National College of Ireland

Stacked Machine Learning for Explainable Financial Fraud Detection Using LIME and LLM

MSc Research Project
M.Sc. in Data Analytics

Varun Sai Yandapalli
Student ID: 23325836

School of Computing
National College of Ireland

Supervisor: Mr. Eric Gyamfi

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Varun Sai Yandapalli
Student ID:	23325836
Programme:	MSc. In Data Anlalytics **Year:** 2024-2025
Module:	MSc. Research Project
Supervisor:	Eric Gyamfi
Submission Due Date:	15th September 2025
Project Title:	Stacked Machine Learning for Explainable Financial Fraud Detection Using LIME and LLM

Word Count: **7050** **Page Count**: 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Yandapalli Varun Sai

Date: 15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Stacked Machine Learning for Explainable Financial Fraud Detection Using LIME and LLM

Varun Sai Yandapalli
Student ID: 23325836

Abstract

Financial fraud detection involves identifying and preventing unauthorized banking activities that can cause significant financial losses to institutions and customers. Traditional models, such as rule-based systems or single classifiers, struggle with evolving fraud patterns, severe class imbalance, limited interpretability, and scalability issues. This study proposes a robust, explainable, and intelligent framework integrating ensemble learning, Explainable AI (XAI), and a Large Language Model (LLM). Logistic Regression, Decision Tree, Random Forest, AdaBoost, and XGBoost were combined in a Stacking ensemble to enhance accuracy and generalization. Data preprocessing included label encoding, SMOTE for imbalance handling, feature scaling, and top-feature selection via Random Forest. Local Interpretable Model-agnostic Explanations (LIME) provided local feature-based explanations, while the Sonar Pro LLM transformed them into concise, human-readable insights. Experimental results showed the Stacking Classifier achieved the highest accuracy 91% by outperforming all base models. There are some key predictors like City, State, and Transaction_Device increased fraud likelihood, while Bank_Branch and Account_Type often reduced it. The integration of LIME and LLM improved transparency, trust, and compliance readiness, making the system both accurate and interpretable for real-world financial fraud prevention.

Keywords: Financial Fraud Detection, Stacking Ensemble, Explainable AI, Machine Learning.

1 Introduction

1.1 Aim of the study

The aim of this study is to design and implement a robust, explainable, and intelligent financial fraud detection system by leveraging stacked ensemble learning, Explainable AI (XAI), and Large Language Models (LLM). The system uses multiple machine learning models combined through a Stacking approach to improve detection accuracy and generalization across diverse transaction patterns. The focus is not only on identifying fraudulent activity but also on ensuring transparency behind each prediction. For interpretability, XAI tools such as Local Interpretable Model-agnostic Explanations (LIME) is applied to understand the impact of features like Transaction_Location, City, and State on the prediction outcomes. A specific record from the test dataset is used in the LIME explanation phase to locally interpret model decisions. This data, along with a custom query, is then sent to the Sonar Pro model, which returns a human-readable explanation highlighting

key factors influencing the decision—whether fraud or not. In this case, the model predicted 0 ("not fraud"), strongly influenced by city and state features with high negative LIME values. The goal is to build a scalable, interpretable, and accurate fraud detection framework that enhances trust and decision support in financial systems.

1.2 Research Question

The research question for this study is:

How effectively can a stacked ensemble ML model combined with XAI and explain financial fraud, as measured by accuracy, precision, recall, and explanation clarity on a real-world transactional dataset?

1.3 Objectives of the Research

The research objectives for this report are:

1. To develop and evaluate a stacked ensemble ML model combining multiple classifiers (e.g., Logistic Regression, Decision Tree, Random Forest, AdaBoost, and XGBoost) to enhance the accuracy and robustness of financial fraud detection on real-world transactional data.
2. To apply LIME for interpreting model predictions and identifying key transactional features influencing fraud classification, thereby increasing transparency and user trust in the decision-making process.
3. To integrate Sonar Pro API for generating natural language explanations of fraud detection outcomes, transforming complex model outputs into accessible insights for non-technical stakeholders and end-users.

1.4 Outline of the Report

This project outline provides a structured overview of the study, covering introduction, literature review, methodology, design, implementation, evaluation, and conclusion with future directions.

1. Chapter 1 – Introduction: This chapter defines the aim, research question, and objectives of the study, emphasizing the need for robust, explainable financial fraud detection.
2. Chapter 2 – Related Work: This chapter reviews traditional, rule-based, and machine learning approaches in fraud detection, highlighting their strengths, weaknesses, and research gaps.
3. Chapter 3 – Research Methodology: This chapter describes the dataset, preprocessing steps, visualization, and feature selection.
4. Chapter 4 – Design Specification: This chapter presents the system architecture, showing the sequential workflow from data collection to preprocessing, model training, evaluation, and explainability layers.
5. Chapter 5 – Implementation: This chapter details the implementation of individual classifiers, stacking ensemble, and integration of LIME with Sonar Pro LLM for interpretability.
6. Chapter 6 – Evaluation: This chapter provides performance results through confusion matrices, accuracy comparisons, classification reports, and LIME visualizations.

7. Chapter 7 – Conclusion and Future Work: This chapter summarizes key findings, identifies limitations, and outlines directions for future research and system improvements.

2 Related Work

2.1 What is Financial Fraud Detection?

Financial fraud detection is considered an action or process of detecting and stopping any unauthorized, deceptive, or malicious methods that are meant to illegitimately gain funding in favour of a victim at the expense of an individual, a business, or an institution (Ezekiel Onyekachukwu Udeh et al., 2024). Examples of such activities are financial fraud, identity theft, money laundering, insurance fraud and fraud created through online transactions. In the contemporary digital economy, the level of the financial transactions volume, velocity, and variety, in particular, online and mobile financial transactions, have expanded extensively, and in this regard, fraud detection has become a crucial element of financial protection (Bansal et al., 2024). The effectiveness of the traditional fraud detection approaches was highly concentrated on the rule-based mechanisms and manual audits that only detected patterns when there was a recognized pattern but still, were unable to highlight more complex or changing fraud schemes (Elumilade et al., 2021). The most modern methods use analytics, machine learning (ML) and deep learning (DL) to identify fine grained anomalies as well as patterns in high dimension financial data at scale. Such systems have the capability of operating in near real time and this closes the door on fraudulent activity tremendously. Furthermore, the implementation of XAI will make the decision process transparent, which enables the institutions to adhere to the rules, including the General Data Protection Regulation (GDPR). On the whole, financial fraud detection does not only preserve monetary properties but also secures customers and their trust, upholds the brand, and guarantees regulatory safety in a universally linked and risky financial ecosystem.

2.2 Traditional Detection Methods

(Islam et al., 2024) have offered rule-based model (RBM) to identify financial frauds and it does not rely on any data resampling methodology, which is of utmost importance because in general traditional classification models falter due to the imbalanced data in the training sample. The paper looks at how to differentiate between valid and fraudulent transactions which is a constant issue in fraud detection by generating rules that can be understandable, as opposed to relying on black-box machine learning techniques. The RBM has been tested with the help of the following performance. The result revealed an advantage of the RBM compared to all the comparison models in that it yielded score of accuracy and precision of 0.99.

An additional recall, which involves an improved system on detection of fraud on insurance claims based on analysis of rules by adding classification using association rule mining (CBA) by (Baumann, 2021) is a proposal on the need to enhance the current expert systems that tend to isolate individual rules. The overall objective of the research was to find more efficient for detecting and involving illegal insurance claims by revealing and utilizing latent connections among pairs of regulations, which are generally disregarded in traditional systems. In order to do this, the method used association rule mining to identify meaningful rule connections, and the genetic optimizer was utilized in assigning optimal weights as

applied to the rules making the process of decision-making more refined. The results indicate an enhanced fraud detection capability compared to the conventional systems, which illustrates the advantage of the use of inter-rule relationships. The possible disadvantage however is the computational complexity brought about by the genetic optimization and the continuous rules update to ensure the adequate performance against the changing pattern of fraud.

(Kumar and Saxena, 2022) developed a low-budget and hardware-free way of detecting fraud related to online transactions by implementing a set of rules that can combine expertise on defending online transactions in order to assess the amount of potential fraud. The paper deals with the growing lack of reliability of numerous currently used text-based authentication techniques including one-time passwords that are susceptible to hacking, card cloning and eavesdropping. The system implemented by recording the user-specific characters of behavior typed speed, time and pattern when the user enters transaction PINs form a biometric profile that identifies legitimate users without their awareness, and at the same time detects impostor. The authentication process against identity uses rule-based system, which looks at anomalies in real time using these behavioral measurements. The results of the experiments showed that security had improved and the risk of fraud when making online purchases was very low. Nonetheless, a possible weakness is that the system is sensitive to a change of behavior due to fatigue, stress, or difference in a device that may cause false positivity or loss of precision in the real world.

(Ahmed et al., 2021) developed an ontology-based financial fraud detection and deterrence model, which focuses on proactive fraud deterrence using an Intimation Rule-Based (IRB) that generates alert information in the form of the rule, bundle, and interrelated items in terms of a severity level. The research is in regard to the increasing problem of online fraud compounded by the mass use of internet banking, coupled with the shortcomings of major reactive means of detecting fraud, which tend to only detect occurrence much later into the malevolence activity. The findings have shown that timeliness, and accuracy of fraud alerts have improved hence making the system more deterrent. One of the major weaknesses of approach, however, is that it relies on quality and completion of underlying ontology; incompleteness of the knowledge base might compromise functionality of the recognition system in detecting newly appeared fraud methods.

(Beigi and Amin Naseri, 2020) developed a hybrid method that integrates data mining with statistical tools to provide a higher score in the detection of fraud with large, complex financial data that are more accurate and reduce the cost of misclassification. It is realized that data sets with a large volume are highly time consuming to physically detect the fraudulent transactions, and in this regard the study has provided a three phased approach to be used in detection of the same: The first phase involves identification of the relevant features in terms of genetic algorithm, the second involves the determination of the most effective resampling strategy through design of experiments (DOE) and response surface methodology and the third being use of the cost sensitive C4.5 decision tree in a AdaBoost ensemble framework. This model was applied to a real life data set and was able to show that misclassification costs had decreased by at least 14 percent in comparison to other conventional models like decision trees, naive Bayes, Bayesian networks, neural networks and artificial immune systems. Although putting an integrated approach into motion will improve the level of detection, a disadvantage of the approach is that tuning the models becomes difficult due to both complexity and the computational load when combining several optimization and learning methods in the process of doing so.

2.3 ML in Financial Fraud Detection

As the digitalization of financial systems has increased at a high rate, fraud schemes have become more advanced, making the techniques of detecting these schemes ineffective. In the field of financial fraud detection, Machine Learning (ML) has become a game-changer as it provides flexible adapted solutions based on data which can track latent patterns and dynamic threats with outstanding precision. (Ali et al., 2022) suggest that a systematic literature review (SLR) can help to analyze the utilization of machine learning (ML) techniques in detecting financial frauds, as the traditional practice of financial frauds verification through manual checks is usually inefficient due to its high costs, inaccuracy, and lengthy periods of time required to complete this task successfully. Through Kitchenham method by clearly stating the protocols to use in selecting articles, 93 articles, which were relevant, were retrieved using major electronic databases of their inclusion and exclusion criteria. The synthesized findings revealed that financial fraud presentations take the greatest number, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) turn out to be the prevailing algorithms that are mostly used. Data imbalance, changing patterns of fraud and scarcity of good quality datasets have been found to be the main challenges. Findings emphasized the power of ML in enhancing accuracy and efficiency in detection compared to traditional approaches but shortcomings on its use lie on several aspects such as generalizability to various types of fraud and dynamic fraudulent actions, as well as its dependency on particular list of data which can also limit their real-life applicability in the world.

To address these inefficiencies associated with the conventional manual method of detecting financial frauds, (Prasad and Kumar, 2023) have suggested machine learning strategy, which aims at enhancing the accuracy, expediency, and affordability of detecting financial frauds. The study used a dataset of 284,807 European cardholder transactions in 2013 where only 492 of the transactions turned out to be fraud and those preprocessed features included Label Encoding, and SMOTE to handle the severe class imbalance, and performs feature reduction using PCA. The accuracy, precision and recall were used to evaluate models and train them. The ANN model provided the highest accuracy that is 98.69 percent of accuracy, precision and recall, which is better than DT and SVM. The research proved that ANNs could be used to detect fraud effectively besides the fact that the issues are also present regarding handling extreme class imbalance, solely relying on one dataset and the possibility of overfitting that might constrain the ability to generalize findings to various time periods, the methods in which fraud occurs or even to an industry setting.

(Kamuangu, 2024) has suggested that current ineffective rule-based and manual systems provide the motivation to use the advancing technologies to detect financial fraud and provide guidance on the future of research in this area. The paper considers the historical background of the issue of financial fraud, describes the inadequacies of the old methods, and the comparison and analysis of different algorithms of ML and AI. It uses the evaluation parameters to do a comparative analysis of the strengths and weaknesses of various models. The insights indicate that although the ML and AI solutions within the applicable framework can significantly improve the fraud detection processes compared to the traditional ones, issues will be related to model adaptation to a changing fraud dynamic, real-world applicability, and data insufficiency. Though the study presents interesting information about the performance and possible enhancement of algorithm, it follows secondary research instead of conducting experiments which reduces the possibility that findings could be tested in real worlds.

(Hernandez Aros et al., 2023) has proposed a comprehensive literature review on financial fraud detection using machine learning techniques, aiming to understand current research trends, commonly used datasets, and model performance. Employing PRISMA and Kitchenham methodologies, 104 articles published between 2012 and 2023 were selected from major databases such as Scopus, IEEE Xplore, Taylor & Francis, SAGE, and ScienceDirect, based on predefined inclusion and exclusion criteria. The review analyzed author contributions, publication sources, country-wise trends, dataset usage, fraud types, and applied ML models with their evaluation metrics. Findings revealed a strong preference for real datasets, with fraud detection being the most prevalent application, and data sourced from stock exchanges in countries like China, Canada, the U.S., Taiwan, and Tehran. Synthetic data usage was minimal, accounting for less than 7% of datasets, and research contributions were concentrated in China, India, Saudi Arabia, and Canada, with limited output from Latin America. While the study provides valuable insights into global research patterns and methodological practices, its reliance on secondary sources means it does not experimentally validate the comparative effectiveness of different models, limiting practical performance assessment.

(Ashtiani and Raahemi, 2022) has proposed a systematic literature review on intelligent methods for detecting fraudulent financial statements (FFS), aiming to address the limitations of traditional manual auditing, which is costly, slow, and often inaccurate. Using the Kitchenham methodology as a structured protocol, 47 relevant articles were extracted, synthesized, and analyzed to explore the application of machine learning and data mining techniques, along with the datasets used in corporate fraud detection. The review found that supervised algorithms are predominantly applied compared to unsupervised methods like clustering, and highlighted the underexplored potential of semi-supervised, bio-inspired, and evolutionary heuristic approaches as shown in Table 1. It also emphasized the promise of leveraging unstructured data—such as textual and audio sources—for anomaly detection, which could provide richer insights despite introducing new challenges. While the study identifies key trends, gaps, and future research opportunities, it is limited by its reliance on secondary literature without empirical validation, and by the fact that the insights drawn may not fully capture evolving real-world fraud patterns.

(Adijat Bello et al., 2023) has proposed an exploration of diverse Machine Learning (ML) techniques to enhance fraud prevention in financial transactions, addressing the shortcomings of traditional rule-based systems in detecting complex and evolving fraud patterns as shown in Table 1. The study examines supervised learning models such as logistic regression, decision trees, and neural networks trained on historical transaction data for high-accuracy classification, alongside unsupervised methods like clustering and anomaly detection for uncovering novel fraud types without labeled data. Advanced approaches, including deep learning models like CNNs and RNNs for sequential data analysis and NLP for detecting suspicious textual content, are also discussed. The integration of these techniques enables real-time monitoring, predictive analytics, and adaptive learning as data volume grows, significantly improving fraud detection capabilities. However, challenges such as ensuring data privacy, maintaining high-quality and diverse training datasets, and addressing the interpretability of complex models remain. While the study highlights the broad applicability and potential of ML in scalable, proactive fraud prevention, it is conceptual in nature and lacks empirical validation, limiting direct assessment of real-world performance.

Table 1: Summary Table

Author(s), Year	Focus of Study	Method / Approach	Key Findings	Limitations
(Ali et al., 2022)	Systematic literature review (SLR) on ML in fraud detection	Kitchenham SLR, 93 studies reviewed	SVM and ANN are most widely used; ML improves accuracy & efficiency vs. manual checks	Data imbalance, limited generalizability, dataset dependency
(Prasad and Kumar, 2023)	Detecting card fraud in European transactions	Preprocessing with Label Encoding, SMOTE, PCA; compared ANN, DT, SVM	ANN achieved 98.69% accuracy, outperforming DT & SVM	Extreme class imbalance, reliance on single dataset, overfitting risk
(Kamuangu, 2024)	Review of ML & AI for fraud detection	Comparative analysis of algorithms	ML/AI improves fraud detection vs. rule-based/manual systems	No experiments, only secondary research; lacks real-world validation
(Hernandez Aros et al., 2023)	Literature review of ML in financial fraud	PRISMA + Kitchenham, 104 articles analyzed	Real datasets preferred, fraud detection dominant, global research patterns mapped	No experimental validation, secondary data dependency
(Ashtiani and Raahemi, 2022)	Detecting fraudulent financial statements (FFS)	Kitchenham SLR, 47 articles reviewed	Supervised ML dominates; potential in semi-supervised, bio-inspired, and unstructured data use	Relies on literature only, no real-world testing; evolving patterns may not be captured
(Adijat Bello et al., 2023)	Exploring diverse ML methods for fraud prevention	Supervised (LR, DT, NN), unsupervised (clustering, anomaly detection), DL (CNN, RNN), NLP	Broad applicability; real-time monitoring and adaptive learning feasible	Conceptual work, lacks empirical validation; data quality and interpretability challenges

3 Research Methodology

3.1 Dataset Description

The dataset in the present study was gathered by Kaggle and is intended to deal with fraud detection in a bank transaction and has plenty of information about transactions and

customers. It consists of 24 columns that record various things concerning the banking operations. Each row is a transaction and has a unique identifier called Transaction_ID, the other details include the Customer_ID, Customer_Name, Gender, Age and geographical details (State, City). It also has banking related features, that is, it contains banking related features such as Bank_Branch, Account_Type, Account_Balance, and transaction details such as Transaction_Date, Transaction_Time, Transaction_Amount, Transaction Type, and Merchant Category. The use of the device can be described using Transaction_Device, Device_Type, and Transaction_Location (e.g., latitude and longitude), and this helps to track digital footprints. The target variable, which is represented by the column Is_Fraud, denotes that a transaction is a fraud (with the value of 1) or honest (with the value of 0). Other contextual information are the identifier of the merchant (Merchant_ID), the currency being used (Transaction_Currency), and other customer communication (Customer_Contact, Customer_Email). It also has the Transaction_Description that has a textual context to each transaction. This is a comprehensive dataset that allows robust feature engineering, exploratory data analysis, and using a ML model in fraud detection, and allow explainability and transparency in the predictions that are made based on XAI techniques.

3.2 Data Preprocessing

Preparing the data is one of the most important stages of making the dataset ready to be used to train the model and make quality predictions. Originally, all the categorical variables are determined with the help of X.select_dtypes and encoded into numerical ones by using Label Encoding. This is done to enable the machine learning algorithm whose learning mode needs to be fed with numerical data to make use of features like Customer_Name, Gender, State, City, Bank_Branch, Account_Type, Transaction_Type, Merchant_Category, and device related features like Transaction_Device and Device_Type. The encoded data was still affected by one of the typical problems of dataset used in fraud detection, class imbalance, where the negative example (fraud) is much less than the positive such as (legitimate). To deal with this, the SMOTE is used with the parameter sampling_strategy='minority'. This method creates artificial samples of the minority class to make the content more equal to avoid bias in training. Once the balanced distribution of classes is attained, the feature scaling becomes the next important thing. MinMaxScaler is applied to all features to ensure its values are in a range that is comparable.

3.3 Data Visualization

Figure 1 is a bar chart that visualizes the distribution of fraud cases across different transaction types. Credit and Transfer transactions slightly lead with just over 2000 fraud cases, while Withdrawal has slightly fewer, just under 1900 cases. This plot highlights that fraudulent activities are not limited to a specific transaction type, underscoring the need for broad-based detection mechanisms across all categories.

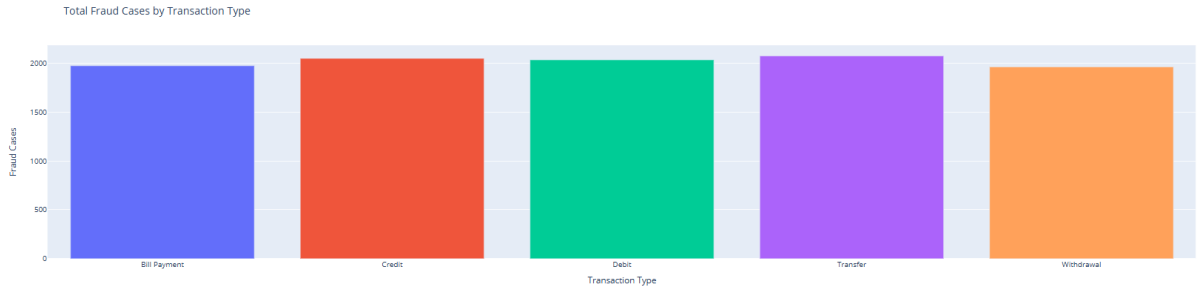

Figure 1: Bar Chart Showing Total Fraud Cases by Transaction Type

Figure 2 presents a bar chart that illustrates the average account balance for customers across different U.S. states. The x-axis represents various states, while the y-axis shows the average account balance ranging between approximately 34.5K to 36.5K. States such as New Mexico, Nebraska, and New York show slightly higher average balances, near 36.5K, whereas states like Idaho, Georgia, and Alabama show relatively lower averages, closer to 34.6K. Despite these small variations, the overall distribution appears relatively uniform, indicating that the average account balance does not vary drastically by state. This suggests a balanced customer profile across regions, which can be important when analyzing location-based fraud trends.

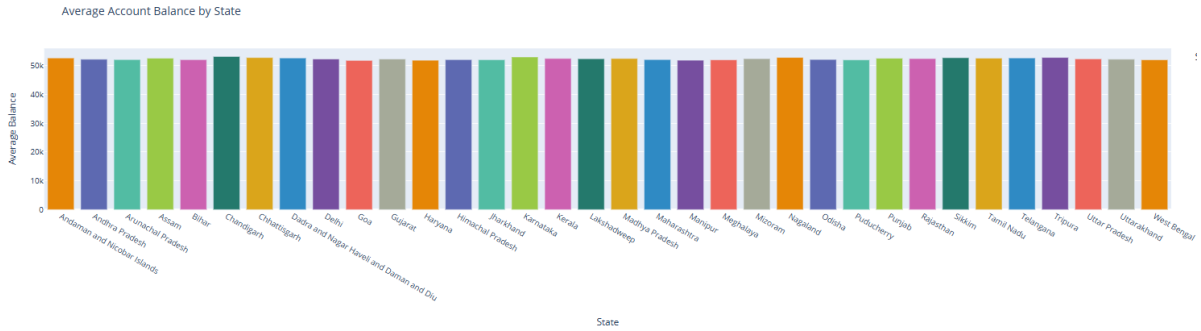

Figure 2: Bar Chart Showing Average Account Balance by State

Figure 3 is a pie chart representing the distribution of device types used for transactions. The chart is divided into four nearly equal segments, indicating that each device type accounts for approximately 25% of the total usage. Specifically, three categories show exact 25% usage, while one slightly leads at 25.1%, and another slightly trails at 24.9%. This balanced distribution suggests that no single device type—such as Smartphone, Laptop, Desktop, or Tablet—dominates user behavior, implying that fraud detection models must account for a wide variety of devices equally when analyzing transaction patterns.

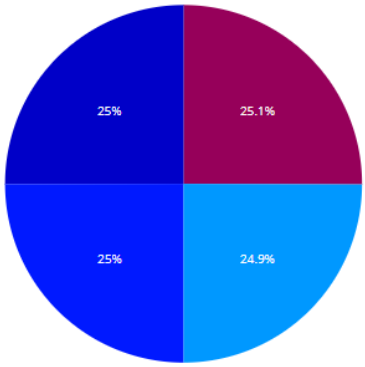

Figure 3: Pie Chart Showing Device Type Usage

Figure 4 is a bar chart that displays the distribution of fraud cases based on gender. The x-axis represents the gender categories—Female and Male, while the y-axis indicates the number of fraud cases, which ranges up to 5000. Both genders show a nearly equal count of fraud cases, with females slightly exceeding 4900 cases and males just below 4900 cases. This minimal variation indicates that gender has little influence on the likelihood of fraud,

suggesting that fraudulent activity is evenly distributed across male and female customers. Such insight is crucial for building unbiased fraud detection models.

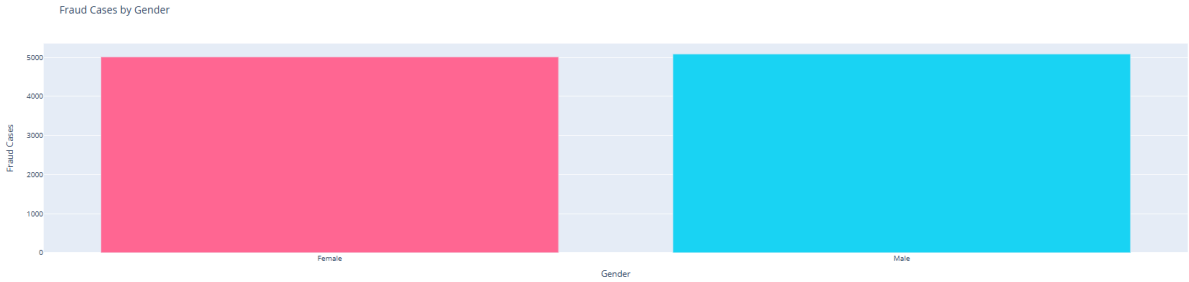

Figure 4: Bar Chart Showing Fraud Cases by Gender

Figure 5 is a box plot that visualizes the distribution of three important numerical features—Is_Fraud, Account_Balance, Transaction_Amount, and Age. Each variable is plotted on the y-axis against their respective value ranges on the x-axis (scaled between 0 and 100). The median age appears close to 45, while transaction amounts and account balances also show median values near 50% of the normalized range. The boxes reflect interquartile ranges (IQR), and the whiskers show minimum and maximum values without outliers. The Is_Fraud feature, being binary, shows two distinct positions, confirming its categorical nature. This plot is helpful for identifying value concentration and spread, spotting skewness or possible anomalies across these variables.

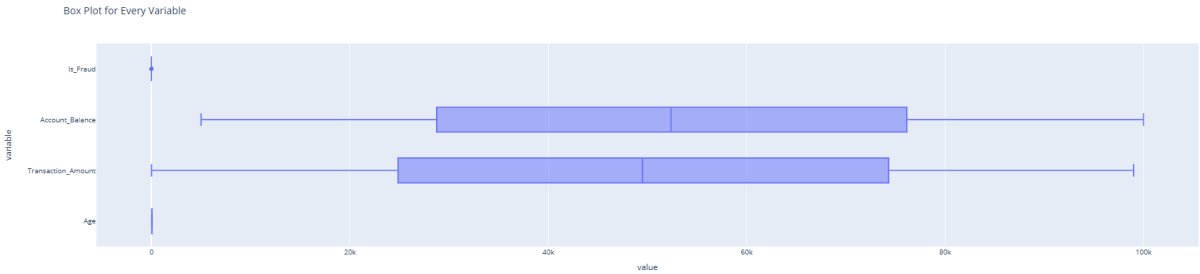

Figure 5: Box Plot for Key Numerical Variables

3.4 Feature Importance

In order to determine the most significant attributes in detecting fraud, the Random Forest classifier would be utilized in finding the significance scores of the features. In this ensemble model, all the features used as input are ranked on their importance in terms of accuracy of predictions. Once the model is trained on the entire dataset, a visualization plot that indicates the most important features is created, which implies the variables that influence the highest probability of fraction determination the most. According to the scores, the 15 most significant features are indicated to go ahead and have the model trained further. Such dimensionality reduction exercise boosts the model efficiency, decreases vulnerability to overfitting, and only the most relevant attributes used in the final predictive models.

Figure 6 is a horizontal bar chart that shows the features importance score based on a Random Forest model. The x-axis represents the importance score that will be between 0.00

and 0.07 and the y-axis is the input features. State is the most prominent characteristic, and its value of around 0.072 domineers that of the rest of the following characteristics, which are Transaction_Type, Transaction_Device, and Account_Type, having values of above 0.06. Attributes such as Gender, Customer_Email, and Customer_Name tend to provide the least contribution, and their score lies below 0.03. It is also a significant chart in the feature selection process because it reminds the most significant attributes (based on the feature influence) that affect the prediction of fraud and will thus enhance the model accuracy and explainability.

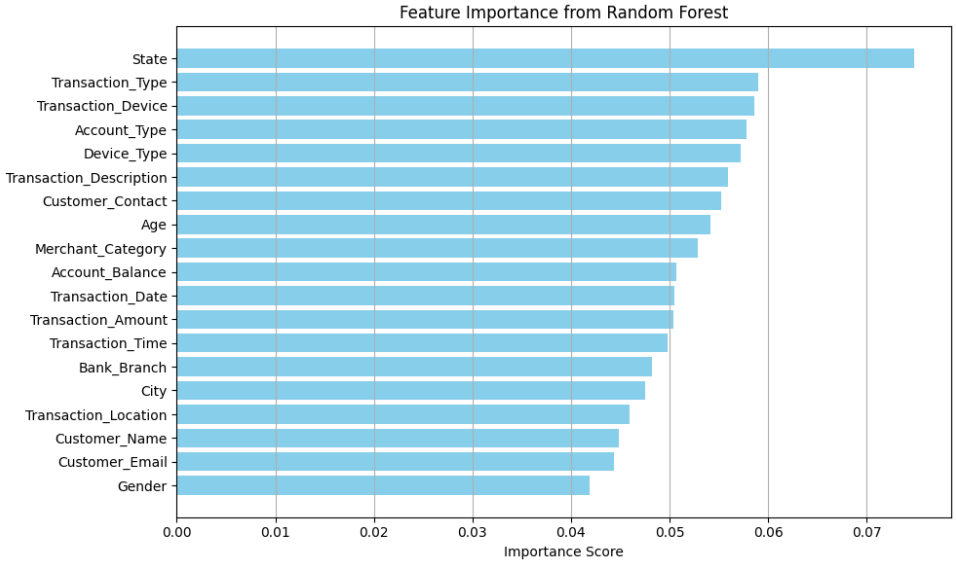

Figure 6: Bar Chart Showing Feature Importance from Random Forest

4 Design Specification

The system architecture diagram for the financial bank fraud detection framework illustrates a comprehensive and sequential workflow designed for robust, explainable, and intelligent fraud prediction as shown in Figure 7. It begins with bank transaction fraud detection dataset, where raw transaction data is sourced, typically from financial institutions. This data undergoes Data Preprocessing, including label encoding, handling imbalanced classes using SMOTE, and feature scaling through MinMax normalization. Next, Feature Importance is determined using a Random Forest model, helping identify the top 15 most influential attributes for predicting fraud. The refined dataset is then split into training and testing sets in the Train-Test Split stage. The Model Training block follows, where multiple machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, AdaBoost, XGBoost—are trained and integrated into a Stacking Ensemble Model to boost accuracy and generalization. Post-training, Model Evaluation is conducted using confusion matrices and performance metrics such as accuracy, precision, recall, and F1-score. To ensure interpretability, the system incorporates an XAI Layer (LIME), which provides insight into which features influenced the predictions and LLM mentioned at the end of it to summarise the output and get reason for the actual output.

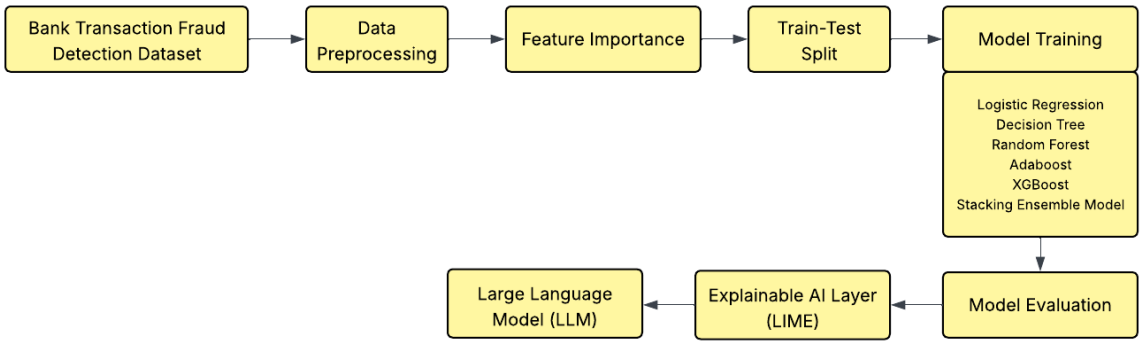

Figure 7: System Architecture Diagram

5 Implementation

5.1 Implementation of Logistic Regression

The Logistic Regression model has been carried out with the help of the LogisticRegression class, L2 regularization penalty (penalty='l2') and regularization strength C parameter with value 0.1. Optimizing is done using the lbfgs solver, and random number is set to 1 so that it can be reproduceable. The model is fit by using lr.fit(X_train, y_train) and the lr.predict(X_test) is used to generate predictions. Logistic Regression can also be helpful in case of binary classification like fraud detection because it is a linear model. It provides probabilities according to sigmoid function and this shows interpretable decision boundaries either fraudulent or not.

5.2 Implementation of Random Forest

The Random Forest Classifier can be initialized like RandomForestClassifier(max_depth=5) to stop the training process when the individual trees are deep. It is also trained by rfmodel.fit(X_train, y_train) and predictions are made by using rfmodel.predict (X_test). This method of ensemble develops numerous decision trees during training and returns the mode of classes to be used in classification. The advantage of the model is that it deals with large-sized feature spaces and is able to capture non-linear relationships. A shallow depth of 5 implies that the emphasis is put on generalizing rather than fitting noise in data, which is since in the case of fraud, there might be a lot of convoluted yet slight patterns.

5.3 Implementation of Decision Tree

The Decision Tree model is made through DecisionTreeClassifier(random_state=1) to make the results of the same. It is trained on the training set by mit.fit and applied to the test set on mit.predict. This model constructs a flow diagram like structure, wherein each internal node makes a choice depending on a feature, and each terminal node makes an outcome. Although they are convenient to work with and are absolutely clear, decision trees are easily overfit. Nonetheless, they are effective in capturing complex decision boundaries and thus providing a good benchmark model in fraud detection tasks.

5.4 Implementation of AdaBoost

The AdaBoost Classifier is incorporated by the means of AdaBoostClassifier(random_state=1). The ada.fit(X_train, y_train) train it and the ada.predict(X_test) is used to make prediction. AdaBoost trains a sequence of weak classifiers that are combined in each iteration using the weights of the incorrectly classified instances. In this manner, it pays more attention to the difficult-to-classify representatives. The AdaBoost lends to fraud detection because it can be applied to enhance the predictive power in a way that model complexity has also improved to a small extent. This assists in parity of anti-bias and variance as well as resistence to overfitting on noisy data.

5.5 Implementation of XGBoost

To prevent overfitting, the XGBoost model will be initialized by XGBClassifier (max_depth=5, random_state=1), where the depth of the trees that should be created is limited to five. Training uses model.fit(X_train, y_train) and to get the predictions, model.predict(X_test) is used. Since it is optimized, it has the effectiveness of gradient-boosting algorithm with parallel processing, tree pruning, and regularization. It is well-suited to detect the presence of fraud since it can effectively deal with missing values and complex patterns existing in both high-dimensional and heterogeneous data.

5.6 Implementation of Stacking

The Stacking Classifier is a meta-ensemble model in which the predictions of a set of base learners are combined to give a better overall accuracy. In this implementation, three base-models are adapted, namely AdaBoostClassifier, RandomForestClassifier, and LogisticRegression, which are initialized with the same seed random_state=1. The models are the same and the output of those models is then taken as an input feature in the meta-model which is an XGBClassifier. The instance of the stacking model is constructed with the help of StackingClassifier() class with classes which are used as the base models and meta_classifier which takes XGBoost model. Once the model is fitted using stac_model.fit(X_train, y_train), one can get the predictions by calling stac_model.predict(X_test.values). Stacking uses the relative merits of various algorithms by enabling the meta-model to learn the best way of combining the results of various algorithms. In general, this ensemble approach tends to perform better than individual models particularly on tasks requiring complex models such as in fraud detection where patternation differences occur.

5.7 Implementation of XAI

The LimeTabularExplainer is further initialized and uses the training data (X_train.values) which is then converted to a NumPy array. The explainer has been set to undertake classification task by setting mode =classification, and this guarantees that the explainer properly interprets class probabilities.

The test set again contains one data point which is chosen on the basis of an index (sample_idx = 1) and passed on to NumPy as an array. This instance is then inputted to the explain_instance method, where the predict_fn is also given. In this method, probabilities of each of the classes are calculated, and LIME can tell how each of the features contributed to the final prediction.

The explanation is subsequently visualized that shows a table of how the individual features are contributing to the decision of the model. The application of LIME provides a significant level of interpretability giving us an insight on why a transaction was classified as being fraud or not which is very much necessary in developing trust in AI-powered financial systems.

Figure 7 shows the output of the LIME interpretation of a single prediction by the stacking classifier on the test instance of the index 5.

LIME Explanation for Instance Index 5

	Feature	Importance
0	0.00 < City <= 0.50	0.162223
1	0.40 < Transaction_Time <= 0.80	0.122009
2	0.27 < Merchant_Category <= 0.47	0.090342
3	Bank_Branch <= 0.28	-0.075547
4	0.00 < State <= 0.33	0.071081
5	Transaction_Date <= 0.27	0.070119
6	0.26 < Transaction_Device <= 0.49	0.042711
7	Account_Type > 0.72	-0.031928
8	Age <= 0.16	0.028502
9	Customer_Name <= 0.27	-0.026514

Figure 7: LIME Explanation for Instance Index 5

5.8 Integration with LLM (Sonar Pro) for Interpretability

To enhance interpretability of the stacking classifier, the LIME Tabular Explainer was implemented. The explainer was initialized using the training dataset along with feature names and class labels in classification mode. A specific test instance (index 1) was selected to generate a local explanation of model predictions. LIME analyzed feature contributions by approximating the classifier's behavior around the chosen instance and provided importance scores for each feature. The output was converted into a DataFrame for tabular representation, displayed within the notebook, and optionally saved as a CSV file. This process improved model transparency and interpretability for fraud detection.

After obtaining feature-level explanations from LIME, the identified feature contributions were passed to the Sonar Pro LLM to generate a concise, human-readable interpretation. The LLM summarized that the most important positive contributors to the fraud prediction were City in the range (0.00–0.50], Transaction_Time between 0.40–0.80, and Merchant_Category between 0.27–0.47. These value ranges historically aligned with higher fraud risk in the training dataset. Negative contributors, such as Bank_Branch ≤ 0.28, Account_Type > 0.72, and Customer_Name ≤ 0.27, slightly reduced the fraud probability. For the selected transaction, the model classified the case as fraud (1) because the positive risk factors outweighed the negatives. This combined LIME+LLM pipeline enabled the translation of numerical feature contributions into actionable insights for financial analysts, improving trust, regulatory compliance, and decision support in real-world fraud detection scenarios.

6 Evaluation

6.1 Case Study 1: Logistic Regression

Figure 8 represents the confusion matrix obtained from the Logistic Regression model used for financial fraud detection. The matrix shows the model's performance across four categories: True Negatives (11,792), where non-fraudulent transactions were correctly identified; False Positives (7,199), where legitimate transactions were wrongly classified as fraud; False Negatives (6,549), where fraudulent transactions went undetected; and True Positives (12,443), where fraudulent activities were accurately identified. While the model demonstrates a balanced ability to detect both fraud and non-fraud, the relatively high count of false positives and false negatives indicates room for optimization, especially in reducing misclassifications to enhance precision and recall.

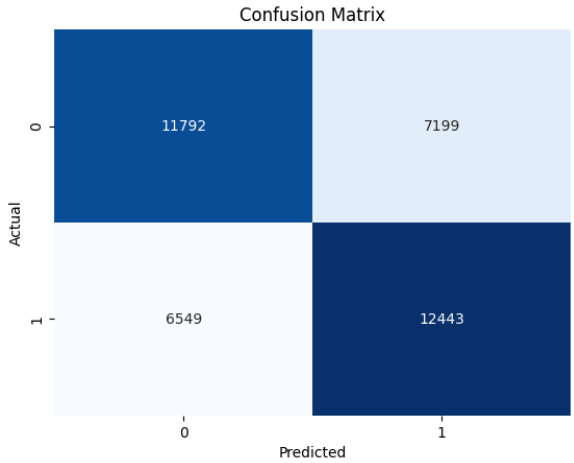

Figure 8: Confusion Matrix

6.2 Case Study 2: Random Forest

Figure 9 shows the confusion matrix for the Random Forest model applied to the financial fraud detection task. The matrix illustrates True Negatives (13,677) where legitimate transactions were correctly classified, False Positives (5,314) where genuine transactions were incorrectly labeled as fraud, False Negatives (4,760) where fraudulent transactions were missed, and True Positives (14,232) representing correct fraud predictions. Compared to logistic regression, this model demonstrates stronger predictive performance, especially with a higher true positive rate and reduced false negatives. The Random Forest model provides a better balance between sensitivity and specificity, making it more reliable for fraud detection.

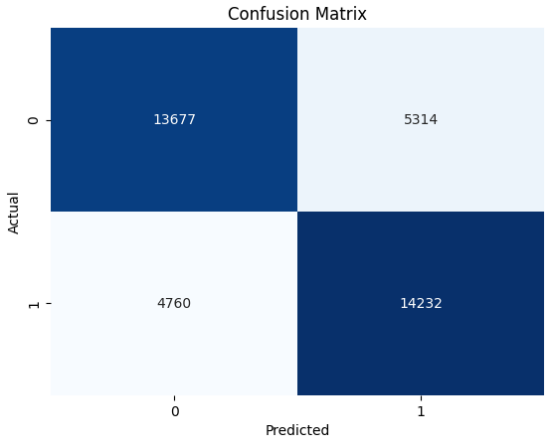

Figure 9: Confusion Matrix

6.3 Case Study 3: Decision Tree

Figure 10 displays the confusion matrix for the Decision Tree model used in the financial fraud detection system. It shows a strong classification performance with 15,326 True Negatives, indicating accurate identification of non-fraudulent transactions, and 16,475 True Positives, representing correctly detected fraudulent transactions. The model also resulted in 3,665 False Positives, where normal transactions were misclassified as fraud, and 2,517 False Negatives, where fraudulent activities were missed. With the lowest false negative rate among the compared models, the Decision Tree shows high recall and overall precision, making it a highly effective and interpretable option for fraud detection.

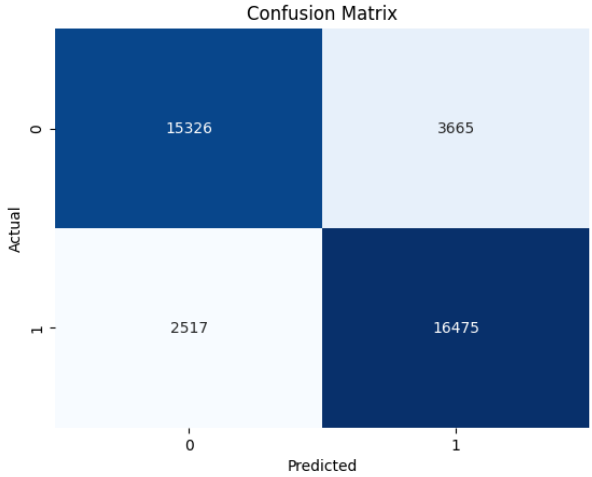

Figure 10: Confusion Matrix

6.4 Case Study 4: AdaBoost

Figure 11 presents the confusion matrix for the AdaBoost model applied in the financial fraud detection process. The matrix includes 13,577 True Negatives, meaning legitimate transactions were correctly classified, and 15,527 True Positives, indicating accurate fraud detection. However, the model also produced 5,414 False Positives, where genuine transactions were flagged as fraud, and 3,465 False Negatives, where fraudulent cases were missed. AdaBoost demonstrates a good balance between sensitivity and specificity, with strong true positive performance. However, the slightly higher false positive rate suggests a tendency to over-predict fraud, requiring fine-tuning for improved precision.

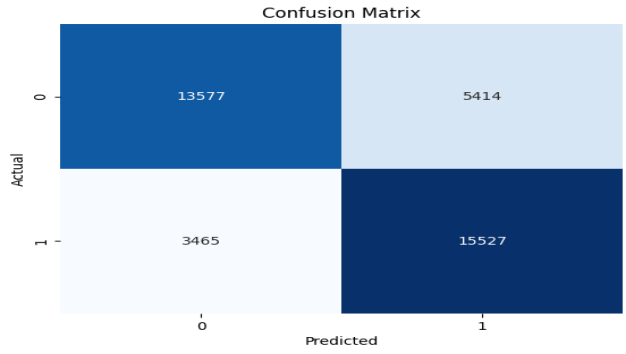

Figure 11: Confusion Matrix

6.5 Case Study 5: XGBoost

Figure 12 delivers high accuracy with 16,801 True Negatives, correctly identifying non-fraudulent transactions, and 16,714 True Positives, successfully detecting fraudulent activities. It has relatively low error rates with only 2,190 False Positives, where genuine transactions were misclassified as fraud, and 2,278 False Negatives, representing missed fraud cases. This balance highlights XGBoost's strong predictive performance, achieving both high precision and recall. The model stands out for its robustness and efficiency, making it one of the most reliable choices for fraud detection in this study.

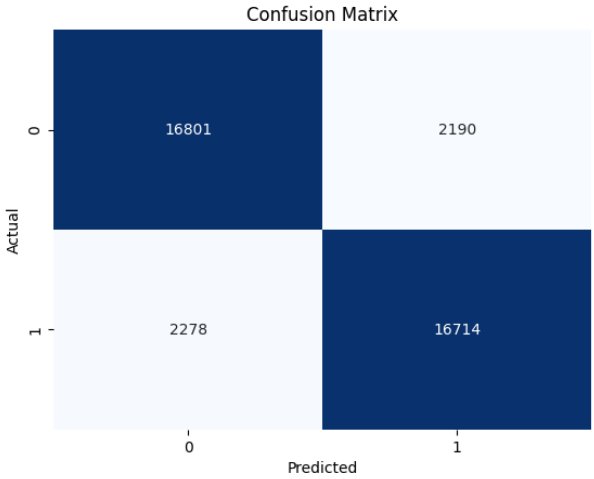

Figure 12: Confusion Matrix

6.6 Case Study 6: Stacking

Figure 13 displays the confusion matrix for the Stacking model, which combines predictions from multiple base learners to improve financial fraud detection. The matrix reveals excellent classification results with 16,994 True Negatives and 17,502 True Positives, showing that the model effectively identifies both legitimate and fraudulent transactions. It records only 1,997 False Positives and 1,490 False Negatives, indicating minimal misclassifications. Among all tested models, Stacking achieves the highest accuracy, with the lowest error rates across both fraud and non-fraud classes. Its superior performance makes it the most balanced and dependable approach in this experiment.

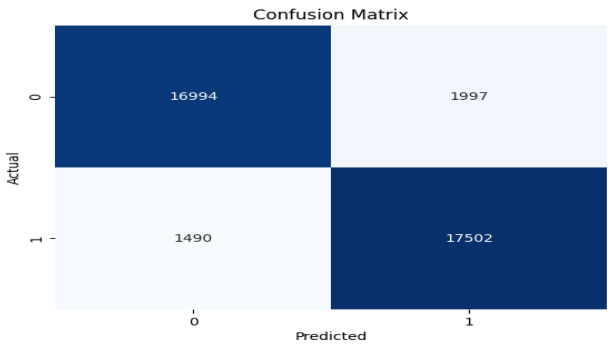

Figure 13: Confusion Matrix

Figure 14 shows the classification report of the Stacked ML model for financial fraud detection. For non-fraud (Class 0), precision is 0.92 and recall 0.89, indicating accurate identification with minimal false fraud alerts. For fraud (Class 1), precision is 0.90 and recall 0.92, reflecting strong capability to capture fraudulent cases with few false positives. The model performs consistently well, effectively distinguishing between fraudulent and legitimate transactions, ensuring reliability for real-world fraud prevention.

	precision	recall	f1-score	support
0	0.92	0.89	0.91	18991
1	0.90	0.92	0.91	18992

Figure 14: Classification Report for Best Model

The Table 2 above summarizes the classification accuracy of various machine learning models used for financial fraud detection. Logistic Regression performs the weakest with 64% accuracy, likely due to its linear nature and inability to capture complex patterns. Tree-based models such as the Decision Tree (84%) and Random Forest (73%) offer better performance due to their capacity for non-linear decision making. Boosting methods further enhance accuracy, with AdaBoost achieving 77% and XGBoost reaching 88%. The Stacking Classifier outperforms all, delivering the highest accuracy at 91%, highlighting the effectiveness of combining diverse models to capture complex fraud patterns more robustly.

Table 2: Model Accuracy Table

Model	**Accuracy (%)**
Logistic Regression	64%
Random Forest Classifier	73%
Decision Tree Classifier	84%
AdaBoost Classifier	77%
XGBoost Classifier	88%
Stacking Classifier (Best)	91%

Figure 15 is a LIME visualization showing the factors influencing the model's fraud prediction for a specific transaction. The prediction probability is 100% fraud (Class 1). Positive contributors such as City (0.50), Transaction_Time (0.80), and Merchant_Category (0.43) increased fraud likelihood, while negative contributors like Bank_Branch (0.24) and Account_Type (0.86) reduced it. This figure demonstrates the explainability aspect, showing exactly which features and value ranges drove the model's decision toward classifying the transaction as fraudulent.

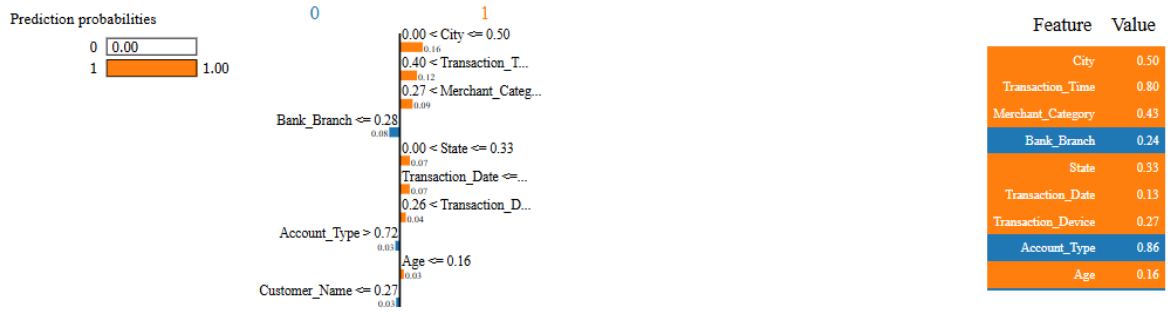

Figure 15: LIME-Based Feature Contribution Visualization for Fraud Prediction

7 Conclusion and Future Work

7.1 Conclusion

This project presents a comprehensive approach to financial fraud detection by combining the strengths of traditional machine learning algorithms with the interpretability of XAI and the narrative clarity of LLM. Through meticulous data preprocessing, feature engineering, and the integration of multiple classifiers into a Stacking ensemble, the system achieved superior predictive performance, with ensemble methods effectively capturing the complex, non-linear patterns inherent in transactional data. The use of LIME provided transparent, feature-level explanations for individual predictions, enabling deeper insights into model reasoning, while the integration of the Perplexity AI Sonar Pro model translated these technical outputs into concise, human-readable explanations. This not only enhanced stakeholder trust but also supported regulatory compliance by offering clear justification for automated decisions. Key predictors—such as Transaction_Location, City, State, and Transaction_Device—were identified as high-impact indicators of fraudulent activity. Overall, the framework demonstrates that combining statistical learning with advanced explainability techniques produces a fraud detection system that is accurate, interpretable, and actionable, making it a valuable asset for financial institutions aiming to detect and prevent fraud while upholding fairness and accountability.

7.2 Limitations and Future Works

Although the project has shown good results, there are a number of limitation associated with the project. The dataset although rich, might not reflect on the diversity of transactions as seen in any global financial systems. Some of the features, such as Customer_Email and Customer_Contact, remained underappreciated because of privacy and sparsity issues, and this can restrict the accuracy of models. The other weakness is that the information may be biased due to the unfair distribution of fraudulent cases which are uncommon in nature. Although imbalance management techniques were used in the form of SMOTE and scaling, some additional advanced methods of imbalance management can be examined. More so, such explainability layers with LIME is local and instance-based and hence may fail in capturing global model behavior and performance over unseen data. The Sonar LLM integration, as brilliant, also requires timely and good quality and API performance, which can introduce latency or dependency risks. Potential future work is to enlarge the dataset with cross-bank transactional data, use deep learning models such as LSTM or transformers to

allow longer fraud patterns, and generalize the models better. Besides, the system can be more adaptive by improving explainability in the form of a global feature attribution of the system, and inclusion of real-time feedback mechanisms. The translation of the model to the cloud-based systems to operate it in the real world in terms of live fraud detection and support the secure, scalable, and compliant solutions would also be crucial to operate it in the real world of the financial ecosystems.

References

Adijat Bello, O., Folorunso, A., Zainab Budale, O., 2023. Machine learning approaches for enhancing fraud prevention in financial transactions.

Ahmed, M., Ansar, K., Muckley, C.B., Khan, A., Anjum, A., Talha, M., 2021. A semantic rule based digital fraud detection. PeerJ Comput. Sci. 7, e649. https://doi.org/10.7717/peerj-cs.649

Ali, A., Abd Razak, S., Othman, S.H., Eisa, T.A.E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., Saif, A., 2022. Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. Appl. Sci. 12, 9637. https://doi.org/10.3390/app12199637

Ashtiani, M.N., Raahemi, B., 2022. Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review. IEEE Access 10, 72504–72525. https://doi.org/10.1109/ACCESS.2021.3096799

Bansal, U., Bharatwal, S., Bagiyam, D.S., Kismawadi, E.R., 2024. Fraud Detection in the Era of AI: Harnessing Technology for a Safer Digital Economy, in: Irfan, M., Elmogy, M., Gupta, S., Khalifa, F., Dias, R.T. (Eds.), Advances in Finance, Accounting, and Economics. IGI Global, pp. 143–164. https://doi.org/10.4018/979-8-3693-6321-8.ch006

Baumann, M., 2021. Improving a Rule-based Fraud Detection System with Classification Based on Association Rule Mining. https://doi.org/10.18420/INFORMATIK2021-091

Beigi, S., Amin Naseri, M.R., 2020. Credit Card Fraud Detection using Data mining and Statistical Methods. J. AI Data Min. 8. https://doi.org/10.22044/jadm.2019.7506.1894

Elumilade, O.O., Ogundeji, I.A., Achumie, G.O., Omokhoa, H.E., Omowole, B.M., 2021. Enhancing fraud detection and forensic auditing through data-driven techniques for financial integrity and security. J. Adv. Educ. Sci. 1, 55–63. https://doi.org/10.54660/.JAES.2021.1.2.55-63

Ezekiel Onyekachukwu Udeh, Prisca Amajuoyi, Kudirat Bukola Adeusi, Anwulika Ogechukwu Scott, 2024. The role of big data in detecting and preventing financial fraud in digital transactions. World J. Adv. Res. Rev. 22, 1746–1760. https://doi.org/10.30574/wjarr.2024.22.2.1575

Hernandez Aros, L., Bustamante Molano, L.X., Gutierrez Portela, F., Moreno Hernández, J.J., Rodríguez Barrero, M.S., 2023. Detection of financial fraud by applying ML techniques a RSL. https://doi.org/10.7910/DVN/CM8NVY

Islam, S., Haque, Md.M., Rezaul Karim, A.N.M., 2024. A rule-based machine learning model for financial fraud detection. Int. J. Electr. Comput. Eng. IJECE 14, 759. https://doi.org/10.11591/ijece.v14i1.pp759-771

Kamuangu, P., 2024. A Review on Financial Fraud Detection using AI and Machine Learning. J. Econ. Finance Account. Stud. 6, 67–77. https://doi.org/10.32996/jefas.2024.6.1.7

Kumar, J., Saxena, V., 2022. Rule-Based Credit Card Fraud Detection Using User's Keystroke Behavior, in: Kumar, R., Ahn, C.W., Sharma, T.K., Verma, O.P., Agarwal, A. (Eds.), Soft Computing: Theories and Applications, Lecture Notes in Networks and Systems. Springer Nature Singapore, Singapore, pp. 469–480. https://doi.org/10.1007/978-981-19-0707-4_43

Prasad, E., Kumar, H., 2023. Enhancing Performance of Financial Fraud Detection Through Machine Learning Model.