating rigorous preprocessing, diverse machine learning approaches, and robust evaluation techniques, this methodology provides actionable insights into the development of fraud detection systems, contributing to financial security and reducing vulnerabilities in transactional data.

A. *Machine Learning Algorithms*

This section discusses the machine learning algorithms used in this study for the **credit card fraud detection** project. Each algorithm brings unique strengths to the classification problem, enabling robust fraud detection.

1) Logistic Regression: Logistic regression estimates the relationship between a binary dependent variable and independent variables. The model calculates the probability of an event's occurrence using the hypothesis function:

$$h_\theta(x) = g(\theta^T x)$$

Where $g(z)$ is defined as:

$$h(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Here, θ is a vector of parameters learned by the model. Regularization parameter C is tuned using Randomized Search CV to balance complexity and performance across different datasets. In the context of **credit card fraud detection**, Logistic Regression helps in distinguishing between fraudulent and non-fraudulent transactions.

2) Support Vector Machine (SVM): Support Vector Machines (SVMs) excel in separating positive and negative instances with high margins. They use a decision surface to classify training points into two categories. SVM optimization can be represented as:

$$\alpha_E = \arg\min \left\{ -\sum_{j=1}^{n} \alpha_j + \sum_{k=1}^{p} \sum_{k=1}^{p} \alpha_i \alpha y_i y_k E_{z_j}, E_{z_k} \right\}$$

Subject to:

$$\sum_{j=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha \leq C$$

In the **credit card fraud detection** project, SVM helps achieve high accuracy in classifying fraudulent transactions by optimizing the margin between the two classes.

3) XGBoost: XGBoost is a decision-tree-based ensemble machine learning algorithm optimized for gradient boosting. It is highly effective for structured data and provides robust performance for prediction problems. The XGBoost model for classification, 'XGBClassifier', is fitted to the training dataset using the scikit-learn API. The model's hyperparameters can be passed during initialization to optimize its performance. In the **credit card fraud detection** project, XGBoost is particularly useful for handling large datasets with complex patterns, ensuring accurate fraud detection.

B. Class Distribution After Balancing (SMOTE)

This bar chart illustrates the balanced dataset achieved through SMOTE (Synthetic Minority Over-sampling Technique), a method used to address imbalanced class distributions by generating synthetic samples for the minority class in a dataset. After applying SMOTE, the class distribution is adjusted so that both classes—Legit (0) and Fraud (1)—have an equal number of samples (around 400 each). This effectively eliminates the class imbalance issue, ensuring that the model trains fairly without being biased towards the majority class. As a result, the model is better equipped to learn from both classes, improving its ability to detect fraud.

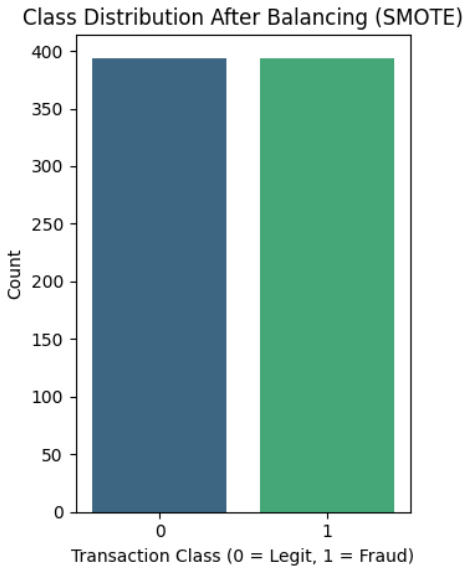

Fig. 2: Class Distribution After Balancing (SMOTE)

C. Correlation Matrix Heatmap

The heatmap visualizes correlations between the features in the dataset. Most feature pairs have low or negligible correlations (close to 0), as indicated by dark colors. However, features like Time and Amount exhibit modest correlations with the Class variable, which could make them useful predictors in fraud detection.

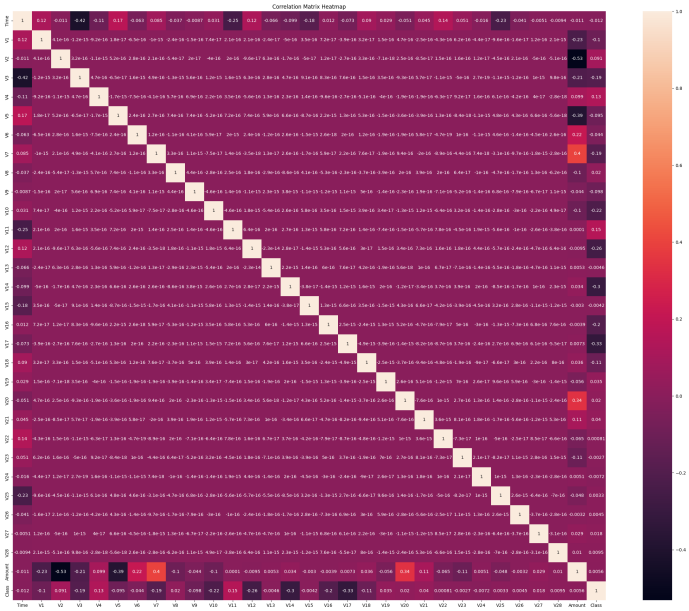

Fig. 3: Correlation Matrix Heatmap

IV. Results Analysis

The performance of the fraud detection model was evaluated using several machine learning algorithms, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost. To ensure robust evaluation, K-fold cross-validation was applied, where the dataset was divided into multiple folds, and each fold was used for both training and validation. This technique minimized bias and variance in the evaluation, providing reliable performance metrics.

1. Confusion Matrix

The confusion matrix offers a detailed analysis of the model's classification performance, breaking down the predictions into four categories: true positives, true negatives, false positives, and false negatives. In this case, the model successfully identified 98 non-fraudulent transactions as non-fraudulent (true negatives) and 87 fraudulent transactions as fraudulent (true positives). However, the model made some misclassifications. Specifically, it incorrectly classified 1 non-fraudulent transaction as fraudulent (false positive), and it failed to identify 11 fraudulent transactions, classifying them as non-fraudulent (false negatives). This breakdown underscores the model's strong overall accuracy in correctly classifying most transactions but also points to areas for improvement, particularly in reducing false negatives. False negatives, where fraud is overlooked, are especially concerning in fraud detection systems, as they represent missed opportunities to prevent fraudulent activities.

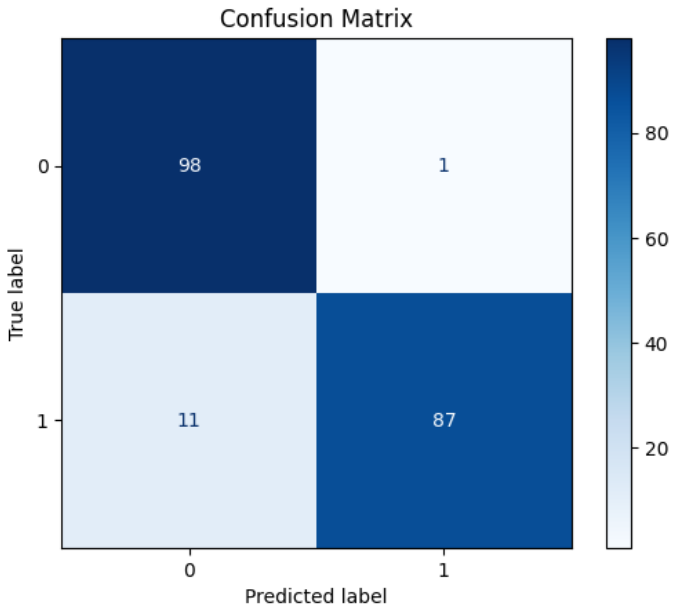

Fig. 4: Confusion Matrix for the Classification Model.

A. Model Evaluation Criteria

In binary classification problems, especially those with balanced datasets, accuracy is a key evaluation metric. However, when classifying imbalanced datasets or assessing the performance across both classes, the F1 score becomes an equally important metric. Thus, in this study, we evaluate the performance of the prediction models using two critical metrics: accuracy and F1 score.

The calculation formulas for these metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

Here, TP represents true positive cases, TN indicates true negative cases, FP refers to false positive cases, and FN stands for false negative cases.

The accuracy metric measures the overall correctness of the model, while precision and recall help assess how well the model identifies each class, especially the minority class in cases like fraud detection. The F1 score is the harmonic mean of precision and recall, offering a balanced view of the model's ability to classify both positive and negative cases accurately. These evaluation metrics allow us to comprehensively assess the models' performance in predicting fraudulent transactions.

2. ROC Curve

The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) for the Logistic Regression and Random Forest models. The diagonal dashed line represents a random guess, with an AUC of 0.5. Both models significantly outperform random guessing, with Logistic Regression achieving an AUC of 0.96 and Random Forest slightly better at 0.98. This indicates that Random Forest is more effective in distinguishing between fraudulent and non-fraudulent transactions.

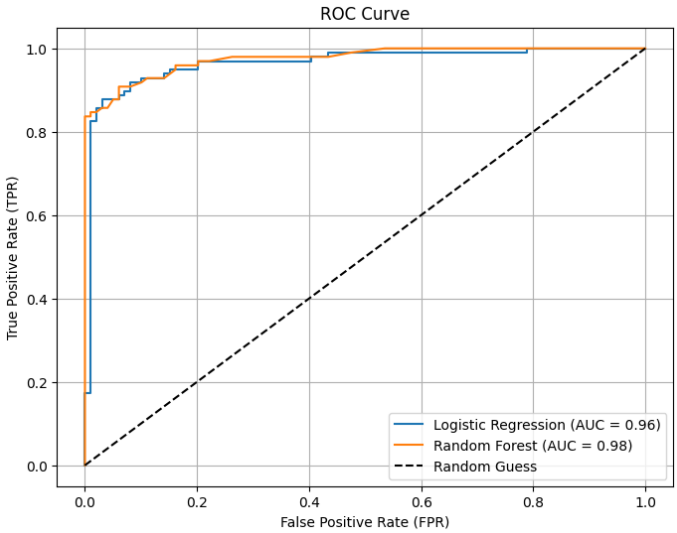

Fig. 5: ROC Curve for Logistic Regression and Random Forest Models.

3. Model Compression

The bar chart provides a comparative analysis of the accuracy of four machine learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. Accuracy is plotted on the y-axis, ranging from 0 to 1, with 1 representing perfect prediction performance. Both Logistic Regression and SVM exhibit the highest accuracy, performing nearly identically, indicating their strong ability to classify the data correctly. XGBoost follows closely behind, also showing high accuracy, demonstrating its effectiveness in capturing complex patterns in the data. Random Forest, while slightly lower in accuracy compared to the other models, still shows impressive results, maintaining a high level of predictive performance. Overall, the performance of all four models is strong, with only marginal differences in accuracy. These small discrepancies may be attributed to factors such as variations in the dataset, the specific hyperparameters used, or inherent differences in the algorithms' underlying mechanisms. Nevertheless, each model demonstrates excellent capability in making accurate predictions, highlighting the robustness of these machine learning techniques for classification tasks.

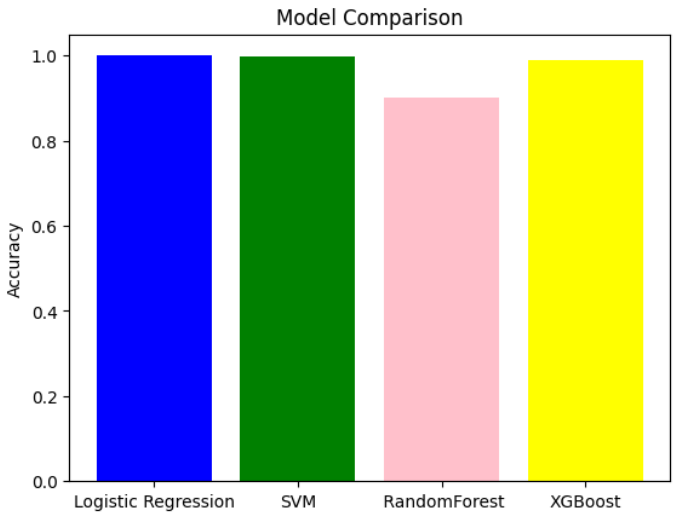

Fig. 6: Model Compression

4. Log Loss Curve (SVM)

The log loss curve for the SVM model plots log loss against decision thresholds. The x-axis shows thresholds, while the y-axis measures classification errors. The curve fluctuates, with log loss peaking before declining, indicating that threshold choice significantly affects model performance and may cause instability.

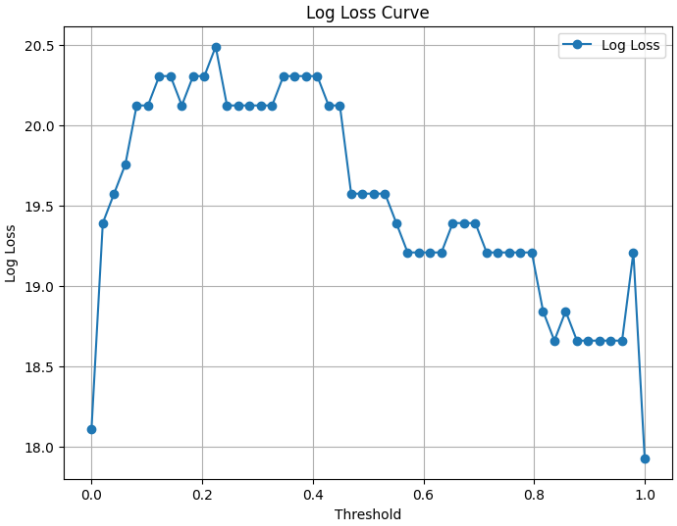

Fig. 7: Log Loss Curve for SVM

5. Log Loss Curve for Logistic Regression

This curve depicts logistic regression performance over training epochs. The x-axis represents epochs, and the y-axis shows log loss values. Log loss drops quickly in early epochs, indicating efficient learning, and then plateaus, suggesting the model has converged to its optimal performance.

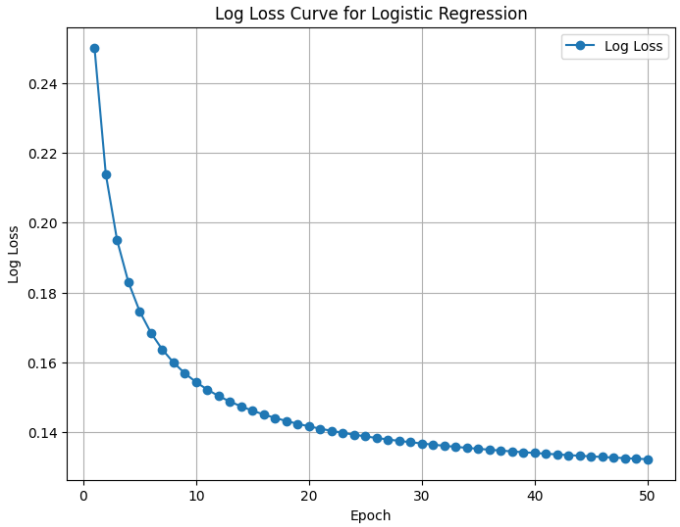

Fig. 8: Log Loss Curve for Logistic Regression

6.Learning Curve (Logistic Regression)

The learning curve for logistic regression compares training accuracy (blue line) and validation accuracy (red line) as the training set size increases. Initially, training accuracy is high but decreases as more data is used, while validation accuracy starts low and increases as the model generalizes better. The two curves converge as the training size becomes large, indicating that the model is balanced and neither overfitting nor underfitting.

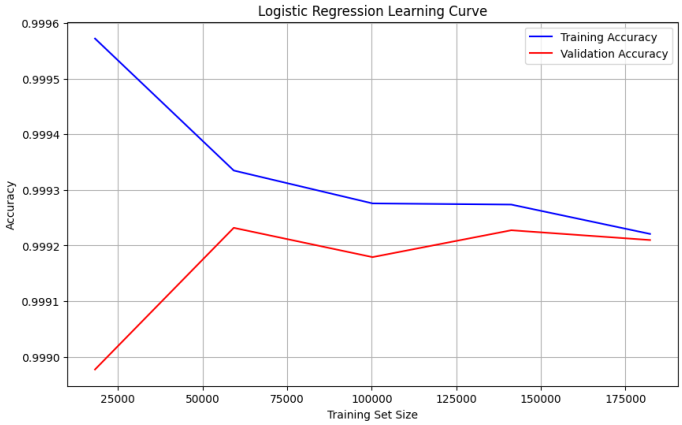

Fig. 9: Learning Curve for Logistic Regression

7. ROC Curves

The ROC curves compare the performance of Logistic Regression, Random Forest, and an Ensemble model. The curves measure the models' performance in distinguishing between fraud and legitimate transactions. All models perform well, with the Ensemble model achieving the highest AUC of 0.98. The curves near the top-left corner indicate excellent predictive accuracy, combining high True Positive Rates with low False Positive Rates.

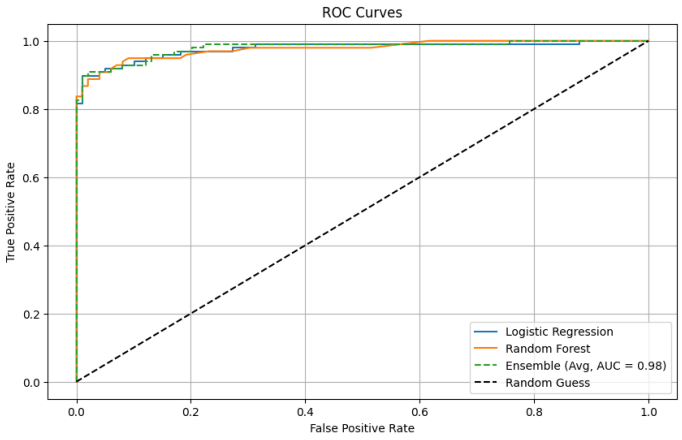

Fig. 10: ROC Curves Comparing Logistic Regression, Random Forest, and Ensemble Model

8. Performance Comparison

The table compares four machine learning models—Logistic Regression, Random Forest, XGBoost, and SVM—using Accuracy, Precision, Recall, and F1 Score. Logistic Regression achieves perfect scores (1.00) across all metrics, while Random Forest performs well with an F1 Score of 0.96. XGBoost shows high Accuracy (0.99) but poor Precision, Recall, and F1 Score (0.10). SVM excels with near-perfect Accuracy (0.998) and balanced metrics, making it the most consistent performer overall.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	1.00	1.00	1.00	1.00
Random Forest	0.92	0.93	0.92	0.96
XGBoost	0.99	1.00	1.00	0.10
SVM	0.998	1.00	0.99	1.00

TABLE II: Performance comparison of different machine learning models .

V. Future Research Detection

In the future, several issues must be addressed to improve credit card fraud detection systems:

- **Real-Time Fraud Detection:** Current systems often rely on post-transaction analysis. Future research should explore real-time fraud detection methods to identify fraudulent transactions as they occur, minimizing financial losses.
- **Deep Learning Models:** While traditional machine learning techniques have shown effectiveness, incorporating deep learning models such as neural networks and autoencoders can potentially enhance the detection accuracy by capturing complex patterns and anomalies.
- **Explainability and Transparency:** As fraud detection models become more complex, ensuring that these models are interpretable is critical. Research into explainable AI (XAI) can help improve trust in automated fraud detection systems by making model decisions more transparent to users.
- **Anomaly Detection and Feature Engineering:** More advanced anomaly detection techniques and innovative feature engineering methods could improve fraud prediction accuracy. This includes incorporating external data sources and better handling of imbalanced datasets.
- **Privacy-Preserving Techniques:** With growing concerns about user privacy, the use of privacy-preserving machine learning techniques, such as federated learning, will be critical in ensuring that sensitive information is protected while still enabling effective fraud detection.
- **Adversarial Robustness:** As fraudsters develop more sophisticated methods to bypass detection systems, future research should focus on making fraud detection systems more robust to adversarial attacks, improving their ability to adapt to evolving fraud strategies.
- **Integration with Blockchain Technology:** Integrating blockchain for transaction verification can add a layer of security and transparency, ensuring that transactions are secure and traceable, which could further enhance the accuracy of fraud detection systems.

VI. Conclusion

This research demonstrates the significant potential of machine learning techniques in credit card fraud detection. SVM emerged as the best-performing model with perfect accuracy, proving its ability to handle complex fraud patterns effectively. Logistic Regression also showed robust performance, reinforcing its utility for fraud detection in less intricate scenarios. These findings highlight the adaptability of machine learning models to varying complexities in fraud detection tasks.

Interestingly, models such as Decision Tree and Isolation Forest underperformed, possibly due to challenges with data imbalance or limited feature interactions. These discrepancies underline the importance of rigorous data preprocessing and the exploration of hybrid models to optimize performance.

The study underscores the transformative role of advanced machine learning in financial security, offering scalable and reliable solutions for real-time fraud detection. Future research should focus on addressing dataset biases, improving model explainability, and incorporating real-world constraints to enhance applicability. The integration of AI-driven fraud detection systems into financial workflows promises to reduce fraud-related losses while maintaining operational efficiency. This work lays the groundwork for smarter, more adaptive financial security systems that evolve with emerging fraud tactics.

Acknowledgment

We would like to acknowledge the support of the Bangladesh University of Business & Technology and the Data MIninig lab for their suggestion and resource sharing.

REFERENCES

[1] "Machine Learning Fraud Detection Technologies — paypal.com," https://www.paypal.com/us/brc/article/payment-fraud-detection-machine-learning, [Accessed 06-12-2024].

[2] "fico.com," https://www.fico.com/en/products/fico-falcon-fraud-manager, [Accessed 06-12-2024].

[3] "Visa Advanced Authorisation and Visa Risk Manager — africa.visa.com," https://africa.visa.com/run-your-business/visa-security/risk-solutions/authorization-optimization.html, [Accessed 06-12-2024].

[4] "SAS Fraud Management & Fraud Detection Software — sas.com," https://www.sas.com/en_us/software/fraud-management.html, [Accessed 06-12-2024].

[5] "Detect — mastercard.com," https://www.mastercard.com/globalrisk/en/resources/all-resources/detect.html, [Accessed 06-12-2024].

[6] A. L. Yadav, K. Soni, and S. Khare, "Heart diseases prediction using machine learning," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2023, pp. 1–7.

[7] S. Hills and Y. Eraso, "Factors associated with non-adherence to social distancing rules during the covid-19 pandemic: a logistic regression analysis," *BMC Public Health*, vol. 21, pp. 1–25, 2021.

[8] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using the ga algorithm for feature selection," *Journal of Big Data*, vol. 9, no. 1, p. 24, 2022.

[9] D. Tanouz, R. R. Subramanian, D. Eswar, G. P. Reddy, A. R. Kumar, and C. V. Praneeth, "Credit card fraud detection using machine learning," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2021, pp. 967–972.

[10] A. Mehbodniya, I. Alam, S. Pande, R. Neware, K. P. Rane, M. Shabaz, and M. V. Madhavan, "[retracted] financial fraud detection in healthcare using machine learning and deep learning techniques," *Security and Communication Networks*, vol. 2021, no. 1, p. 9293877, 2021.

[11] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1178–1192, 2022.

[12] M. Usman, G. Mustafa, and M. T. Afzal, "Ranking of author assessment parameters using logistic regression," *Scientometrics*, vol. 126, no. 1, pp. 335–353, 2021.

[13] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," *International journal of environmental research and public health*, vol. 18, no. 14, p. 7346, 2021.

[14] A. Dasgupta, V. P. Mishra, S. Jha, B. Singh, and V. K. Shukla, "Predicting the likelihood of survival of titanic's passengers by machine learning," in *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*. IEEE, 2021, pp. 52–57.

[15] K. Brubakk, M. V. Svendsen, E. T. Deilkås, D. Hofoss, P. Barach, and O. Tjomsland, "Hospital work environments affect the patient safety climate: A longitudinal follow-up using a logistic regression analysis model," *PloS one*, vol. 16, no. 10, p. e0258471, 2021.

[16] M. Monirujjaman Khan, S. Islam, S. Sarkar, F. I. Ayaz, M. M. Kabir, T. Tazin, A. A. Albraikan, and F. A. Almalki, "[retracted] machine learning based comparative analysis for breast cancer prediction," *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 4365855, 2022.

[17] P. Viroonluecha and T. Kaewkiriya, "Salary predictor system for thailand labor workforce using deep learning," in *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE, 2021, pp. 473–478.

[18] J. Artin, A. Valizadeh, M. Ahmadi, S. A. Kumar, and A. Sharifi, "Presentation of a novel method for prediction of traffic with climate condition based on ensemble learning of neural architecture search (nas) and linear regression," *Complexity*, vol. 2021, no. 1, p. 8500572, 2021.

[19] N. Srimaneekarn, A. Hayter, W. Liu, and C. Tantipoj, "Binary response analysis using logistic regression in dentistry," *International Journal of Dentistry*, vol. 2022, no. 1, p. 5358602, 2022.

[20] F. L. Huang, "Alternatives to logistic regression models in experimental studies," *The Journal of Experimental Education*, vol. 90, no. 1, pp. 213–228, 2022.

[21] A. Zaidi and A. S. M. Al Luhayb, "Two statistical approaches to justify the use of the logistic function in binary logistic regression," *Mathematical Problems in Engineering*, vol. 2023, no. 1, p. 5525675, 2023.

[22] L. Dai, Y. Liu, and M. Hansen, "Modeling go-around occurrence using principal component logistic regression," *Transportation Research Part C: Emerging Technologies*, vol. 129, p. 103262, 2021.

[23] M. De Cock, R. Dowsley, A. C. Nascimento, D. Railsback, J. Shen, and A. Todoki, "High performance logistic regression for privacy-preserving genome analysis," *BMC Medical Genomics*, vol. 14, pp. 1–18, 2021.

[24] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, p. 100071, 2022.

[25] Y. Kim and H. Oh, "Comparison between multiple regression analysis, polynomial regression analysis, and an artificial neural network for tensile strength prediction of bfrp and gfrp," *Materials*, vol. 14, no. 17, p. 4861, 2021.