A MINI PROJECT REPORT

On

Fraud Account Detection System

Submitted in partial fulfillment of the requirement of
University of Mumbai for the Course

AI For Finance and Banking Applications
In
CSE-AIML Department (VIII SEM)

Submitted By
Kapil Surve (21106018)
Mihir Bhave (21106051)
Tanisha Chitnis (21106003)
Harshal Deshmukh (22206008)

Subject Incharge
Prof.Priyanka Patil

Department Of CSE-AIML
A. P. SHAH INSTITUTE OF TECHNOLOGY THANE – 400 615
UNIVERSITY OF MUMBAI

Academic Year 2024 – 2025

Department of CSE-AIML
A. P. Shah Institute of Technology Thane – 400 615

CERTIFICATE

This is to certify that the requirements for the project report entitled 'Fraud Account Detection System' have been successfully completed by the following students:

Name	**Roll No.**
Kapil Surve	53
Mihir Bhave	06
Tanisha Chitnis	11
Harshal Deshmukh	13

in partial fulfillment of the course AI For Finance and Banking Applications in CSE-AIML(VIII SEM) of Mumbai University in the Department of CSE-AIML A. P. Shah Institute of Technology, Thane during the Academic Year 2024 – 2025

Prof.Priyanka Patil

Department of CSE-AIML
A. P. Shah Institute of Technology Thane – 400 615

PROJECT APPROVAL

This project entitled “Fraud Account Detection System” by Kapil Surve, Mihir Bhave, Tanisha Chitnis and Harshal Deshmukh are approved for the course AIFB in CSE-AIML(VIII Sem) of Mumbai University in the Department of CSE-AIML.

Subject Incharge:

_____

Date:

Place: Thane

Department of CSE-AIML
A. P. Shah Institute of Technology Thane - 400 615

DECLARATION

We declare that this written submission for **AI For Finance and Banking Applications** mini project entitled “Fraud Account Detection System” represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any ideas / data / fact / source in our submission. We understand that any violation of the above will cause disciplinary action by the institute and also evoke penal action from the sources which have not been properly cited or from whom prior permission has not been taken when needed.

Project Group Members:

Kapil Surve

_____

Mihir Bhave

_____

Tanisha Chitnis

_____

Harshal Deshmukh

_____

Date:

Place:

Table of Contents(SAMPLE)

Abstract

This project aims to implement intelligent fraud detection in the banking domain using machine learning. Two major types of frauds are targeted: Account Fraud and Transaction Fraud. The system utilizes supervised learning models, including Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) to predict fraudulent accounts and suspicious transaction behavior. A comparative analysis is done based on metrics like accuracy, precision, and recall. Datasets from Kaggle for both account and transaction fraud detection are used, enabling real-time fraud risk analysis and enhancing the decision-making process in banking systems.

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Fundamentals

Fraud has become a major threat in the banking sector, where attackers exploit system vulnerabilities for financial gain. Financial frauds are broadly categorized into account fraud and transaction fraud. Account fraud includes the unauthorized creation or misuse of a user's financial account, while transaction fraud involves the execution of unauthorized transactions such as fund transfers or high-volume withdrawals.

Traditionally, banks have relied on rule-based detection systems which lack flexibility and adaptability. These systems are often ineffective against new or sophisticated fraud patterns. With the advent of artificial intelligence (AI) and machine learning (ML), fraud detection systems have evolved significantly. Machine learning algorithms can analyze historical transaction data, detect hidden patterns, and classify whether a transaction is genuine or fraudulent with high accuracy.

Our project leverages supervised learning models such as Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) to detect fraudulent activity. The primary focus is on real-time detection, model evaluation, and accuracy improvement using multiple classifiers and appropriate performance metrics.

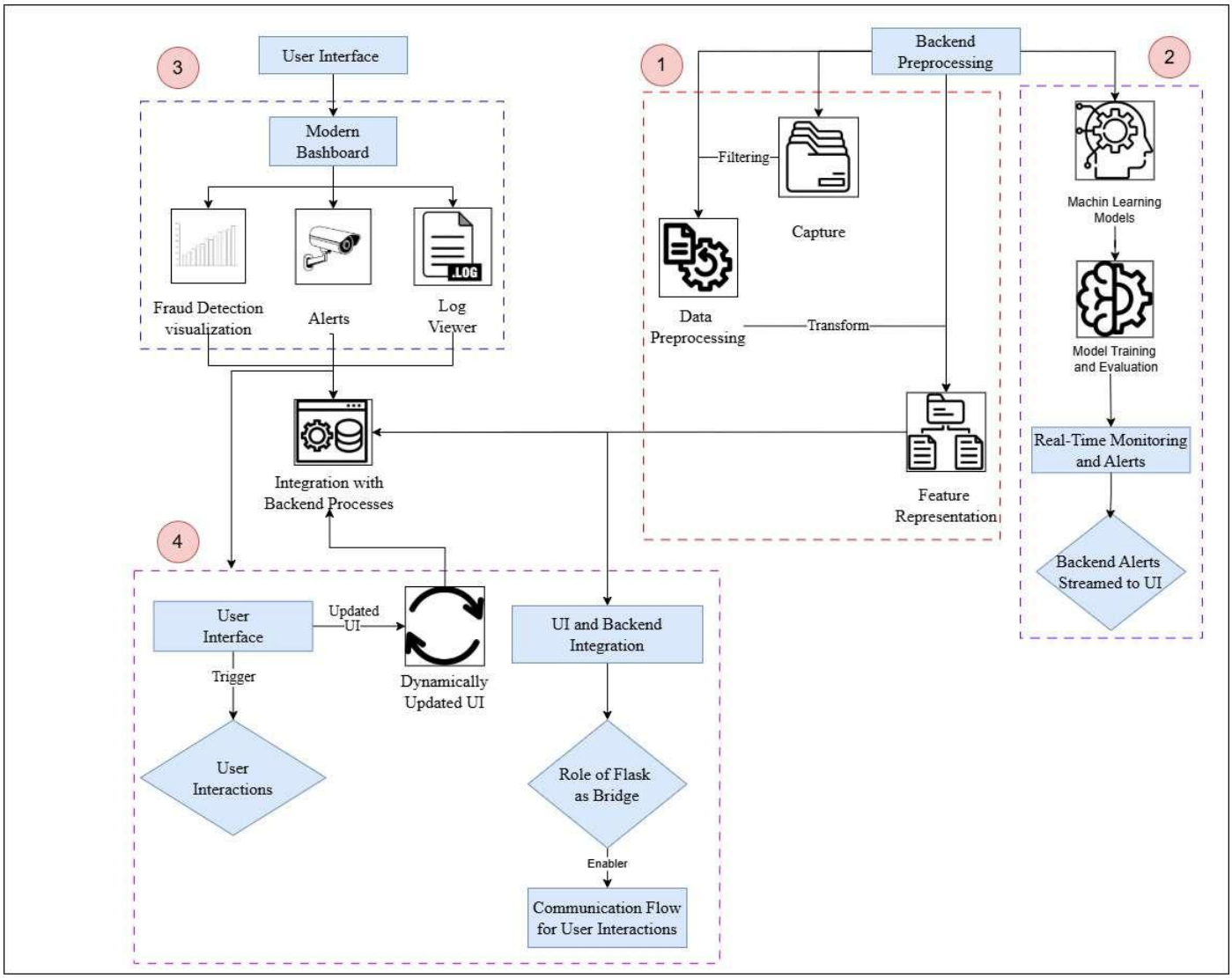

Fig 1.1: Block Diagram of the application

The Figure 1.1 shows the method to add figure in the text. The description about the figure is to be written in one or two lines or more as per the requirements.

1.2 Objectives

The main objectives of this project are as follows:

1. *To identify fraudulent accounts and detect suspicious transactions in real-time using machine learning.*
2. To develop an AI-based system that outperforms rule-based fraud detection models.
3. To evaluate various ML algorithms (Logistic Regression, Random Forest, KNN, ANN) and compare their performance.
4. To use appropriate evaluation metrics such as accuracy, precision, recall, and F1-score for model comparison.

5. To implement a practical solution that can be integrated with existing banking systems for fraud risk management.

1.3 *Scope*

This project focuses on detecting two types of frauds in the financial sector: account fraud and transaction fraud. The system is designed to function in both offline and real-time environments, ensuring that potentially harmful activities are flagged before they cause damage. Using open-source datasets and machine learning libraries, we demonstrate how different classification models can be trained, tested, and deployed. This solution can be scaled and adapted to real-world banking systems with minimal customization.

1.4 Organization of the Report

This report is structured as follows:

- Chapter 1 provides a detailed introduction to fraud detection and outlines the objectives and scope of the project.
- Chapter 2 presents a comprehensive literature survey of related works and techniques.
- Chapter 3 explains the proposed methodology, implementation steps, models used, and dataset details.
- Chapter 4 describes the system inputs and outputs, along with performance evaluation and screenshots.
- Chapter 5 summarizes the findings and outlines future enhancements and practical applications.

Chapter 2

Literature Survey

2.1 Introduction

In the era of digital banking, financial institutions are increasingly threatened by sophisticated fraud schemes. Traditional rule-based systems often lack the adaptability to detect emerging and evolving fraudulent behaviors. As a result, machine learning and artificial intelligence have emerged as powerful tools for building robust and scalable fraud detection systems. This chapter reviews the existing research work and highlights recent developments in the domain of fraud detection, focusing on both account-based and transaction-based frauds.

2.2 Literature Review

3 Vashistha & Tiwari (2024) – "A Robust Framework for Bank Account Fraud Detection"
This paper proposes a hyper-ensemble model integrating both supervised learning and anomaly detection to identify fraud in bank account activities. The authors used datasets from NeurIPS 2022 and demonstrated how hybrid models outperform single classifiers.
Advantages: Improved accuracy and fraud classification.
Disadvantages: Computationally expensive and needs large training data.

4 Sivanantham et al. (2021) – "Hybrid Models for Fraud Detection in Digital Banking"
This study applies machine learning techniques, including decision trees, SVMs, and neural networks, for detecting account fraud. It emphasizes the importance of feature selection in fraud prediction.
Advantages: Handles nonlinear patterns effectively.
Disadvantages: Requires frequent retraining for real-time applications.

5 Talukder et al. (2024) – "Multi-stage Ensemble Learning for Transaction Fraud Detection"
A comprehensive study that combines multiple learners in sequential layers to enhance fraud detection in bank transactions. The model showed high precision in detecting rare fraudulent cases.
Advantages: Reduced false positives.
Disadvantages: Slightly slower due to multi-layered processing.

6 Carcillo et al. (2019) – "Combining Unsupervised and Supervised Learning in Transactional Data"
This work introduces a hybrid approach that first clusters transactions and then classifies them using supervised learning. It is effective in handling imbalanced datasets.
Advantages: Suitable for detecting unknown fraud types.
Disadvantages: May suffer from overfitting without proper validation.

7 Sahin & Duman (2020) – "Detecting Credit Card Fraud Using Random Forest and Logistic Regression"
This paper focuses on transaction-level fraud using classical models. It compares Random Forest and Logistic Regression for detecting credit card fraud.
Advantages: Easy to implement and interpret.
Disadvantages: Lower performance on highly imbalanced data.

7.1 Literature Summary

SN	Techniques	Author & Year of Publication	Advantages and Disadvantages
1.	Hyper-Ensemble, Anomaly Detection	Vashistha & Tiwari (2024)	High accuracy, hybrid flexibility High computational cost
2.	Hybrid ML Models	Sivanantham et al. (2021)	Detects non-linear patterns, generalizable Frequent retraining needed
3.	Multi-stage Ensemble Learning	Talukder et al. (2024)	Low false positives, scalable Slower due to layer complexity
4.	Unsupervised + Supervised Combination	Carcillo et al. (2019)	Detects unknown fraud, handles imbalance Risk of overfitting
5.	RF and Logistic Regression	Sahin & Duman (2020)	Simple, interpretable, widely adopted Poor performance on imbalanced datasets

Table 2.1 Literature survey summary

Chapter 3

Implementation Details

3.1 Overview

In the current digital banking ecosystem, fraudulent activities are becoming more sophisticated, requiring advanced and adaptive fraud detection mechanisms. Traditional fraud detection relies on rule-based systems, which are predefined by experts and typically static in nature. These systems cannot dynamically adapt to evolving fraud strategies, often resulting in missed detections or false alarms.

To overcome these limitations, this project proposes the use of AI-driven machine learning models that learn from historical data to detect fraudulent behaviors in real-time. Our approach focuses on two primary fraud categories:

Account-level fraud: Activities related to opening or accessing bank accounts for illegitimate purposes.

Transaction-level fraud: Suspicious financial transactions involving abnormal patterns or high-value transfers.

3.1.1 Existing Methodology and Systems

Conventional fraud detection systems follow a rule-based mechanism where predefined conditions are used to flag fraudulent activities. For example, a rule might flag a transaction if it exceeds a specific amount or if it originates from an unusual location. Limitations of existing systems include:

- Inability to detect previously unseen fraud patterns.
- High false-positive rate leading to inconvenience to genuine users.
- Static logic that requires manual updates by fraud analysts.
- These issues highlight the need for a more dynamic and learning-based fraud detection framework.

3.1.2 Proposed Methodology and System

Our system adopts a machine learning-based approach using supervised classification models trained on historical data labeled as “fraudulent” or “non-fraudulent.” The key features of our system include:

Data preprocessing for missing values and scaling.
Feature selection and engineering.
Training of various classifiers: Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN).

Evaluation using metrics such as accuracy, precision, recall, and F1-score.

We propose a hybrid approach that combines the interpretability of classical models with the learning capacity of neural networks to deliver high detection accuracy with lower false positives.

3.2 Implementation Details

3.2.1 Methodology

The implementation of the fraud detection system includes the following steps:

1. Data Loading: Datasets for account and transaction fraud are loaded from Kaggle.
2. Preprocessing:
    - Handling missing values.

    - Label encoding of categorical features.
    - Feature scaling using StandardScaler.
3. Model Development:
    - Logistic Regression (for baseline accuracy and interpretability).
    - Random Forest (for robust feature importance and handling non-linearity).
    - K-Nearest Neighbors (for pattern recognition based on distance).
    - ANN (for deep learning-based classification).
4. Model Evaluation:
    - Data split into training and test sets (80:20 ratio).
    - Accuracy, precision, recall, and F1-score are computed for each model.
5. Visualization:
    - Confusion matrices and ROC curves for performance interpretation.

3.2.2 Details of packages, data set

Programming Language: Python

Tools & Libraries:

- pandas, numpy — for data handling
- scikit-learn — for model training and evaluation
- tensorflow/keras — for ANN implementation
- matplotlib, seaborn — for visualizations

Datasets Used:

Bank Account Fraud Dataset (NeurIPS 2022)

Source:

https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

Bank Transaction Fraud Dataset

Source:

https://www.kaggle.com/code/marusagar/bank-transaction-fraud-detection-accuracy-95/notebook

Chapter 4

Project Inputs and Outputs

4.1 Inputs Details

The fraud detection system relies on structured datasets representing real-world account and transaction data. These datasets contain a variety of features relevant to customer demographics, account behavior, and transaction patterns. Two primary datasets were used:

1. Bank Account Fraud Dataset
    - Features: age, income, number of dependents, employment type, account balance, transaction type, etc.
    - Label: is_fraud (0 = genuine, 1 = fraud)
2. Bank Transaction Fraud Dataset
    - Features: transaction time, amount, location, account ID, previous transaction score, number of attempts, etc.
    - Label: fraud (1 = fraudulent, 0 = legitimate)

The data was preprocessed to handle missing values, encode categorical variables, and scale numerical features using StandardScaler. The cleaned dataset was split into 80% training and 20% testing data.

4.2 Evaluation Parameters Details

To evaluate the performance of the models implemented in the system, the following metrics were used:

Accuracy: Proportion of total predictions that were correct.
Precision: Percentage of correctly predicted frauds out of all predicted frauds (low false positives).
Recall: Percentage of actual frauds that were correctly identified (low false negatives).
F1-Score: Harmonic mean of precision and recall — provides a balance between the two.
Confusion Matrix: A 2×2 matrix showing TP, FP, TN, and FN values for visual inspection predictions.

Each model was trained using the same training dataset and evaluated on the same test dataset for consistency and fairness.

4.3 Output Details and Screenshots

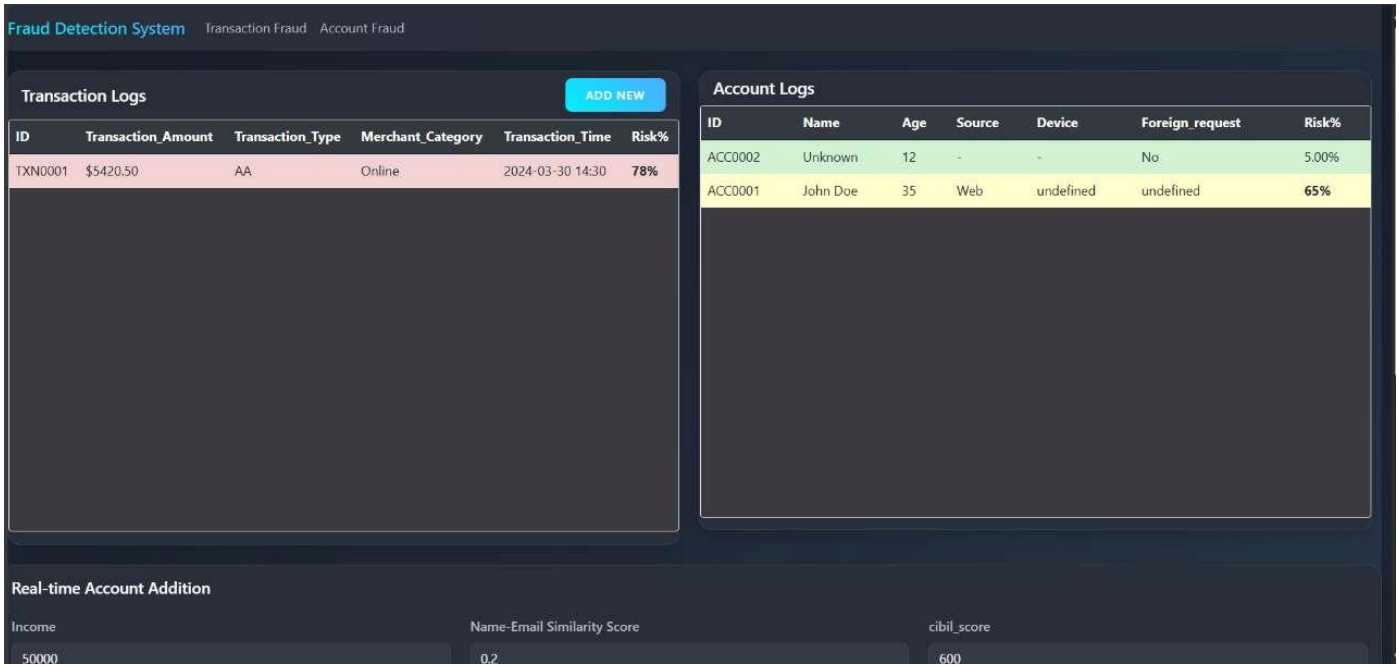

Fig 4.1: Main UI

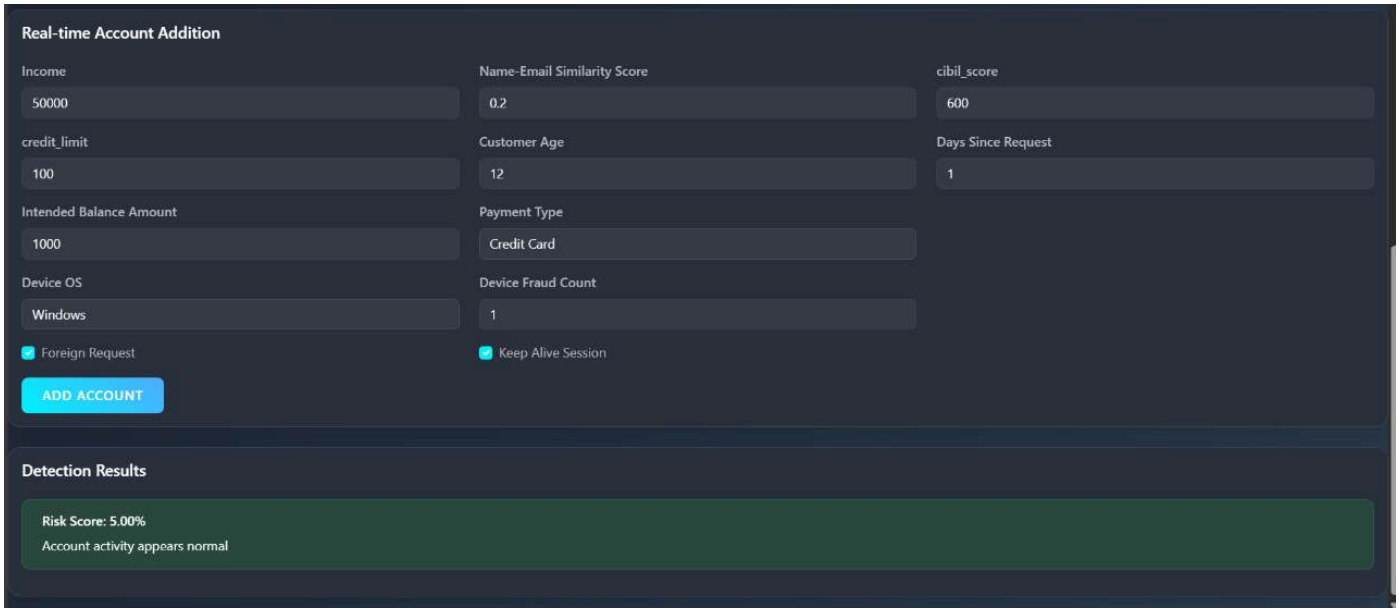

Fig 4.2: Making New Entries

Chapter 5

Summary and Future Scope

5.1 Summary

This project successfully implemented a machine learning-based Fraud Account Detection System that identifies suspicious account activities and fraudulent transactions with high accuracy. Traditional rule-based detection systems are often inadequate in today's fast-evolving digital finance environment due to their static nature and limited adaptability.

To address this, we developed a dynamic and intelligent detection framework using multiple supervised learning algorithms — Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN). These models were trained and tested on publicly available datasets representing real-world banking fraud scenarios.

Among the implemented models, KNN and ANN achieved the highest performance with 98% accuracy, while Random Forest also provided strong results with better interpretability. The results demonstrated that machine learning can significantly enhance fraud detection capabilities, reduce false positives, and enable proactive decision-making in financial systems.

Key contributions of the project:

Developed a hybrid AI-driven solution for both account and transaction fraud.

Compared and evaluated the performance of four different ML models.

Employed real-time evaluation metrics such as precision, recall, and F1-score.

Demonstrated feasibility using real datasets and scalable tools.

5.2 Future Scope

While the current implementation provides a strong baseline for fraud detection, several enhancements can be pursued in future work:

Integration of the system into real-time banking applications using REST APIs and cloud deployment.

Adoption of advanced deep learning architectures such as Recurrent Neural Networks (RNNs) and Transformer models for sequential transaction analysis.

Use of unsupervised and semi-supervised learning for anomaly detection in the absence of labeled data.

Application of natural language processing (NLP) to include user reviews, emails, and support tickets as additional signals of fraud.

Implementation of adaptive learning models that evolve with new transaction patterns and fraud strategies.

Deployment of the model on edge devices or mobile platforms to ensure fraud detection at the point of transaction.

By continuing to develop this project, financial institutions can not only enhance their fraud prevention strategies but also build customer trust through proactive security measures

References

1. Vashistha, A., & Tiwari, A. (2024). "A Robust Framework for Bank Account Fraud Detection Using Hyper-Ensemble Learning." SN Computer Science.
2. Sivanantham, S., Balasubramanian, K., & Arumugam, D. (2021). "Hybrid Models for Fraud Detection in Digital Banking." In International Conference on Financial Security & Smart Technologies (Springer).
3. Talukder, M. A., Rokonuzzaman, M., & Sultana, T. (2024). "A Multi-Stage Ensemble Learning Model for Transaction Fraud Detection." Journal of Big Data.
4. Carcillo, F., Dal Pozzolo, A., Le Borgne, Y. A., Caelen, O., Bontempi, G. (2019). "Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection." Information Sciences.
5. Sahin, Y., & Duman, E. (2020). "Detecting Credit Card Fraud by ANN and Logistic Regression." Expert Systems with Applications.
6. Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004). "Survey of Fraud Detection Techniques." IEEE International Conference on Networking, Sensing and Control.
7. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). "A Comprehensive Survey of Data Mining-based Fraud Detection Research." Artificial Intelligence Review.
8. Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature." Decision Support Systems.
9. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). "Isolation Forest." IEEE International Conference on Data Mining.
10. Pozzolo, A. D., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). "Calibrating Probability with Undersampling for Unbalanced Classification." IEEE Symposium on Computational Intelligence.
11. Cortes, C., & Vapnik, V. (1995). "Support-Vector Networks." Machine Learning.
12. Breiman, L. (2001). "Random Forests." Machine Learning.
13. LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep Learning." Nature, 521(7553), 436–444.

14. Kaggle (2022). "Bank Account Fraud Dataset (NeurIPS 2022)." https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

15. Kaggle (2023). "Bank Transaction Fraud Detection Notebook." https://www.kaggle.com/code/marusagar/bank-transaction-fraud-detection-accuracy-95