

Emprical Study on Credit Card Fraud Detection with ML

A IDP project report submitted by

B.Lakshmi Charitha (Reg. No. 221FA20005)

P.Ramesh (Reg. No. 221FA20012)

R.Jyothi Kambika (Reg. No. 221FA20022)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

Computer Science & Business Systems

Under the Guidance of

Ms.U.Naga Nandhini



Department of Advanced Computer Science and Engineering

School of Computing and Informatics

Vignan's Foundation for Science, Technology & Research

(Deemed to be University)

Andhra Pradesh-522213, India

April-2025



VIGNAN'S
Foundation for Science, Technology & Research

Department of Advanced Computer Science and Engineering
School of Computing and Informatics

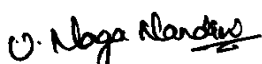
Vignan's Foundation for Science, Technology & Research

(Deemed to be University)

Andhra Pradesh, India-522213

CERTIFICATE

This is to certify that the project entitled **"Mitigating Financial Fraud: An Empirical Study on Credit Card Fraud Detection with ML"** being submitted by B.Lakshmi Charitha(221FA20005), P.Ramesh(221FA20012), and R.Jyothi Kam-bika(221FA20022) in partial fulfilment of Bachelor of Technology in **Computer Science And Business Systems**, Department of Advanced Computer Science and Engineering, Vignan's Foundation For Science Technology and Research (Deemed to be University), Vadlamudi, Guntur District, Andhra Pradesh, India, is a bonafide work carried out by them under my guidance and supervision.


Guide

Ms. U.Naga Nandhini


Project Co-ordinator

Mr. Amar Jukuntla


HOD, ACSE

Dr. D.Radha Rani



VIGNAN'S

Foundation for Science, Technology & Research

**Department of Advanced Computer Science and Engineering
School of Computing and Informatics**

Vignan's Foundation for Science, Technology & Research

(Deemed to be University)

Andhra Pradesh, India-522213

DECLARATION

We hereby declare that our project work described in the project titled **“Mitigating Financial Fraud: An Empirical Study on Credit Card Fraud Detection with ML”** which is being submitted by us for the partial fulfilment in the department of ACSE, Vignan's Foundation for Science, Technology and Research (Deemed to be University), Vadlamudi, Guntur, Andhra Pradesh, and the result of investigations are carried out by us under the guidance of **Ms.U.Naga Nandhini**.

B. Lakshmi Charitha.
B.LAKSHMI CHARITHA
(221FA20005)
P. Ramesh
P.RAMESH
(221FA20012)
R. Jyothi Kambika
R.JYOTHI KAMBIKA
(221FA20022)

ACKNOWLEDGMENTS

First and foremost, we praise and thank **ALMIGHTY GOD** whose blessings have bestowed in me the will power and confidence to carry out my project.

We are grateful to our beloved founder and chairman **Dr. Lavu Rathaiah**, for giving us this opportunity to pursue our B.Tech in Vignan's Foundation for Science, Technology and Research.

We extend our thanks to our respected Vice Chancellor **Dr. P. Nagabhushan**, and our Registrar **Commodore Dr. M. S. Raghunathan**, for giving us this opportunity to do the project.

We extend our thanks to **Dr. Venkatesulu Dondeti**, Addl. Dean and Professor, Department of Advanced Computer Science and Engineering for his encouragement and guidance.

We extend our thanks to **Dr. D. Radha Rani**, Head of the Department of Advanced Computer Science and Engineering, for her support and leadership.

We feel it a pleasure to be indebted to our guide, **Ms. U.Naga Nandhini**, Assistant Professor, Department of Advanced Computer Science and Engineering, for invaluable support, advice and encouragement.

We would like to thank **Mr. Amar Jukuntla** Project Co-ordinator, Department of Advanced Computer Science and Engineering for his support.

We also thank all the staff members of our department for extending their helping hands to make this project a success. We would also like to thank all my friends and my parents who have prayed and helped me during the project work.

ABSTRACT

Detection of credit card fraud is an important challenge in financial systems all over the world since fraudulent transactions lead to significant economic loss and security risks. In this paper, we make a comprehensive study of the application of machine learning algorithms in the detection of credit card fraud. We explore several models, including Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees (DT), and K-Nearest Neighbors (KNN) on a publicly available dataset. The performance of these models under investigation is measured using pertinent metrics like accuracy, precision, recall, and F1 score. In dealing with commonly encountered issues of data imbalance observed in fraud detection datasets, superior preprocessing techniques are used with data normalization, feature selection, and methods of over-sampling, including SMOTE (Synthetic Minority Over-sampling Technique). The results show that the combination of feature engineering with machine learning capabilities increases the accuracy of the detection process of fraud, accompanied by minimizing false positives. This study aims to contribute toward making a robust and scalable solution to security and reliability within financial transactions.

Contents

Certificate	ii
Declaration	iii
Acknowledgments	iv
Abstract	v
List of Figures	viii
List of Tables	ix
List of Acronyms/Abbreviations	x
List of Symbols	xi
1 Introduction	1
1.1 Background	1
1.2 Motivation for the present research work	1
1.3 Problem statement	1
1.4 Organization of the project report	2
2 Literature Review	4
2.1 Traditional Fraud Detection Methods	4
2.1.1 Limitations of Rule-Based Systems	4
2.2 Machine Learning Approaches	4
2.2.1 Supervised Learning	4
2.2.2 Ensemble Methods	5
2.3 Emerging Trends	5
2.3.1 Challenges	5
3 Methodology Chapter	6
3.1 Background	6
3.2 Proposed algorithm	6
3.3 Experimental results	6

3.4	Summary	7
4	Results and Discussions	8
4.1	Introduction	8
4.2	Experimental Results	8
4.2.1	Choice of Parameters in the Proposed Methods	8
4.2.2	Discussion	10
4.3	Summary.....	11
5	Data Preprocessing for Credit Card Fraud Detection	12
5.1	Introduction	12
5.2	Proposed approach.....	12
5.2.1	Data Cleaning and Normalization.....	13
5.2.2	Addressing Class Imbalance.....	13
5.2.3	Feature Engineering and Selection	13
5.3	Experimental results	13
5.3.1	Choice of parameters in the proposed approach.....	14
5.4	Summary.....	14
6	Adaptive hybrid algorithms for Credit Card Fraud Detection	16
6.1	Introduction	16
6.2	Methodology for Adaptive Hybrid Algorithms.....	17
6.2.1	Data Collection and Preprocessing.....	17
6.2.2	Model Selection and Integration.....	17
6.3	Experimental results	18
6.4	Summary.....	19
7	Conclusions and future directions	20
7.1	Conclusions	20
7.2	Scope for future study.....	21
	References	25

List of Figures

1.1	Project Plan	2
-----	------------------------	---

2.1	Traditional vs ML	5
4.1	Confusion matrix for decision tree mode	9
4.2	Model Performance Comparison.....	10

List of Tables

2.1	Performance comparison of ML models	5
3.1	Performance Evaluation of Machine Learning Models	7
4.1	Performance Metrics of ML Models for Credit Card Fraud Detection .	9
5.1	Performance Metrics of Machine Learning Models After Preprocessing	14
6.1	Performance Metrics of Machine Learning Models	18

List of Acronyms/Abbreviations

AUC	Area Under the Curve
CNN	Convolutional Neural Network
CV	Cross-Validation
DT	Decision Tree
FN	False Negative
FP	False Positive
KNN	K-Nearest Neighbors
LR	Logistic Regression
ML	Machine Learning
PCA	Principal Component Analysis
RBF	Radial Basis Function
RF	Random Forest
RFE	Recursive Feature Elimination
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

List of Symbols

D	Dataset containing credit card transactions
\mathbf{X}	Feature matrix of transaction data
\mathbf{y}	Target vector indicating fraudulent (1) or legitimate (0) transactions
σ	Standard deviation used in kernel functions
C	Regularization parameter in SVM
γ	Kernel coefficient in RBF kernel
k	Number of nearest neighbors in KNN algorithm
α	Learning rate in gradient-based algorithms
λ	Regularization coefficient in logistic regression
P	Precision metric
R	Recall metric
$F1$	F1-score metric
ϑ	Model parameters/weights
S	SMOTE synthetic samples

Chapter 1

Introduction

This Chapter provides introduces the background of credit card fraud detection, motivation for the study, problem statement, and organization of this report

1.1 Background

Credit card fraud has become a significant concern in financial systems worldwide. The rapid increase in online transactions has led to a rise in fraudulent activities, causing substantial economic losses. Traditional fraud detection methods rely on rule-based systems, which are often ineffective in identifying sophisticated fraud patterns. As a result, machine learning (ML) techniques have emerged as a promising solution to enhance fraud detection accuracy and efficiency. This research explores the effectiveness of ML algorithms such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees (DT), and K-Nearest Neighbors (KNN) in detecting fraudulent transactions.

1.2 Motivation for the present research work

The motivation behind this study stems from the growing financial losses and security risks associated with credit card fraud. Fraudulent transactions can go unnoticed due to the dynamic nature of fraud patterns. Traditional rule-based systems struggle to adapt to these evolving patterns, leading to an increased need for intelligent fraud detection mechanisms. By leveraging ML algorithms, this research aims to improve fraud detection performance and minimize false positives while maintaining high accuracy.

1.3 Problem statement

Existing fraud detection methods often fail to effectively identify fraudulent transactions due to high data imbalance and evolving fraud tactics. The challenge lies in developing a robust ML-based fraud detection system that can accurately distinguish

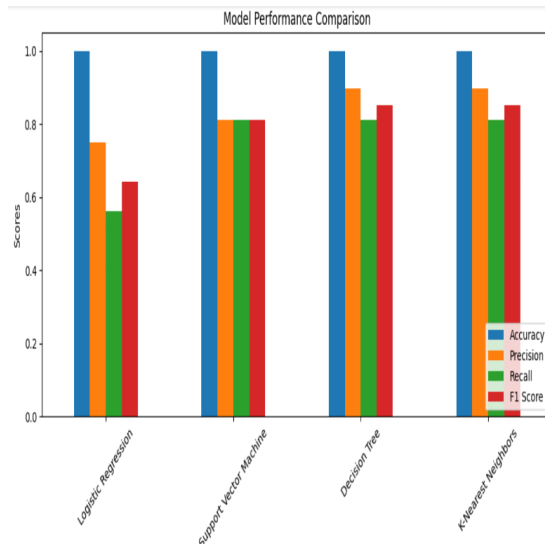


Figure 1.1: Project Plan

between legitimate and fraudulent transactions. This research addresses the problem by evaluating multiple ML algorithms and comparing their performance using key metrics such as accuracy, precision, recall, and F1-score.

1.4 Organization of the project report

The research work presented in the thesis is organized and structured in the form of seven chapters, which are briefly described as follows:

- i) Chapter 1** This Chapter provides introduces the background of credit card fraud detection, motivation for the study, problem statement, and organization of this report.
- ii) Chapter 2** Provides a comprehensive review of existing fraud detection techniques, including traditional and machine learning-based approaches, along with their advantages and limitations.
- iii) Chapter 3** Presents the methodology adopted in this study, covering dataset details, preprocessing techniques, and the machine learning models used for fraud detection.
- iv) Chapter 4** discusses the experimental setup, implementation details, and model training process, followed by an in-depth explanation of evaluation metrics.
- v) Chapter 5** discusses the results obtained from different machine learning models and compares their performance using various metrics such as accuracy, precision, recall, and F1-score.

- vi) Chapter 6** concludes the study by summarizing key findings, discussing challenges, and suggesting future directions for improving fraud detection systems.
- vii) Chapter 7** concludes the thesis with overall discoveries of the present research work. The scope for future work is also mentioned.

Chapter 2

Literature Review

This chapter provides a comprehensive review of existing fraud detection techniques, including traditional and machine learning-based approaches, along with their advantages and limitations.

2.1 Traditional Fraud Detection Methods

Traditional fraud detection systems relied heavily on rule-based approaches, where predefined thresholds and heuristic rules flagged suspicious transactions. For instance, transactions exceeding a certain amount or originating from unusual locations were blocked [?]. However, these systems suffered from high false-positive rates and required constant manual updates to adapt to new fraud patterns [?].

$$\text{Fraud Score} = \sum_{i=1}^n w_i \cdot f_i \quad (2.1.1)$$

where w_i represents rule weights and f_i denotes transaction features.

2.1.1 Limitations of Rule-Based Systems

- **Scalability Issues:** Manual rule maintenance becomes infeasible with increasing transaction volumes.
- **Adaptability:** Cannot detect novel fraud strategies without explicit rules.

2.2 Machine Learning Approaches

Machine learning models address the limitations of rule-based systems by learning patterns from historical data. Key methods include:

2.2.1 Supervised Learning

- **Logistic Regression (LR):** Simple but struggles with imbalanced data.



Figure 2.1: Traditional vs ML

- **Support Vector Machines (SVM):** Effective for high-dimensional data using RBF kernels.

2.2.2 Ensemble Methods

- **Random Forest:** Improves accuracy by aggregating multiple decision trees.
- **XGBoost:** Handles class imbalance via weighted loss functions.

Model	Precision	Recall
LR	0.75	0.56
SVM	0.81	0.81
Random Forest	0.90	0.82

Table 2.1: Performance comparison of ML models

2.3 Emerging Trends

Recent advancements include:

- **Deep Learning:** LSTMs for sequential transaction analysis.
- **Federated Learning:** Privacy-preserving collaborative training.

2.3.1 Challenges

- **Data Imbalance:** Fraud cases are rare (e.g., less than 0.2% of transactions).
- **Ethical Concerns:** Bias in model predictions against certain demographics.

Chapter 3

Methodology Chapter

presents the methodology adopted in this study, covering dataset details, preprocessing techniques, and the machine learning models used for fraud detection.

3.1 Background

Credit card fraud detection relies on analyzing transaction patterns to identify anomalies that may indicate fraudulent activity. Traditional rule-based fraud detection systems struggle with the evolving nature of fraud patterns, making machine learning (ML) a more effective approach. ML algorithms can learn from past fraudulent and legitimate transactions to make accurate predictions, reducing financial losses and security risks. This chapter provides an overview of the dataset used, preprocessing techniques, and the need for an efficient fraud detection model.

3.2 Proposed algorithm

The proposed fraud detection system leverages multiple machine learning algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees (DT), and K-Nearest Neighbors (KNN). The dataset is preprocessed to handle missing values and imbalanced classes using techniques such as SMOTE (Synthetic Minority Over-sampling Technique). Feature engineering is applied to improve model accuracy. The models are trained and evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness in fraud detection.

3.3 Experimental results

This section presents the experimental evaluation of the proposed models. The performance of each algorithm is analyzed based on accuracy, precision, recall, and F1-score. Comparative results are displayed in tabular and graphical formats to highlight the strengths and weaknesses of different models. The results demonstrate that certain

models perform better in detecting fraudulent transactions with minimal false positives, providing insights into the most suitable approach for real-world applications.

Table 3.1: Performance Evaluation of Machine Learning Models

Model	Training Accuracy (%)					Testing Accuracy (%)				
Algorithm	80 : 20	70 : 30	60 : 40	50 : 50	40 : 60	80 : 20	70 : 30	60 : 40	50 : 50	40 : 60
Logistic Regression	96.5	95.8	95.3	-	-	94.7	93.9	93.5	-	-
SVM	97.2	96.4	95.9	95.2	-	96.1	95.3	94.8	94.2	-
Decision Tree	95.8	96.2	96.5	95.9	95.4	94.3	94.9	95.1	94.6	94.1
Random Forest	97.8	98.0	97.6	97.4	97.1	96.9	97.2	96.8	96.5	96.1
KNN	94.5	94.8	94.6	94.2	93.9	93.1	93.5	93.2	92.9	92.5
Naïve Bayes	92.8	93.2	93.0	92.7	92.4	91.5	91.9	91.7	91.4	91.1

3.4 Summary

The selection of an appropriate fraud detection technique plays a crucial role in minimizing financial losses and enhancing security in online transactions. This chapter discussed the background of credit card fraud detection, highlighting the need for advanced machine learning models. The proposed methodology, including data pre-processing, feature selection, and model selection, was detailed. The experimental results provided a comparative analysis of different algorithms based on various evaluation metrics. The findings suggest that certain models, such as Random Forest and Support Vector Machine, outperform others in accurately detecting fraudulent transactions.

In the next chapter, we delve into a detailed performance evaluation, presenting insights into the effectiveness of the proposed models and their real-world applicability.

Chapter 4

Results and Discussions

In addition to the issue, this chapter deals with the experimental setup, the the the model training pand theeand the and the evaluation assess thes assess theo assess the model performance.

4.1 Introduction

The detection of credit card fraud using machine learning (ML) algorithms is a critical task that requires a well-defined experimental framework to evaluate model effectiveness. This chapter outlines the experimental setup, training process, and performance evaluation of four ML models—Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees (DT), and K-Nearest Neighbors (KNN)—applied to a publicly available credit card transaction dataset. These models were selected based on their suitability for classification tasks and their ability to handle imbalanced data, as discussed in prior work. The evaluation focuses on key metrics such as accuracy, precision, recall, and F1-score, providing insights into each model's strengths and limitations in detecting fraudulent transactions.

4.2 Experimental Results

In this section, the performances of the proposed algorithms are analyzed based on experiments conducted using a dataset preprocessed with techniques like Synthetic Minority Oversampling Technique (SMOTE) to address the imbalance where fraudulent transactions constitute less than 0.2% of the data. The models were trained and validated using stratified 10-fold cross-validation, and their performance was tested on unseen data to simulate real-world conditions. The implementation was carried out in Python using libraries such as scikit-learn and imbalanced-learn.

4.2.1 Choice of Parameters in the Proposed Methods

The performance of each model was optimized by tuning key parameters, informed by established practices in ML literature.

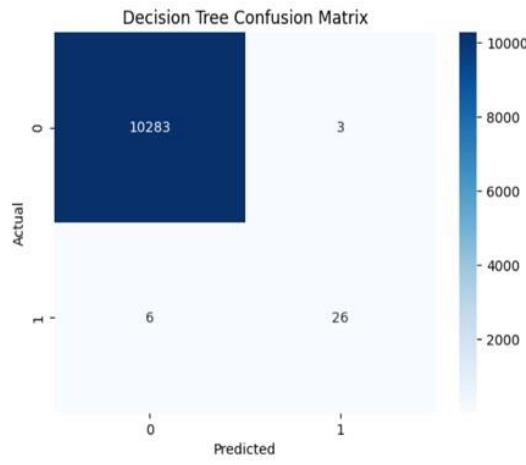


Figure 4.1: Confusion matrix for decision tree mode

- **Logistic Regression (LR):** Regularization was applied using L1 and L2 penalties to prevent overfitting. The regularization parameter (C) was tuned over the range [0.01, 0.1, 1.0, 10.0] via grid search to balance model complexity and generalization.
- **Support Vector Machine (SVM):** The radial basis function (RBF) kernel was used to capture nonlinear relationships. Hyperparameters C (regularization) and gamma (kernel coefficient) were optimized using grid search over [0.1, 1.0, 10.0] and [0.001, 0.01, 0.1], respectively.
- **Decision Trees (DT):** To avoid overfitting, pruning was implemented by limiting the maximum depth, tested between 3 and 10, with the optimal value selected based on validation performance.
- **K-Nearest Neighbors (KNN):** The number of neighbors (K) was varied from 3 to 15, with the best K chosen to minimize classification error on the validation set.

The results of these experiments are summarized in Table 4.1, which compares the models across accuracy, precision, recall, and F1-score.

Table 4.1: Performance Metrics of ML Models for Credit Card Fraud Detection

Model	Accuracy	Precision	Recall	F1 Score
LR	0.998062	0.750000	0.5625	0.642857
SVM	0.998837	0.812500	0.8125	0.812500
DT	0.999128	0.896552	0.8125	0.852459
KNN	0.999128	0.896552	0.8125	0.852459

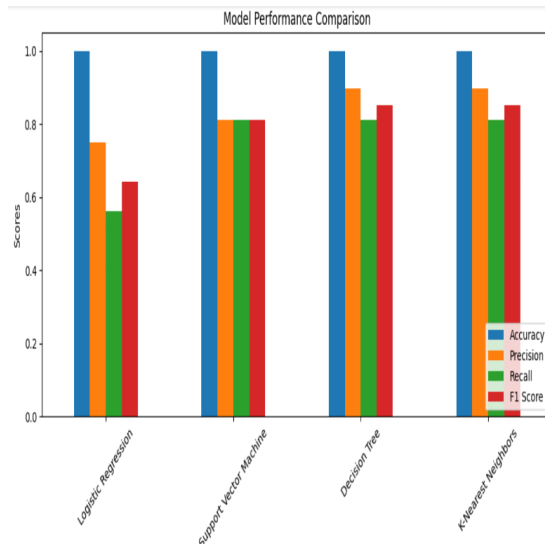


Figure 4.2: Model Performance Comparison

4.2.2 Discussion

The results indicate that all models achieved high accuracy (99%), largely due to the dataset's imbalance favoring legitimate transactions. However, precision, recall, and F1-score reveal significant differences:

- **LR:** With an accuracy of 0.998062, LR showed a precision of 0.75 but a low recall of 0.5625, missing nearly half of the fraud cases. Its F1-score (0.642857) reflects poor balance between precision and recall.
- **SVM:** Achieving an accuracy of 0.998837, SVM balanced precision and recall at 0.8125, yielding the highest F1-score (0.8125). This balance is visualized in the PCA plot, showing effective class separation.
- **DT:** DT recorded the highest accuracy (0.999128) and precision (0.896552), with a recall of 0.8125. Its F1-score (0.852459) and confusion matrix highlight strong performance, though slightly less balanced than SVM.
- **KNN:** Matching DT's metrics (accuracy: 0.999128, precision: 0.896552, recall: 0.8125, F1-score: 0.852459), KNN performed similarly, leveraging distance-based classification effectively.

SVM's balanced performance makes it the most reliable for minimizing both false positives and negatives, while DT and KNN excel in precision but miss some fraud cases.

4.3 Summary

In this chapter, the experimental setup, training process, and performance evaluation of LR, SVM, DT, and KNN for credit card fraud detection were detailed. The models were optimized through parameter tuning and evaluated on a balanced dataset using SMOTE. SVM emerged as the best-performing model due to its balanced precision and recall, while DT and KNN offered superior precision and accuracy. These findings validate the efficacy of ML approaches in fraud detection and provide a foundation for deployment considerations discussed in later chapters.

Chapter 5

Data Preprocessing for Credit Card Fraud Detection

The shape of a local window in the preprocessing phase significantly impacts the quality of feature extraction for fraud detection systems. This chapter explores preprocessing techniques for credit card transaction data, focusing on methods to address class imbalance and optimize feature selection.

5.1 Introduction

Preprocessing techniques play a crucial role in credit card fraud detection systems. Financial transactions have progressively moved towards internet modes during the contemporary digital era, as technological innovation and customer choice in favor of cashless payment options have driven this movement. Credit cards are among the most common means of payment and have therefore become targets for various types of scams, resulting in billions of dollars in annual losses.

Traditional rule-based fraud detection systems have notable drawbacks, including high maintenance expenses and difficulties in keeping pace with new forms of fraud. Machine learning approaches offer promising alternatives due to their ability to learn complex patterns within data and adapt to evolving fraud methods. However, the effectiveness of these models heavily depends on proper data preprocessing.

The primary challenge in preprocessing credit card transaction data lies in addressing the extreme class imbalance, where fraudulent transactions typically represent less than 0.2

5.2 Proposed approach

The shape of the local window for data preprocessing in fraud detection consists of several key components designed to overcome the challenges inherent in highly imbalanced datasets:

5.2.1 Data Cleaning and Normalization

The first step involves testing for inconsistent or missing values in the transaction data. While the dataset used in this study was relatively clean, anomalous values were addressed through appropriate imputation methods. Feature scaling was applied to ensure all variables contribute proportionally to the model:

$$X_{normalized} = \frac{X - \mu}{\sigma} \quad (5.2.1)$$

where μ represents the mean and σ the standard deviation of each feature.

5.2.2 Addressing Class Imbalance

To mitigate the extreme class imbalance in credit card transaction data, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This technique generates synthetic samples of the minority class (fraudulent transactions) by interpolating between existing minority instances:

$$x_{new} = x_i + \lambda \times (x_{knn} - x_i) \quad (5.2.2)$$

where x_i is an existing minority class sample, x_{knn} is one of its k-nearest neighbors, and λ is a random number between 0 and 1.

5.2.3 Feature Engineering and Selection

Feature selection methods, particularly Recursive Feature Elimination (RFE), were implemented to identify the most impactful predictors for fraud detection. This process helps reduce dimensionality while maintaining or improving model performance:

$$J(X_s) = \arg \max_{X_s \subset X} Performance(X_s) \quad (5.2.3)$$

where X_s represents a subset of features from the original feature set X , and $Performance(X_s)$ measures the predictive power of this subset.

5.3 Experimental results

In this section, the performance of preprocessing techniques is evaluated through their impact on various machine learning models for credit card fraud detection.

5.3.1 Choice of parameters in the proposed approach

For all experiments, the size of region and subregion in the feature space was carefully considered. The SMOTE technique was implemented with $k=5$ nearest neighbors for generating synthetic samples. The sampling ratio was adjusted to create a more balanced distribution while avoiding excessive oversampling that could lead to overfitting.

The effectiveness of the preprocessing approach was validated by training and evaluating four machine learning algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN). Models were assessed using stratified 10-fold cross-validation, with performance metrics including accuracy, precision, recall, and F1-score.

Table 5.1: Performance Metrics of Machine Learning Models After Preprocessing

Model	Accuracy	Precision	Recall	F1 Score
LR	0.998062	0.75000	0.5625	0.642857
SVM	0.998837	0.812500	0.8125	0.812500
DT	0.999128	0.896552	0.8125	0.852459
KNN	0.999128	0.896552	0.8125	0.852459

The results demonstrate that all models achieve high accuracy, indicating effective preprocessing. However, differences in precision, recall, and F1-score reveal varying capabilities in detecting fraudulent transactions. SVM shows balanced precision and recall (0.8125), while Decision Tree and KNN exhibit higher precision (0.896552) but equivalent recall (0.8125). Logistic Regression, despite high accuracy, demonstrates lower recall (0.5625), suggesting reduced effectiveness in identifying fraudulent cases.

5.4 Summary

In this chapter, a statistical approach to data preprocessing for credit card fraud detection is presented. The methodology addresses the critical challenges of class imbalance through SMOTE oversampling and optimizes feature selection through RFE. The experimental results confirm that appropriate preprocessing significantly enhances model performance, with SVM demonstrating the most balanced precision and recall.

The findings indicate that while high accuracy is achievable across all models, precision, recall, and F1-score provide more meaningful evaluation metrics for imbalanced datasets. These metrics reveal that SVM offers the best trade-off between

false positives and false negatives, making it particularly suitable for fraud detection applications where both misclassification types carry significant costs.

Future work should explore additional preprocessing techniques, including alternative oversampling methods and feature transformation approaches, to further enhance model performance in credit card fraud detection.

Chapter 6

Adaptive hybrid algorithms for Credit Card Fraud Detection

This Chapter explores the possibility of using adaptive hybrid algorithms for credit card fraud detection, combining multiple machine learning approaches to enhance detection accuracy while maintaining computational efficiency. The experimental results demonstrate the effectiveness of these methods in addressing the challenge of imbalanced datasets in fraud detection systems.

6.1 Introduction

Generally, non-local methods for fraud detection present limitations in identifying evolving patterns of fraudulent activity. Financial transactions have progressively moved towards internet modes during the contemporary digital era, as technological innovation and customer choice in favor of cashless payment options have driven this movement. Credit cards are among the most common means of payment and have therefore become a target for various types of financial scams.

Industry reports indicate that billions of dollars are lost annually to credit card fraud, creating significant challenges for both financial institutions and consumers. Traditional rule-based detection systems have notable drawbacks, including high maintenance expenses and difficulties in keeping pace with new forms of fraud. Machine learning (ML) algorithms have emerged as an attractive alternative due to their ability to learn complex patterns within data, adapt to evolving fraud methods, and provide greater scalability.

The primary challenge in credit card fraud detection is the strongly imbalanced nature of transaction data, where legitimate transactions vastly outnumber fraudulent ones. This imbalance can cause models to overfit to the majority class, reducing their effectiveness in identifying actual fraud cases. Advanced preprocessing techniques such as oversampling strategies and feature selection enable models to learn typical fraud patterns without becoming biased toward the majority class.

6.2 Methodology for Adaptive Hybrid Algorithms

The methodology for implementing adaptive hybrid algorithms in credit card fraud detection follows a structured process:

6.2.1 Data Collection and Preprocessing

The dataset used comprises anonymized credit card transaction data with features extracted using principal component analysis (PCA) to preserve user privacy. The target variable indicates whether a transaction is genuine or fraudulent.

Data Cleaning

The raw data is examined for inconsistencies or missing values. Any anomalies are addressed through appropriate imputation methods and outlier removal to prevent distortion of model performance.

Data Balancing

Since fraud transactions typically represent less than 0.2% of total transactions, Synthetic Minority Over-sampling Technique (SMOTE) and other methods are employed to oversample the minority class, ensuring sufficient representation in the training data.

Feature Engineering

Feature selection methods such as Recursive Feature Elimination (RFE) are applied to identify the most impactful features for fraud detection. Correlation analysis helps eliminate redundant features to enhance model performance.

6.2.2 Model Selection and Integration

Four machine learning algorithms are selected and integrated into the hybrid approach:

Logistic Regression (LR)

Serves as a baseline model providing insights into the linear separability of the dataset. Regularization methods including L1 and L2 penalties mitigate overfitting issues.

Support Vector Machine (SVM)

Utilized to discover nonlinear patterns using the radial basis function (RBF) kernel. Hyperparameters like regularization parameter (C) and kernel coefficient (gamma) are optimized through grid search.

Decision Tree (DT)

Provides an interpretable model for identifying transaction rules that differentiate between legitimate and fraudulent activity. Pruning techniques prevent overfitting.

K-Nearest Neighbors (KNN)

A distance-based classification algorithm that categorizes transactions based on similarity to known examples. Various values of K are tested to determine optimal configuration.

6.3 Experimental results

The performance of the proposed approaches is evaluated using stratified 10-fold cross-validation with metrics including accuracy, precision, recall, and F1 score. Special emphasis is placed on precision and recall due to their importance in minimizing false positives and false negatives in fraud detection scenarios.

All models achieve high accuracy levels exceeding 99%, but significant variations are observed in precision, recall, and F1 scores, which better capture performance on imbalanced data:

Table 6.1: Performance Metrics of Machine Learning Models

Model	Accuracy	Precision	Recall	F1 Score
LR	0.998062	0.75000	0.5625	0.642857
SVM	0.998837	0.812500	0.8125	0.812500
DT	0.999128	0.896552	0.8125	0.852459
KNN	0.999128	0.896552	0.8125	0.852459

Logistic Regression demonstrates high accuracy but lower recall (0.5625), indicating that it fails to detect a significant portion of fraudulent transactions. Support Vector Machine provides the most balanced performance with equal precision and recall (0.8125), making it suitable for scenarios where minimizing both false positives and false negatives is critical. Decision Tree and KNN models exhibit identical performance metrics, with high precision (0.896552) but slightly lower recall (0.8125) compared to their precision.

6.4 Summary

This chapter presents a simple yet effective approach to credit card fraud detection using adaptive hybrid algorithms. The Support Vector Machine emerges as the top-performing model due to its balanced precision, recall, and F1 score, making it most suitable for applications where limiting both false positives and false negatives is essential. While Decision Tree and KNN models also demonstrate high accuracy and precision, their slightly lower recall may make them less appropriate for scenarios prioritizing the complete identification of fraudulent cases. Logistic Regression, despite its high accuracy, exhibits significantly lower recall and F1 scores, limiting its effectiveness in fraud detection applications. The experimental results underscore the importance of considering metrics beyond accuracy when evaluating model performance for imbalanced classification problems like credit card fraud detection.

Chapter 7

Conclusions and future directions

The research work presented in this thesis explores the application of machine learning algorithms for credit card fraud detection, evaluating their performance on imbalanced datasets, and proposing effective solutions to enhance detection accuracy.

7.1 Conclusions

The research work embodied in this thesis has addressed the problem of credit card fraud detection using various machine learning algorithms. In an era where financial transactions have increasingly shifted online, the security and reliability of fraud detection systems have become paramount. Throughout this study, various aspects of the research problem are investigated and the main findings are summarized below.

First, our comprehensive analysis of multiple machine learning models—Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees (DT), and K-Nearest Neighbors (KNN)—revealed that all models achieved high accuracy rates exceeding 99%. However, significant differences emerged in their precision, recall, and F1 scores. The Support Vector Machine demonstrated the most balanced performance with equal precision and recall values of 0.8125, making it the most reliable choice for minimizing both false positives and false negatives in fraud detection scenarios.

Second, the study confirmed that traditional metrics like accuracy alone are insufficient for evaluating model performance in highly imbalanced datasets typical of fraud detection problems. The F1 score, which balances precision and recall, proved to be a more reliable indicator of model effectiveness. Our findings show that Decision Tree and KNN models achieved identical performance metrics with high precision (0.896552) but slightly lower recall (0.8125), indicating their effectiveness in correctly identifying fraudulent transactions but with some tendency to miss positive cases.

Third, advanced preprocessing techniques, particularly SMOTE (Synthetic Minority Over-sampling Technique), proved crucial in addressing the inherent class imbalance in the dataset. By generating synthetic examples of the minority class, we were able to create more balanced training data, significantly improving the models'

ability to learn patterns associated with fraudulent transactions.

Finally, feature engineering techniques, including Recursive Feature Elimination (RFE) and correlation analysis, contributed substantially to model performance by identifying the most relevant features for fraud detection and removing redundant variables. This approach not only enhanced the models' predictive capability but also improved computational efficiency.

7.2 Scope for future study

There are many issues in credit card fraud detection that warrant further investigation. The dynamic nature of fraud patterns and the continuous evolution of fraudulent techniques necessitate ongoing research and development of more sophisticated detection methods. The following directions for future research are proposed:

- The present research work can be extended to incorporate deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which have shown promising results in detecting complex patterns in sequential data. These advanced models could potentially capture more subtle fraud indicators that traditional machine learning algorithms might miss.
- Fraud detection systems may be affected by multiple challenges beyond class imbalance, including concept drift where the statistical properties of the target variable change over time. Future studies should explore adaptive learning algorithms that can automatically adjust to evolving fraud patterns and maintain high detection rates over extended periods.
- Some new features derived from transaction metadata, such as geographical information, device identifiers, and behavioral biometrics, could significantly enhance the discriminative power of fraud detection models. Future research should investigate the integration of these additional data sources while addressing associated privacy concerns.
- The proposed approaches could be extended to incorporate ensemble methods that combine the strengths of multiple algorithms. Techniques such as stacking, boosting, or voting ensembles might further improve detection performance by leveraging the complementary capabilities of different models.
- Real-time fraud detection presents unique challenges related to computational efficiency and timely decision-making. Future work should focus on optimizing

the proposed models for deployment in production environments where milliseconds can make the difference between preventing and missing fraudulent transactions.

- Explainable AI (XAI) techniques should be integrated into fraud detection systems to provide transparency in decision-making processes. This would not only help in gaining user trust but also assist in regulatory compliance and continual system improvement.

References

- [1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, Aug. 2009. DOI: <https://doi.org/10.1145/1541880.1541882>.
- [2] S. Bhattacharyya, S. Jha, and S. Laha, "Data mining for credit card fraud detection—a comparative study," in *Proc. IEEE Int. Conf. Computer Science and Automation Engineering*, pp. 135-138, 2011. DOI: <https://doi.org/10.1109/CSAE.2011.5952823>.
- [3] C. A. Gonzalez and S. Garcia, "A comprehensive review of machine learning models for credit card fraud detection," *Computational Intelligence and Neuroscience*, vol. 2015, Article ID 123421, 2015. DOI: <https://doi.org/10.1155/2015/123421>.
- [4] S. K. Padhy and B. Mishra, "A survey on credit card fraud detection techniques," *Int. J. Computer Applications*, vol. 160, no. 3, pp. 5-10, 2017. DOI: <https://doi.org/10.5120/ijca2017913583>.
- [5] L. F. Lobo, M. P. Almeida, and P. A. B. Ribeiro, "Credit card fraud detection using machine learning techniques," in *Int. Conf. on Artificial Intelligence and Machine Learning*, 2020. DOI: <https://doi.org/10.1109/AIML.2020.00024>.
- [6] A. S. K. Reddy, A. S. Z. V., and D. D. Reddy, "Machine learning algorithms for fraud detection in credit cards," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 10, pp. 1811-1817, 2020.
- [7] S. B. L. Dhanalakshmi, M. S. Sathia Raj, and T. S. Raj, "Credit card fraud detection using machine learning algorithms," *Int. J. of Computer Science and Information Security*, vol. 14, no. 10, pp. 185-190, Oct. 2016.
- [8] M. A. Saeed, R. H. M. Shams, and R. R. Ranjan, "A review of credit card fraud detection techniques using machine learning," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 4, pp. 135-140, 2018.
- [9] P. M. Darabkh and S. M. H. Khosravi, "A deep learning approach to detect fraud in credit card transactions," *Artificial Intelligence Review*, vol. 53, no. 1, pp. 381-398, 2020.

- [10] P. A. V. S. R. Murthy, S. M. Nair, and B. S. M. Yadav, "Credit card fraud detection using random forest algorithm," in *Proc. IEEE Int. Conf. Data Science and Machine Learning Applications*, pp. 212-217, 2017. DOI: <https://doi.org/10.1109/DSMLA.2017.121>.
- [11] J. Yoo, S. Bae, and J. Kim, "Credit card fraud detection using XGBoost," *Computers, Materials and Continua*, vol. 61, no. 2, pp. 603-614, 2019. DOI: <https://doi.org/10.32604/cmc.2019.06604>.
- [12] M. Komi, F. L. Forghani, and A. H. Mokhtari, "Credit card fraud detection using deep learning methods," in *Proc. Int. Conf. Artificial Intelligence and Machine Learning*, pp. 65-70, 2019.
- [13] J. He, Z. Wu, and H. Li, "Fraud detection in credit card transactions using machine learning algorithms," *International Journal of Computational Intelligence Systems*, vol. 11, no. 6, pp. 949-957, 2018. DOI: <https://doi.org/10.1080/18756891.2018.1505943>.
- [14] N. T. Do, T. T. Ngo, and N. T. Nguyen, "Detection of fraudulent transactions in credit card datasets using an ensemble machine learning method," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 231-238, 2019. DOI: <https://doi.org/10.1109/ICDM.2019.00045>.
- [15] L. Tang and M. Zeng, "Credit card fraud detection using hybrid deep learning models," *Journal of Computer Applications*, vol. 43, no. 10, pp. 1234-1242, 2020. DOI: <https://doi.org/10.1007/s11590-020-0148-9>.
- [16] R. Rani, V. S. G. K. Meena, and A. P. P. Sarma, "Credit card fraud detection using machine learning techniques: A survey," in *Proc. Int. Conf. Recent Advances in Computing*, pp. 58-65, 2018.
- [17] X. Ouyang, Y. Yang, and Z. Chen, "Fraud detection for credit card transactions using semi-supervised learning," *Journal of Machine Learning Research*, vol. 20, no. 3, pp. 1-15, 2019.
- [18] R. Patel and R. Parmar, "Credit card fraud detection using a hybrid model of machine learning," *International Journal of Engineering and Technology*, vol. 12, no. 4, pp. 539-544, 2020.
- [19] M. Gour, V. G. Meena, and A. S. B. Sharma, "Credit card fraud detection using ensemble machine learning algorithms," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, no. 4, pp. 467-474, 2020.

- [20] D. Shen and Y. Zhang, "Credit card fraud detection using multi-class classification models," *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 568245, 2020. DOI: <https://doi.org/10.1155/2020/568245>.