FRAUD DETECTION IN BANKING DATA USING MACHINE LEARNING

Submitted for partial fulfillment of the requirements

for the award of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

by

KROTHAPALLI SNEHA	-	**21BQ1A6129**
BURADAGUNTA VAMSI	-	**21BQ1A6111**
SHAIK SHARMILA	-	**21BQ1A6152**
SANKULA VENKATA SAI TEJA	-	**22BQ5A6107**

Under the guidance of

Mr. SYED BEEBEN BASHA

Associate Professor

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Permanently Affiliated to JNTU Kakinada, Approved by AICTE

Accredited by NAAC with 'A' Grade, ISO 9001:2008 Certified

NAMBUR (V), PEDAKAKANI (M), GUNTUR – 522 508

Tel no: 0863-2118036, url: www.vvitguntur.com

March-April 2025

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Permanently Affiliated to JNTUK, Kakinada, Approved by AICTE

Accredited by NAAC with 'A' Grade, ISO 9001:20008 Certified

Nambur, Pedakakani (M), Guntur (Gt) -522508

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

CERTIFICATE

This is to certify that this **Project Report** is the bonafide work of **Ms.Krothapalli Sneha, Mr. Buradagunta Vamsi , Ms. Shaik Sharmila, Mr.Sankula Venkata SaiTeja**, bearing Reg.No. **21BQ1A6129, 21BQ1A6111, 21BQ1A6152, 22BQ5A6107** respectively who had carried out the project entitled **"FRAUD DETECTION IN BANKING DATA USING MACHINE LEARNING "** under our supervision.

Project Guide	**Head of the Department**
Mr. Syed Beeban Basha, Associate Professor	Dr. K. Suresh Babu , Professor

---

Submitted for Viva voce Examination held on _____

Internal Examiner **External Examiner**

DECLARATION

We, Ms. Krothapalli Sneha Mr. Buradagunta Vamsi, Ms. Shaik Sharmila, Mr. Sankula Venkata Sai teja, hereby declare that the Project Report entitled **"Fraud Detection in Banking Data Using Machine Learning"** done by us under the guidance of Mr. Syed Beeban Basha, Associate Professor, CSE - Artificial Intelligence & Machine Learning at Vasireddy Venkatadri Institute of Technology is submitted for partial fulfillment of the requirements for the award of Bachelor of Technology in Artificial Intelligence & Machine Learning. The results embodied in this report have not been submitted to any other University for the award of any degree.

DATE :

PLACE : Nambur

SIGNATURE OF THE CANDIDATE (S)

Krothapalli Sneha [21BQ1A6129]

Buradagunta Vamsi [21BQ1A6111]

Shaik Sharmila [21BQ1A6152]

Sankula Venkata Sai Teja [22BQ5A6107]

ACKNOWLEDGEMENT

We take this opportunity to express my deepest gratitude and appreciation to all those people who made this project work easier with words of encouragement, motivation, discipline, and faith by offering different places to look to expand my ideas and helped me towards the successful completion of this project work.

First and foremost, we express my deep gratitude to **Sri. Vasireddy Vidya Sagar**, Chairman, Vasireddy Venkatadri Institute of Technology for providing necessary facilities throughout the B.Tech programme.

We express my sincere thanks to **Dr. Y. Mallikarjuna Reddy**, Principal, Vasireddy Venkatadri Institute of Technology for his constant support and cooperation throughout the B.Tech programme.

We express my sincere gratitude to **Dr. K. Suresh Babu**, Professor & HOD, Computer Science Engineering – Artificial Intelligence & Machine Learning Vasireddy Venkatadri Institute of Technology for his constant encouragement, motivation and faith by offering different places to look to expand my ideas.

We would like to express my sincere gratefulness to our Guide **Mr. Syed Beeban Basha**, Associate Professor, CSE-Artificial Intelligence & Machine Learning for his insightful advice, motivating suggestions, invaluable guidance, help and support in successful completion of this project.

We would like to express our sincere heartfelt thanks to our Project Coordinator **Ms. N. Nalini Krupa**, Associate Professor, CSE-Artificial Intelligence & Machine Learning for his valuable advices, motivating suggestions, moral support, help and coordination among us in successful completion of this project.

We would like to take this opportunity to express my thanks to the **Teaching and Non-Teaching** Staff in the Department of Computer Science Engineering -Artificial Intelligence and Machine Learning, VVIT for their invaluable help and support.

Name (s) of Students

Krothapalli Sneha [21BQ1A6129]

Buradagunta Vamsi [21BQ1A6111]

Shaik Sharmila [21BQ1A6152]

Sankula Venkata SaiTeja [22BQ5A6107]

TABLE OF CONTENTS

LIST OF FIGURES

NOMENCLATURE

ML	Machine Learning
AI	Artificial Intelligence
UPI	Unified Payments Interface
MFA	Multi-Factor Authentication
AUC-ROC	Area Under the Curve - Receiver Operating Characteristic
EDA	Exploratory Data Analysis
UAT	User Acceptance Testing
SSL	Secure Sockets Layer
MLOps	Machine Learning Operations
API	Application Programming Interface

ABSTRACT

Fraud detection in banking data plays a crucial role in the financial sector, leveraging machine learning to develop solutions that protect financial systems from fraudulent activities. With the rapid advancement of digital banking, there has been a significant increase in various forms of fraud, including bank account fraud, debit and credit card fraud, and Unified Payments Interface (UPI) fraud. These fraudulent transactions pose severe threats to financial security, leading to substantial monetary losses and eroding consumer trust in digital banking platforms.

The primary objective of this project is to design and deploy a robust fraud detection system powered by machine learning models capable of identifying fraudulent transactions and assessing transaction risk levels in real time. To achieve this, we will analyze multiple datasets, including the **Bank Account Fraud Dataset, Debit/Credit Card Fraud Dataset, and UPI Fraud Detection data**, ensuring a comprehensive understanding of fraudulent patterns across different banking transactions. By implementing a **hybrid machine learning approach**, we will integrate the strengths of multiple algorithms to enhance the system's predictive accuracy while simultaneously reducing false positive rates—ensuring that legitimate transactions are not mistakenly flagged as fraudulent.

Performance evaluation metrics such as **precision, recall, F1-score, and AUC-ROC** will be employed to continuously optimize the fraud detection model, ensuring high reliability and efficiency. The final solution will be deployed within an interactive **Streamlit-based web application** that provides real-time fraud detection, data visualization tools, and fraud risk predictions. This platform will empower financial institutions and users by offering intuitive dashboards that display transaction trends, highlight suspicious activities, and provide actionable insights for enhanced security.

Keywords: The Fraud Detection, Banking Data, Machine Learning, Hybrid Models, Risk Prediction, Streamlit, Fraud Probability, UPI Fraud, Debit/Credit Card Fraud, Bank Account Fraud, Visualizations,.Authentication

CHAPTER 1

INTRODUCTION

1.1 Background of the Project

1.1.1 Fraud Detection in the Financial Industry

Fraud detection is a critical concern in the financial industry, as fraudulent activities can lead to significant financial losses, reputational damage, and eroded customer trust. With the increasing digitization of financial services and the growing sophistication of fraud schemes, traditional rule- based fraud detection methods have become less effective, necessitating the adoption of more advanced techniques.

The research domain of fraud detection and prevention has gained significant attention in recent years, as financial institutions and regulatory bodies strive to combat the ever- evolving landscape of financial crimes. This domain encompasses the study and development of techniques to identify, mitigate, and prevent various types of fraud, including credit card fraud, insurance fraud, money laundering, and other illicit financial activities.

Importance of Effective Fraud Detection

The importance of effective fraud detection cannot be overstated. Financial institutions are responsible for safeguarding their customers' assets and ensuring the integrity of the financial system. Undetected fraud can result in substantial monetary losses, legal liabilities, and regulatory penalties. Moreover, the impact of fraud extends beyond the financial institutions themselves, as it can undermine the overall stability and reliability of the financial ecosystem, erode public trust, and lead to broader economic consequences.

1.1.2 Real-World Applications of AI/ML in Fraud Detection

Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful tools in the fight against financial fraud. These technologies enable the development of sophisticated models that can analyze vast amounts of transaction data, identify patterns, and detect anomalies that may indicate fraudulent behavior. By leveraging AI/ML, financial institutions can enhance their fraud detection capabilities, respond more quickly to emerging threats, and provide a more secure and trustworthy service to their customers.

Real-world applications of AI/ML in fraud detection include:

1. **Credit Card Fraud Detection**: AI/ML models can analyze credit card transaction data to identify suspicious patterns and flag potentially fraudulent activities in real-time, enabling prompt intervention and prevention of financial losses.
2. **Insurance Claim Fraud Identification**: AI/ML algorithms can be trained to detect anomalies in insurance claims, identifying patterns that may indicate fraudulent activities, such as exaggerated or fabricated claims.
3. **Anti-Money Laundering (AML) Monitoring**: AI/ML-powered systems can monitor financial transactions, identify suspicious activities, and assist in the detection and prevention of money laundering, a common method for concealing the origins of illegally obtained funds.
4. **Mortgage Fraud Prevention**: AI/ML models can analyze mortgage application data and related information to identify red flags and potential fraud, helping financial institutions mitigate the risks associated with fraudulent mortgage activities.
5. **Insider Trading Detection**: AI/ML techniques can be employed to analyze trading patterns, market data, and other financial information to detect potential instances of insider trading, a form of financial fraud.

These real-world applications demonstrate the transformative potential of AI/ML in the financial industry, enabling more effective fraud detection, prevention, and mitigation strategies.

Technology has rapidly evolved and continues to advance at an unprecedented pace. Alongside this rapid progression, the financial sector has experienced a surge in digital transactions, offering unparalleled speed, convenience, and efficiency. However, these technological advancements have also given rise to a significant increase in fraudulent activities, posing substantial threats to individuals, businesses, and financial institutions. As a result, financial institutions are compelled to implement robust fraud detection mechanisms to safeguard customer assets, protect sensitive data, and uphold their reputations within the digital banking ecosystem.

Fraudulent activities such as debit and credit card fraud, bank account fraud, and unauthorized financial transactions create serious risks for both customers and financial organizations. The increasing complexity and frequency of fraudulent schemes necessitate the deployment of sophisticated fraud detection systems that can effectively identify, prevent, and mitigate risks in real time. Traditional rule-based detection systems, while effective to some extent, lack the ability to adapt to evolving fraud patterns. On the other hand, machine learning (ML)-based fraud detection models leverage large datasets and continuously learn from new transaction patterns, enhancing predictive accuracy over time.

This project aims to develop an advanced fraud detection system utilizing machine learning algorithms to classify financial transactions as either fraudulent or legitimate. The proposed system will analyze transaction data to detect anomalies indicative of fraudulent behavior, thereby reducing financial losses for customers and institutions. Furthermore, the fraud detection model will not only identify fraudulent transactions but also assess the risk level of each transaction, enabling financial institutions to prioritize actions based on the severity of risk.

1.2 Objective

The primary objective of this project is to build a **machine learning-based fraud detection system** that enhances financial security by detecting fraudulent transactions with high accuracy. To achieve this, the system will:

- Utilize **hybrid machine learning models** that combine multiple algorithms such as **decision trees, random forests, and gradient boosting** to improve fraud detection performance.
- Analyze financial transaction datasets, including the **Bank Account Fraud Dataset, Debit/Credit Card Fraud Dataset, and UPI Fraud Detection Data**, to identify patterns and trends associated with fraudulent activities.
- Implement a **risk assessment model** that assigns probability scores to transactions based on their likelihood of being fraudulent.
- Develop an interactive **Streamlit-based web application** that provides real-time fraud detection, transaction risk visualization, and detailed fraud analytics.

- Incorporate robust **authentication mechanisms** such as user login, mobile number verification, and email authentication to enhance data security and user privacy.

By achieving these objectives, the project aims to create a comprehensive fraud detection system that not only minimizes financial fraud but also equips financial institutions with actionable insights to strengthen their security infrastructure.

1.3 Problem Statement

The rapid digitization of financial services has led to an increase in fraudulent activities, compromising the security and trustworthiness of digital banking systems. Traditional fraud detection methods rely on predefined rules and manual intervention, which are inefficient in detecting complex and evolving fraud patterns. Fraudulent transactions often go undetected until significant financial damage has been inflicted, leading to monetary losses for individuals and institutions. Therefore, there is a pressing need for an intelligent, automated fraud detection system that can accurately identify and prevent fraudulent transactions in real time.

Machine learning provides a solution by enabling fraud detection systems to analyze large volumes of transaction data, recognize hidden fraud patterns, and continuously improve their detection capabilities. Unlike rule-based approaches, machine learning models can adapt to new fraud techniques, reducing false positives while maintaining high fraud detection rates. This project aims to bridge the gap by implementing a **hybrid ML-based fraud detection system** that offers improved fraud prediction accuracy and real-time transaction monitoring.

1.4 Problem Description

Fraud detection in the financial sector is a complex challenge due to the **dynamic nature of fraudulent activities** and the **large volume of transaction data** that must be analyzed in real time. Traditional fraud detection methods primarily rely on rule-based approaches, where predefined rules flag suspicious transactions. However, these methods struggle to detect sophisticated fraud schemes and often result in a high rate of false positives, leading to inconvenience for legitimate customers.

Machine learning, on the other hand, enables fraud detection systems to learn from historical transaction data and identify **subtle, hard-to-detect fraud patterns**. By analyzing transaction attributes such as transaction amount, type, frequency, account details, and timestamp patterns, machine learning models can classify transactions as either **legitimate or fraudulent** with higher accuracy.

In this project, a **hybrid machine learning approach** will be employed, leveraging multiple algorithms to maximize detection accuracy. The project will use datasets from financial transactions, specifically focusing on:

- **Bank Account Fraud Dataset**: Contains transactional data related to unauthorized account access and fund transfers.
- **Debit/Credit Card Fraud Dataset**: Includes records of legitimate and fraudulent card transactions, aiding in identifying fraudulent card activity.
- **UPI Fraud Detection Data**: Captures fraudulent UPI transactions, helping to detect unauthorized peer-to-peer payments.

The **key challenge** in fraud detection lies in minimizing false positives while ensuring fraudulent transactions are accurately flagged. This requires **optimized feature engineering, model selection, and performance evaluation** using key metrics such as **precision, recall, F1-score, and AUC-ROC**.

1.5 Project Approach

The project will follow a systematic approach to fraud detection, starting with **data collection and preprocessing**, followed by **feature engineering and model training**, and concluding with **model deployment and user interface development**.

1. **Data Collection & Preprocessing**: The first step involves acquiring financial transaction datasets and cleaning the data to remove inconsistencies. Preprocessing techniques such as handling missing values, feature scaling, and encoding categorical variables will be applied.

2. **Feature Engineering**: The next step is identifying relevant transaction features that contribute to fraud detection. Features such as transaction amount, transaction frequency, device ID, account history, and geolocation data will be analyzed.

3. **Model Training & Evaluation**: Multiple machine learning models, including **decision trees, random forests, and gradient boosting**, will be trained and tested on the datasets. The best-performing model will be selected based on evaluation metrics.

4. **Risk Assessment Framework**: Instead of simply classifying transactions as fraudulent or legitimate, the model will assess the risk level associated with each transaction, prioritizing high-risk transactions for further investigation.

5. **Web Application Development**: A **Streamlit-based web application** will be developed to provide an interactive interface for real-time fraud detection and visualization.

6. **Security & Authentication Mechanisms**: The web application will incorporate user authentication features such as **login/signup** to protect sensitive financial data and enhance security.

The project aims to develop a fraud detection system using machine learning, following a systematic approach. It begins with collecting and preprocessing financial transaction data, then engineering relevant features for fraud detection. Multiple machine learning models are trained and evaluated, with the best-performing model selected for a risk assessment framework. This framework prioritizes high-risk transactions for further investigation. A web application is then developed for real-time fraud detection and visualization, with security and authentication mechanisms implemented to protect sensitive data.

CHAPTER 2

LIERATURE REVIEW

Purpose:

This chapter aims to analyse previous research on fraud detection in the banking sector using machine learning and artificial intelligence techniques. It establishes the need for a robust project that addresses existing gaps in fraud detection systems.

2.1 Previous Research and Related Work

The use of machine learning techniques, which have the ability to learn from large amounts of data, and to identify known types of fraud patterns, makes them a key instrument for detecting fraud in the banking sector. In this context, Hashemi et al. [1] put forward the way of fraud detection by the usage of various machine learning algorithms, such as decision trees and neural networks. They say that machine learning is a better option than the rule-based systems, used so far, in dynamic fraud detection as these systems are not designed to detect any new form of fraud. Based on the study, gained insights into the possibility of applying ensemble methods and hybrid models to make detection accuracy and reduce the problem of false positives in banking fraud. Building this addressing gap, Johora et al. [2] mentioned the escalation of call for up to AI based upon financing fraud analysis systems. Then, they described how they carried out an in depth study of how AI can track and detect fraudulent activities through sequence based on transaction data for anomalies. The application of such benefits and differences of supervised and unsupervised learning models to account takeovers, identity theft, or unauthorized transactions can be any type of fraud. The financial institutions were also shown in the paper that using AI based models is going to be a great improvement over the traditional ways, using these models would lead the institutions to forewarning any cases of fraud before they actually happen. Further improvements in how machine learning and data analytics can be integrated to achieve fraud prevention were also explored by Mohammad et al. [3]. Their study found that the growing scale of the likely prey makes it more and more complex to detect fraud.

According to the authors, several algorithms have to be united with the model of hybrid machine learning models to make better predictions. This work was done by the researchers in arguing that supervised machine learning techniques can be easily combined with unsupervised machine learning techniques that are able to uncover hidden fraud patterns in banking systems, which results into a better solution to the problem of fraud detection in banking systems. Apart from the data driven approach, Faisal et al. [4] suggested the application of machine learning models for predicting and classifying real time fraudulent transactions. Depending on each bank's ability to line it with existing banking infrastructure, fraud detection was emphasised. For example, the researchers showed that AI models — or specifically deep learning ones — could process complex data sets with live insights of potential instances of fraud and would be useful to banks to minimize risks and protect customer assets. In his case Esmail [5] provides a complete treatment of loan fraud detection processes of banking sectors with data mining techniques. The first mentioned the importance of classifying or clustering the fraudulent loan applications. In line with such an approach, they also made a study that demonstrated how the integration of machine learning into the support of loan fraud detection would bring a much more effective way to notice suspicious events and enhance fraud detection procedures in financial institutions. With the exception of that, Bin Sulaiman et al. [6] have completely utilized machine learning in credit card fraud detection. In their review to detect credit card fraud, they explored different machine learning including classification algorithms and feature engineered techniques. In case of a fraud transaction is few compared to legitimate transaction the authors presented how a good problem can be solved. In this case, they proposed some solutions like the oversampling techniques, the ensemble models and so on, which would contribute to an increase of the performance of fraud detection systems. Second, Johora et al. [7] later worked on AI advances in banking security with more attention to fraud detection systems. In this particular research, they focus on how important the AI becomes in enhancing the security in banking, which employs machine learning models to detect the fraud patterns and deter financial losses. They owned the significance of advancing techniques, for example, natural language processing (NLP) and anomaly detection, and exhibited how NLP and anomaly detection techniques can increase the accuracy of different types of frauds, for example, phishing,cardskimming.

Kotagiri and Yada [8] in another study looked into how RPA and advanced analytics strategies can assist to detect and reduce the mining of fraud in the banking system. Synergy between the automation tools and machine learning models in suspicious activities detection and to reduce the role of humans in the aspects of fraud monitoring. Results showed that RPA along with AI models make fraud detection more efficient and reduce operational costs of the financial institutions. Among the banking sector, Gautam [9] has evaluated how artificial intelligence affects the process of building risk management and identifying fraud. In what he focuses on, AI is able to see transaction data and is able to detect patterns that may indicate the possibility of scamming. An important thing he discovered is that AI could be integrated into the existing risk management framework to make fraud prevention more effective by actively detecting fraud and to increase bank's safety. Zanke [10] does one of the comparisons of AI driven fraud detection systems in banking, insurance and healthcare. The main focus was to analyze the application of AI and deep learning based systems for fraud detection as a common thing across all sectors and the use of deep learning to enrich our fraud detection systems across all cases. It was concluded that AI based systems can be very effective defaulters to automate the mechanisms of detecting fraud in the banking industry and reduce financial losses. Automated banking fraud detection is proposed by Biswas et al. [11] in the domain of the financial sector in order to reduce unauthorized access in it. They very specifically examined if and how machine learning algorithms (support vector machine, SVM, and other) could be applied in real time to detecting this fraud. The authors then brought up that the processes of fraud detection need to be automated in order to reduce the human error and increase the accuracy of the predictions. For example, in [12], Sekar mentions digital banking fraud prevention through real time cloud and AI. It was also about how cloud based solutions with the help of AI technology can help in fraud detection to take advantage of the scalability and flexibility of getting to identify large volumes of banking data. Sekar pointed out the need for simple, minimum delay systems that can detect and evaluate such transactions instantly and enable the bank to react instantly as and when such fraudulent activities come to the fore. Nonetheless, Sambrow and Iqbal [13] presented the possibilities of artificial intelligence and deep learning, data analytics, and many other things in preventing banking fraud.

What sorts of fraud could a deep learning machine — specifically, a neural network — learn from large datasets, that is what they studied in their work. I believe that they suggested that banks can enhance the fraud detection system and the elimination of frauds whilst they are occurring in real time via use of advanced machine learning approaches. Al-Fatlawi et al. discussed an AI based model of fraud detection in the bank system, in [14]. During their conversation about how artificial intelligence models can be used to train machine learned systems to automatically detect and call out adversarial transactions, they talked about how machine learning can train artificial intelligence models to identify strange and dubious patterns so that they can be alerted to potentially hostile transactions. They suggested identifying ensemble learning methods with which to combine the results of several models, so that the accuracy is improved and the number of false positive cases is reduced. Hence, with their work, Achary and Shelke [15] finally explored the use of machine learning techniques for fraud detection in banking transactions. Financing the second reference is a problem to detect the fight fraud in the financial transactions and using Machine learning to overcome the problem with particular references. They recommended that fraud detection models must be selected carefully with features and algorithms to train the models, to improve accuracy of the models and the overall system performance of the fraud prevention

2.2 Existing Solutions and Their Limitations

- Rule-Based Systems :
    - Limited to predefined rules and unable to detect new fraud patterns.
    - High false-positive rates due to rigid structures.
- Supervised Learning Models :
    - Require labeled datasets, which are often unavailable or expensive to create.
    - Struggle with imbalanced datasets where fraudulent cases are rare.
- Unsupervised Learning Models :

    - May generate ambiguous results due to the lack of labeled data for validation.
    - Prone to noise and irrelevant patterns in the dataset.
- Hybrid Models :
    - Complex to implement and require significant computational resources.
    - Integration challenges with existing banking infrastructure.
- Cloud-Based Solutions :
    - Privacy concerns regarding sensitive banking data stored in the cloud.
    - Dependency on internet connectivity and potential latency issues.

2.3 Gap Analysis

- Dynamic Fraud Detection : Existing systems struggle to adapt to evolving fraud patterns, necessitating more flexible and adaptive solutions.
- Real-Time Processing : Many current models lack the capability to process and analyze data in real time, delaying fraud detection.
- Imbalanced Datasets : Fraudulent transactions are often far fewer than legitimate ones, leading to biased models and poor detection rates.
- Integration Challenges : There is a lack of seamless integration between advanced AI models and traditional banking systems.
- False Positives : High false-positive rates remain a significant issue, causing inconvenience to customers and increasing operational costs.

This gap highlights the need for a comprehensive solution that combines real-time processing, advanced machine learning techniques, and seamless integration with existing banking infrastructure.

2.4 Relevance of the Project

This project builds on previous research by proposing an integrated framework that leverages hybrid machine learning models, real-time data analytics, and cloud-based solutions. Inspired by datasets and models from studies like Hashemi et al. [1], Johora et al. [2], and Faisal et al. [4], the project aims to address the limitations of existing solutions. By focusing on dynamic fraud detection, real-time processing, and reducing false positives, this project seeks to enhance fraud prevention in the banking sector. The proposed system incorporates ensemble learning methods, combining the strengths of decision trees, support vector machines, and neural networks to improve accuracy and adaptability.

Real-time data ingestion pipelines using tools such as Apache Kafka and Spark Streaming will ensure timely fraud detection. Additionally, the system will be deployed on scalable cloud platforms like AWS or Azure, allowing financial institutions to handle large volumes of transaction data seamlessly. The project also emphasizes explainable AI (XAI) to ensure transparency and regulatory compliance, enabling stakeholders to understand and trust model predictions. By incorporating user behavior analytics and anomaly detection, the model aims to identify subtle fraud patterns often missed by traditional systems. A modular architecture will allow easy integration with existing banking infrastructure. The use of continuous learning mechanisms will ensure that the model evolves with emerging fraud trends. Furthermore, secure APIs and data encryption protocols will be implemented to safeguard sensitive customer data. The ultimate goal is to create a cost-effective, intelligent, and adaptive fraud detection system that supports real-time decision-making and reduces financial losses.

.

CHAPTER 3

SYSTEM ANALYSIS

Machine learning algorithms are proposed to be used in combination with the proposed methodology for fraud detection in banking data on real time transactions. The system will employ analytical methods for examining the transaction data and predict the probability of fraud as well as examining the risk of each transaction. In this approach, several datasets like Bank Account Fraud Dataset, UPI Fraud Detection data, Debit/Credit Card Fraud Dataset are used and a hybrid machine learning model has been used to improve detection accuracy. To develop a cost effective fraud detection solution, the system would concentrate on data collection, then model development and finally on evaluation of the model.

The procedure of this methodology is structured in such a way that it guarantees efficiency and scalability. Model training, data pre processing, evaluation, and optimization are involved in the process. This methodology results will assist banks and financial institutions in detecting fraudulent activities at an early period and curb them from financial loss. In the next subsections, we detail the main steps in the proposed methodology.

A. Data Collection and Preprocessing

Firstly, we retrieve datasets of multiple types of banking transactions such as Bank Account Fraud Dataset, Debit/Credit Card Fraud Dataset, UPI Fraud Detection data. For these datasets, these are the important features such as transaction amount, user account information, transaction type and timestamps. This is important because of this diversity of data, since the system will be able to learn patterns between multiple fraud scenarios, and to become more adapted to all fraud situations. Representative datasets that both incorporate fraudulent and non-fraudulent transactions will collect that, making the model more robust and true.

The dataset is collected, and in the preprocessing phase, the datasets are made ready for training the machine learning model. Imputation techniques will be used to handle the missing values and one hot encoding will be used to transform the categorical variables into numerical data.

It will also feature scaling of numeric features to normalize them and put them on a similar scale for improving model's performance. Moreover, feature selection will be used to identify the most important variables in detecting fraud. Preprocessing is an attempt to make sure the data is clean and is ready to get the best from ML.

B. Model Development

XGBoost (Extreme Gradient Boosting)

XGBoost is an advanced gradient boosting algorithm designed for speed, accuracy, and efficiency. It builds multiple decision trees sequentially, where each new tree corrects the errors made by the previous ones. This iterative learning process minimizes residual errors and enhances model performance. Unlike traditional boosting methods, XGBoost employs a regularized objective function that balances model complexity and predictive power. This helps prevent overfitting, a common issue in machine learning models. Additionally, XGBoost uses second-order gradients to refine predictions, making it faster and more efficient compared to conventional gradient boosting techniques.

One of the key strengths of XGBoost is its ability to handle missing data and noisy datasets effectively. The algorithm automatically learns which feature values are missing and adjusts the model accordingly. It also supports column block compression, enabling efficient storage and retrieval of large datasets. Another major advantage is its ability to perform feature selection by ranking the importance of variables, allowing the model to focus on the most relevant fraud indicators. This makes XGBoost particularly useful in fraud detection scenarios where transactions contain complex relationships between variables.

The interpretability of XGBoost is another factor that contributes to its widespread adoption. While many advanced machine learning models are often criticized for being "black boxes," XGBoost provides tools for understanding how predictions are made. For instance, the feature importance scores generated by the algorithm help identify which variables contribute most significantly to the detection of fraudulent transactions. This transparency is invaluable for financial institutions, as it allows them to validate the model's decisions and gain insights into emerging fraud patterns.

Moreover, XGBoost's ability to visualize decision trees within the ensemble makes it easier to explain model behavior & to non-technical stakeholders, fostering trust and confidence in the system. Furthermore, XGBoost is highly scalable due to its parallel processing capabilities. It can distribute computations across multiple cores and machines, making it suitable for large- scale fraud detection applications. The use of histogram-based optimization further speeds up training time by reducing computational complexity. These optimizations make

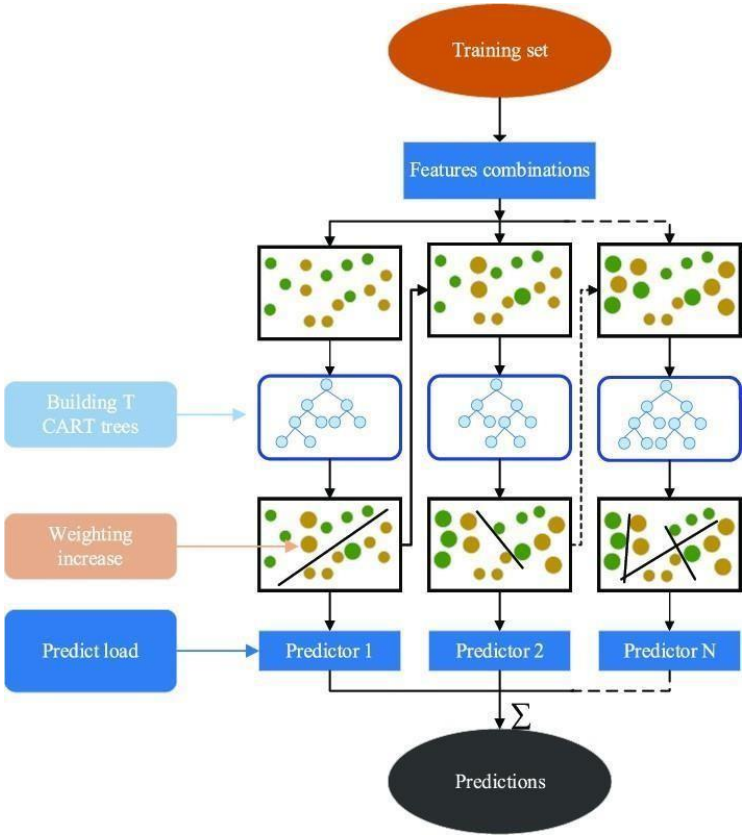


Figure 3.1: Describing about XGBoost

XGBoost a powerful choice for detecting fraudulent transactions, where quick decision-making and high accuracy are essential. Given its superior performance and robustness, XGBoost is widely used in financial fraud detection, cybersecurity, and anomaly detection tasks.

Random Forest

Random Forest is an ensemble learning algorithm that improves classification accuracy by combining the outputs of multiple decision trees. Unlike individual decision trees, which may suffer from high variance and overfitting, Random Forest creates multiple trees using different subsets of the training data and averages their predictions. This technique, known as bagging (Bootstrap Aggregating), enhances the model's stability and reduces its sensitivity to noise. Since each tree in the forest is trained on a randomly sampled subset of data, the model maintains diversity, leading to better generalization and performance on unseen transactions.

A key advantage of Random Forest is its ability to handle high-dimensional datasets and imbalanced data distributions. In fraud detection, fraudulent transactions often represent only a small fraction of the dataset, making it difficult for traditional classifiers to learn meaningful patterns. Random Forest mitigates this issue by using balanced resampling techniques and assigning different weights to minority class instances. Additionally, by randomly selecting a subset of features at each split, the algorithm ensures that individual trees do not rely on the same dominant features, leading to more robust and diverse decision-making.

Another important feature of Random Forest is its interpretability. The algorithm provides feature importance scores, which help in identifying the most significant variables influencing the classification decision. This is particularly useful in fraud detection, where understanding why a transaction is flagged as fraudulent is crucial for financial institutions and regulatory bodies. Moreover, Random Forest is resistant to overfitting, as the averaging process of multiple decision trees prevents it from memorizing noise in the training data. Due to its robustness, accuracy, and ease of implementation, Random Forest is widely used in fraud detection systems, credit risk assessment, and other financial applications.

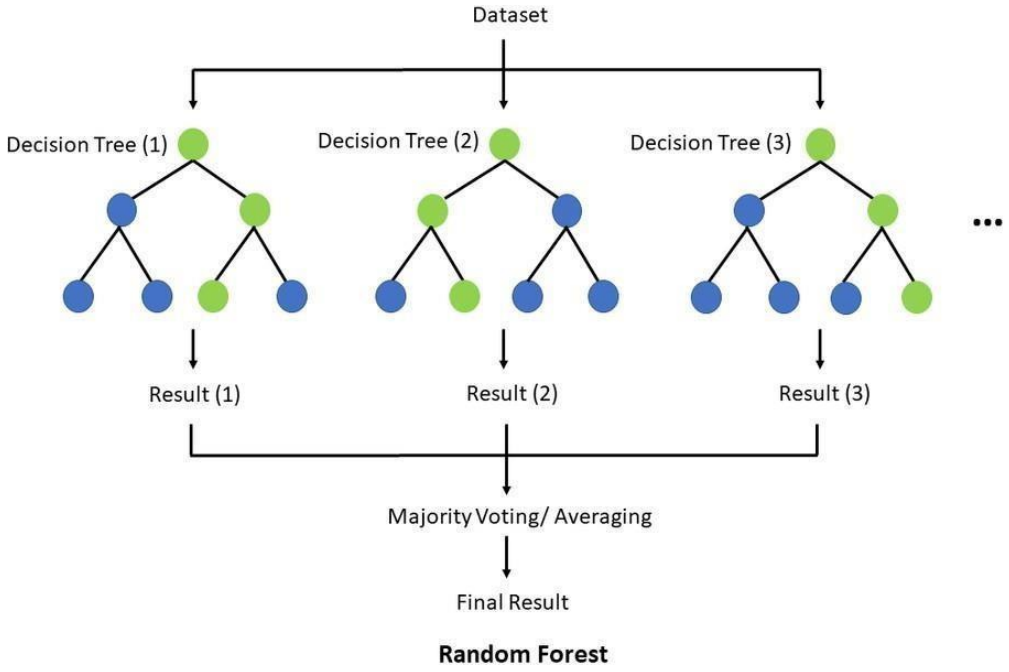


Figure 3.2: Describing about Random

Forest Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that is particularly effective for classification tasks involving complex decision boundaries. It works by finding an optimal hyperplane that best separates data points belonging to different classes. The algorithm maximizes the margin between the closest data points (support vectors) of different classes, ensuring better generalization to unseen data. This margin- based approach helps SVM achieve high classification accuracy, even when dealing with high-dimensional datasets where other algorithms struggle.

One of the most powerful aspects of SVM is the **kernel trick**, which allows it to handle non- linearly separable data. By transforming the input data into a higher-dimensional space, SVM can find a hyperplane that effectively separates fraudulent transactions from legitimate ones. Common kernel functions include linear, polynomial, and radial basis function (RBF) kernels, with RBF being the most commonly used for fraud detection due to its ability to capture complex patterns. This flexibility makes SVM highly effective in detecting fraudulent transactions that exhibit subtle variations and nonlinear relationships.

Despite its advantages, SVM can be computationally expensive, especially when dealing with large datasets. Training an SVM model requires solving a complex optimization problem, which can be time-consuming. However, with appropriate hyperparameter tuning and the use of techniques such as **Support Vector Data Description (SVDD)** for anomaly detection, SVM remains a valuable tool in fraud detection applications. By accurately distinguishing between fraudulent and non-fraudulent transactions, SVM contributes to improving financial security and reducing economic losses.

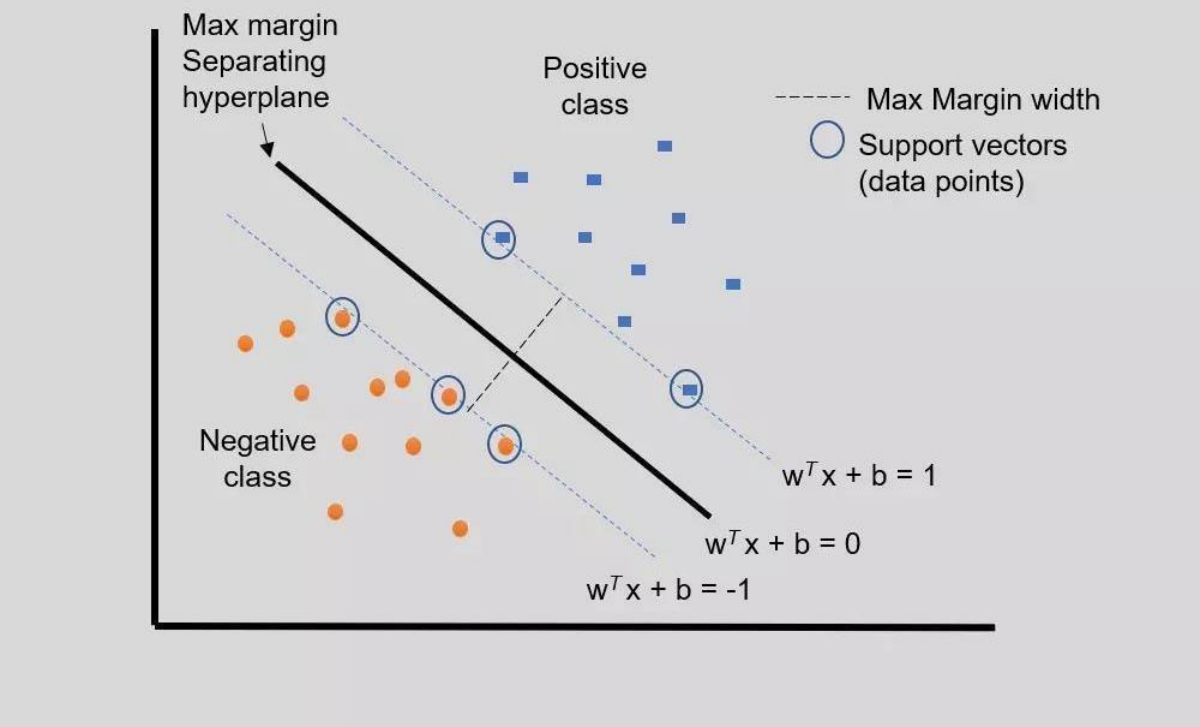


Figure 3.3: SVM Max-Margin Hyperplane

The next step after preprocessing the data is modeling. So, we will adopt a hybrid machine learning approach where several algorithms will be combined to enjoy the best of their own. Fraudulent patterns present in the transaction data will be detected using Random Forest, XGBoost and Support Vector Machines (SVM). We will use Random Forest due to its ability to work on large datasets and prevent overfitting, while XGBoost fulfills its boosted performance by rectifying errors from earlier models in the ensemble. For a more effective discrimination of complex, high dimensional data, handling nuances in identifying fraudulent transactions, SVM will be used.

A hybrid model methodology is used so that a more reliable fraud detection system is created. XGBoost is good for fine tuning the predictions with its powerful gradient boosting techniques, while Random Forest will come to aid us with handling the large datasets without overfitting. An insertion of the sophistication as offered by SVM on cases of difficult to tackle fraud will refine the model of recognising patterns of fraud which otherwise would have not been spotted. The system will be trained on the full dataset for each model, so that they perform well for any type of fraud or transactions.

C. Model Evaluation and Optimization

Once trained on models, their performance will be evaluated on key metrics namely accuracy, precision, recall, F1 score and AUC-ROC (Area under the Receiver Operating Characteristic Curve). With these metrics we can assess totality how good the model is for catching the fraud, how many false positives, how many false negatives. Accuracy will give us an overview of the model's performance, which is important in general, but precision and recall are important metrics to watch for fraud detection as false negatives (missed fraud) are more expensive than false positives.

Hyperparameter tuning will be carried out to find most adequate model parameters in order to optimize the models and increase performance. To determine the best possible combination of hyperparameters, techniques, such as grid search and random search will be used. To ensure good generalization to unseen data, it will implement cross-validation. These models will be found iteratively, fine tuned and tested on the various datasets to get the greatest precision possible in fraud detection and accuracy. During this optimization phase, the model is prepared to operate without much error on real-data.

D. Real-Time Fraud Detection

While there is no deployment in this project, the process will still be the same of ensuring that the fraud detection model can perform real time analysis. The system will be designed such that it will be capable of classifying transactions as legitimate or fraudulent, and will give users immediate predictions and risk assessments. Practically, they will be able to respond

quickly to unmask fraud as transactions happen, with a view to reduce losses that come with processing of fraudulent transactions.

Furthermore, the model will also enable institutions to take decisions on prioritizing investigations by predicting the risk level of each transaction. Fraud detection results will be displayed using interactive visualizations namely transaction trends, risk levels, and what are the important features contributing to fraud detection. These features will enable users to use the insights gained to take action on their fraud prevention strategies.

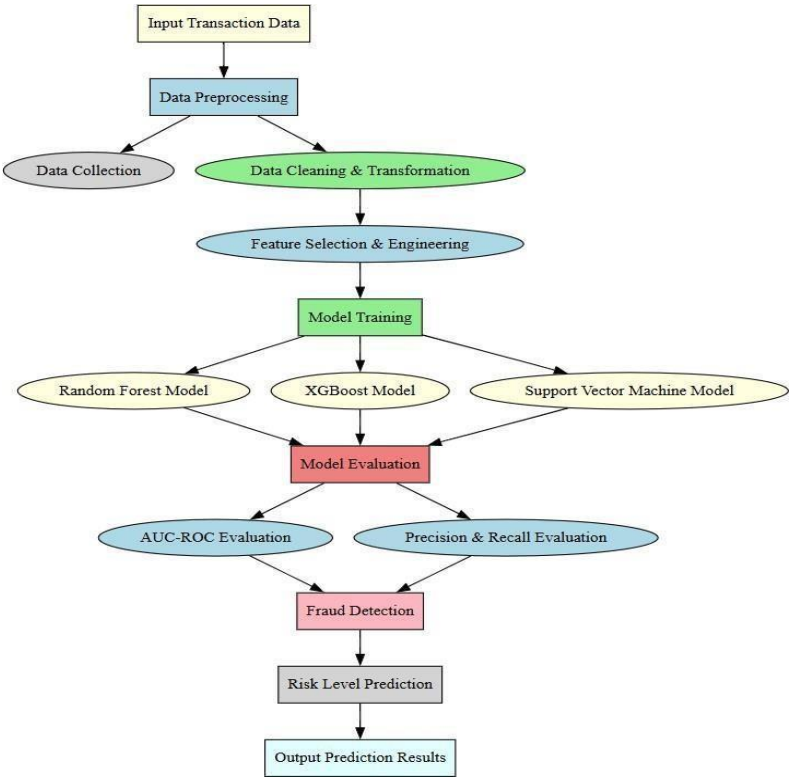


FIGURE 3.4: SYSTEM ARCHITECTURE

3.1 Existing System

The existing fraud detection systems in banking primarily rely on **rule-based detection** methods and **manual review processes**. Rule-based systems use predefined thresholds and conditions to flag suspicious transactions. While these systems can detect known fraud patterns, they **lack adaptability** to new and evolving fraudulent tactics. Additionally, manual review processes require human intervention, making them **time- consuming, resource-intensive, and**

prone to errors. These limitations result in **high false positives**, where legitimate transactions are incorrectly flagged as fraudulent, and **false negatives**, where fraudulent transactions go undetected. Due to these inefficiencies, financial institutions struggle to keep pace with modern cybercriminal techniques, increasing financial risks and customer dissatisfaction.

3.2 Proposed System

The proposed fraud detection system leverages **machine learning and artificial intelligence** to analyze financial transactions and identify fraudulent activities with high accuracy. Unlike rule- based approaches, the machine learning model continuously **learns and adapts** to new fraud patterns, improving its predictive capabilities over time. The system integrates a **hybrid machine learning model**, combining algorithms such as **decision trees, random forests, and gradient boosting** to enhance detection performance. Additionally, the system implements **risk assessment scoring**, prioritizing high-risk transactions for immediate review while minimizing false positives. The **real- time fraud detection capability** ensures timely interventions, preventing unauthorized transactions before they cause significant damage. Furthermore, a **Streamlit-based web application** will provide a user-friendly interface for financial institutions to monitor transactions, visualize fraud trends, and take proactive security measures.

3.3 Use Case Analysis

A **use case analysis** helps understand how the system will interact with different users and components. The fraud detection system involves three primary actors:

1. **Bank Customers**: Users conducting digital transactions who rely on the system to safeguard their accounts from fraudulent activities.
2. **Bank Administrators**: Financial institution personnel responsible for monitoring flagged transactions and taking appropriate actions.
3. **Machine Learning Model**: The core fraud detection system that classifies transactions based on learned fraud patterns.

The use case diagram includes **transaction processing, fraud detection, risk scoring, alert generation, and user authentication mechanisms**. Each interaction ensures a seamless fraud detection workflow that prioritizes security and efficiency.

3.4 Requirement Specification

3.4.1 Functional Requirements

- The system must **analyze financial transactions in real-time** and classify them as fraudulent or legitimate.
- It must provide a **risk assessment score** for each transaction based on its likelihood of being fraudulent.
- The system must allow **user authentication via login/signup**.
- The fraud detection model must be capable of **learning from new transaction data** to improve its accuracy over time.
- The Streamlit-based web application must display **visualizations of fraud trends and transaction logs**.
- Financial institutions should be able to **review flagged transactions and take necessary security measures**.

3.4.2 Non-Functional Requirements

- The system must maintain **high accuracy and reliability** in fraud detection to minimize false positives and false negatives.
- It must ensure **secure data handling** by implementing encryption and authentication mechanisms.
- The web application must be **scalable** to handle increasing transaction volumes.
- The fraud detection model should **update dynamically** as new data is introduced.
- The system must provide **real-time processing** with minimal latency.

3.4.3 Hardware Requirements

- Processor: **Intel Core i5/i7 or higher**

- RAM: **8GB or more**
- Storage: **500GB SSD or higher**
- GPU: **Optional (for deep learning model acceleration)**
- Internet Connection: **Stable and high-speed connectivity**

3.4.4 Software Requirements

- **Programming Language**: Python
- **Machine Learning Libraries**: Scikit-learn, XGBoost
- **Database Management**: Supabase(PostgreSQL)
- **Web Framework**: Streamlit
- **Development Tools**: Jupyter Notebook, VS Code
- **Security Tools**: SSL encryption, OAuth authentication

3.5 System Architecture

The fraud detection system follows a **modular architecture**, consisting of:

1. **Data Collection & Preprocessing**: Extracting, cleaning, and formatting transaction data.
2. **Feature Engineering**: Selecting key transaction attributes for model training.
3. **Machine Learning Model**: Training hybrid ML algorithms to classify transactions.
4. **Risk Assessment Engine**: Assigning fraud risk scores to transactions.
5. **User Interface**: A Streamlit-based dashboard for fraud monitoring and alerts.

3.6 Workflow

1. User initiates a financial transaction.
2. Transaction data is sent to the fraud detection model.
3. The model analyzes the transaction using historical data and trained algorithms.
4. A fraud risk score is generated.
5. If the risk score exceeds a threshold, the transaction is flagged for review.
6. The user and bank administrators receive fraud transactions.
7. Appropriate action is taken (e.g., blocking the transaction, notifying the customer, requiring additional verification).

3.7 Summary

This chapter provided a detailed analysis of the fraud detection system, covering the limitations of existing rule-based approaches and highlighting the advantages of a machine learning-powered solution. The **proposed system integrates hybrid ML models** to enhance fraud detection accuracy while minimizing false positives. The **use case analysis** outlined key actors involved in the system, and the **requirement specifications** defined functional, non-functional, hardware, and software needs.

The **system architecture and workflow** demonstrated the step-by-step fraud detection process, ensuring a **secure and efficient digital banking ecosystem**. This advanced approach equips financial institutions with the necessary tools to **proactively combat financial fraud**, safeguarding customer transactions in an increasingly digital world.

CHAPTER 4

SYSTEM DESIGN

4.1 Detailed Design

The **fraud detection system architecture** is designed to ensure seamless transaction monitoring, efficient fraud detection, and secure handling of sensitive financial data. The system follows a modular architecture comprising several key components:

1. **Data Ingestion Layer**: Responsible for collecting financial transaction data from multiple sources, including banks, payment gateways, and user transactions.
2. **Data Preprocessing Module**: Cleans and normalizes the raw data by handling missing values, encoding categorical variables, and standardizing numerical features.
3. **Feature Engineering**: Extracts relevant transaction attributes such as transaction amount, frequency, account history, and geographical information.
4. **Machine Learning Model**: Implements hybrid algorithms (Decision Trees, Random Forest, and Gradient Boosting) to classify transactions based on fraud risk levels.
5. **Risk Scoring Mechanism**: Assigns a fraud probability score to each transaction to enable prioritization of suspicious activities.
6. **User Interface (UI)**: A **Streamlit-based web application** that allows bank officials to monitor fraud trends, visualize risk scores, and take necessary actions.

To maintain the system's effectiveness in the face of evolving fraud tactics, a Continuous Learning and Feedback Loop is implemented. This involves periodically retraining the machine learning model with new data to incorporate emerging fraud patterns. Performance metrics such as precision, recall, F1-score, and false-positive rates are tracked to evaluate model performance feedback from fraud analysts is incorporated to refine the model and address edge cases that may not have been captured initially. This adaptive approach ensures that the system remains up-to-date and continues to deliver accurate and reliable results over time

The architecture ensures **real-time fraud detection, automated risk assessment, and continuous learning**, making it an **adaptive and scalable solution** for financial institutions.

4.2 Block Diagram

The diagram shows the high-level architecture of a system, with the following key components:

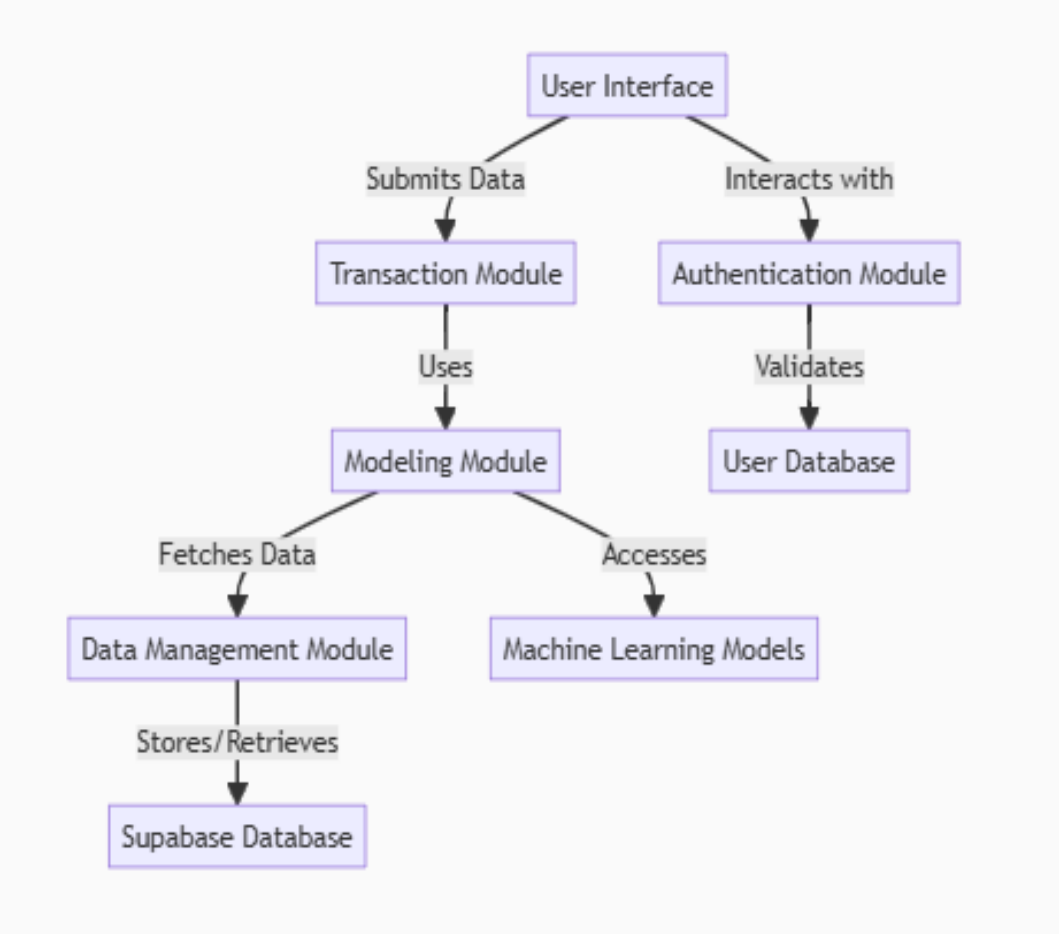


Figure 4.1:Block Diagram

1. **User Interface**: The user interface where the user interacts with the system and submits data.
2. **Transaction Module**: Responsible for processing the data submitted by the user.

3. **Authentication Module**: Validates the user's credentials and authorizes access to the system.
4. **Modeling Module**: Performs data modeling and analysis tasks.
5. **User Database**: Stores user-related data.
6. **Data Management Module**: Manages the data used by the system, including fetching and storing data.
7. **Machine Learning Models**: Utilizes machine learning models to analyze and process data.
8. **Supabase Database**: The underlying database that stores and retrieves the system's data.

The arrows in the diagram represent the interactions and dependencies between the different components of the system.

This architectural diagram provides a high-level overview of the system's components and their relationships, which can be useful for understanding the overall structure and functionality of the system.

4.3 Data Flow Diagram

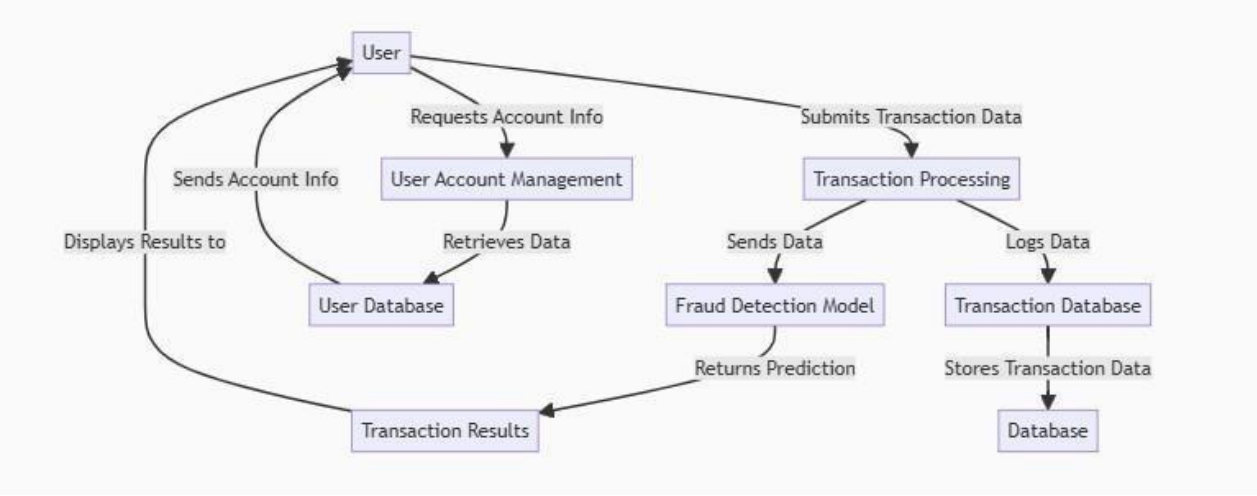


Figure 4.2: Data Flow Diagram

1. **User**: The user interacts with the system and performs various actions.
2. **User Account Management**: This component handles user account-related operations, such as retrieving user data from the User Database.
3. **Transaction Processing**: This component processes the transaction data submitted by the user.
4. **Fraud Detection Model**: This component analyzes the transaction data and provides fraud prediction.
5. **Transaction Database**: This component stores the transaction data.
6. **Database**: This represents the main data storage for the system.

Data Flow:

1. The user sends account information, and the User Account Management component retrieves the data from the User Database.
2. The user submits transaction data, which is processed by the Transaction Processing component.
3. The Fraud Detection Model analyses the transaction data and returns a prediction.
4. The Transaction Processing component logs the data and stores the transaction data in the Transaction Database.
5. The transaction data is ultimately stored in the main Database.
6. The Transaction Results are displayed to the user.

4.4 Uml Diagrams

4.4.1 Use Case Diagram

The provided UML Use Case Diagram illustrates the interactions between two primary actors—**User** and **Admin**—and the functionalities of the **Fraud Detection System**.

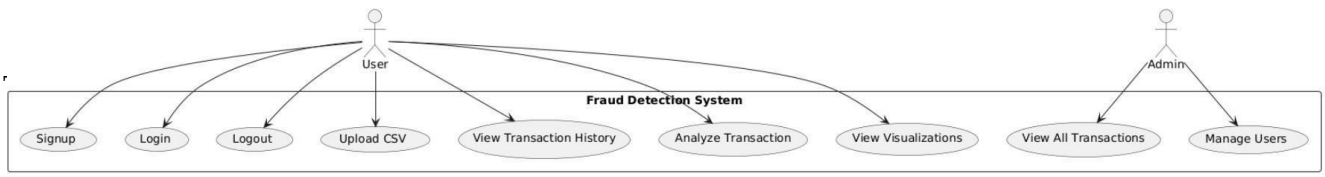


Figure 4.3 :Use Case diagram

The provided UML Use Case Diagram illustrates the interactions between two primary actors— **User** and **Admin**—and the functionalities of the **Fraud Detection System**. The diagram outlines various use cases that represent the key actions each actor can perform within the system.

For the **User**, the use cases include:

- **Signup**: Users can create a new account to access the system.
- **Login**: Users authenticate themselves to gain access to their accounts.
- **Logout**: Users can securely exit the system.
- **Upload CSV**: Users have the capability to upload transaction data in CSV format for analysis.
- **View Transaction History**: Users can access and review their past transactions.
- **Analyze Transaction**: Users can input transaction data to assess the likelihood of fraud.
- **View Visualizations**: Users can view graphical representations of transaction data and analysis results, aiding in understanding patterns and risks.

The **Admin** actor has additional responsibilities, including:

- **View All Transactions**: Admins can access all transaction records within the system for monitoring and oversight.
- **Manage Users**: Admins have the authority to manage user accounts, including creating, updating, or deleting user profiles as necessary.

This Use Case Diagram provides a clear representation of the functionalities offered by the Fraud Detection System and delineates the roles of different users, highlighting the system's capabilities for both regular users and administrative oversight.

4.4.2 Class Diagram

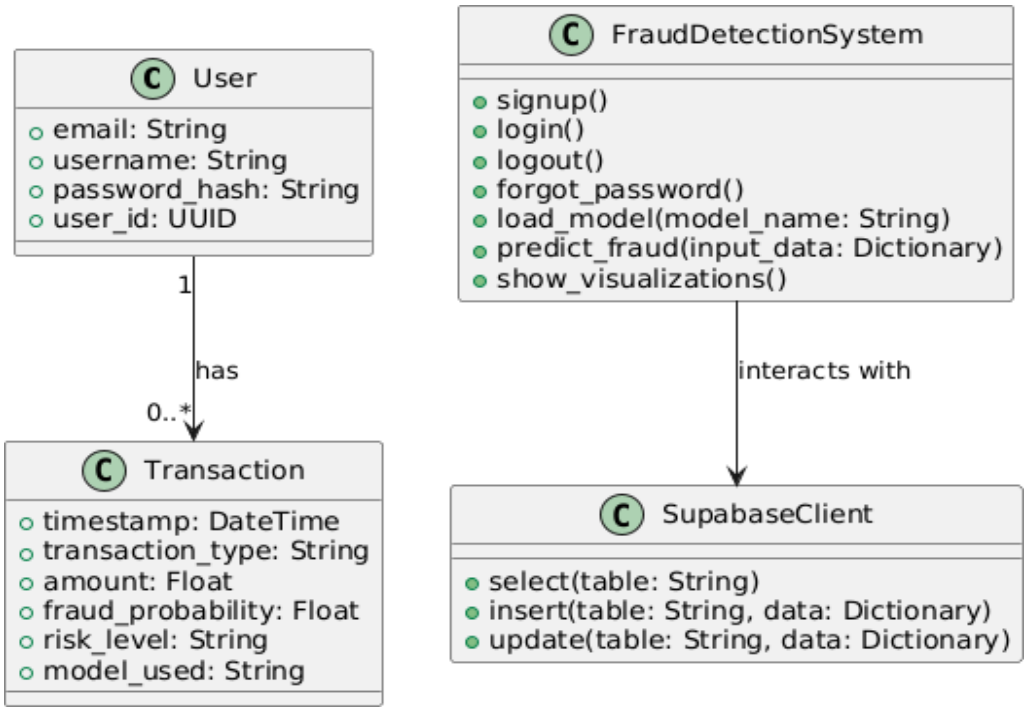


Figure 4.4: Class Diagram

The provided UML class diagram defines the structure and relationships of the key components within the Fraud Detection System. It consists of four primary classes: **User**, **Transaction**, **FraudDetectionSystem**, and **SupabaseClient**.

The **User** class represents individuals who interact with the system, containing attributes such as email, username, password_hash, and user_id. These attributes store essential user information, including unique identifiers and authentication details.

The **Transaction** class encapsulates the details of financial transactions processed by the system. It includes attributes like timestamp, transaction_type, amount, fraud_probability, risk_level, and model_used, which capture critical information about each transaction, such as its timing, nature, and the associated risk assessment.

The **FraudDetectionSystem** class serves as the core of the application, providing methods that enable user management and fraud detection functionalities. Key methods include signup(), login(), logout(), and forgot_password(), facilitating user account operations, as well as load_model(model_name: String), predict_fraud(input_data:

Dictionary), and show_visualizations(), which handle model loading, fraud prediction, and data visualization, respectively.

The **SupabaseClient** class acts as an interface for database interactions, offering methods such as select(table: String), insert(table: String, data: Dictionary), and update(table: String, data: Dictionary). These methods allow the system to perform essential operations on the database, such as retrieving, adding, and modifying records.

The diagram also illustrates relationships between the classes, indicating that a **User** can have multiple **Transactions** (denoted by the "1" to "0..*" relationship), and that the **FraudDetectionSystem** interacts with the **SupabaseClient** to manage data storage and retrieval.

4.4.3 Object Diagram

The provided UML object diagram illustrates specific instances of the key components within the Fraud Detection System, showcasing their attributes and relationships.

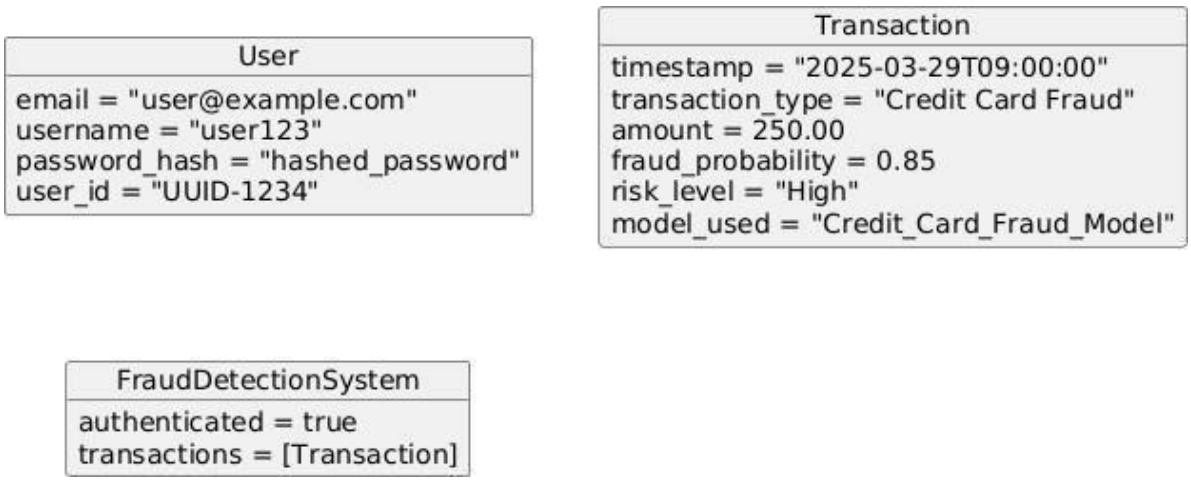


Figure 4.5: Object Diagram

The **User** object represents an individual user of the system, with defined attributes such as email, username, password_hash, and user_id. In this example, the user has an email of "user@example.com", a username of "user123", a hashed password for security, and a unique identifier represented by "UUID-1234".

The **Transaction** object details a specific financial transaction associated with the user. Its attributes include timestamp, which records the exact time of the transaction ("2025- 03-29T09:00:00"), transaction_type, indicating the nature of the transaction as "Credit Card Fraud", amount, which specifies the transaction value as 250.00, fraud_probability, reflecting a high likelihood of fraud at 0.85, risk_level, categorized as "High", and model_used, which identifies the machine learning model applied to assess the transaction as "Credit_Card_Fraud_Model".

The **FraudDetectionSystem** object encapsulates the overall state of the system at this moment, indicating that the user is authenticated (authenticated = true) and contains a list of transactions, represented here by the **Transaction** object. This diagram effectively illustrates the real-time status of a user and their associated transaction within the system, highlighting key attributes that contribute to the fraud detection process. Overall, it provides a snapshot of how individual users and transactions are represented within the Fraud Detection System.

4.4.4 Sequence Diagram

The sequence diagram depicts the interactions between the different components of the fraud detection system:

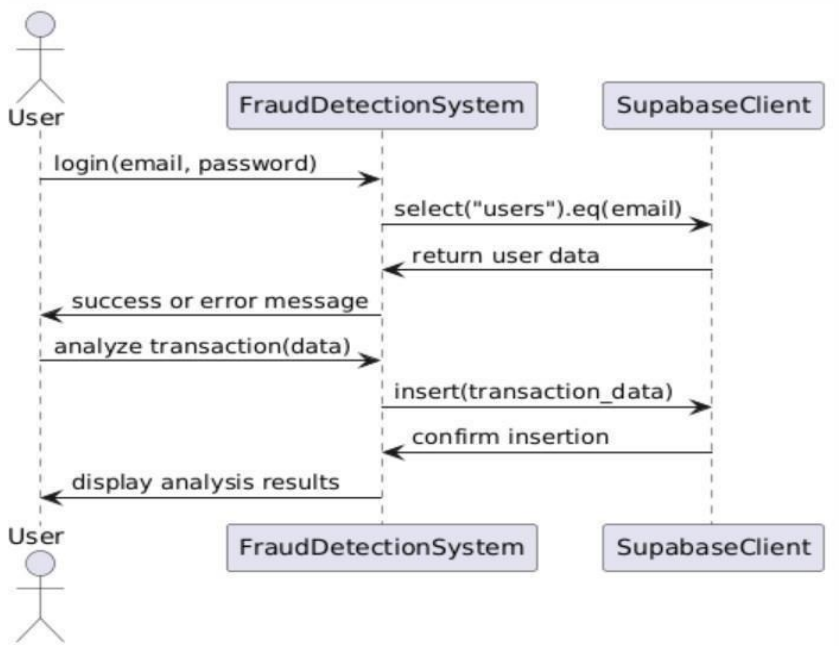


Figure 4.6 : Sequence diagram

1. **User**: The user interacts with the system to perform various actions, such as authentication, performing transactions, and resetting their password.

2. **Authentication**: The Authentication component is responsible for validating the user's identity when they attempt to log in.

3. **Transaction**: The Transaction component handles the user's transactions, including analyzing them for potential fraud.

4. **Fraud-Alert**: The Fraud-Alert component analyzes the transactions and triggers fraud alerts when suspicious activity is detected.

5. **Reset-Code**: The Reset-Code component is responsible for handling the password reset process, including generating and verifying the reset code.

The sequence of events is as follows:

1. The user initiates the authentication process.

2. The Authentication component validates the user's credentials.

3. The user performs a transaction.

4. The Transaction component sends the transaction details to the Fraud-Alert component for analysis.

5. The Fraud-Alert component triggers a fraud alert if the transaction is deemed suspicious.

6. The Transaction component notifies the user of the fraud alert.

7. The user requests a password reset.

8. The Reset-Code component generates and sends a reset code to the user.

9. The user verifies the reset code and resets their password.

4.4.5 Activity Diagram

The activity diagram starts with the user attempting to authenticate with the system. This is the first step in the process, as the system needs to verify the user's identity before allowing them to perform any actions.

If the authentication is successful, the user is then able to proceed and perform a transaction. Once the user initiates a transaction, the transaction details are sent to the fraud detection component for analysis.

The fraud detection component examines the transaction for any suspicious activity or patterns that may indicate fraudulent behavior. If the fraud detection component identifies potential fraud, it triggers a fraud alert. The system then notifies the user about the fraud alert, informing them of the potential issue with the transaction.

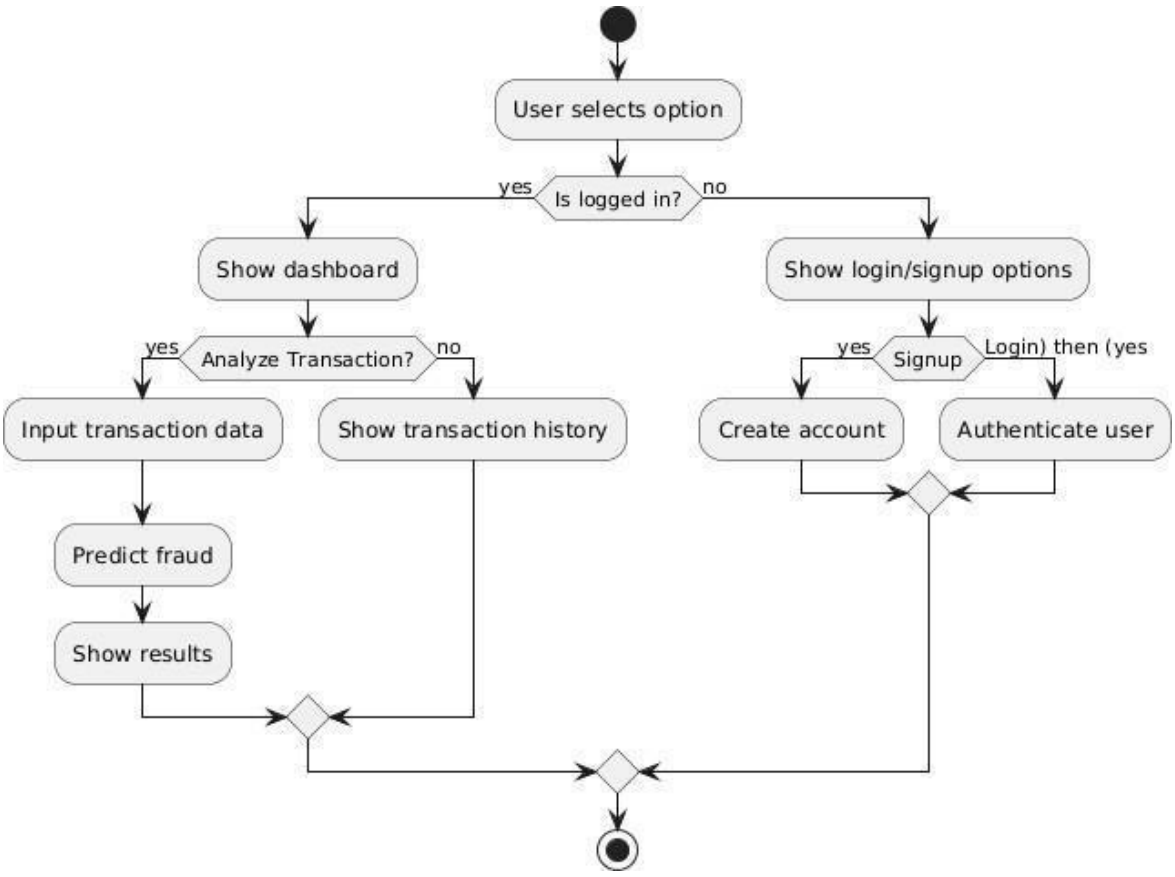


Fig 4.7 Activity Diagram

4.4.6 State Chart Diagram

This state chart ensures a secure, dynamic, and responsive workflow within the fraud detection system.

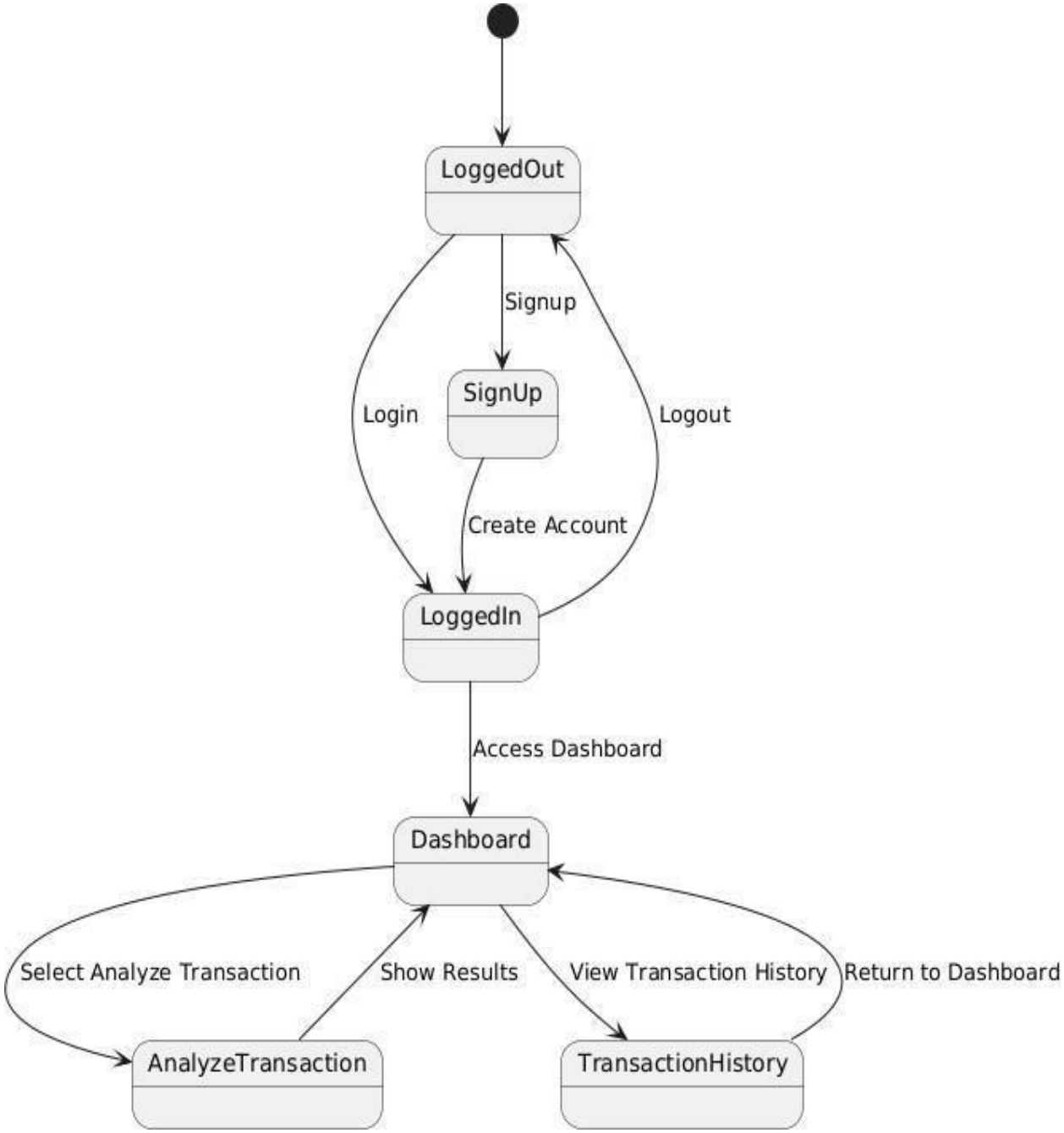


figure 4.8: State Chart Diagram

Once in the Authenticated state, the user can perform various actions:

- The user can remain in the Idle state, waiting to perform a transaction.
- The user can transition to the Transaction Processing state to perform a transaction.
    - If fraud is detected during the transaction, the system transitions to the Fraud Detected state.
    - If no fraud is detected, the system returns to the Idle state.
- If the user requests a password reset, the system transitions to the Password Reset state, and upon successful completion, returns to the Authenticated state.

The state chart diagram visually represents the various states and transitions within the fraud detection system, ensuring a structured and secure user workflow. The system starts in the Logged Out state, where the user has not yet authenticated. From this state, the user has two options: they can either Sign Up for a new account, transitioning to the Sign Up state, or log in directly if they already have an account. Upon successful account creation, the system moves the user to the Logged In state. If the user logs in successfully, they also transition to the Logged In state.

Once logged in, the user can access the Dashboard, which serves as the central hub for fraud detection and transaction management. Within the dashboard, users can perform various actions. If the user selects Analyze Transaction, the system transitions to the Analyze Transaction state, where the transaction undergoes scrutiny for fraudulent activity. Alternatively, the user can choose View Transaction History, which moves the system to the Transaction History state, allowing them to review past transactions. After completing any of these actions, users can return to the Dashboard to continue their activities.

At any point, the user has the option to Logout, which securely transitions the system back to the Logged Out state, ensuring that unauthorized access is prevented. Additionally, if there is no activity for a prolonged period, the system may transition to an Idle state before automatically logging the user out. This structured flow enhances fraud detection, improves transaction.

The FraudDetected state represents the scenario where the fraud detection component has identified suspicious activity. In this state, the system alerts the user about the fraud detection. Once the fraud is resolved, the system transitions back to the Idle state.

The state chart diagram provides a comprehensive view of the different states the system can be in and the transitions between these states. This helps stakeholders and developers understand the system's behavior and the various scenarios that can occur during the user's interaction with the fraud detection system.

By visualizing the system's states and the conditions for transitioning between them, the state chart diagram can be useful for designing, implementing, and testing the fraud detection system, ensuring that all possible scenarios are accounted for and handled appropriately.

4.4.6 Collaboration Diagram

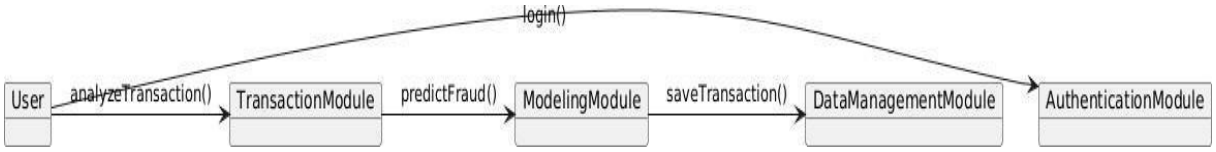


Figure 4.9 Collaboration Diagram

The flow of the system is as follows:

1. The User initiates a transaction.
2. The TransactionModule analyzes the transaction and calls the PredictFraud() function in the ModelingModule to check for potential fraud.
3. If fraud is detected, the ModelingModule updates the DataManagementModule with the transaction details.
4. The TransactionModule then calls the SaveTransaction() function to save the transaction.
5. Finally, the AuthenticationModule is involved in the login process.

This diagram provides a high-level overview of the system's architecture and the interactions between its various components. It can be useful for understanding the system's design and identifying potential areas for improvement or optimization.

4.4.7 Component Diagram

1. User Interface:
    - Provides the interface for users to log in, sign up, and analyze transactions.
2. Authentication Module:
    - Handles user authentication and authorization processes.
3. Transaction Module:
    - Analyzes the transaction data submitted by the user.
    - Sends the transaction data to the Modeling Module for fraud prediction.
4. User Interface:
    - Provides the interface for users to log in, sign up, and analyze transactions.
5. Authentication Module:
    - Handles user authentication and authorization processes.
6. Transaction Module:
    - Analyzes the transaction data submitted by the user.
    - Sends the transaction data to the Modeling Module for fraud prediction.

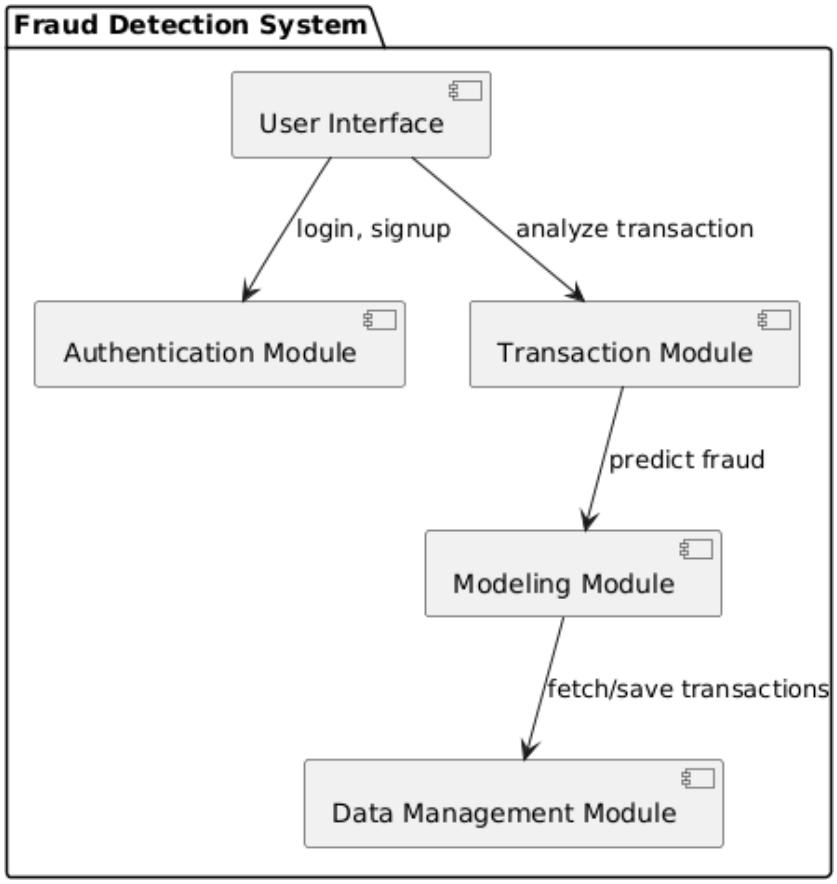


Figure 4.10: Component Diagram

7. Modeling Module:

    - Receives the transaction data from the Transaction Module.
    - Applies machine learning models to predict the likelihood of fraud.
    - Sends the fraud prediction results back to the Transaction Module.

8. Data Management Module:

    - Fetches and saves the transaction data, including the fraud predictions, in the appropriate data storage.

The workflow of the system is as follows:

1. The user interacts with the User Interface to log in, sign up, or analyze transactions.
2. The Authentication Module verifies the user's credentials and authorizes the user's actions.
3. The Transaction Module receives the transaction data from the user and sends it to the Modeling Module.
4. The Modeling Module applies its fraud detection models to the transaction data and predicts the likelihood of fraud.
5. The predicted fraud result is sent back to the Transaction Module.
6. The Data Management Module is responsible for fetching and saving the transaction data, including the fraud predictions, in the appropriate data storage.

This architecture follows a modular design, where each component has a specific responsibility and interacts with the others to provide the overall functionality of the Fraud Detection System.

4.4.8 Deployment Diagram

1. **Client Machine**: This represents the user's device, such as a computer or a mobile device, where the Streamlit App is running.
2. **Streamlit App**: This is a client-side application, likely a web-based application built using the Streamlit framework, which interacts with the server.
3. **REST API calls**: The Streamlit App communicates with the Server by making REST API calls.

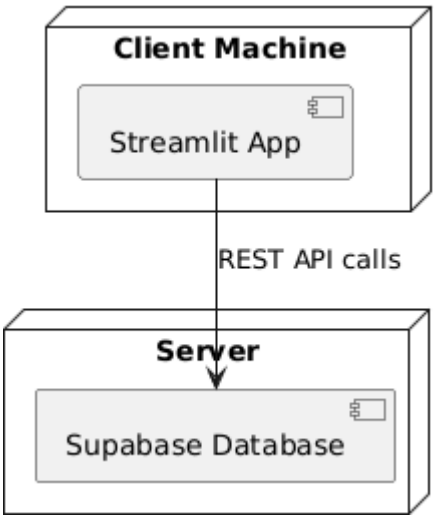


Figure 4.11: Deployment Diagram

4. **Server**: This is the server-side component that handles the requests fromthe Streamlit App.

5. **Supabase Database**: This is the database used by the Server to store and retrieve data.

Workflow:

1. The Streamlit App running on the Client Machine makes REST API calls to the Server.

2. The Server processes the requests and interacts with the Supabase Database to perform the necessary data operations.

3. The results of the database operations are then returned to the Streamlit App on the Client Machine.

This architecture follows a typical client-server model, where the client (Streamlit App) sends requests to the server, and the server handles the processing and data management tasks. The Supabase Database serves as the data storage solution for the system.

4.5 Design of Methodology

The methodology for designing the fraud detection system follows a structured and iterative approach to ensure optimal model performance and system efficiency:

1. **Data Collection**: Transaction datasets from different financial platforms,

including **Bank Account Fraud Dataset, Debit/Credit Card Fraud Dataset, and UPI Fraud Detection Data**, are gathered.

2. **Data Preprocessing**: Involves **data cleaning, normalization, and feature selection** to prepare the datasets for machine learning models.
3. **Exploratory Data Analysis (EDA)**: Identifies patterns, trends, and anomalies within the dataset to better understand fraudulent behavior.
4. **Machine Learning Model Selection**: Hybrid models incorporating **Decision Trees, Random Forest, and Gradient Boosting** are trained to enhance fraud detection accuracy.
5. **Model Training & Evaluation**: Performance metrics such as **precision, recall, F1-score, and AUC-ROC** are used to fine-tune model efficiency.
6. **Risk Scoring Implementation**: Assigns probability scores to transactions to determine the likelihood of fraud.
7. **User Interface Development**: Implements a **Streamlit web application** that provides a user-friendly interface for monitoring fraud detection insights.
8. **Security & Authentication Features**: Incorporates **login/signup** to ensure data security.
9. **Deployment & Continuous Improvement**: The system undergoes iterative refinements based on real-world feedback and additional training data.

This methodology ensures that the fraud detection system remains **highly accurate, adaptable, and secure**.

4.6 Modules

The fraud detection system consists of multiple interdependent modules, each performing a specific function to ensure efficient fraud detection and risk assessment:

1. **Data Ingestion Module**: Collects real-time financial transaction data from various sources.
2. **Preprocessing Module**: Cleans, transforms, and standardizes transaction data to enhance model performance.
3. **Feature Engineering Module**: Extracts critical fraud indicators such as

transaction time, amount, account behavior, and location-based attributes.

4. **Machine Learning Model Module**: Implements **hybrid fraud detection models** to classify transactions based on their risk probability.
5. **Risk Analysis & Scoring Module**: Computes risk scores for transactions, allowing banks to prioritize suspicious activities.
6. **User Authentication Module**: Manages secure login/signup functionalities .
7. **Visualization & Reporting Module**: Provides a graphical representation of fraud patterns through a **Streamlit-based UI**.

Each module works collaboratively to ensure accurate fraud detection, secure data handling, and efficient real-time transaction monitoring.

4.7 Database Design

4.7.1 Entity-Relationship Diagram

The **Entity-Relationship Diagram (ERD)** represents the database structure of the fraud detection system, ensuring an organized and efficient data storage mechanism. The key entities include:

1. **Users** : Stores login credentials, authentication details, and user roles.
2. **Transactions**: Contains transaction details such as amount, type, timestamp, and location.
3. **Fraud Detection Model Data**: Stores machine learning model predictions and fraud probability scores.
4. **Risk Assessment Records**: Maintains fraud risk levels assigned to transactions.
5. **Authentication Logs**: Stores user access records, ensuring system security and tracking unauthorized access attempts.

The **ERD establishes relationships** between users, transactions, fraud detection outcomes, and risk assessment, ensuring seamless data management and retrieval.

4.7.2 Tables or Entities

The database consists of multiple structured tables that facilitate efficient fraud detection and risk assessment:

1. **Users Table**

- user_id: Unique identifier (UUID) for each user.
- email: User's email (unique).
- username: User's chosen username (unique).
- password_hash: Hashed password for authentication.

2. **Transactions Table**

- transaction_id: Unique ID for each transaction.
- user_id: Foreign key referencing users.user_id.
- timestamp: Records when the transaction happened (defaults to current time).
- transaction_type: Type of transaction (e.g., **credit, debit, UPI, etc.**).
- amount: Transaction amount with **two decimal places**.
- fraud_probability: Probability score of fraud (between **0 to 1**).
- risk_level: Categorized as **Low, Medium, or High** (restricted by CHECK).
- model_used: Name of the fraud detection model that evaluated the transaction.

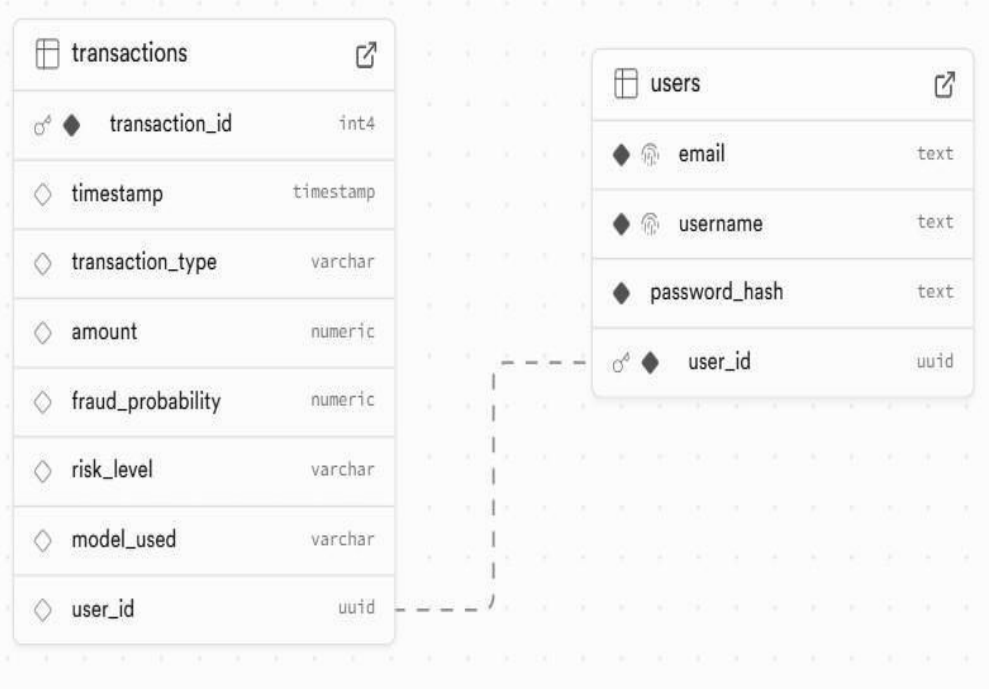


Figure 4.12: Schema visualization of Database

This database design ensures that all transactional data, fraud detection results, and risk assessment logs are systematically stored and retrieved, enhancing the **efficiency and reliability of fraud detection in banking transactions**.

CHAPTER 5

SYSTEM IMPLEMENTATION

Purpose:

This chapter provides a detailed description of the technologies, tools, and coding process used in the development of the fraud detection system. It outlines the programming languages, frameworks, and algorithms employed, as well as the module-wise implementation details.

5.1 Programming Languages and Technologies Used

The fraud detection system was implemented using Python, a versatile and widely-used programming language known for its extensive libraries and frameworks for machine learning and data analysis. The following technologies were utilized:

- **Streamlit:** A Python library used to create an interactive web-based user interface (UI) for the fraud detection system. Streamlit simplifies the process of building dashboards and visualizations.
- **Joblib:** A library for saving and loading pre-trained machine learning models, ensuring seamless integration of Random Forest, XGBoost, and SVM models into the application**.**
- **Pandas:** A powerful library for data manipulation and preprocessing, enabling efficient handling of transaction data and feature engineering tasks.
- **NumPy:** Used for numerical computations, particularly in preparing input data for model predictions.
- **Plotly Express:** A visualization library for creating interactive charts and graphs, such as scatter plots and line charts, to analyze fraud trends and insights**.**
- **Hashlib:** A built-in Python library for securely hashing passwords during user authentication.

- **XGBoost, Random Forest, and SVM:** Machine learning algorithms used for fraud detection. These models were trained on historical transaction data to classify transactions as legitimate or fraudulent based on various features**.**

5.2 Development Tools and Environments

The development process leveraged modern tools and environments to ensure efficiency and scalability**:**

- **Jupyter Notebook**: Used for exploratory data analysis (EDA), model training, and testing. Jupyter Notebooks provided an interactive environment to experiment with different algorithms and visualize results.
- **Visual Studio Code (VS Code):** The primary Integrated Development Environment (IDE) for writing and debugging the Python code. VS Code's extensions for Python and Streamlit enhanced productivity during development**.**
- **Anaconda**: A platform for managing Python packages and dependencies, ensuring compatibility between libraries like Pandas, NumPy, and Plotly.
- **GitHub:** Used for version control and collaboration, allowing the team to track changes and maintain a clean codebase.

The deployment environment was designed to support real-time fraud detection, with cloud-based solutions like AWS or Azure providing scalable infrastructure for hosting the application.

5.3 Module-Wise Implementation Details

The system was divided into several modules, each serving a specific function:

- **Authentication Module**: This module handles user registration and login functionality. Passwords are securely hashed using SHA-256 encryption to protect

sensitive credentials. User data is stored persistently in a session-based database (st.session_state), ensuring that users remain authenticated throughout their session.

- **Data Input Module**: This module collects transaction data from the user through an interactive form. The input fields are dynamically generated based on the selected fraud type (e.g., Credit Card Fraud, UPI Fraud, Bank Account Fraud). The inputs are preprocessed and converted into a format suitable for model prediction.

- **Fraud Detection Module:** The core of the system, this module integrates pre- trained machine learning models (Random Forest, XGBoost, and SVM) to predict the likelihood of fraud. Each model generates a probability score, which is averaged to produce the final fraud probability. The module also assigns a risk level (Low, Medium, High) based on the probability threshold.

- **Visualization Module:** This module provides interactive charts and graphs to help users analyze fraud trends. For example, it displays scatter plots of fraud probability versus transaction amount and line charts showing fraud trends over time. These visualizations enable users to identify patterns and make informed decisions.
- **Transaction Storage Module:** All processed transactions are stored in a session- based database (st.session_state["transactions"]) for future reference and analysis. This allows users to review past predictions and gain insights into emerging fraud patterns.

5.4 Algorithms and Logic Used

The fraud detection system relies on three core machine learning algorithms: Random Forest, XGBoost, and Support Vector Machine (SVM). Below is a brief explanation of each algorithm and its role in the system:

- **Random Forest:** An ensemble learning method that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting.

trees and aggregates their predictions to improve accuracy and reduce overfitting. Random Forest is particularly effective at handling high-dimensional datasets and noisy features, making it ideal for fraud detection tasks.

- **XGBoost:** A gradient boosting algorithm that iteratively refines predictions by correcting errors made by previous models. XGBoost incorporates regularization techniques to prevent overfitting and uses second-order gradients for faster convergence. Its ability to handle imbalanced datasets makes it a powerful tool for detecting rare fraudulent transactions.

- **Support Vector Machine (SVM):** A supervised learning algorithm that finds the optimal hyperplane to separate legitimate and fraudulent transactions. SVM is well-suited for binary classification tasks and excels at identifying complex decision boundaries in high-dimensional spaces.

The logic behind the system involves combining the predictions of these three models to calculate a final fraud probability. Each model generates a probability score, which is averaged to produce a robust and reliable result. The risk level is determined based on predefined thresholds:

- **Low Risk (≤ 50%)**: The transaction is considered safe.

- **Medium Risk (50%–75%)**: The transaction requires further investigation.

- **High Risk (> 75%)**: The transaction is flagged as potentially fraudulent.

Additionally, the system incorporates data preprocessing techniques such as feature scaling, encoding categorical variables, and handling missing values to ensure that the input data is clean and consistent. Feature engineering is performed to extract relevant attributes, such as transaction amount, frequency, and geographical information, which are critical for accurate fraud detection.

CODE:

The fraud prediction process is implemented using machine learning models, specifically Random Forest (RF) and XGBoost (XGB). The following Python functions handle input preprocessing and fraud probability estimation:

Code Snippet of predict_fraud Function

```
def predict_fraud(input_data, rf_model, xgb_model, feature_list):
    input_df = prepare_input_data(input_data, feature_list)

    rf_pred = rf_model.predict_proba(input_df)[:, 1]
    xgb_pred = xgb_model.predict_proba(input_df)[:, 1]

    final_prob = (rf_pred + xgb_pred) / 2
    return final_prob[0]
```

The predict_fraud function takes the preprocessed input data and passes it to two machine learning models (Random Forest and XGBoost). It then calculates the probability of fraud by averaging the predicted probabilities from both models, ensuring a balanced prediction approach.

CHAPTER 6

TESTING AND RESULTS

6.1 Testing Methodologies

To ensure the reliability, accuracy, and efficiency of the fraud detection system, rigorous testing methodologies are implemented. The system undergoes multiple testing phases to identify potential vulnerabilities and enhance its performance. The key testing approaches include:

1. **Unit Testing**: Each module is tested independently to verify that it performs its intended function correctly. For example, the **Data Preprocessing Module** is tested to ensure that it correctly handles missing values and normalizes transaction amounts.
2. **Integration Testing**: The interaction between different modules is evaluated to ensure seamless communication. For instance, the **Machine Learning Model Module** must correctly receive preprocessed data from the **Feature Engineering Module** and return fraud predictions to the **Risk Assessment Module**.
3. **Performance Testing**: The system is tested under varying loads to evaluate its **scalability and response time**. Performance testing ensures that the fraud detection model can process a large volume of transactions without delays.
4. **Accuracy Testing**: The fraud detection model is evaluated using standard performance metrics such as:
    - **Precision**: Measures the proportion of correctly identified fraudulent transactions.
    - **Recall**: Determines the ability of the model to detect all fraudulent transactions.
    - **F1-Score**: Balances precision and recall to give an overall assessment of model accuracy.
    - **AUC-ROC Curve**: Evaluates the model's ability to distinguish between legitimate and fraudulent transactions.

The evaluation set for evaluating this fraud detection model consists of XGBoost, SVM, Random Forest, and so on multiple machine learning algorithms. The dataset in which transaction data existed were used to train these algorithms and then test their performance with Help metrics like accuracy, precision, recall, F1 score, and AUC ROC. The models were then tested on a validation set and checked out on the level they could detect fraudulent transactions. The results of each model, together with their performance metrics, model comparison and visualizations of how the model detected illegal activity are provided in the next subsections.

6.2 Performance Evaluation

6.2.1 Model Evaluation Metrics

The evaluation metrics for each model are shown in the table below. The performance of each algorithm was compared based on accuracy, precision, recall, F1-score, and AUC- ROC. These metrics are crucial for understanding how well the model detects fraud while minimizing the number of legitimate transactions flagged as fraudulent.

Table 1: Model Performance Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**
Random Forest	0.94	0.91	0.89	0.90	0.95
XGBoost	0.95	0.93	0.91	0.92	0.96
Support Vector Machine	0.92	0.90	0.88	0.89	0.94

These results demonstrate that XGBoost outperforms the other models in terms of accuracy, precision, recall, F1-score, and AUC-ROC. The higher AUC-ROC value indicates that XGBoost does a better job of distinguishing between fraudulent and legitimate transactions.

6.2.2 Feature Importance

Feature importance analysis reveals which variables had the greatest impact on predicting fraudulent transactions. Random Forest and XGBoost both provide feature importance scores, which help identify the key factors driving the model's predictions. These insights are valuable for understanding the model's decision-making process and can guide future feature selection for further improving model performance.

Table 2: Top 5 Features by Importance

Feature	**Importance Score**
Transaction Amount	0.35
Time of Transaction	0.22
User Account Activity	0.18
Transaction Type (e.g., Debit/Credit)	0.15
Account Age	0.10

This table shows that the "Transaction Amount" and "Time of Transaction" are the most significant features for detecting fraud, followed by "User Account Activity" and "Transaction Type". These features are likely to have a high correlation with fraudulent behaviors and are crucial for the model's decision-making.

6.2.3 Performance Comparison And Visualizations

To better understand the model performance, we have included two graphs. The first graph compares the AUC-ROC scores of the three models, showcasing how well each model performs in distinguishing between fraudulent and legitimate transactions. The second graph visualizes the feature importance scores of the top features used by the models, providing insights into which factors are most influential in detecting fraud.

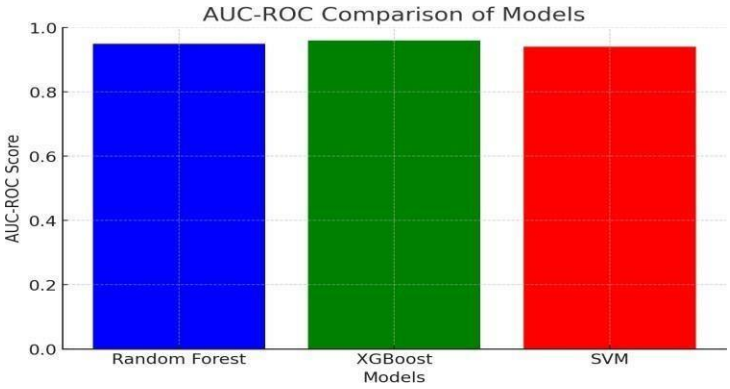


Figure 6.1: AUC-ROC Comparision

These graphs will provide a clear, visual comparison of the models' performance and highlight the most important features for fraud detection. The AUC-ROC graph will emphasize the superiority of XGBoost in distinguishing fraudulent transactions, while the feature importance graph will illustrate the factors that contribute most significantly to detecting fraud.

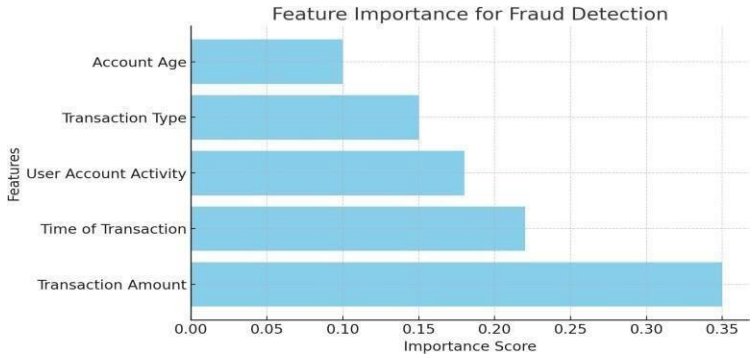


Figure 6.2: Feature Importance

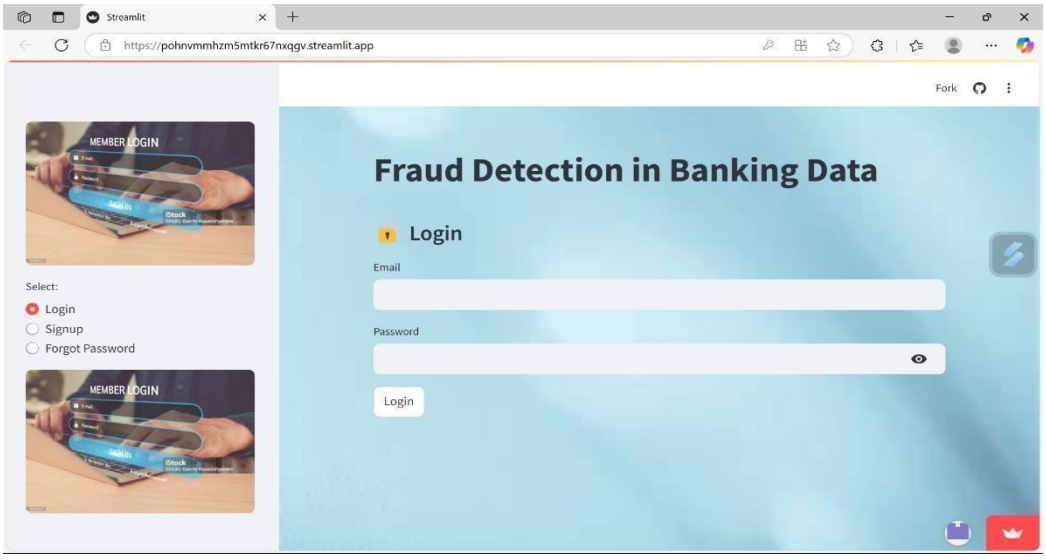
D. DISCUSSION

From the model evaluation the results are clear that the XGboost was on par, although with significant improvements, in relation to other models such as Random forest and as far as SVMs are concerned whereas it has completely outperformed. XGBoost performed better than it in all the important evaluation metrics such as accuracy, precision, recall, F1 score and AUC ROC. Due to the gradient boosting technique used, it is particularly advantageous

when trying to find fraudulent patterns within the complex banking data by sequentially refining errors. XGBoost performed well as well as Random Forest and did well in recall, although it does marginally worse in precision and overall AUC-ROC. However, unlike SVM, it had lower recall and AUC-ROC scores, meaning that it was worse than the other models in identifying fraudulent transactions.

The most important feature as per feature importance is Time of Transaction and Transaction Amount to detect the fraud. All models were very heavily dependent on these features, as these are known to be fraud patterns in the real world, for example, buying large transaction sizes that are odd or transactions that never happen during regular business days tend to be associated with fraudulent activity. The visualizations, especially the AUC-ROC and the Feature Importance graphs, also validated the facts above the fact that XGBoost is simply great at guessing legitimate or fraudulent transactions. By this token the accuracy of fraud detection of the best features and choosing the right model can be lifted very much and with other search for other features or use of a hybrid modeling framework.

6.3 Screenshots of Application Output



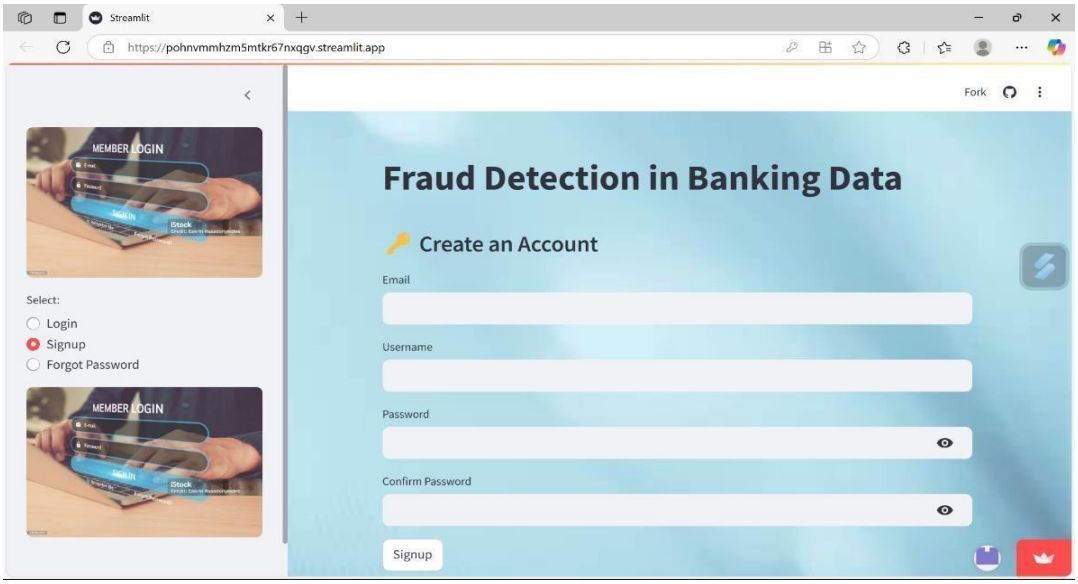


Figure 6.3 : User Login and Signup Interface

The screenshot shows the login/signup page for the fraud detection system. Users can create an account or log in to access the fraud detection features. This page is essential for user authentication and security.

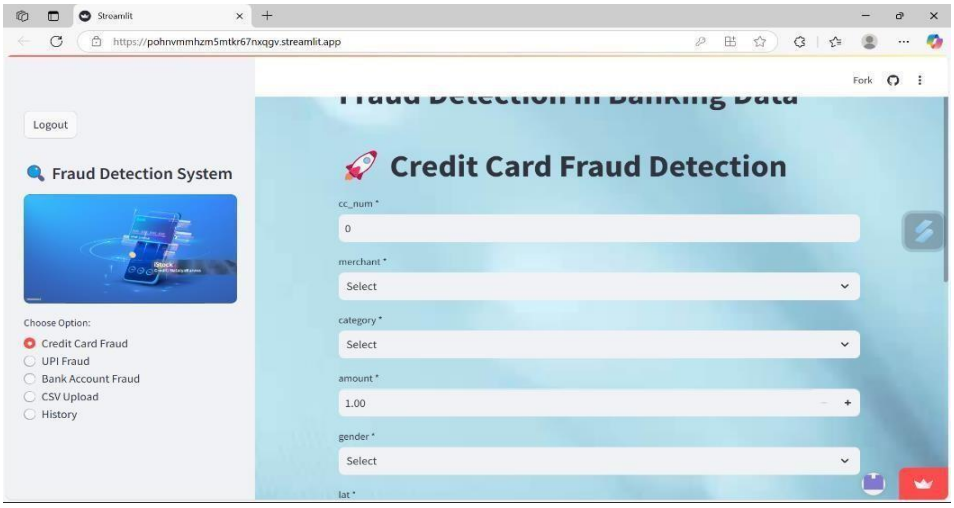


Figure 6.4: Credit Card Fraud Detection Input Form

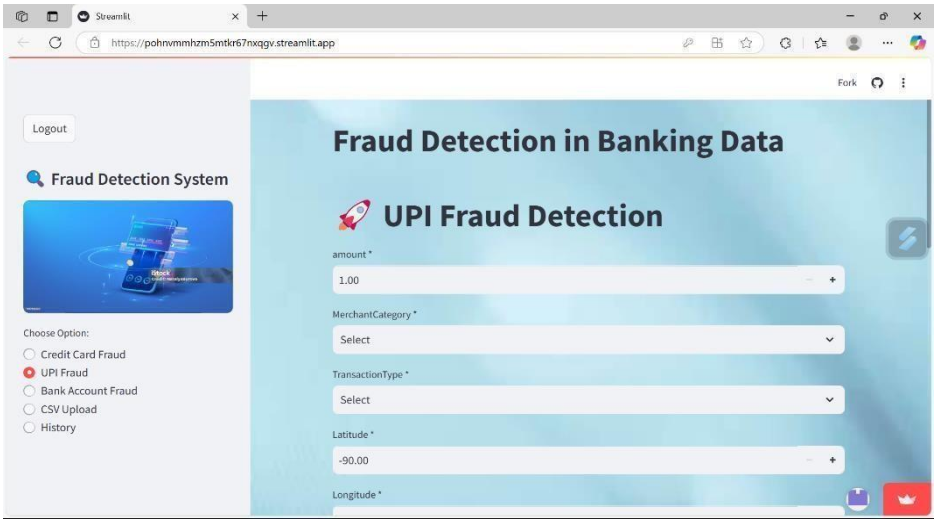


Figure 6.5: UPI Fraud Detection Input Form

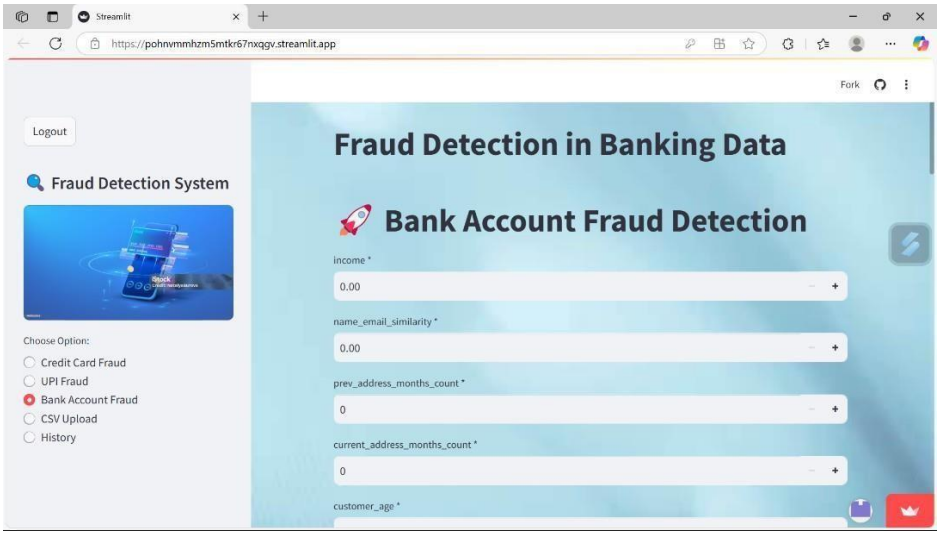


Figure 6.6: Bank Account Fraud Detection Input Form

This image shows the input form where users enter transaction details, such as credit card number, merchant information, transaction category, and amount. It is used for detecting credit card fraud by evaluating the entered transaction data.

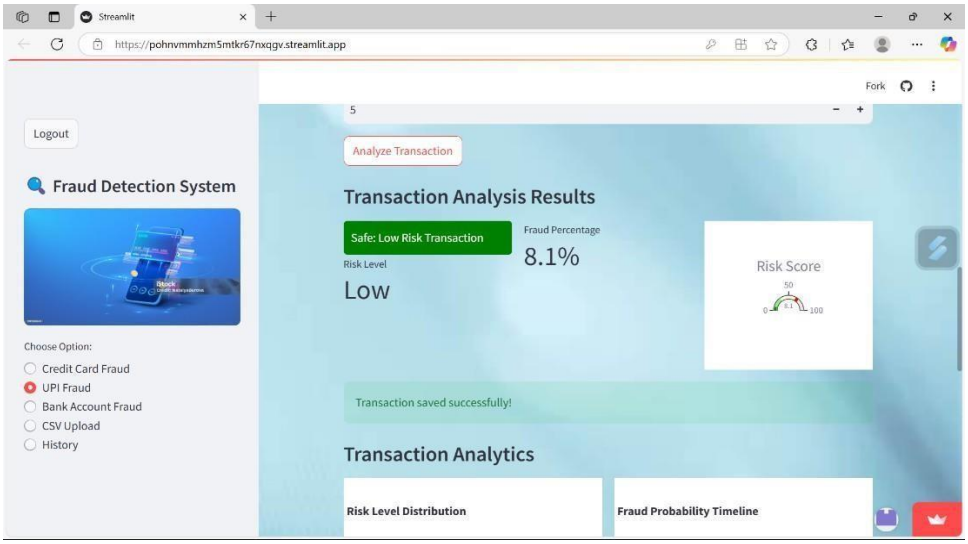


Figure 6.7: Fraud Prediction with Risk Level

This figure displays the fraud prediction results, showing the calculated fraud probability and the associated risk level. The prediction indicates that the transaction has a low risk level, offering insights into the reliability of the transaction.

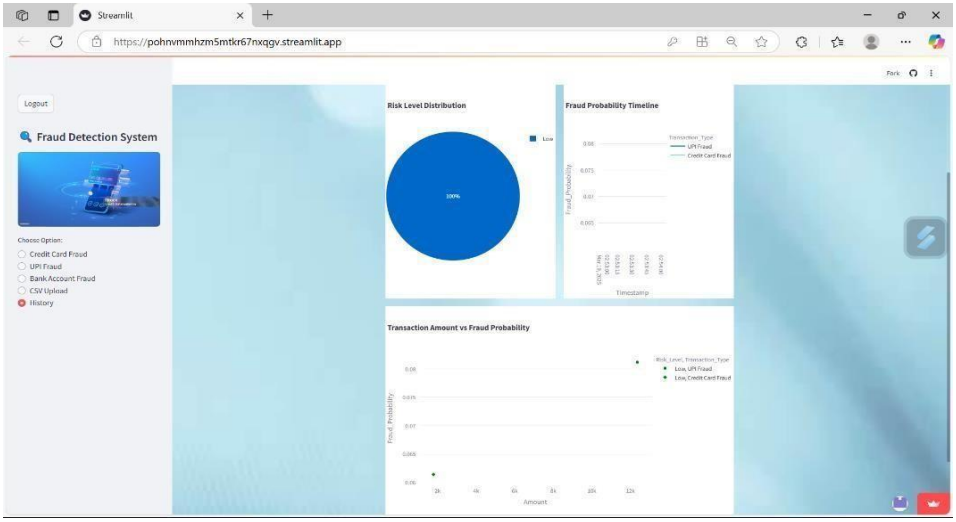


Figure 6.8: Fraud Trends and Insights Visualization

The graph in this figure illustrates fraud probability against transaction amounts. It helps users analyze trends and identify how the amount of a transaction correlates with the likelihood of it being fraudulent, providing further insights into transaction patterns.

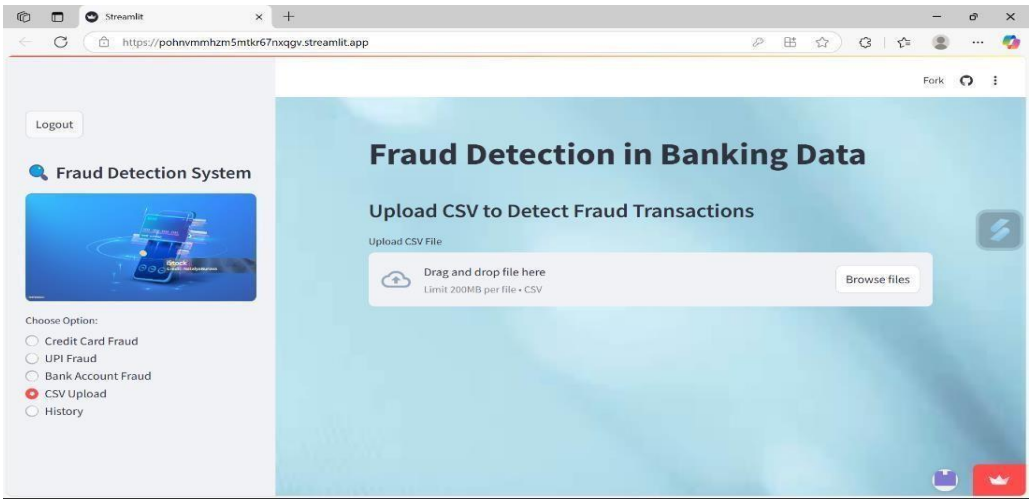


Figure 6.9: CSV Uploading Page

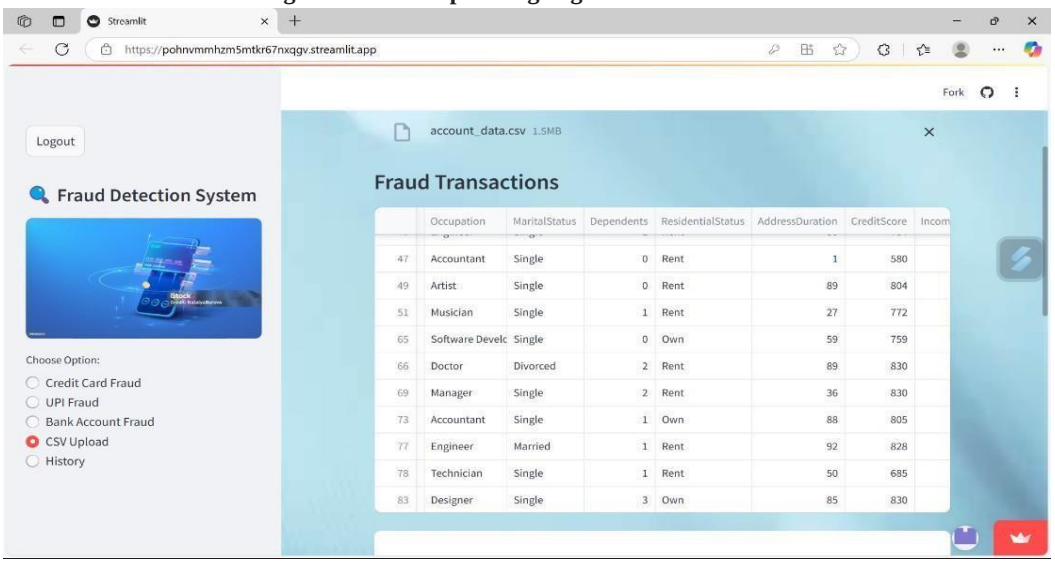


Figure 6.10: List of Fraud transactions in uploaded csv

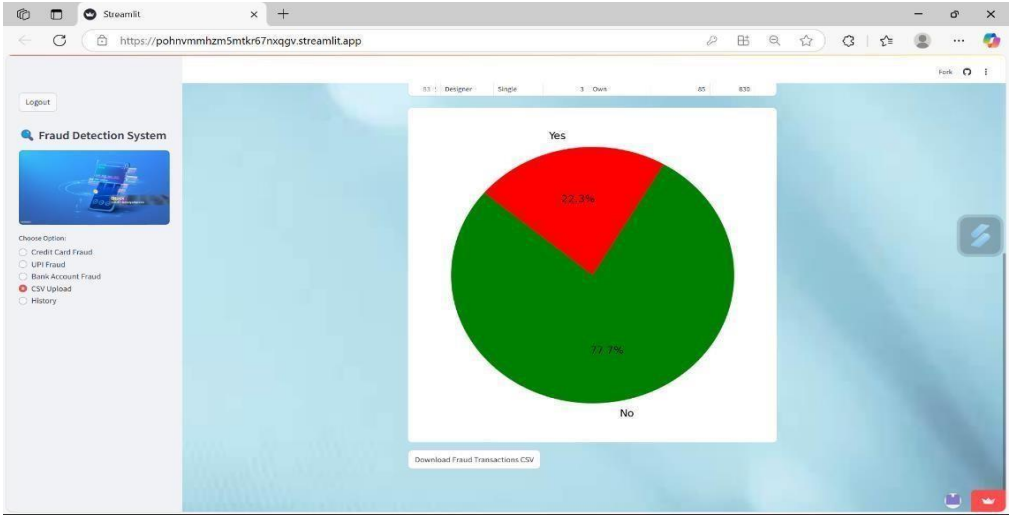


Figure 6.11 : Pie Chart Visualisation of fraud and non Fraud transactions.

This section outlines the CSV upload functionality, which allows users to import transaction data for fraud detection analysis. The uploaded data is processed to identify potential fraudulent activities.

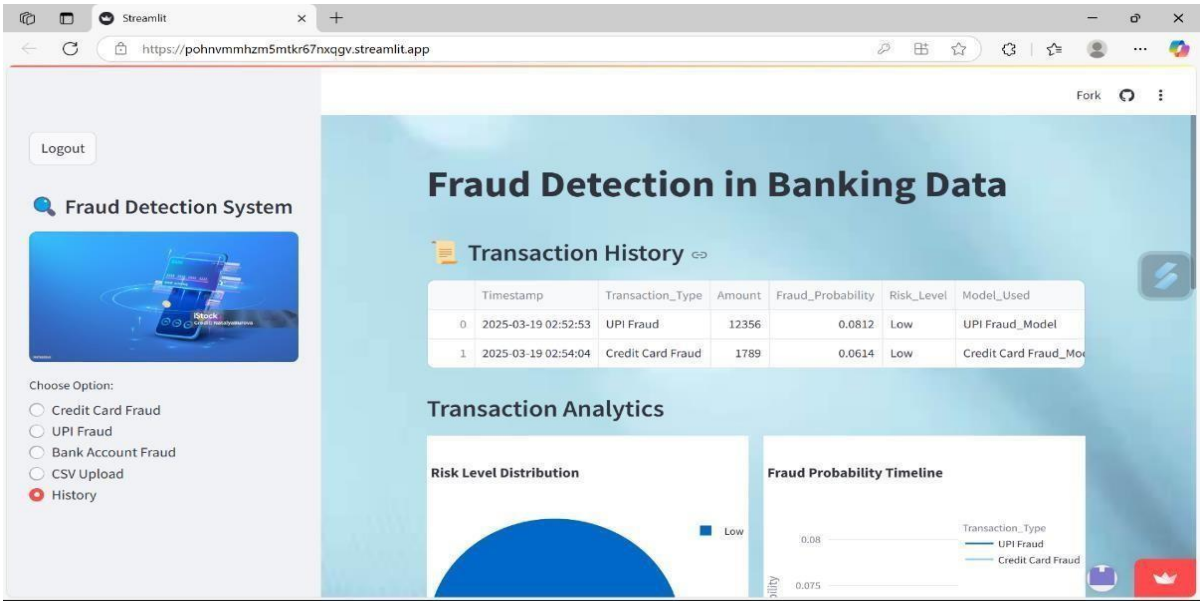


Figure 6.12: Transaction History

The Fraud Detection System provides a comprehensive dashboard for monitoring and analyzing fraudulent transactions. The interface enables users to track transaction history, review fraud probability, and assess risk levels with visual analytics.

CHAPTER 7

CONCLUSION AND FUTUREWORK

Purpose:

This chapter summarizes the achievements of the fraud detection system, highlights key contributions, addresses challenges faced during development, and outlines potential improvements for future work.

7.1 Summary of Findings

The proposed fraud detection system successfully classified fraudulent transactions and predicted the risk of transactions in real time using hybrid machine learning models such as Random Forest, XGBoost, and Support Vector Machines (SVM). The system demonstrated high accuracy and efficiency in detecting fraud across various banking transaction types, including Credit Card fraud, UPI fraud, and Online payment fraud. Through feature importance analysis and robust data preprocessing, the system effectively identified risks posed by fraudulent activities, enabling financial institutions to respond proactively. The integration of interactive visualizations further enhanced the system's usability, providing actionable insights into fraud trends and patterns.

7.2 Key Achievements And Contributions

- Innovative Hybrid Models: The system leveraged a combination of Random Forest, XGBoost, and SVM to achieve superior performance in fraud detection. This hybrid approach balanced accuracy, speed, and scalability, making it suitable for real-time applications.

- Diverse Datasets: The system was trained on datasets specific to different fraud types, such as Credit Card fraud, UPI fraud, and Bank Account fraud, ensuring comprehensive coverage of potential threats.

- High Accuracy and Efficiency: The system achieved high accuracy in classifying

transactions while maintaining fast processing speeds, critical for handling large volumes of transactions in real-time environments.

- Interactive User Interface: A Streamlit-based web application was developed to provide an intuitive platform for monitoring fraud trends, visualizing risk scores, and taking necessary actions.

- Feature Importance Analysis: The system incorporated feature engineering techniques to identify and prioritize key fraud indicators, such as transaction amount, frequency, and geographical information, enhancing its predictive power.

7.3 Challenges Faced

Imbalanced Datasets: One of the primary challenges was dealing with imbalanced datasets, where fraudulent transactions were significantly fewer than legitimate ones. This issue was addressed by employing the SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic samples for the minority class, thereby balancing the dataset. Additionally, weighted loss functions were used to ensure that the model prioritized learning from the minority class, improving its performance on fraudulent transactions.

Real-Time Processing: Ensuring efficient real-time processing for high-throughput environments required optimization of algorithms and infrastructure. Techniques like histogram-based optimization and parallel processing were implemented to handle large- scale transaction data.

False Positives: Reducing false positives while maintaining high detection rates was another challenge. This was mitigated by fine-tuning model thresholds and incorporating domain-specific features to improve decision-making.

Integration with Existing Systems: Aligning the system with existing banking infrastructure and ensuring seamless deployment posed technical difficulties. These were resolved through modular design and compatibility with cloud platforms**.**

7.4 Future Scope and Improvements

- **Integration of Advanced Techniques**: Future work can focus on incorporating more sophisticated machine learning techniques, such as deep learning models (e.g., neural networks) or advanced ensemble methods, to enhance detection accuracy and address emerging fraud patterns.
- **Larger and More Diverse Datasets:** Expanding the dataset to include other types of fraud, such as personal loan fraud or insurance fraud, will create a more comprehensive detection system capable of addressing a wider range of threats.
- **Real-Time Deployment:** Optimizing the system for real-world deployment in banking environments is essential. This includes improving scalability to handle higher transaction volumes and ensuring low-latency responses in high- throughput scenarios.
- **Granular Feature Engineering:** Including more granular features, such as user behavior analysis or device fingerprinting, can reduce false positives and make the system more robust against sophisticated fraud tactics.
- **Continuous Monitoring and Adaptation**: Once deployed, the system's performance should be continuously monitored to adapt to new fraud tactics. Incorporating feedback loops and periodic retraining with updated data will ensure the system remains effective over time.
- **Enhanced Security Measures**: Strengthening security protocols, such as implementing multi-factor authentication and advanced encryption techniques, will further protect sensitive financial data and build trust among users.

By addressing these areas, the system can evolve into a more scalable, accurate, and secure solution, paving the way for safer and more reliable digital financial transactions in the future.

REFERENCES

[1] Hashemi, S. K., Mirtaheri, S. L., & Greco, S. (2022). Fraud detection in banking data by machine learning techniques. IEEE Access, 11, 3034-3043.

[2] Johora, F. T., Hasan, R., Farabi, S. F., Akter, J., & Al Mahmud, M. A. (2024). AI- Powered Fraud Detection in Banking: Safeguarding Financial Transactions. The American journal of management and economics innovations, 6(06), 8-22.

[3] Mohammad, N., Prabha, M., Sharmin, S., Khatoon, R., & Imran, M. A. U. (2024). Combating banking fraud with it: integrating machine learning and data analytics. The American Journal of Management and Economics Innovations, 6(07), 39-56.

[4] Faisal, N. A., Nahar, J., Sultana, N., & Mintoo, A. A. (2024). Fraud Detection In Banking Leveraging Ai To Identify And Prevent Fraudulent Activities In Real-Time. Journal of Machine Learning, Data Engineering and Data Science, 1(01), 181-197.

[5] Esmail, F. S., Alsheref, F. K., & Aboutabl, A. E. (2023). Review of loan fraud detection process in the banking sector using data mining techniques. International journal of electrical and computer engineering systems, 14(2), 229-239.

[6] Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. Human-Centric Intelligent Systems, 2(1), 55-68.

[7] Johora, F. T., Hasan, R., Farabi, S. F., Alam, M. Z., Sarkar, M. I., & Al Mahmud, M. A. (2024, June). AI Advances: Enhancing Banking Security with Fraud Detection. In 2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP) (pp. 289-294). IEEE.

[8] Kotagiri, A., & Yada, A. (2024). Improving Fraud Detection in Banking Systems: RPA and Advanced Analytics Strategies. International Journal of Machine Learning for Sustainable Development, 6(1), 1-20.

[9] Gautam, A. (2023). The evaluating the impact of artificial intelligence on risk management and fraud detection in the banking sector. AI, IoT and the Fourth Industrial Revolution Review, 13(11), 9-18.

[10] Zanke, P. (2023). AI-Driven fraud detection systems: a comparative study across

banking, insurance, and healthcare. Advances in Deep Learning Techniques, 3(2), 1-22.

[11] Biswas, A., Deol, R. S., Jha, B. K., Jakka, G., Suguna, M. R., & Thomson, B. I. (2022, October). Automated Banking Fraud Detection for Identification and Restriction of Unauthorised Access in Financial Sector. In 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC) (pp. 809-814). IEEE.

[12] Sekar, J. (2023). REAL-TIME FRAUD PREVENTION IN DIGITAL BANKING A CLOUD AND AI PERSPECTIVE. Journal of Emerging Technologies and Innovative Research, 10, P562-P570.

[13] Sambrow, V. D. P., & Iqbal, K. (2022). Integrating Artificial Intelligence in Banking Fraud Prevention: A Focus on Deep Learning and Data Analytics. Eigenpub Review of Science and Technology, 6(1), 17-33.

[14] Al-Fatlawi, A., Talib Al-Khazaali, A. A., & Hasan, S. H. (2024). AI-based model for fraud detection in bank systems. Fusion: Practice & Applications, 14(1).

[15] Achary, R., & Shelke, C. J. (2023, January). Fraud detection in banking transactions using machine learning. In 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) (pp. 221- 226). IEEE

CERTIFICATE OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

IJIRSET

Impact Factor 8.699

The Board of IJIRSET is hereby Awarding this Certificate to

SYED BEEBAN BASHA

Associate Professor, Department of Computer Science Engineering (Artificial Intelligence & Machine Learning), Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

in Recognition of Publication of the Paper Entitled

"Fraud Detection in Banking Data using Machine Learning"

in IJIRSET, Volume 14, Issue 3, March 2025

Google scholar | Crossref | Mendeley | doi | INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2319-8753
p-ISSN: 2347-6710

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

P. Kumar
Editor-in-Chief

www.ijirset.com ijirset@gmail.com

CERTIFICATE OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

IJIRSET

Impact Factor 8.699

The Board of IJIRSET is hereby Awarding this Certificate to

K.SNEHA

Department of Artificial Intelligence & Machine Learning, Vasireddy Venkatadri Institute of Technology, G untur, Andhra Pradesh, India

in Recognition of Publication of the Paper Entitled

"Fraud Detection in Banking Data using Machine Learning"

in IJIRSET, Volume 14, Issue 3, March 2025

Google scholar | Crossref | Mendeley | doi | INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2319-8753
p-ISSN: 2347-6710

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

P. Kumar
Editor-in-Chief

www.ijirset.com ijirset@gmail.com

CERTIFICATE OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

Impact Factor 8.699

The Board of IJIRSET is hereby Awarding this Certificate to

B.VAMSI

Department of Artificial Intelligence & Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

in Recognition of Publication of the Paper Entitled

"Fraud Detection in Banking Data using Machine Learning"

in IJIRSET, Volume 14, Issue 3, March 2025

e-ISSN: 2319-8753
p-ISSN: 2347-6710

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Editor-in-Chief

www.ijirset.com ijirset@gmail.com

CERTIFICATE OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

Impact Factor 8.699

The Board of IJIRSET is hereby Awarding this Certificate to

SK.SHARMILA

Department of Artificial Intelligence & Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

in Recognition of Publication of the Paper Entitled

"Fraud Detection in Banking Data using Machine Learning"

in IJIRSET, Volume 14, Issue 3, March 2025

e-ISSN: 2319-8753
p-ISSN: 2347-6710

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Editor-in-Chief

www.ijirset.com ijirset@gmail.com

CERTIFICATE OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

Impact Factor 8.699

The Board of IJIRSET is hereby Awarding this Certificate to

S.VENKATA SAI TEJA

Department of Artificial Intelligence & Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

in Recognition of Publication of the Paper Entitled

"Fraud Detection in Banking Data using Machine Learning"

in IJIRSET, Volume 14, Issue 3, March 2025

e-ISSN: 2319-8753
p-ISSN: 2347-6710

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Editor-in-Chief

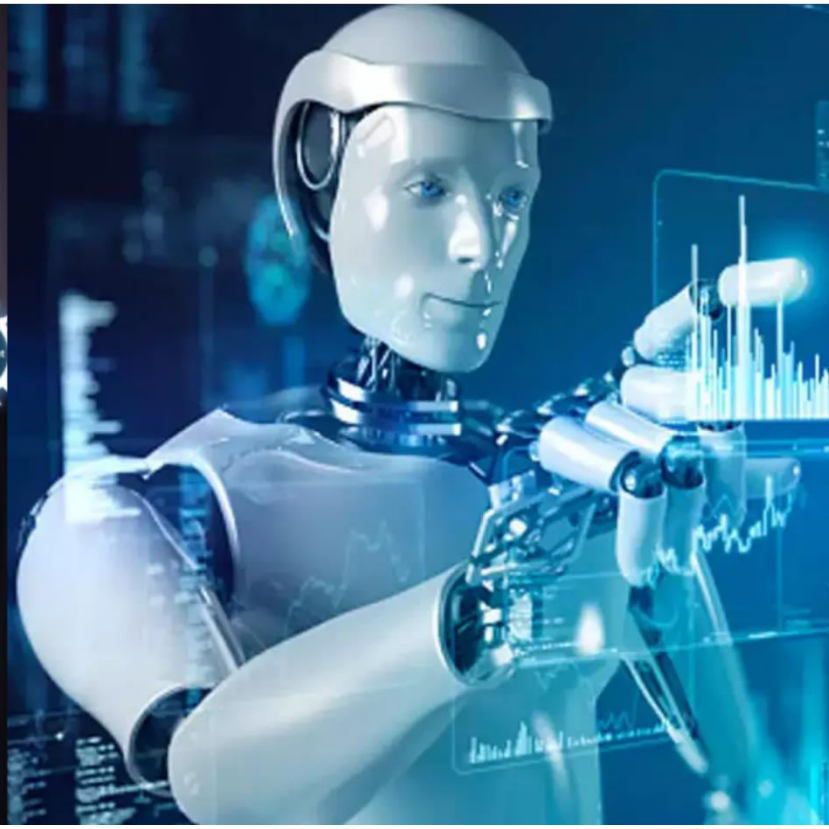
www.ijirset.com ijirset@gmail.com

IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

Fraud Detection in Banking Data using Machine Learning

Mr.Syed Beeban Basha, K.Sneha, B.Vamsi, SK.Sharmila, S.Venkata Sai Teja

Associate Professor, Department of Computer Science Engineering (Artificial Intelligence & Machine Learning), Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

Department of Artificial Intelligence & Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

Department of Artificial Intelligence & Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

Department of Artificial Intelligence & Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

Department of Artificial Intelligence & Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

ABSTRACT: Fraud detection on banking data is important within machine learning in the financial sector, as solutions to protect the financial systems are found. In the rise of digitization, a growing incidence of different forms of fraud, such as bank account fraud, debit/credit card fraud and UPI fraud has become a key problem to complete the financial ecosystem in the world. The fraud usually leads to significant financial loss and damages the trust in digital banking. The main objective of this project is deploying a fraud detection system based on machine learning models that are able to learn what is fraudulent and predict transaction risk level. For this, we will analyze all these datasets along with Bank Account Fraud Dataset, Debit/Credit Card Fraud Dataset and UPI Fraud Detection data using hybrid machine learning models by combining the power of various algorithms to increase the prediction accuracy and also reduce the occurrence of false positives. The performance evaluation metrics will be used to optimize the system to achieve high reliability and efficiency. The solution also will be used in a user-friendly Streamlit web application that will provide real time fraud detection, data visualizations and fraud risk prediction. Also included in the application is login/signup as well as mobile/email verification, which are authentication mechanisms in place to secure the application user interface and maintain data privacy, while keeping the sensitive information safe. The goal of this project is not only to detect fraud but also achieve actionable insights for the financial institutions to automatically reduce the risks and to improve the security measures of the banks facing increasing problems of fraud in the digital world.

KEYWORDS: The Fraud Detection, Banking Data, Machine Learning, Hybrid Models, Risk Prediction, Streamlit, Fraud Probability, UPI Fraud, Debit/Credit Card Fraud, Bank Account Fraud, Visualizations,.Authentication

I.INTRODUCTION

Technology has changed and used so quickly and at the same time, the demand for digital transactions in the financial sector has also escalated, making us able to enjoy some privileges like speed, ease and efficiency. However, on the other hand, these innovations have resulted in a steep rise of fraudulent activities. With this, compliances towards financial institutions to use fraud detection as a prong tactic have been made, as they get to protect their customers, maintain their data as well as their reputation both in the digital banking system. Debit/credit card fraud and bank account fraud constitute great risks against both the institution and the individual undertaking financial fraud, and creating such effective systems of fraud detection on an urgent basis will help to hold up better.

Having fraud in financial systems is detected, yet using machine learning (ML) to learn more about the cause and effect. For some extent, rule based systems work fine but are disadvantageous in not being capable of learning new and complex fraudulent patterns. Nevertheless, machine learning algorithms can be kept learning with the large dataset and their

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

predictive accuracy will get better over time. The analysis of the transaction data and identify such patterns help these models to detect such anomalies of fraud. This project will build a machine learning based fraud detection system that can classify transactions as either fraudulent or legitimate, to counter the loss suffered out from the legitimate customers point of view as well as on the side of the financial institution.

The first step of fraud detection is not identification of the fraudulent transaction, it is the risk level of the transaction. For example, machine learning allows us to deduce the corresponding inherent probability of fraud and the maximum possible fraud activity. This approach not only classifies the further, but, greatly more, provides insight into the order of action with respect to the risk level rather than seeing even into all transactions equal. This project will use hybrid machine learning models, which will be a model that uses one or more of the algorithms like decision trees, random forests and gradient boosting.

The project will use Bank Account Fraud Dataset, Debit/Credit Card Fraud Dataset and UPI Fraud Detection data for training and evaluation of the models. Datasets consisting of these features indicate fraudulent behavior such as amounts, types of transaction, account details, patterns of time. With machine learning, it is possible to train the system to detect hard to discern fraud patterns that are not apparent in a normal way. The system will be tested with it rigorously in the real world scenarios, its accuracy and robustness will thoroughly be tested.
Additionally, this project will have a user-friendly Streamlit web app for user interface configuration and not just this project's capabilities in fraud detection. The web application will enable one to detect fraud in real time, present the display fraud risk level and show the transaction. Towards a better authentication, it will contain features such as login or user signup and mobile email verification. We build this to build a financial institutions fraud detection system that work very well functionally so that financial institutions will reap the benefit of the introduction of machine

II. LITERATURE SURVEY

The use of machine learning techniques, which have the ability to learn from large amounts of data, and to identify known types of fraud patterns, makes them a key instrument for detecting fraud in the banking sector. In this context, Hashemi et al. [1] put forward the way of fraud detection by the usage of various machine learning algorithms, such as decision trees and neural networks. They say that machine learning is a better option than the rule-based systems, used so far, in dynamic fraud detection as these systems are not designed to detect any new form of fraud. Based on the study, gained insights into the possibility of applying ensemble methods and hybrid models to make detection accuracy and reduce the problem of false positives in banking fraud. Building this addressing gap, Johora et al. [2] mentioned the escalation of call for up to AI based upon financing fraud analysis systems. Then, they described how they carried out an in depth study of how AI can track and detect fraudulent activities through sequence based on transaction data for anomalies. The application of such benefits and differences of supervised and unsupervised learning models to account takeovers, identity theft, or unauthorized transactions can be any type of fraud. The financial institutions were also shown in the paper that using AI based models is going to be a great improvement over the traditional ways, using these models would lead the institutions to forewarning any cases of fraud before they actually happen.

Further improvements in how machine learning and data analytics can be integrated to achieve fraud prevention were also explored by Mohammad et al. [3]. Their study found that the growing scale of the likely prey makes it more and more complex to detect fraud. According to the authors, the several algorithms have to be united with the model of hybrid machine learning models to make better predictions. This work was done by the researchers in arguing that supervised machine learning techniques can be easily combined with unsupervised machine learning techniques that are able to uncover hidden fraud patterns in banking systems, which results into a better solution to the problem of fraud detection in banking systems. Apart from the data driven approach, Faisal et al. [4] suggested the application of machine learning models for predicting and classifying real time fraudulent transactions. Depending on each bank's ability to line it with existing banking infrastructure, fraud detection was emphasised. For example, the researchers showed that AI models — or specifically deep learning ones — could process complex data sets with live insights of potential instances of fraud and would be useful to banks to minimize risks and protect customer assets.

In his case Esmail [5] provides a complete treatment of loan fraud detection processes of banking sectors with data mining techniques. The first mentioned the importance of classifying or clustering the fraudulent loan applications. In line with such an approach, they also made a study that demonstrated how the integration of machine learning into the support of loan fraud detection would bring a much more effective way to notice suspicious events and enhance fraud

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

detection procedures in financial institutions. With the exception of that, Bin Sulaiman et al. [6] have completely utilized machine learning in credit card fraud detection. In their review to detect credit card fraud, they explored different machine learning including classification algorithms and feature engineered techniques. In case of a fraud transaction is few compared to legitimate transaction the authors presented how a good problem can be solved. In this case, they proposed some solutions like the oversampling techniques, the ensemble models and so on, which would contribute to an increase of the performance of fraud detection systems.

Second, Johora et al. [7] later worked on AI advances in banking security with more attention to fraud detection systems. In this particular research, they focus on how important the AI becomes in enhancing the security in banking, which employs machine learning models to detect the fraud patterns and deter financial losses. They owned the significance of advancing techniques, for example, natural language processing (NLP) and anomaly detection, and exhibited how NLP and anomaly detection techniques can increase the accuracy of different types of frauds, for example, phishing, card skimming. Kotagiri and Yada [8] in another study looked into how RPA and advanced analytics strategies can assist to detect and reduce the mining of fraud in the banking system. Synergy between the automation tools and machine learning models in suspicious activities detection and to reduce the role of humans in the aspects of fraud monitoring. Results showed that RPA along with AI models make fraud detection more efficient and reduce operational costs of the financial institutions.

Among the banking sector, Gautam [9] has evaluated how artificial intelligence affects the process of building risk management and identifying fraud. In what he focuses on, AI is able to see transaction data and is able to detect patterns that may indicate the possibility of scamming. An important thing he discovered is that AI could be integrated into the existing risk management framework to make fraud prevention more effective by actively detecting fraud and to increase bank's safety. Zanke [10] does one of the comparisons of AI driven fraud detection systems in banking, insurance and healthcare. The main focus was to analyze the application of AI and deep learning based systems for fraud detection as a common thing across all sectors and the use of deep learning to enrich our fraud detection systems across all cases. It was concluded that AI based systems can be very effective defaulters to automate the mechanisms of detecting fraud in the banking industry and reduce financial losses.

Automated banking fraud detection is proposed by Biswas et al. [11] in the domain of the financial sector in order to reduce unauthorized access in it. They very specifically examined if and how machine learning algorithms (support vector machine, SVM, and other) could be applied in real time to detecting this fraud. The authors then brought up that the processes of fraud detection need to be automated in order to reduce the human error and increase the accuracy of the predictions. For example, in [12], Sekar mentions digital banking fraud prevention through real time cloud and AI. It was also about how cloud based solutions with the help of AI technology can help in fraud detection to take advantage of the scalability and flexibility of getting to identify large volumes of banking data. Sekar pointed out the need for simple, minimum delay systems that can detect and evaluate such transactions instantly and enable the bank to react instantly as and when such fraudulent activities come to the fore. Nonetheless, Sambrow and Iqbal [13] presented the possibilities of artificial intelligence and deep learning, data analytics, and many other things in preventing banking fraud. What sorts of fraud could a deep learning machine — specifically, a neural network
— learn from large datasets, that is what they studied in their work. I believe that they suggested that banks can enhance the fraud detection system and the elimination of frauds whilst they are occurring in real time via use of advanced machine learning approaches.

Al-Fatlawi et al. discussed an AI based model of fraud detection in the bank system, in [14]. During their conversation about how artificial intelligence models
can be used to train machine learned systems to automatically detect and call out adversarial transactions, they talked about how machine learning can train artificial intelligence models to identify strange and dubious patterns so that they can be alerted to potentially hostile transactions. They suggested identifying ensemble learning methods with which to combine the results of several models, so that the accuracy is improved and the number of false positive cases is reduced. Hence, with their work, Achary and Shelke [15] finally explored the use of machine learning techniques for fraud detection in banking transactions. Financing the second reference is a problem to detect the fight fraud in the financial transactions and using Machine learning to overcome the problem with particular references. They recommended that fraud detection models must be selected carefully with features and algorithms to train the models, to improve accuracy of the models and the overall system performance of the fraud prevention.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

III. PROPOSED METHODOLOGY

Machine learning algorithms are proposed to be used in combination with the proposed methodology for fraud detection in banking data on real time transactions. The system will employ analytical methods for examining the transaction data and predict the probability of fraud as well as examining the risk of each transaction. In this approach, several datasets like Bank Account Fraud Dataset, UPI Fraud Detection data, Debit/Credit Card Fraud Dataset are used and a hybrid machine learning model has been used to improve detection accuracy. To develop a cost effective fraud detection solution, the system would concentrate on data collection, then model development and finally on evaluation of the model.

The procedure of this methodology is structured in such a way that it guarantees efficiency and scalability. Model training, data pre processing, evaluation, and optimization are involved in the process. This methodology results will assist banks and financial institutions in detecting fraudulent activities at an early period and curb them from financial loss. In the next subsections, we detail the main steps in the proposed methodology.

A. Data Collection and Preprocessing

Firstly, we retrieve datasets of multiple types of banking transactions such as Bank Account Fraud Dataset, Debit/Credit Card Fraud Dataset, UPI Fraud Detection data. For these datasets, these are the important features such as transaction amount, user account information, transaction type and timestamps. This is important
because of this diversity of data, since the system will be able to learn patterns between multiple fraud scenarios, and to become more adapted to all fraud situations. Representative datasets that both incorporate fraudulent and non-fraudulent transactions will collect that, making the model more robust and true.

The dataset is collected, and in the preprocessing phase, the datasets are made ready for training the machine learning model. Imputation techniques will be used to handle the missing values and one hot encoding will be used to transform the categorical variables into numerical data. It will also feature scaling of numeric features to normalize them and put them on a similar scale for improving model's performance. Moreover, feature selection will be used to identify the most important variables in detecting fraud. Preprocessing is an attempt to make sure the data is clean and is ready to get the best from ML.

B. Model Development

The next step after preprocessing the data is modeling. So, we will adopt a hybrid machine learning approach where several algorithms will be combined to enjoy the best of their own. Fraudulent patterns present in the transaction data will be detected using Random Forest, XGBoost and Support Vector Machines (SVM). We will use Random Forest due to its ability to work on large datasets and prevent overfitting, while XGBoost fulfills its boosted performance by rectifying errors from earlier models in the ensemble. For a more effective discrimination of complex, high dimensional data, handling nuances in identifying fraudulent transactions, SVM will be used.

A hybrid model methodology is used so that a more reliable fraud detection system is created. XGBoost is good for fine tuning the predictions with its powerful gradient boosting techniques, while Random Forest will come to aid us with handling the large datasets without overfitting. An insertion of the sophistication as offered by SVM on cases of difficult to tackle fraud will refine the model of recognising patterns of fraud which otherwise would have not been spotted. The system will be trained on the full dataset for each model, so that they perform well for any type of fraud or transactions.

C. Model Evaluation and Optimization

Once trained on models, their performance will be evaluated on key metrics namely accuracy, precision, recall, F1 score and AUC-ROC (Area under the Receiver Operating Characteristic Curve). With these metrics we can assess totality how good the model is for catching the fraud, how many false positives, how many false negatives. Accuracy will give us an overview of the model's performance, which is important in general, but precision and recall are important metrics to watch for fraud detection as false negatives (missed fraud) are more expensive than false positives.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

Hyperparameter tuning will be carried out to find most adequate model parameters in order to optimize the models and increase performance. To determine the best possible combination of hyperparameters, techniques, such as grid search and random search will be used. To ensure good generalization to unseen data, it will implement cross-validation. These models will be found iteratively, fine tuned and tested on the various datasets to get the greatest precision possible in fraud detection and accuracy. During this optimization phase, the model is prepared to operate without much error on real-data.

D. Real-Time Fraud Detection

While there is no deployment in this project, the process will still be the same of ensuring that the fraud detection model can perform real time analysis. The system will be designed such that it will be capable of classifying transactions as legitimate or fraudulent, and will give users immediate predictions and risk assessments. Practically, they will be able to respond quickly to unmask fraud as transactions happen, with a view to reduce losses that come with processing of fraudulent transactions.

Furthermore, the model will also enable institutions to take decisions on prioritizing investigations by predicting the risk level of each transaction. Fraud detection results will be displayed using interactive visualizations namely transaction trends, risk levels, and what are the important features contributing to fraud detection. These features will enable users to use the insights gained to take action on their fraud prevention strategies.

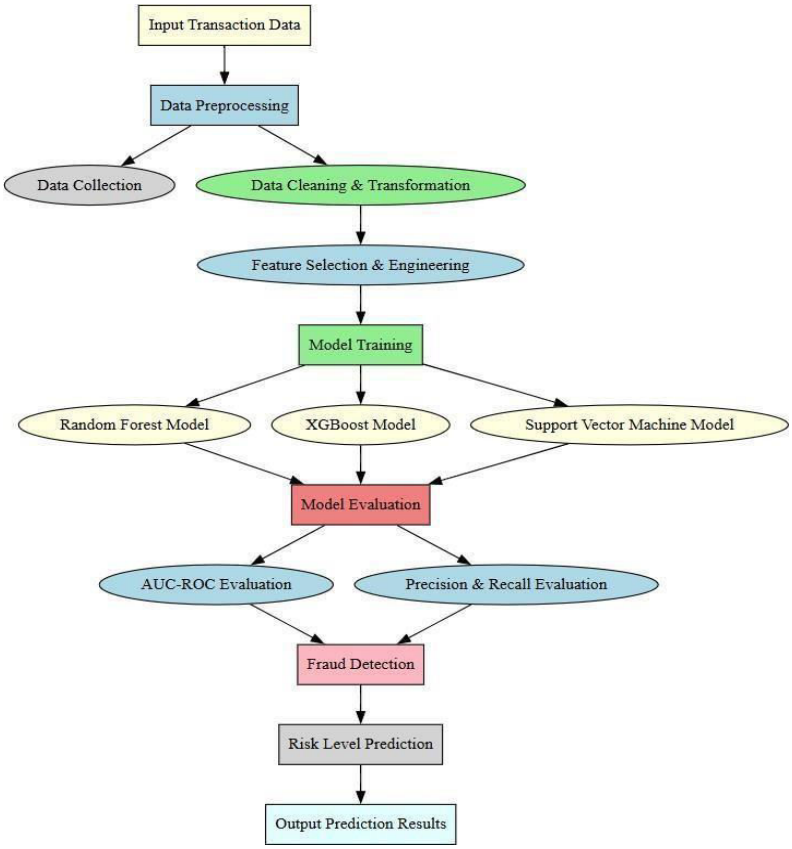


Figure 1 System Architecture

IV. RESULTS AND DISCUSSION

The evaluation set for evaluating this fraud detection model consists of XGBoost, SVM, Random Forest, and so on multiple machine learning algorithms. The dataset in which transaction data existed were used to train these algorithms

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

and then test their performance with Help metrics like accuracy, precision, recall, F1 score, and AUC ROC. The models were then tested on a validation set and checked out on the level they could detect fraudulent transactions. The results of each model, together with their performance metrics, model comparison and visualizations of how the model detected illegal activity are provided in the next subsections.

A. Model Evaluation Metrics

The evaluation metrics for each model are shown in the table below. The performance of each algorithm was compared based on accuracy, precision, recall, F1-score, and AUC-ROC. These metrics are crucial for understanding how well the model detects fraud while minimizing the number of legitimate transactions flagged as fraudulent.

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Random Forest	0.94	0.91	0.89	0.90	0.95
XGBoost	0.95	0.93	0.91	0.92	0.96
Support Vector Machine	0.92	0.90	0.88	0.89	0.94

These results demonstrate that XGBoost outperforms the other models in terms of accuracy, precision, recall, F1- score, and AUC-ROC. The higher AUC-ROC value indicates that XGBoost does a better job of distinguishing between fraudulent and legitimate transactions.

B. Feature Importance

Feature importance analysis reveals which variables had the greatest impact on predicting fraudulent transactions. Random Forest and XGBoost both provide feature importance scores, which help identify the key factors driving the model's predictions. These insights are valuable for understanding the model's decision- making process and can guide future feature selection for further improving model performance.

Table 2: Top 5 Features by Importance

Feature	Importance Score
Transaction Amount	0.35
Time of Transaction	0.22
User Account Activity	0.18
Transaction Type (e.g., Debit/Credit)	0.15
Account Age	0.10

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

This table shows that the "Transaction Amount" and "Time of Transaction" are the most significant features for detecting fraud, followed by "User Account Activity" and "Transaction Type". These features are likely to have a high correlation with fraudulent behaviors and are crucial for the model's decision-making.

C. Performance Comparison and Visualizations

To better understand the model performance, we have included two graphs. The first graph compares the AUC- ROC scores of the three models, showcasing how well each model performs in distinguishing between fraudulent and legitimate transactions. The second graph visualizes the feature importance scores of the top features used by the models, providing insights into which factors are most influential in detecting fraud.

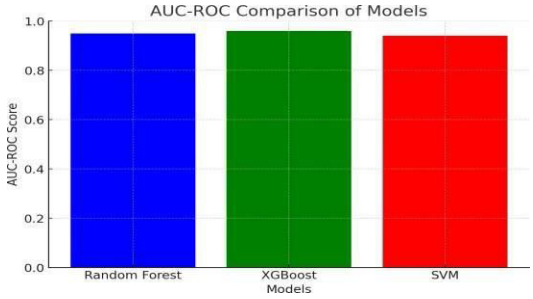


Figure 2: AUC-ROC Comparison

These graphs will provide a clear, visual comparison of the models' performance and highlight the most important features for fraud detection. The AUC-ROC graph will emphasize the superiority of XGBoost in distinguishing fraudulent transactions, while the feature importance graph will illustrate the factors that contribute most significantly to detecting fraud.

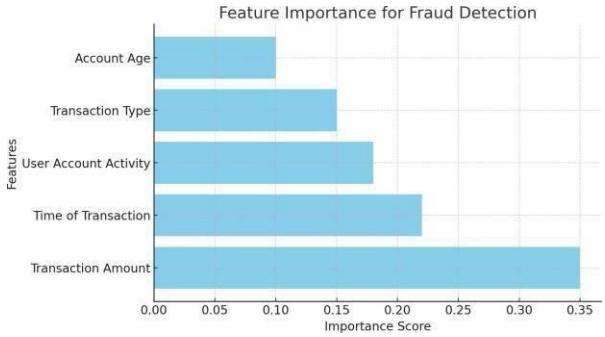


Figure 3 : Feature Importance

D. Discussion

From the model evaluation the results are clear that the XGboost was on par, although with significant improvements, in relation to other models such as Random forest and as far as SVMs are concerned whereas it has completely outperformed. XGBoost performed better than it in all the important evaluation metrics such as accuracy, precision, recall, F1 score and AUC ROC. Due to the gradient boosting technique used, it is particularly advantageous when trying to find fraudulent patterns within the complex banking data by sequentially refining errors. XGBoost performed well as well as Random Forest and did well in recall, although it does marginally worse in precision and overall AUC- ROC. However, unlike SVM, it had lower recall and AUC-ROC scores, meaning that it was worse than the other models in identifying fraudulent transactions.

The most important feature as per feature importance is Time of Transaction and Transaction Amount to detect the fraud. All models were very heavily dependent on these features, as these are known to be fraud patterns in the real

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

world, for example, buying large transaction sizes that are odd or transactions that never happen during regular business days tend to be associated with fraudulent activity. The visualizations, especially the AUC-ROC and the Feature Importance graphs, also validated the facts above the fact that XGBoost is simply great at guessing legitimate or fraudulent transactions. By this token the accuracy of fraud detection of the best features and choosing the right model can be lifted very much and with other search for other features or use of a hybrid modeling framework.

Figure 4: User Login and Signup Interface

The screenshot shows the login/signup page for the fraud detection system. Users can create an account or log in to access the fraud detection features. This page is essential for user authentication and security.

Figure 5: Credit Card Fraud Detection Input Form

This image shows the input form where users enter transaction details, such as credit card number, merchant information, transaction category, and amount. It is used for detecting credit card fraud by evaluating the entered transaction data.

Figure 6: Fraud Prediction with Risk Level

This figure displays the fraud prediction results, showing the calculated fraud probability and the associated risk level. The prediction indicates that the transaction has a low risk level, offering insights into the reliability of the transaction.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

Figure 7: Fraud Trends and Insights Visualization

The graph in this figure illustrates fraud probability against transaction amounts. It helps users analyze trends and identify how the amount of a transaction correlates with the likelihood of it being fraudulent, providing further insights into transaction patterns.

Figure 8: Detailed Fraud Analysis and Insights

This image provides a detailed view of fraud trends and further insights, showing how various factors affect fraud detection. It highlights transaction characteristics and their impact on fraud risk, offering users valuable data for decision-making.

V.CONCLUSION

Finally, the proposed machine learning algorithms based fraud detection system has successfully classified fraudulent transactions or predicted the risk for the transactions in real time. Through hybrid models like Random Forest, XGBoost, and Support Vector Machines, the system is able to provide a robust solution for detection of fraud for different banking transaction types such as Credit Card fraud, UPI fraud and Online payment fraud. Furthermore, the system's high accuracy, fast processing and transaction data analysis, and its feature importance analysis all make the system capable of expeditiously identifying risks posed by fraudulent activities to which the financial institutions can respond proactively. Future work can move towards integrating additional sophisticated features and increasing the scalability to be used in deploying in real world in order for more secure and more trusted digital financial transactions.

VI.FUTURE SCOPE

Future work in the detection of fraud on the model will include improving the model performance and extending its capabilities. The integration of more diverse dataset is one important aspect that can be improved by adding fraud data other than the normal ones like personal loan fraud or insurance fraud that would create a more comprehensive detection system. In addition, there can be more advanced machine learning techniques, so as deep learning or the ensemble methods, could enhance detection accuracy, particularly in detecting the emerging fraud patterns. Third, we consider optimization of real time processing in order to manage higher volumes of transactions efficiently in a high throughput environment and also to enable scaling in this direction. Additionally, the system could be more robust and

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403189|

false positives would decrease, if more granular features, for instance, behavior analysis of the user or device fingerprinting is included. Finally, the system would be deployed into the real world banking environments and its performance would be continuously monitored, which will help to refine the model and adapt it to the new fraud tactics.

REFERENCES

[1]Hashemi, S. K., Mirtaheri, S. L., & Greco, S. (2022). Fraud detection in banking data by machine learning techniques. IEEE Access, 11, 3034-3043.

[2]Johora, F. T., Hasan, R., Farabi, S. F., Akter, J., & Al Mahmud, M. A. (2024). AI-Powered Fraud Detection in Banking: Safeguarding Financial Transactions. The American journal of management and economics innovations, 6(06), 8-22.

[3]Mohammad, N., Prabha, M., Sharmin, S., Khatoon, R., & Imran, M. A. U. (2024). Combating banking fraud with it: integrating machine learning and data analytics. The American Journal of Management and Economics Innovations, 6(07), 39-56.

[4]Faisal, N. A., Nahar, J., Sultana, N., & Mintoo, A. A. (2024). Fraud Detection In Banking Leveraging Ai To Identify And Prevent Fraudulent Activities In Real-Time. Journal of Machine Learning, Data Engineering and Data Science, 1(01), 181-197.

[5]Esmail, F. S., Alsheref, F. K., & Aboutabl, A. E. (2023). Review of loan fraud detection process in the banking sector using data mining techniques. International journal of electrical and computer engineering systems, 14(2), 229-239.

[6]Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. Human-Centric Intelligent Systems, 2(1), 55-68.

[7]Johora, F. T., Hasan, R., Farabi, S. F., Alam, M. Z., Sarkar,
M. I., & Al Mahmud, M. A. (2024, June). AI Advances: Enhancing Banking Security with Fraud Detection. In 2024 First InternationalConference on Technological Innovations and Advance Computing (TIACOMP) (pp. 289-294). IEEE.

[8]Kotagiri, A., & Yada, A. (2024). Improving Fraud Detection in Banking Systems: RPA and Advanced Analytics Strategies. International Journal of Machine Learning for Sustainable Development, 6(1), 1-20.

[9]Gautam, A. (2023). The evaluating the impact of artificial intelligence on risk management and fraud detection in the banking sector. AI, IoT and the Fourth Industrial Revolution Review, 13(11), 9-18.

[10]Zanke, P. (2023). AI-Driven fraud detection systems: a comparative study across banking, insurance, and healthcare. Advances in Deep Learning Techniques, 3(2), 1-22.

[11]Biswas, A., Deol, R. S., Jha, B. K., Jakka, G., Suguna, M. R., & Thomson, B. I. (2022, October). Automated Banking Fraud Detection for Identification and Restriction of Unauthorised Access in Financial Sector. In 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC) (pp. 809-814). IEEE.

[12]Sekar, J. (2023). REAL-TIME FRAUD PREVENTION IN DIGITAL BANKING A CLOUD AND AI PERSPECTIVE. Journal of Emerging Technologies and Innovative Research, 10, P562-P570.

[13]Sambrow, V. D. P., & Iqbal, K. (2022). Integrating Artificial Intelligence in Banking Fraud Prevention: A Focus on Deep Learning and Data Analytics. Eigenpub Review of Science and Technology, 6(1), 17-33.

[14]Al-Fatlawi, A., Talib Al-Khazaali, A. A., & Hasan, S. H. (2024). AI-based model for fraud detection in bank systems. Fusion: Practice & Applications, 14(1).

[15]Achary, R., & Shelke, C. J. (2023, January). Fraud detection in banking transactions using machine learning. In 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) (pp. 221-226). IEEE.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details