Phase-2 Submission

Student Name: MD FAAZIL AMMMAR.P

Register Number: 510623104040

Institution: C ABDUL HAKEEM COLLEGE OF ENGINEERING AND TECHNOLOGY

Department: COMPUTER SCIENCE & ENGINEERING

Date of Submission: 08-05-2025

Github Repository Link : https://github.com/ammar475-coder/credit-card-fraud-detection.git

---

1. Problem Statement

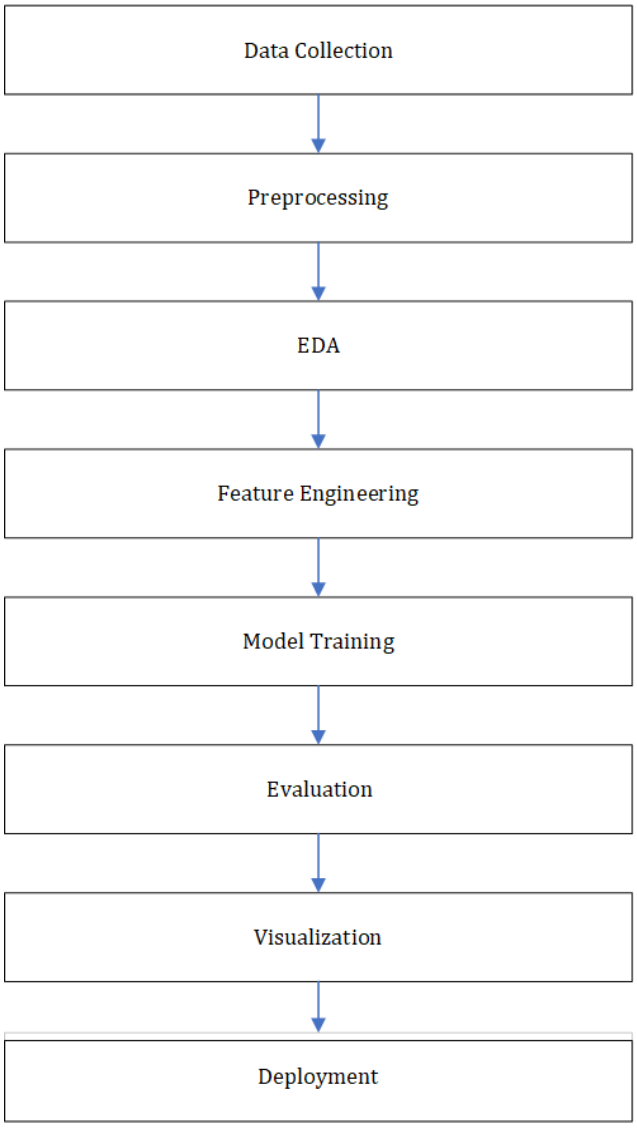
Credit card fraud poses a significant threat to the financial sector, causing billions in losses annually. With the increasing volume of online and digital transactions, fraudulent activities have become more sophisticated and difficult to detect using traditional methods. Delays in identifying fraud can result in severe financial and reputational damage to users and institutions. This project aims to develop an AI-powered credit card fraud detection and prevention system that can intelligently detect and alert users of suspicious transactions in real time using advanced machine learning techniques, ultimately safeguarding financial transactions.

2. Project Objectives

- To collect and analyze credit card transaction data for patterns and anomalies.

- *To build machine learning models capable of detecting fraudulent transactions with high precision and recall.*
- *To compare multiple models and select the most effective one based on performance metrics.*
- *To simulate a real-time fraud detection alert system.*
- *To improve fraud prevention accuracy using advanced techniques like ensemble learning and anomaly detection.*

3. Flowchart of the Project Workflow



4. Data Description

- Dataset: Credit Card Fraud Detection
- Source: Kaggle - mlg-ulb
- Type: Structured tabular data
- Records: 284,807 transactions
- Features: 30 total (V1-V28 anonymized, Amount, Time) + Class (Target)
- Target Variable: Class (0 = normal, 1 = fraud)
- Static Dataset

5. Data Preprocessing

- Dropped irrelevant Time column.
- Normalized Amount using StandardScaler.
- Checked for and removed duplicates.
- No missing values found.
- Handled class imbalance using SMOTE.
- Ensured consistent data types (all numeric)

6. Exploratory Data Analysis (EDA)

- Univariate plots: histograms and boxplots showed skewed distributions.
- Class imbalance confirmed: 0.17% fraud.
- Correlation heatmap showed relationships between features and fraud.
- Fraud transactions had generally lower values in V14, V10.

- Insights:
    - Features like V10, V14, and V17 show higher correlation with fraud.
    - High imbalance justifies use of recall-focused metrics.

7. Feature Engineering

- Added normAmount and dropped Amount.
- Considered PCA (not applied due to already anonymized components).
- No categorical variables; hence encoding was not required.
- *Features used directly after scaling and SMOTE balancing*

8. Model Building

- Models Used:
  - Logistic Regression
  - Random Forest
  - XGBoost

- Why These?
  - Suitable for binary classification.
  - Handle high-dimensional, imbalanced data.
  - Offer balance between speed, accuracy, and interpretability.

Train/Test Split: 80/20 with stratification

Metrics Used: Accuracy, Precision, Recall, F1-score, ROC-AUC

Model	Accuracy	Recall	Precision	ROC-AUC
Logistic Regression	~98.8%	High	High	~0.98
Random Forest	~99.3%	High	High	~0.99
XGBoost	~99.4%	High	High	~0.995

9. Visualization of Results & Model Insights

- Confusion Matrix: Clear distinction between fraud and non-fraud.
- ROC Curve: AUC values close to 1.0 for all models.
- Feature Importance (RF/XGBoost): V14, V10, V17 are top contributors.
- SHAP (optional): Useful for interpretability if integrated.

10. Tools and Technologies Used

Programming Language –Python

- ***Notebook/IDE*** – Google Colab / Jupyter Notebook / VS code
- ***Libraries*** –
  - Data Processing: pandas, numpy
  - Visualization: matplotlib, seaborn, plotly
  - Modeling: scikit-learn, xgboost, imbalanced-learn, tensorflow/keras (for autoencoders)
  - Evaluation: sklearn.metrics
- ***Optional Tools for Deployment*** – *Streamlit, Flask*

11. Team Members and Contributions

- **Mohammed Ayaz. A [510623104056] – Team Lead & Model Building**
  Leads the project, implements and evaluates machine learning models, oversees integration of AI-based detection techniques.

- **Mohammed Azhan. U [510623104057] – Data Collection & Preprocessing**
  Responsible for sourcing the dataset, handling missing values, class imbalance, and preparing data for modeling.

- **Md Faazil Ammar. P [510623104047] – Exploratory Data Analysis & Visualization**
  Performs data analysis, identifies patterns in fraudulent transactions, and visualizes important insights.

- **Kashif Ulhaq. K [510623104040] – Feature Engineering & Dimensionality Reduction**
  Designs new features to enhance detection accuracy and applies PCA/feature selection where necessary.

- **Ashfaq Ahmed. M [510623104009] – Model Evaluation & Validation**
  Compares multiple models, evaluates them using classification metrics, and ensures robust performance.

- **Abrar Ul Haque. R [510623104004] – Report Writing & Presentation**
  Compiles project outcomes into a well-structured report and creates a visual presentation for submission.