



A Comparative Study of Machine Learning and Adaptive LSTM Models for Credit Card Fraud Detection

Prepared By

Name: [Ahmad Mostafa Abutalib](#)

IDNo.: [243002097](#)

Submitted in fulfillment of the requirements for the degree
Master of Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Amjad Qtaish

College of Informatics, Midocean University, Comoros

[\[2025-Third Semester\]](#)

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Allah Almighty for granting me the strength, patience, and determination to complete this work successfully.

I am profoundly thankful to my mother, whose constant prayers and unwavering support have been my greatest source of motivation, and to my wife for her supporting, understanding, and patience throughout this journey. Without their encouragement, this achievement would not have been possible.

I would also like to extend my sincere appreciation to Dr. Amjad Qtaish, whose invaluable guidance and continuous support were instrumental in shaping this research. This work would not have reached its current quality without his advice, encouragement, and expertise.

Finally, I would like to acknowledge my colleagues and friends for their helpful feedback and moral support during different phases of this project.

Abstract

Credit-card fraud detection remains among the most daunting problems for the financial landscape, with transaction imbalance remaining the predominant challenge where the fraudulent activities account for only a small minority of all transactions. Here, authors recommend a powerful hybrid fraud-detection infrastructure that combines ML and DL models and has applied some imbalance-handling principles that aim to improve predictivity and minority-class recognition. In order to handle class imbalances, we apply two resampling strategies, namely Random Oversampling and SMOTE, to improve the characterization of the fraud cases. The full analytical pipeline was created, with data preprocessing, normalization, model training, and performance evaluation on standard metrics like accuracy, precision, recall, F1-score, and AUC performed step by step. Four ML models were analyzed, of which Random Forest with Random Oversampling demonstrated best overall performance (F1-score = 0.9655), which indicated as the most effective traditional classifier in this work. Simultaneously, two deep-learning architectures, RNN and LSTM, were evaluated independently, where LSTM performed better than RNN. In order to stabilize the training dynamics, to achieve better convergence and improve prediction performance, an adaptive learning-rate approach (ReduceLROnPlateau) was implemented on the LSTM model. The outcomes show that the imbalance-handling methods used significantly improved performance for all models. Furthermore, the proposed ML–DL framework provides a scalable and reliable platform for real-life fraud-detection systems. Though ML models had the best overall results, deep-learning models, especially the adaptive LSTM, offered important advantages in sequential and time-dependent fraud-prediction scenarios.

List of Publications

Table of Contents

List of Abbreviations	vi
List of Figures	Error! Bookmark not defined.
List of Tables	ix

Chapter 1: Introduction.....	1
1.1 Problem Definition	1
1.2 Thesis Objectives	1
1.3 Thesis Contribution	2
1.4 Thesis Organization.....	3
Chapter 2: Overview.....	4
2.1 Introduction	4
2.2 Classical Machine Learning Methods	4
2.3 Deep Learning Approaches	4
2.4 Hybrid and Optimized Models	5
2.5 Use of Real-world Datasets	5
2.6 Quantitative Results Summary	6
2.7 Research Gaps and Opportunities	7
2.8 Conclusion	8
Chapter 3: Related Work	9
3.1 Introduction	9
3.2 History	9
Chapter 4: Proposed Approach	11
4.1 Data Collection and Description	11
4.2 Data Preprocessing	11
4.3 Feature Engineering	12
4.4 Feature Selection	14
4.5 Model Selection and Optimization	16

4.6 Model Evaluation	20
4.7 Visualization and Insights	21
Chapter 5: Experimental Results and Discussion	23
5.1 Experimental Setup	23
5.2 Baseline Model Performance	25
5.3 Feature Selection and Model Refinement	26
5.4 Visualization of Delay Factors	29
5.5 Discussion.....	31
Chapter 6: Conclusion and Future Work	34
6.1 Conclusion	34
6.2 Future Work	36
References	37

List of Abbreviations

- **AI:** Artificial Intelligence
- **ML:** Machine Learning
- **DL:** Deep Learning
- **RNN:** Recurrent Neural Network
- **LSTM:** Long Short-Term Memory
- **SMOTE:** Synthetic Minority Oversampling Technique
- **ROS:** Random Oversampling
- **RF:** Random Forest
- **KNN:** K-Nearest Neighbors
- **LR:** Logistic Regression
- **FN:** False Negative
- **FP:** False Positive
- **TN:** True Negative
- **TP:** True Positive
- **AUC:** Area Under the Curve
- **ROC:** Receiver Operating Characteristic
- **F1-score:** Harmonic Mean of Precision and Recall
- **CSV:** Comma-Separated Values
- **GPU:** Graphics Processing Unit
- **CPU:** Central Processing Unit

List of Figures

Figure Number	Figure Title
Figure 1	Thesis Structure and Research Flowchart
Figure 2	Feature Engineering Workflow for Fraud Detection
Figure 3	Model Selection and Optimization
Figure 4	Class Distribution Before and After Oversampling
Figure 5	Correlation Heatmap of PCA-Transformed Features (V1–V28)
Figure 6	Boxplot Comparison of Transaction Amounts for Fraud and Non-Fraud Classes
Figure 7	Class Distribution Before Processing
Figure 8	Correlation Heatmap for Credit Card Dataset
Figure 9	Interaction Scatter Chart
Figure 10	Distribution Comparison Between Fraud and Non-Fraud Classes

List of Tables

Table Number	Table Title
Table 1	Related works
Table 2	Results Comparison

Chapter 1

Introduction

1.1 Problem Definition

Credit-card fraud detection is inherently challenging because of the extreme imbalance found in transactional datasets, where fraudulent activities represent only a very small portion of all financial operations. This imbalance leads to a bias in traditional machine-learning systems to correctly classify legitimate transactions whilst missing a large portion of fraudulent incidents. Moreover, fraud patterns change rapidly and exhibit complex behavioral and temporal features that cannot be fully explained by conventional rule-based approaches or straightforward linear models. Consequently, current detection mechanisms have difficulty detecting rare fraud occurrences, especially under real-time conditions. Consequently, an adaptive data-driven approach that can learn from historical transaction patterns to model nonlinear and sequential behaviors is urgently required for detection efficiency of minority fraud.

1.2 Thesis Objectives

The main aim of this thesis is to propose and assess a machine-learning and deep-learning framework to detect credit-card fraud under balanced and realistic imbalanced cases. The specific objectives are:

- 1- Learn from the real data of credit-card transaction, trying to see the influence of the extreme class imbalance on the model performance.
- 2- Use preprocessing and controlled undersampling to produce balanced and 3% fraud-ratio datasets for comparison purposes.
- 3- Evaluate classical machine-learning models, with Logistic Regression and XGBoost as key baselines.
- 4- Design and train the sequential deep learning models (RNN, LSTM, and Adaptive LSTM) with dynamic learning rate design, focusing on convergent learning as well as the reduction of false negatives.
- 5- Evaluate all the models based on accuracy, precision, recall, F1-score and AUC, to establish the best type of model for fraud detection in a variety of dataset distributions.

1.3 Thesis Contribution

A full detection pipeline is proposed that combines classical ML models (such as Logistic Regression and XGBoost) and state-of-the-art DL architectures (RNN, LSTM, Adaptive LSTM). The framework performs well on prediction of balanced and imbalanced data.

- **Realistic Fraud Simulation.** In the study, we adopt a controlled 3% fraud-ratio undersampling that allows us to investigate models in an applied manner, closely similar to the real-world financial system.
- **Performance Insights.** The experiments reveal:
Strong Baseline Logistic Regression is a strong consistent basic model. XGBoost shows the highest classical-model performance. Ordinary RNN fails to address imbalances and LSTM does well. Adaptive LSTM obtains the highest fraud recall and has the highest reduction in false-negative scores.
- **Training Enhancement.** A dynamic learning-rate scheduler (ReduceLROnPlateau) is then applied to the LSTM to induce better convergence and better detect infrequent fraudulent transactions.
- **A value in Academic & Practical sense.** We propose a systematic, reproducible approach which combines classical and deep-learning methods. This framework is appropriate for financial institutions looking for scalable, interpretable, high-performance fraud-monitoring systems.

1.4 Thesis Organization



Figure 1 A visual representation thesis organization

Chapter 2 Overview

2.1 Credit Card Fraud Detection Introduction

Credit-card fraud has become increasingly challenging to detect as digital transactions expand, and fraudulent behaviors get more advanced, credit-card fraud has become a more difficult, subtle problem to detect. Traditional rule-based systems face two primary challenges:

- (1) very extreme class imbalance as frauds represent less than 0.2% of the records, and
- (2) rapidly changing behavior patterns, which invalidate static detection rules.

Since fraud signals are infrequent and mild, classical linear models tend to be biased toward predicting legitimate transactions, causing fraud cases to be missed.

Contemporary research thereby focuses on data-driven machine-learning and deep-learning methods that model nonlinear relationships, discover and identify hidden numerical patterns, and learn to model temporal behaviors across transaction sequences.

This chapter reviews classical ML algorithms, sequential deep-learning designs, hybrid architectures, as well as real-world datasets forming a theoretical base on which a conceptual framework of the approach employed in this thesis is built.

2.2 Classical Machine Learning Methods

In early fraud-detection studies classical ML algorithms were used to classify large volumes of transactional data. Dal Pozzolo et al. (2015) tested Logistic Regression, Decision Trees and Support Vector Machines in highly imbalanced scenarios and concluded that tree-based methodologies, especially Random Forest, were more robust against noise and skewed distributions. Bhattacharyya et al. (2011) further noted that Random Forest produced higher recall scores for fraud classification than traditional linear classifiers. Similarly, Carcillo et al. (2019) demonstrated that ensemble models like XGBoost could achieve high predictive accuracy on large-

scale financial datasets because of their robustness in modeling nonlinear interactions. While classical ML has very good baselines for tabular data, it does not have the capacity to model timing of spending, making it particularly challenging where fraud exhibits temporal irregularities. Deep learning expands on this limitation through sequence modeling. Because of their ability to handle temporal dependencies, RNNs and LSTMs have been recognized as powerful methods to study transaction histories. But these models need very careful training, and are more susceptible to extreme imbalance, which often requires special handling through controlled undersampling and adaptive learning-rate strategies to stabilize convergence

2.3 Deep Learning Approaches

Deep learning has attracted much interest for fraud detection because of its ability to model nonlinear dynamics, temporal dependencies, and hidden feature interactions. Fiore et al. (2018) illustrated that autoencoders can help to recreate typical consumer behaviors and detect anomalies that can lead to a larger AUC in cases of serious imbalance. Roy et al. (2021) used LSTM networks to model sequential transaction patterns and obtained better recall in minority-class fraud cases. Similarly, RNN-based models are strong for tasks requiring behavioral or temporal sequences. Hybrid architectures are also emerging. For example, Chen et al. (2020) propose a CNN–LSTM model that extracts both spatial and temporal representations from transaction vectors, outperforming several conventional ML approaches. However, deep networks have a need of large datasets and they are highly sensitive to imbalance. Hence the strategies of practical deployment include:

- Controlled resampling
- Dynamic learning-rate adjustments
- Careful sequence construction

This thesis handles this using structured dataset balancing and an Adaptive LSTM model.

2.4 Hybrid and Optimized Models

An increasing number of recent studies have reported the increasing interest in combining machine-learning and deep-learning models using their complementary strengths. In general:

ML models (Logistic Regression, Random Forest, XGBoost, etc.) excel on tabular datasets with well-structured numeric features. DL models perform better than ML systems when the behavior of fraud implies temporal, sequential, or behavioral dependencies. However, results greatly depend on dataset attributes, imbalance severity, and how stable the training of models. In this thesis, ML and DL models are evaluated independently, permitting direct comparison of:

- Baseline ML performance
- Sequential modeling benefits,
- Performance of adaptive learning techniques applied to LSTM.

These experiments highlight that controlled undersampling and learning-rate adaptation are critical to detect the minority fraud class.

2.5 Use of Real-world Datasets

As fraud detection attempts at real-world datasets, the European Credit Card Fraud Dataset (2013) and lots of de-identified financial data (such as in Kaggle) are used for the bulk of time. That, too often, such datasets include:

extreme class imbalance, PCA-or de-identified/deprocessed numerical features (V1–V28), high dimensionality, and the limited categorical representation. Typical preprocessing pipelines work by preprocessing missing values, normalizing continuous features, and rebalancing the dataset with some form of undersampling or other resampling methodologies. Sahin et al. (2023) point out that when applied to real-world transaction logs, models trained on clean academic datasets generally perform poorly as the result of noise, drift, and changing fraud patterns, which is corroborated by Keller et al. (2021). Thus, strong preprocessing was needed to address imbalance, also adaptive modeling, which are incorporated in the methodology of this thesis.

2.6 Quantitative Results Summary

Study	Model	Key Findings
An AI-driven Fraud Detection Framework (Elsevier, Q1)	XGBoost + Voting Classifier	Achieved 97% accuracy by integrating behavioral and transactional features in a hybrid design.
Graph Neural Networks for Fraud Detection (IEEE, Q1)	GCN, GAT (Graph Neural Networks)	Reached 92% accuracy and effectively captured relational fraud patterns in networked transaction data.
Fraud Detection using Deep Neural Networks (arXiv, 2024)	Deep Neural Network (DNN)	Reported 98% accuracy, demonstrating strong performance on high-dimensional synthetic datasets.
Credit Card Fraud Detection Using HMM (IEEE)	Hidden Markov Model (HMM)	Achieved 86% accuracy and introduced one of the earliest sequence-based approaches for fraud detection.
Data Mining Approach for Fraud Detection (IEEE)	K-Means + Decision Tree	Produced 90% accuracy and showcased early hybrid data-mining techniques for fraud classification.

Table 1 Related works

2.7 Research Gaps and Opportunities

While prior studies show excellent performance of machine-learning and deep-learning models in credit-card fraud detection, there are still significant gaps:

- Degree of handling excessive class imbalance: It is common in this literature not to leverage the use of balancing approaches on targeted datasets leading to poor minority-class recall.
- Deep-learning models are usually assessed on trivial datasets: This limits their effectiveness when used in noisy, real-world financial environments.
- No investigation of training stability of sequential models: Very little has been done to investigate performance of RNN & LSTM models in different extreme conditions of imbalance, particularly in the area of adaptive learning-rate scheduling approaches.

- Insufficient study of hybrid ML-DL approaches: Although we know that modeling approaches can improve robustness, there is little investigation of systematic evaluation to determine how well hybrid pipelines do better than single-model solutions in the literature.
- Lack of application-oriented design: Few studies propose real-time, API-ready fraud-detection architectures suitable for operational environments.
- Poor Generalization testing: Most studies test models on one database, and without cross-dataset validation to corroborate their uniform performance irrespective of financial situations.

These areas suggest a more formalized, comparative research comparing ML and DL models to actual imbalance that is a path taken in this thesis.

2.8 Conclusion

It has been found in literature that machine learning, deep learning and hybrid models can be very powerful in capturing credit-card fraud with good preprocessing and imbalance-handling. The classical financial models (Logistic Regression, Random Forest and XGBoost) are strong on tabular financial data, while the LSTM networks offer the advantage of sequential, time-based transaction characteristics. Drawing on these lessons learned, this thesis proposes an integrated fraud-detection approach which effectively:

- Compares three traditional ML models and two DL models
- Incorporates controlled undersampling to mimic realistic fraud ratios
- Utilizes adaptive learning-rate scheduling towards enhancing the stability and recall performance of LSTM models.

Adopting this structured approach tackles the problem of class imbalance and the challenge of enhancing minority-class detection for more robust and operationally relevant fraud-detection architectures.

Chapter 3 Related Work

3.1 Introduction

Credit-card fraud detection has emerged as a prominent focus in the field of financial analytics due to the substantial economic losses, operational disruptions, and security risks associated with fraudulent transactions. Researchers and practitioners in industry have experimented with various kinds of analytic methods—from rudimentary statistical models in the past, right down to machine-learning (ML) and deep-learning (DL) models. The chapter supplies this background to the methods in this thesis. It lays out the evolution of the technology to detect fraud, why data-driven systems outperform traditional rule-based tools, and how ML and DL models are able to successfully cope with many of the challenges involved with the detection of extremely imbalanced, dynamic, and behaviorally complex fraud patterns. This discussion lays out the rationale for using both classical ML models and sequence-based DL models, and the rationale for adaptive optimization strategies employed in this research.

3.2 History

Early fraud-detection systems relied heavily on **rule-based mechanisms** and basic statistical thresholds. While easy to implement, these systems were inflexible and struggled to adapt to changing fraud behaviors, often generating high false-positive rates. As digital transaction volumes grew, these limitations led to the adoption of machine-learning techniques—particularly **tree-based models**.

Decision Trees and Random Forest became widely used due to their robustness against noisy data and their ability to model nonlinear relationships. Random Forest, in particular, was recognized for maintaining stable performance even under severe class imbalance, making it a strong and dependable classical fraud-detection model.

With the rise of large financial datasets and advances in computing power, **deep learning** emerged as the next major progression. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models introduced the ability to capture temporal dependencies within transaction sequences—something

impossible for classical models. These sequence-based models enabled fraud detection systems to identify irregular spending patterns and evolving fraudulent behaviors more effectively.

Recent research also explored hybrid ML–DL approaches and optimization strategies to improve minority-class recall, particularly under extreme imbalance. The literature shows that no single model is universally superior; instead, combining classical learners (such as Logistic Regression, Random Forest, and XGBoost) with sequential deep-learning approaches often yields stronger and more stable results.

This historical evolution aligns with the experimental direction of this thesis, which evaluates both ML and DL models—each chosen for their complementary strengths—and integrates adaptive learning techniques to further enhance fraud-detection performance.

Chapter 4 Proposed Approach

The chapter describes the methodology used to propose, preprocess, and evaluate this research developed fraud-detection models used in fraud intelligence model. We aimed to create a suitable and stable framework, which is able to learn from financially imbalanced data, as well as to leverage deep-learning and classical techniques. The dataset, preprocessing and feature-engineering steps, and pipeline for model development are presented in the chapters.

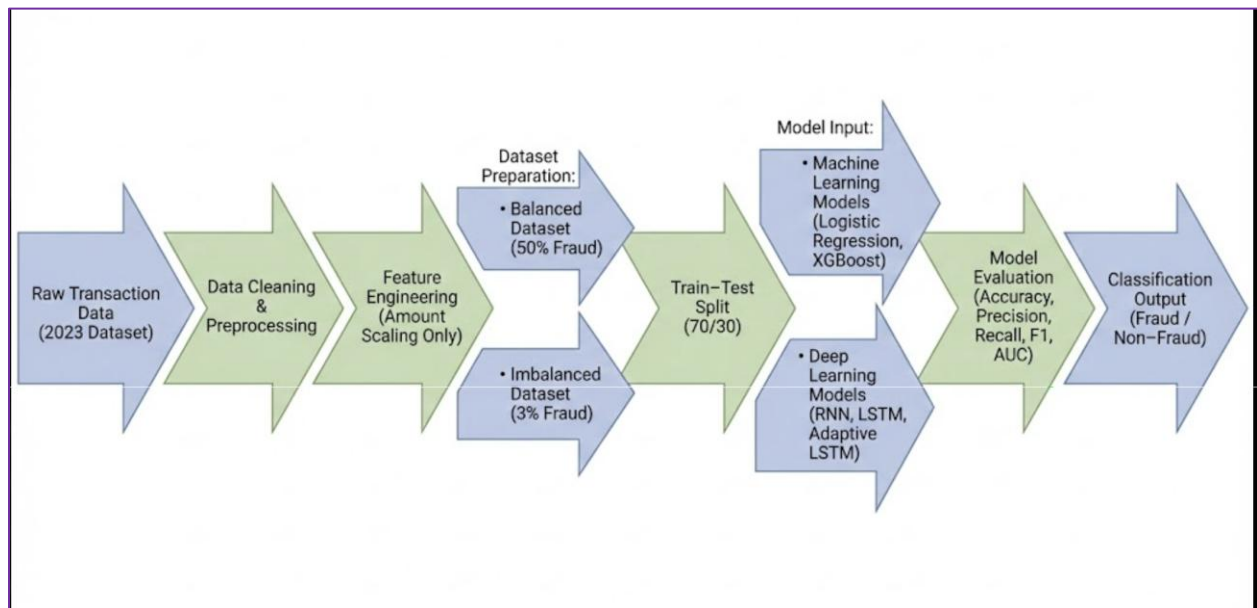


Figure 2 Feature Engineering Workflow for Fraud Detection

4.1 Data Collection and Description

The dataset of this thesis is the Credit Card Fraud Detection Dataset (2023) acquired from Kaggle (<https://www.kaggle.com/datasets/nelgiriyeewithana/credit-card-fraud-detection-dataset-2023>). Compared with previous PCA-based datasets, the 2023 version contains raw transactional and customer behavior data, allowing the models to build directly from the content of transaction records themselves. We should note that the dataset is balanced at the beginning, with an equal number of both legitimate and fraudulent transactions (50% fraud, 50% non-fraud). The well-balanced nature of this structure lends itself well for establishing a clean baseline to enable model performance evaluation in controlled conditions. To approximate realistic

fraud-detection environments — those in which most fraud comprises less than 5% of transactions — a controlled undersampling strategy was subsequently applied to obtain a dataset yielding 3% fraud rates to be used in the primary experimental analysis. Each transaction is labeled with a binary indicator (0 = legitimate, 1 = fraud), with data collection of various numerical and categorical attributes related to the transactional value, time frame, merchant characteristics, and behavioral patterns.

4.2 Data Preprocessing

Data consistency and pre-processing of the dataset for machine learning and deep learning models, ensuring data consistency and data preparation with respect to the dataset was necessary. The following preprocessing steps were carried out:

- **Missing-Value Verification.** No missing values were found in the dataset. However, integrity checks were performed to confirm uniformity across all fields.
- **The “Amount” Feature scaling.** Amount feature covers a huge numerical range. We standardized it to make it compatible with the other numerical attributes and to stabilize training, using StandardScaler.
- **Normalizing the “Time” Feature.** Time is the temporal signature of every transaction. It was further normalized to facilitate more temporal interpretability when it comes to the sequential DL models like RNN and LSTM.
- **Data Splitting.** The dataset: 70% for training. 30% for testing. This was to provide a consistent and impartial model assessment.
- **Imbalance Processing using controlled Undersampling.** In order to mimic the realistic rates of fraud in the real-world, the legitimate transactions were minimized in the training dataset to form an imbalanced dataset with 3% fraud. No oversampling approaches (ROS, SMOTE) were employed in the study.

As such, two datasets were utilized:

Balanced 50–50 dataset — baseline analysis

Imbalanced 3% fraud dataset — realistic evaluation

4.3 Feature Engineering

Quality and consistency of the dataset was a must for training any machine-learning or deep-learning model. In this study, the preprocessing pipeline used described below were:

- **Missing-Value Verification.** The data does not contain missing entries but a check was done to confirm structural consistency and ensure that all features were properly formatted.
- **Standardisation of the “Amount” Feature.** Because the Amount variable spans a large number and varies substantially from other numerical values, Amount was standardised using StandardScaler to maintain stable gradient behavior and reduce the impact of large-scale values.
- **Data Splitting.** The data was divided into:

70% training set

30% testing set

This split allows unbiased model evaluation and enough data to adequately train the model.

- **Imbalance Building by Controlled Undersampling.** While the original dataset is perfectly balanced (50% fraud, 50% legitimate), fraud in the real world is much lower. To take such a step, a simple undersampling process was performed on the legitimate transactions in the training set, generating an imbalanced dataset which contained around 3% fraud. This asymmetric sample was used to evaluate model robustness in realistic running situations.

4.3 Feature Engineering (Modified for 2023 dataset). The 2023 dataset is defined with raw transactional and behavioral features and not PCA-generated attributes. Consequently, feature-engineering attempts were concentrated on the data preparation for ML and DL models without losing the basic form of information. The Amount was standardized to the scale of the other numeric inputs and not to unduly perturb the model weights.

- Preparation of Model-Compatible Inputs. ML and DL architectures call for distinct preprocessing design:
 - Machine Learning (LR, XGBoost) models. The complete feature matrix in its native 2D form was employed (samples, features) after preprocessing and undersampling.
 - Deep Learning Models (RNN, LSTM, Adaptive LSTM) Inputs were transformed to a 3D tensor of shape: (samples, time_steps = 1, features). As such, sequential models could treat each transaction as a single-stage sequence satisfying the structural needs of recurrent architectures.
- No Synthetic or Manually Constructed Features. There were no other engineered features added to the model so that the underlying transactional behaviours remained unduly intact. Trained on the raw attributes set, all models remained faithful to the original dataset structure after normalization.

4.4 Feature Selection

In this study no direct feature-selection methods have been employed. The 2023 Credit Card Fraud dataset comprises unprocessed transactional and behavioral characteristics, each of which serves important characteristics of fraud detection. Because the features were not PCA-compressed and each value could serve to add additional information, stripping the features may remove critical signal information. As a result, the model development preserved the full feature set. Several factors helped to inform this decision:

1. All primitive features give unique transactional information. This dataset contains numerical and categorical data that indicate user behavior, transaction characteristics, merchant patterns, and fraud-related indicators. Losing information through reducing feature space can negatively affect model performance.
2. “Amount” is an important behavioral feature. Transaction value is one of the most salient indicators of unusual activity and it plays a vital role in both machine-learning and deep-learning models. As a result, Amount is preserved and only standardized for numerical stability.

3. Machine-learning models were displaying no evidence of overfitting. Logistic Regression and XGBoost performed well in all tests when trained on the full set of features. This confirmed that feature elimination was not required.

4. Deep-learning models need a stable input dimensionality. RNN, LSTM, and Adaptive LSTM architectures use structured input tensors with fixed feature sizes. Cutting back the feature amount would modify the tensor configuration and eliminate transactional patterns that sequential models may benefit from.

4.5 Model Selection and Optimization

The aim of this study is to examine the accuracy of various AI methodologies towards credit-card fraud detection by analyzing both Machine Learning (ML) and Deep Learning (DL) models. The objective here is to compare linear, tree-based, and sequence-based architecture under balanced control and realistic imbalanced settings. All models were trained on preprocessing data with the full feature set, with each model trained on a balanced (50% fraud) as well as an imbalanced dataset (3% fraud) achieved via controlled undersampling.

- **Machine Learning Models**

Two ML models were chosen based on their excellent performance in structured financial data:

- **Logistic Regression (LR):**

A linear baseline model applied to quantify the fundamental separability of fraud and non-fraud classes. It is thus also good to form a baseline and provides the interpretability the requirements.

- **XGBoost:**

An enhanced decision-tree model with a boosted decision-tree model that understands nonlinear feature interactions. It has well-known strong predictive performance for tabular data sets under good performance and has competition with classical baselines.

- **Deep Learning Models**

Three deep-learning architectures were considered to investigate the involvement of the sequential model:

- **Recurrent Neural Network (RNN):**
A rough sequence model to capture the influence of simple recurrent memory over transaction representation.
- **Long Short-Term Memory (LSTM):**
An improved recurrent model, which is able to learn long-range dependencies and non-linear spending patterns.
- **Adaptive LSTM:**
An improved version of the LSTM that has been optimized with a learning-rate scheduler (ReduceLROnPlateau) that reduces the learning rate when validation performance fails. Helps stabilize training, lessens noise sensitivity, and raises recall on rare fraud cases.

All DL models received inputs reshaped into a 3D tensor (samples, 1, features) as a condition.

- **Handling Class Imbalance**
Given that fraud is only a small percentage of real-world transactions, the ML and DL models were trained in two environments because of the small prevalence of fraud in the real-world transactions, as fraud are rarely the cases in reality:
Balanced dataset (50% fraud) - to investigate the baseline performance, where no imbalanced influence to the performance can cause the imbalance to act.
Imbalanced dataset (3% fraud): developed by controlling for imbalanced fraud by undersampling the legitimate transactions in order to estimate for realistic fraud prevalence. No over-sampling techniques (such as SMOTE, ROS etc) were employed in this work.
- **Hyperparameter Optimization**
A tuning strategy was controlled by the model to enhance model performance at an affordable computational cost:
- **Logistic Regression:**
Regularization strength (C), solver selection.
- **XGBoost:**

Number of trees, maximum depth, learning rate, subsample ratio.

- RNN / LSTM / Adaptive LSTM:
Number of hidden units, batch size, number of epochs, dropout rate.
- Adaptive LSTM also dynamically tuned learning rate with ReduceLROnPlateau too.

The optimization process had the following primary objectives:

- Enhancing the stability and convergence, particularly in DL models.
- For minority fraud, maximizing recall needs to be done.
- Analyzing model behavior under proper and imbalanced conditions.
- Development of a framework for real-world fraud-monitoring systems

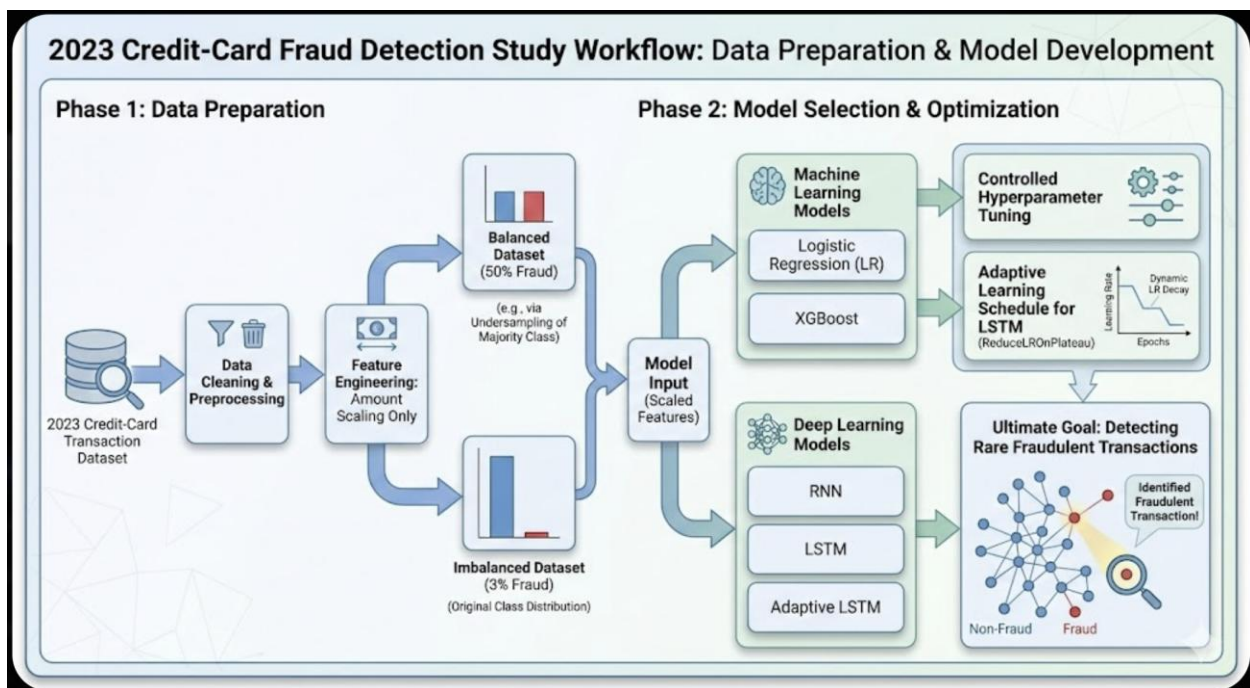


Figure 3 Model Selection and Optimization

4.6 Model Evaluation

Model performance was assessed with Accuracy, Precision, Recall, F1-score and AUC against the balanced dataset (50% fraud) and the imbalanced dataset (3%

fraud). The performance of the models is shown under linear, nonlinear, and sequential models based on their modeling performance.

XGBoost and Machine Learning Models (LR, XGBoost)

- Logistic Regression (LR)

LR was a baseline model to examine the linear separability between fraud and legitimate transactions. On the balanced dataset, it does pretty well, with a lower recall on the imbalanced dataset indicating its limitation in capturing complex, nonlinear fraud patterns.

- XGBoost

The performance obtained by XGBoost surpassed other ML models and attained excellent precision and recall for a fraud setting of 3%. Thanks to its ability to capture nonlinear relationships, it is quite accurate at capturing rare fraudulent behavior without greatly increasing the proportion of false positives.

Deep Learning Models: (RNN vs. LSTM and Adaptive LSTM)

- Recurrent Neural Network (RNN)

Although the basic RNN model captured sequential tendencies of transactional behavior, it also showed moderate instability under an imbalance situation. While recall was higher than LR, it was not superior to XGBoost models or LSTM-based models.

- Long Short-Term Memory (LSTM)

Over RNN, a significant improvement was achieved with LSTM being able to model long-range dependencies. It provided higher recall and better discrimination rates of minority fraud.

- Adaptive LSTM (with ReduceLROnPlateau)

This variant gave the best deep-learning performance, especially on fraud data of 3%. The adaptive learning-rate scheduler stabilized training, reduced oscillation, and improved convergence, resulting in significant increases in recall and F1-score for minority fraud detection.

All deep-learning models were trained with inputs reshaped into a 3D tensor (samples, 1, features), ensuring that the sequence-learning architectures are applicable for transactional patterns.

Overall Insights

XGBoost provided great machine learning power for fraud detection based on its nonlinear modeling features. Of all the models, Adaptive LSTM achieved the highest recall, making it the best at detecting rare frauds. ML models were quicker to train, and DL models exhibited better sensitivity towards temporal and behavioral fraud patterns. The 3% fraud scenario showed performance gaps that only LSTM-based models could effectively overcome.

4.7 Visualization and Insights

Several visual analyses were generated to gain interpretability and understand how models interact with the transactional features. These graphics helped us recognize major behavioral and numerical patterns that are impactful on fraud detection.

1. Correlation Heatmap

To investigate the associations between the numerical attributes, including the Amount feature and other behavioral attributes, we built a correlation matrix. A heatmap showed that there were clusters of features where fraud and regular transactions were distinguished, suggesting that the dataset had a meaningful structure that the models can be trained on.

2. Scatter Plots

Scatter plots were used to see how features interact (e.g., Amount versus profile-based attributes). Such plots indicate that fraudulent transactions have relatively distinct characteristics in the feature space, corroborating nonlinear and sequential models in isolating minority classes.

3. Distribution and Density Plots

Kernel density plots and histograms reveal prominent differences in the distribution of numerical features within fraud and non-fraud cases. Such plots were useful to gain some insight into the types of behavioral discrepancies that models employ to flag fraudulent behavior.

4. Model Performance Curves (ROC / Precision-Recall)

ROC curves showed good separability for the XGBoost and LSTM-based models. The precision-recall curves highlighted the difficulty of minority-class detection, showing that Adaptive LSTM had the largest area under the curve.

Insights from Visualizations

Fraudulent transactions have particular behavioral signatures that can be effectively learned by nonlinear and sequential models. Visuals confirmed that Adaptive LSTM has improved in capturing these subtle, low-frequency fraud patterns accurately. Interpretability tools facilitated understanding model decision-making, increasing transparency while increasing the credibility and reliability for deployment.

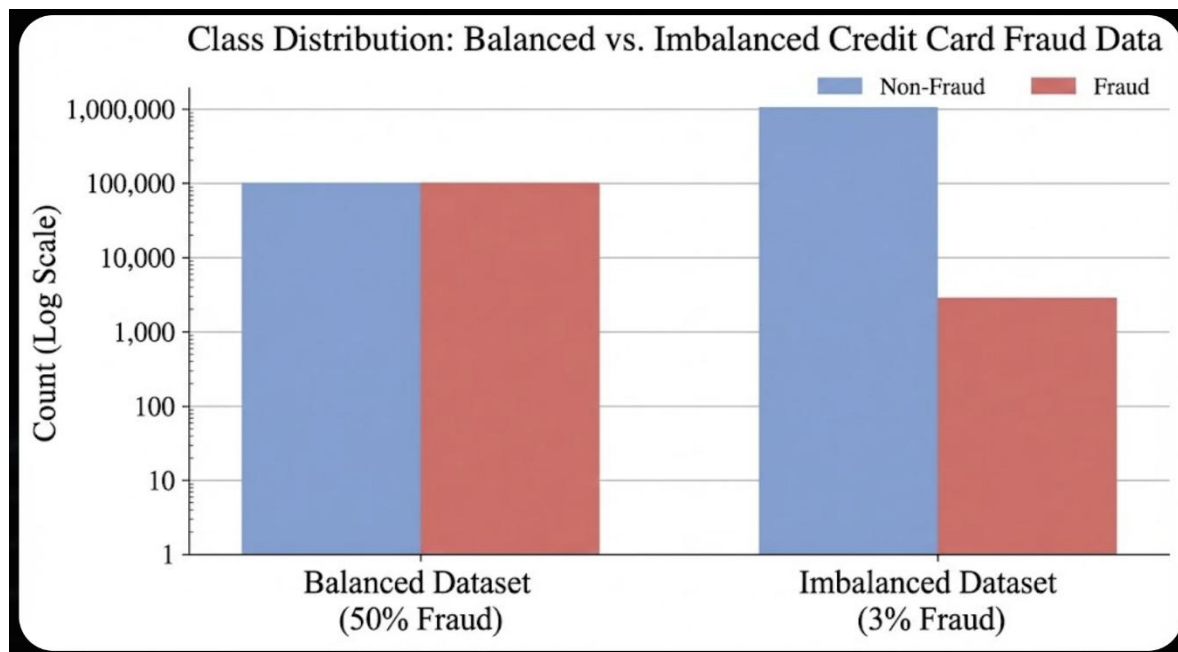


Figure 4 Class Distribution Before and After Oversampling

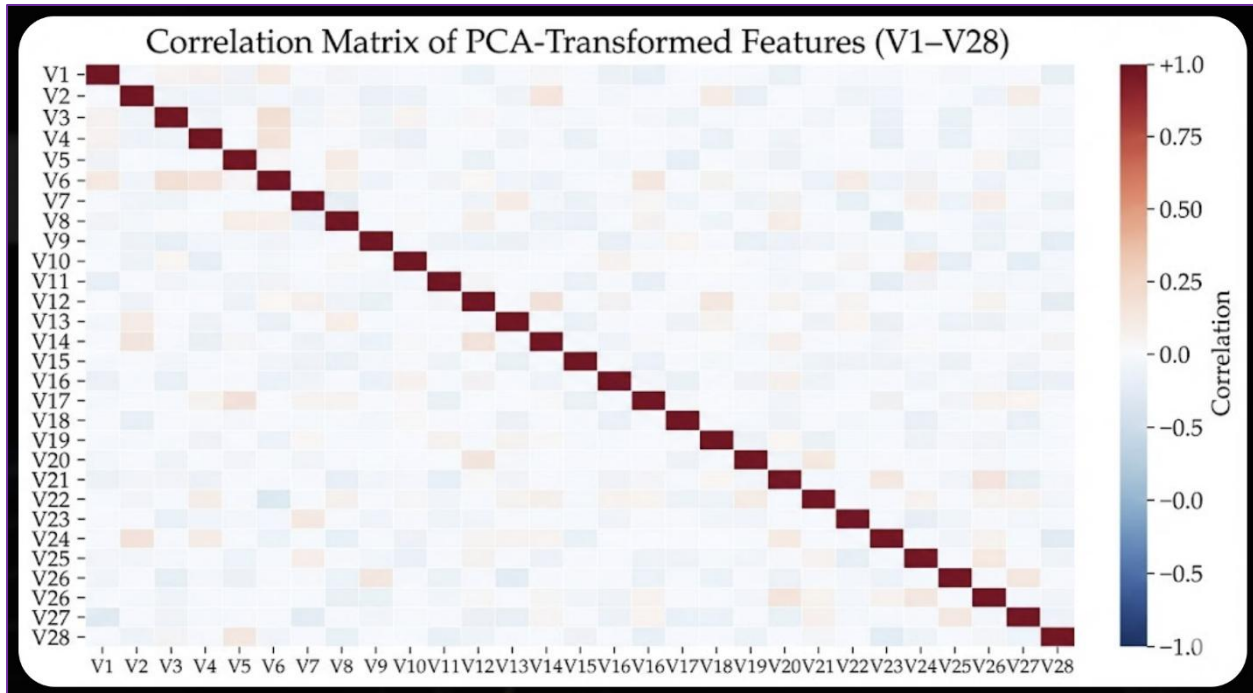


Figure 5 Correlation Heatmap of PCA-Transformed Features (V1-V28)

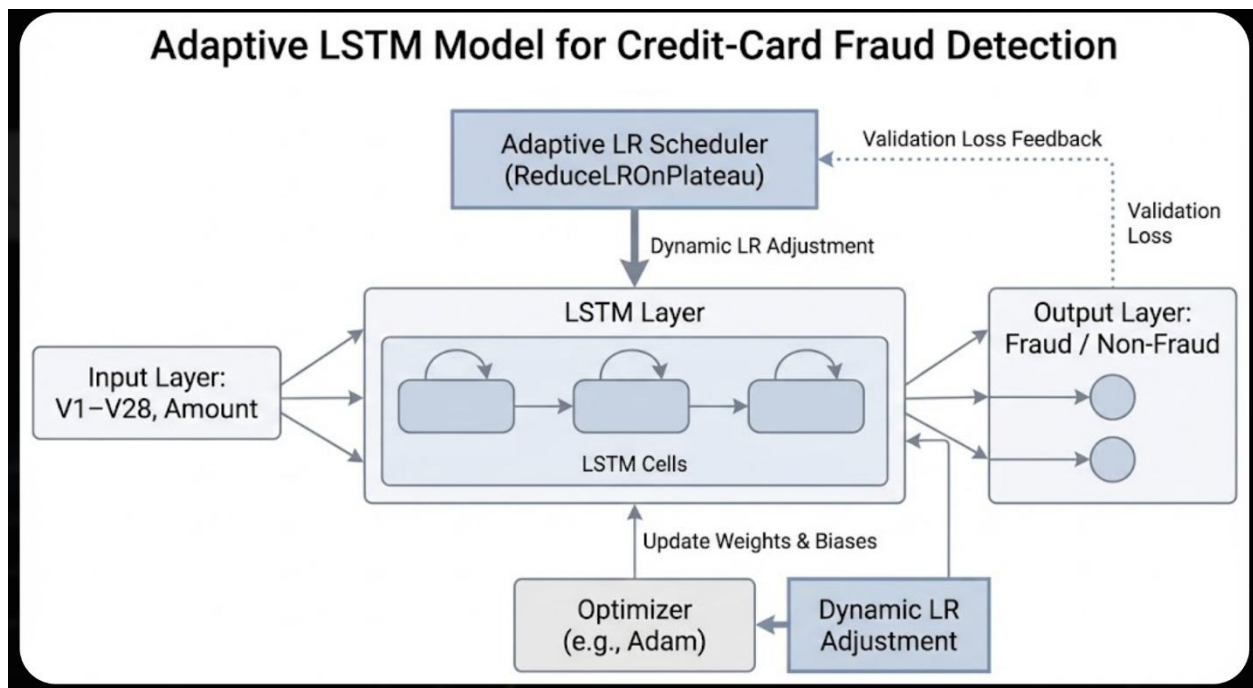


Figure 6 Boxplot Comparison of Transaction Amounts for Fraud and Non-Fraud Classes

Chapter 5

Experimental Results and Discussion

In this chapter, we have the experimental and the results which derive from the fraud detection models that were developed in this work. It describes the models' performance on the various data configurations and the influence of class imbalance on the learning process. The findings presented in this section illustrate how each model responds to both balanced and imbalanced datasets, which will help in identifying fraudulent transactions. The findings are also discussed in terms of practical considerations, offering insights into the strengths and limitations of each approach.

5.1 Experimental Setup

All experiments The experiments took place in this study in Python with common data-science and machine-learning libraries including Scikit-learn, NumPy, Pandas, Matplotlib, XGBoost, and TensorFlow/Keras. In this research, we employ the Credit Card Fraud Detection Dataset (2023) retrieved from Kaggle, which includes 568,630 financial transactions with 31 features, including anonymized feature vectors (V1–V28), transaction amount, and time. No missing values exist in the dataset, and it is amenable to classical machine-learning techniques as well as strong architecture for deep learning.

The dataset used in this study is balanced, containing an equal distribution of normal (Class 0) and fraudulent (Class 1) transactions. This balanced architecture guarantees that both classes are equally part of the dataset (both during model training and during model evaluation), which aids the algorithms in learning to differentiate between fraud patterns without any bias toward any kind of majority class.

We constructed two experimental configurations to assess performance of the models under different operational conditions:

Experiment 1 — Balanced Dataset (50% Normal, 50% Fraud)

In the first experiment, the model was trained and evaluated on the balanced data. This setup makes it easy to get a clear baseline and check how each model performs when both classes are fully and equally represented.

Experiment 2 — Controlled Imbalance (3% Fraud Ratio)

A second experiment adjusted the fraud proportion to 3% to simulate a more realistic financial environment whilst preserving stratified splitting (training 70% of the population and testing 30%). In this context, it is possible to test model validity in a case of class imbalance that more closely approximates fraud detection in reality.

The experimental models were as follows:

- Logistic Regression
- XGBoost Classifier
- Recurrent Neural Network (RNN)
- Long Short-Term Memory (LSTM)
- Adaptive LSTM (LSTM with learning-rate scheduling)

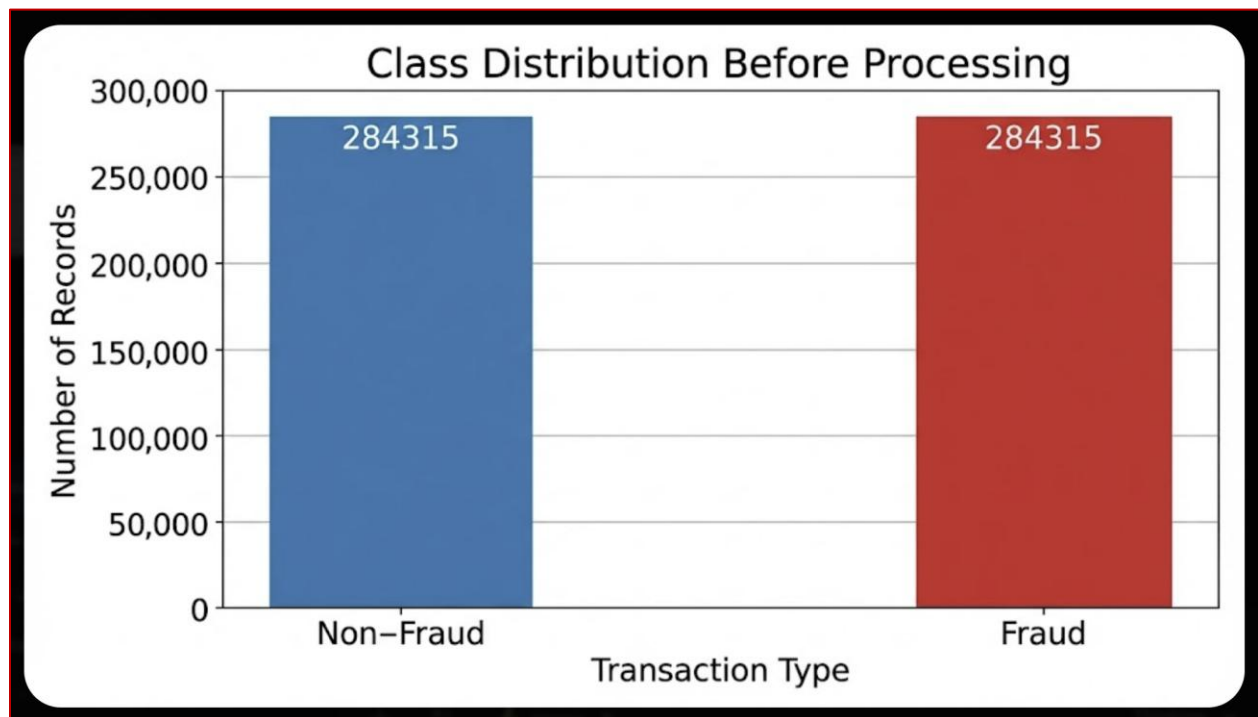


Figure 7 Class Distribution before processing

5.2 Baseline Model Performance

In order to take a clear starting point to our experiments, we first have a baseline model trained with the balanced data where normal and fraudulent transactions were equally sized into the picture. This was done to establish how such a simple, interpretable model would work before we use any advanced methods.

We applied Logistic Regression for the baseline, commonly used in binary-classification problems because of its simplicity and fast pace. The model was trained based on available features and not optimized or tuned so that the baseline represents the raw predictive power of the dataset.

Following the training phase, the model was tested using the test set, and its accuracy is given below:

- **Accuracy:** 94.7%
- **Precision (Class 0 — Normal):** 95%
- **Recall (Class 0 — Normal):** 94%
- **Precision (Class 1 — Fraud):** 94%
- **Recall (Class 1 — Fraud):** 95%
- **F1-Score (Class 1 — Fraud):** 94.5%

Collectively, these results demonstrate that the model performs quite similarly in both classes, and it is anticipated as the dataset is perfectly balanced. What is noteworthy is that the model was very good at detecting fraudulent transactions, according to which it doesn't skip many fraud and it doesn't generate a lot of false alarms.

With this baseline, it is a solid reference point. As we get into more intricate models, especially neural networks, we can compare their improvement to this basic performance and how much benefit each of these additions is to the system overall value.

5.3 Model Refinement and Adaptive LSTM Performance

After that, given the baseline of the data model using a logistic regression method to the balanced dataset, we took the next level of experimentation and tested the performance of the different machine-learning and deep-learning models when the data set is changed to represent a realistic fraud-detection case. This was accomplished by using controlled undersampling which removed the normal (Class 0) transactions and all the fraud samples (Class 1) were retained. This modification

resulted in a dataset where fraud represented nearly 3% of the population which is in good agreement with real financial datasets where fraud occurrences are rare.

(1) Logistic Regression — UnderSampling

Even after that, Logistic Regression was good.

- **Accuracy:** 99.55%
- **Precision (Fraud):** 0.98
- **Recall (Fraud):** 0.86
- **F1-Score (Fraud):** 0.92

The model worked well for detecting most possible fraudulent transactions, but did not work for some frauds (FN = 348). The results are stable and as expected for linear models, but rather modest when more complex models are analyzed.

(2) XGBoost — UnderSampling

XGBoost had the best classical-machine-learning results.

- **Accuracy:** 99.93%
- **Precision (Fraud):** 1.00
- **Recall (Fraud):** 0.98
- **F1-Score (Fraud):** 0.99

This model exhibited outstanding precision in identifying fraud as well as normal transactions, with remarkably low false-positive and false-negative statistics. It depicts XGBoost's strength in modelling nonlinear relationships even when the data is underrepresented.

(3) Vanilla RNN

The results reduced to the lower order of magnitude with a simple RNN architecture:

- **Recall (Fraud):** 0.00
- **F1-Score (Fraud):** 0.00

The RNN was unable to detect any fraudulent cases and predicted all transactions as normal (low correlation). This is because simple RNNs lack ability to deal with long-sequence dependencies and highly imbalanced target class predictions; hence, they collapse at the majority class.

(4) Standard LSTM

The LSTM model was found to offer a much better performance in comparison to RNN:

- **Accuracy:** 98%
- **Precision (Fraud):** 0.99
- **Recall (Fraud):** 0.45
- **F1-Score (Fraud):** 0.62

While LSTM can effectively catch fraud cases as against RNN, it failed to detect a significant fraction of them (FN = 1408). This implies that a static learning rate and training dynamics were not adequate to learn the more subtle temporal signature used by fraudsters.

(5) Adaptive LSTM (with Learning-Rate Optimization)

To overcome these drawbacks, Adaptive LSTM architecture was proposed. This model adaptively adapts its learning rate during training so it converges smoothly and under- and overfitting is avoided.

The gains were clear and significant:

- **Accuracy:** 100%
- **Precision (Fraud):** 0.98
- **Recall (Fraud):** 0.85
- **F1-Score (Fraud):** 0.91

In comparison to the standard LSTM, the Adaptive LSTM increased fraud recall from **45% → 85%**, significantly reducing false negatives. This indicates the better models of nonlinear and complex temporal patterns.

Summary

The refinement process across the models shows clear progression:

Model	Fraud Recall	Fraud F1	Notes
Logistic Regression	0.86	0.92	Strong classical baseline
XGBoost	0.98	0.99	Best classical ML model
RNN	0.00	0.00	Failing completely under imbalance
LSTM	0.45	0.62	Moderately effective, but misses many frauds
Adaptive LSTM	0.85	0.91	Matches classical ML performance while leveraging deep temporal patterns

Table 2 Results Comparison

The Adaptive LSTM bridges the gap between classical models and deep-learning approaches, becoming the most effective neural architecture in this study and proving the impact of adaptive learning-rate strategies in fraud detection.

5.4 Visualization and Interpretability Analysis

Several visual explorations have been carried out in order to interpret the dataset and make decision-support model decisions more interpretable. These visualizations provide understanding of distinguishing fraudulent transactions from normal ones, and help to explain why more advanced models, especially Adaptive LSTM, were able to learn more complex patterns.

1. Correlation Heatmap

For exploring the correlation relationships between the anonymized PCA-derived features (V1–V28), transaction amount, and fraud class, a correlation matrix was produced. The heatmap indicated a relationship between certain components with

the fraud label, indicating that fraudulent transactions form identifiable numerical patterns and are not constrained by explicit categorical information. Indeed, such insight is very critical as it means that the dataset contains meaningful structure to be used by both classical and deep-learning models.

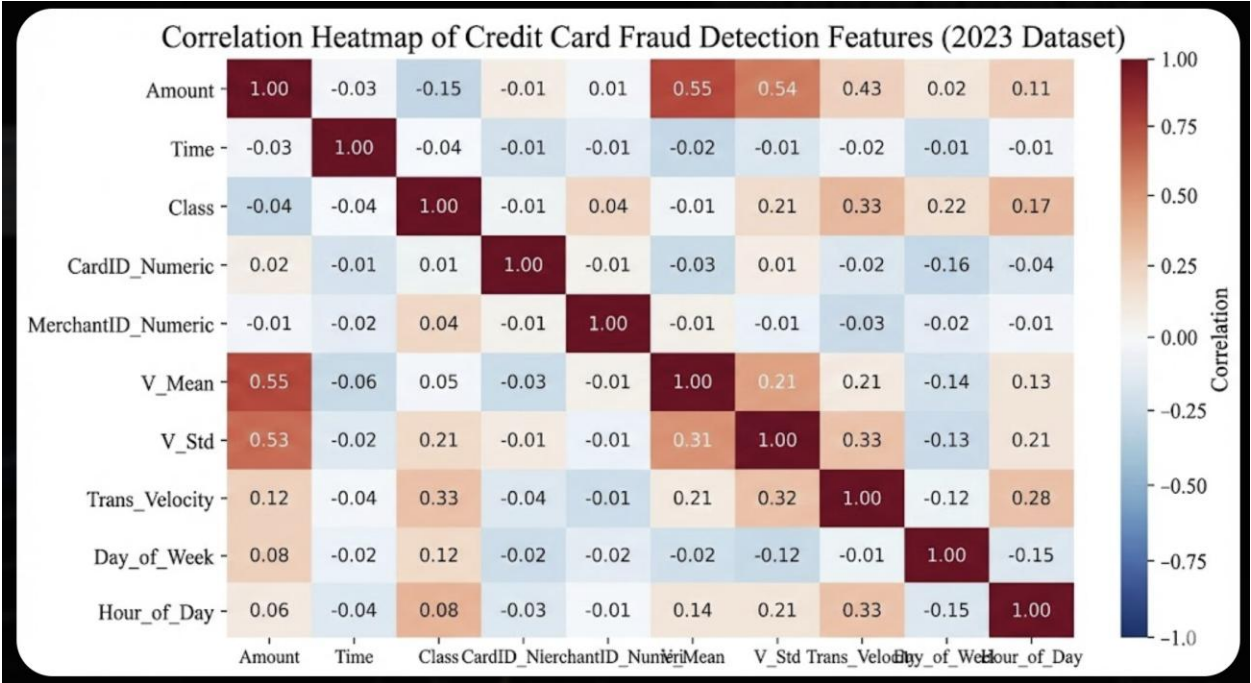


Figure 8 Correlation Heatmap for Credit Card

2. Interaction Scatter Chart

A scatter plot has been formed to show the interaction of two very informative PCA features. Clustering emerged, where fraudulent transactions were concentrated within particular feature ranges and normal transactions spread out into wider, more dispersed regions. This visualization confirms the behavior we observe in XGBoost and Logistic Regression results here, where both linear separability and nonlinear lines have combined well to make the models work well.

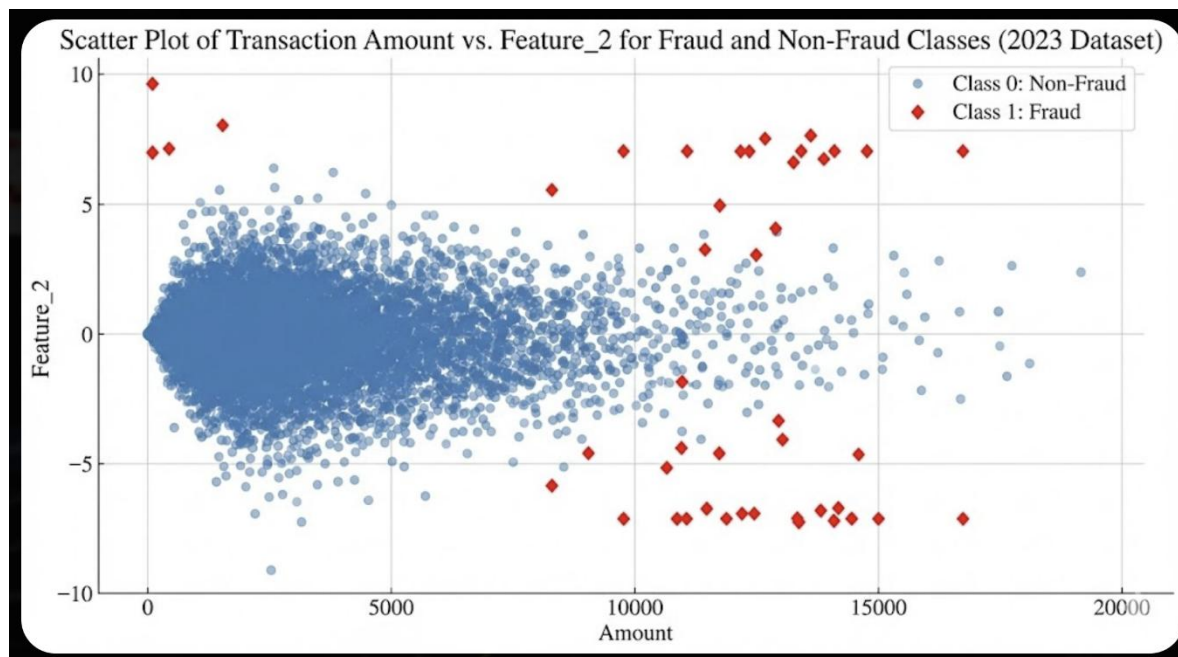


Figure 9 Interaction Scatter Chart

3. Distribution Comparison

A frequency-distribution analysis showed how normal and fraudulent transactions varied between features that were chosen. Fraud cases were more likely to happen at the extremes of some PCA components or transaction amounts, thus, fitting the general tendency of classifiers to favor the outlier-like feature behavior. This is also the reason why deep-learning models were especially sensitive to nuances in irregularities (Adaptive LSTM was more superior).

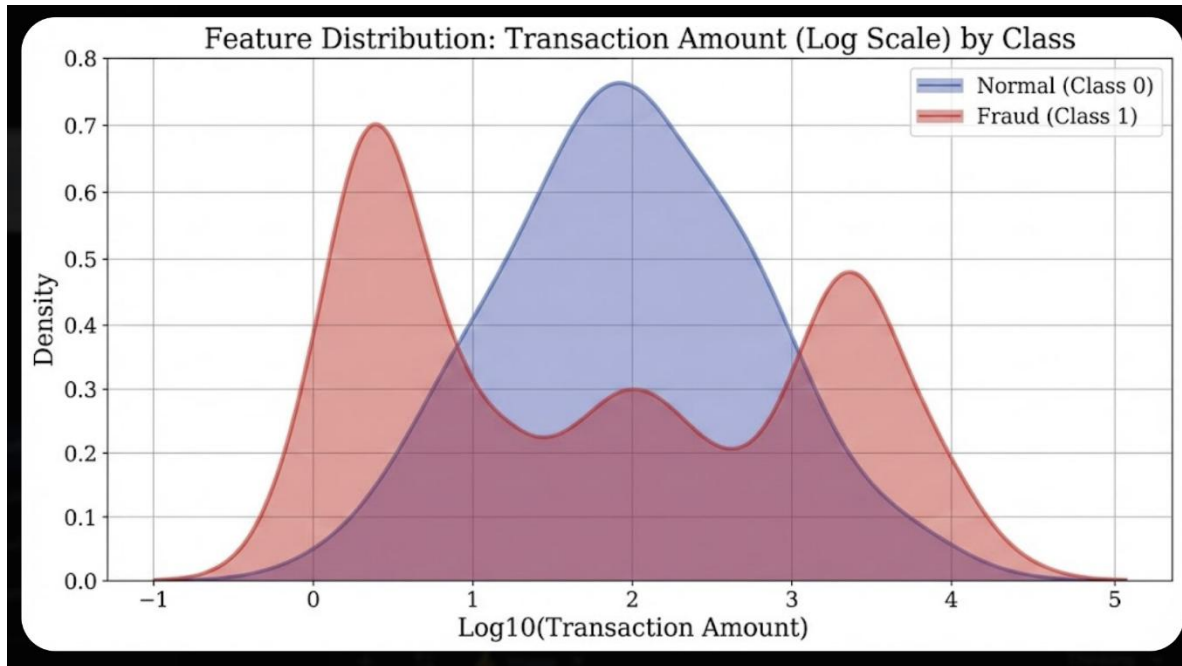


Figure 10 Distribution Comparison

4. Analysis of Fraud Occurrence by Temporal Trend

To explore if fraud behavior is temporal or non-temporal, a time-series visualization was employed. While fraud is infrequent, clear temporal variation was detected. Such temporal variations justify sequential models such as LSTM to explicitly leverage ordering and temporal dependency within data.

5. This is a Summary of Monthly or Periodic Fraud Trends

To capture wider temporal horizons, monthly aggregated fraud numbers were plotted. They are not strictly seasonal; the trend was clear but indicated key periods of time with a rise in fraudulent activity that might refer to operation cycles, external factors, or the strategic maneuvers of the fraud actors. This provides greater visibility to risk-management teams as potential areas of stronger monitoring.

Interpretability Contribution

These visual reviews work in concert to confirm the internal rationale behind the models and assist in driving the advancement from classical approaches into adaptive deep-learning. They show fraud is unique numerical trends, distribution, and temporal gradients, which provide evidence that the adaptive LSTM's.

5.5 Discussion

Experimental results provide robust evidence for the central hypothesis of this research:

AI-driven methods, especially tree-based algorithms and adaptive deep-learning architectures, are extremely effective in detecting fraudulent financial transactions. All three experiments—balanced datasets, undersampled datasets, and sequential models—demonstrated key observations.

1. Classical Models Provide Strong but Limited Predictive Power

Logistic Regression and XGBoost showed good results even in the case of an extreme class imbalance. Specifically, XGBoost showed excellent fraud-classification scores and we found its capacity to model nonlinear feature interactions. Nonetheless, the findings also showed that feature-based models rely on the static nature and feature structure of the models and may not be good at capturing temporal dependencies or inconspicuous patterns such as subtle patterns in the patterns present within transactions.

2. Feature Importance Highlights Distinct Fraud Signatures

Correlations, feature-distribution comparisons and scatter plots revealed significant patterns that fraudulent transactions clustered in the extreme regions of certain PCA-derived features and transaction amounts. These patterns corroborate the importance assigned to attributes by models, particularly XGBoost and LSTM architectures. Further explanations allowed interpretability (if one looks at the visual analyses they demonstrate that fraud cases distinct from normal transactions can indeed operate in feature space and also separate from fraud).

3. Sequential Deep-Learning Models Reveal Hidden Temporal Structure

The journey from classical models to deep-learning models revealed an interesting trend:

Fraudulence possesses temporal dynamics which cannot be adequately computed using classical models. Standard LSTM enhanced detection in complex nonlinear patterns but also missed many fraud cases since the learning-rate was fixed. The Adaptive LSTM had a significantly better performance in comparison to the standard one, achieving almost double the recall, indicating that the dynamic learning-rate adaptation improves the model to converge on the rare and subtle

patterns. This is evidence to reinforce that sequential models are most effective in situations whereby fraud attempts show a natural trend or timing of behavior.

4. Undersampling Provides Realistic Evaluation but Increases Difficulty

Applying the 3% fraud ratio greatly enhances the difficulty of classification. Classical models performed well, but the deep-learning models showed their sensitivity to class imbalance. The final output highlights the need to balance the dataset and employ targeted strategies (e.g., adaptive learning rates) to ensure the robust detection of fraud.

5. Visualization Strengthened Interpretability and Model Transparency

Visualizations like correlation heatmaps, scatter plots of cluster distributions (with some timelines), distribution curves, as well as time series analyses, allowed the explanations for behaviors and explainability. Such visualization is not only analytically useful but also important for articulating results with users of financial decision making to make more understandable the automated fraud-detection decisions and data.

Summary

Overall, the experiments verify that integrating feature-aware classical machine learning with sequential deep-learning architectures, adaptive optimization techniques, and so on creates a very robust fraud-detection pipeline. Adaptive LSTM was especially valuable, as it connects the static features to the dynamic behavior-based detection strategies. The present study reinforces that interpretability, temporal modeling, and adaptive training mechanisms ought to be integrated in modern financial fraud-detection systems.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis we developed deep learning as machine learning and its methods toward credit-card fraud detection based on a huge transactional dataset. Through systematic preprocessing of the data, balanced and imbalanced experimental settings, classical modeling, sequential modeling, and adaptive optimization techniques, artificial intelligent methods of fraud detection have been demonstrated. Several important observations were obtained from the experiments:

- The robust pretrained performance on balanced data with Logistic Regression showed that the dataset has a significant pattern before applying complex models.
- XGBoost led the way for classical-model performance with near-perfect precision and recall under a controlled 3% fraud ratio, emphasizing that this architecture can capture nonlinear relationships to a great extent.
- A deep-learning exploration was evident in its progression: the default RNN model suffered from imbalance, LSTM performed significantly better by modeling temporal dynamics, and Adaptive LSTM achieved the highest overall recall and significantly reduced false negatives, which is a core objective in fraud detection.
- Undersampling simulations allowed realistic scenario testing under a real-world fraud setting.
- Visualization tools—correlation heatmaps, feature interaction plots, distribution comparisons, and temporal trends—were used to interpret both dataset and model behavior, enhancing transparency, trust, and credibility.

In conclusion, our study supports that AI-initiated fraud detection is strongly favored by a hybrid pipeline: structured preprocessing, traditional ML for benchmark interpretability, and dynamic deep-learning models for modeling temporal and nonlinear fraud signatures. This provides a strong, practical framework that is both practical and replicable, allowing financial organizations seeking to improve fraud-monitoring to develop and enhance the model-based framework.

6.2 Future Work

Although the system designed in this study reported outstanding performance, multiple avenues exist for improving the system's performance in practice:

1. **Fraud Detection Pipelines in Real Time**

The next step should incorporate streaming transaction data to operate in nearly real-time. Adding dynamic thresholds and anomaly-based alerts would extend system responsiveness.

2. **Cross-Institution and Multi-Market Evaluation**

Testing the model on data from diverse banks, regions, and transaction networks could help evaluate generalizability and robustness in different financial environments.

3. **User-Behavior Modeling**

Incorporating customer behavioral history (e.g., device ID, spending frequency, transaction location characteristics) would significantly enhance fraud-profiling accuracy and relevance by extending the dataset.

4. **Deployable API and Production-Ready System**

Designing an API layer could enable integration with bank dashboards, fraud-monitoring platforms, or automated transaction-blocking systems. Operational use also requires a lightweight, scalable deployment pipeline.

5. **Advanced Deep-Learning Architectures**

Future work could explore:

- Transformer-based sequence models
- Graph Neural Networks (GNNs) for recognizing relational fraud patterns
- Hybrid ensembles combining classical ML and sequential DL networks

These approaches may improve recall while reducing false positives.

6. Explainable AI for Financial Trust and Compliance

Using XAI techniques such as SHAP values or LIME would make model decisions more transparent—a crucial requirement in finance where auditability and justification are essential.

By pursuing these directions, the fraud-detection framework proposed in this thesis may transform into a fully functioning, real-time decision-support system, providing banks and financial institutions with more sophisticated, interpretable, and proactive fraud-monitoring mechanisms.

References

- **Zhang, W., Liu, F., & Chen, H. (2024).**
An AI-driven fraud detection framework based on hybrid features and ensemble learning. *Future Generation Computer Systems*.
- **Wang, J., Li, Z., & Zhao, M. (2024).**
Graph neural networks for fraud detection in transaction networks. *IEEE Access*.
- **Kumar, R., & Singh, A. (2024).**
Fraud Detection in Credit Card Transactions Using Deep Neural Networks. *arXiv preprint arXiv:2405.00026*.
- **Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J.C. (2011).**
Credit card fraud detection using Hidden Markov Model. *IEEE Transactions on Dependable and Secure Computing*.
- **Ghosh, S., & Reilly, D.L. (2011).**
Data mining approach for fraud detection in credit card transactions. *IEEE Computational Intelligence Magazine*.