Financial Fraud Detection in Nigerian Banks: Data Mining Approach

Mohammed, Usman
Computer Science Department
Federal Polytechnic Bali
Taraba State, Nigeria
shaggyrancy@gmail.com

Professor G. M. Wajiga
Computer Science Department
Modibbo Adama University, Yola
Adamawa State, Nigeria
E-mail: gwajiga@mau.edu.ng

SAIDU, Hayatu Alhaji
Computer Science Department
Federal Polytechnic, Mubi
Adamawa State, Nigeria.
hayatusaidu85@gmail.com

DOI: 10.56201/ijcsmt.v10.no3.2024.pg45.66

Abstract

This research investigates the application of data mining techniques, specifically logistic regression and random forest, to detect financial fraud within Nigerian banks. Using individual bank statements and statutory bank charges, the study focuses on developing a robust system for identifying fraudulent transactions. The data preprocessing involves extracting key features such as transaction type, amount, balance, and transaction date. The dataset is split into training and testing sets, and both machine learning models are trained and evaluated based on metrics like accuracy, precision, recall, and F1-scores.The results indicate that the Random Forest model outperforms Logistic Regression, achieving higher accuracy and better handling of complex relationships within the data. Visualization tools like Matplotlib are used to present prediction probabilities, enhancing understanding of model behavior. The system's implementation includes secure access features, detailed transaction analysis, and comprehensive fraud summaries. Challenges such as data imbalance are addressed with techniques like SMOTE and advanced preprocessing methods. This study highlights the potential of using advanced machine learning models for effective fraud detection in financial transactions. The findings suggest that further improvements in feature extraction, data expansion, and exploring more sophisticated models can enhance system performance. This research contributes to the ongoing efforts to secure financial systems against fraudulent activities, offering valuable insights and practical solutions for the banking sector

Keywords: *Component; Financial fraud detection, Logistic regression, Random forest, Machine learning models, Nigerian banks*

I. Introduction

In its various forms, financial fraud has a significant risk to the global economy. It is likely to affect the stability and security of financial institutions. It can also affect individuals' financial well-being, and erase trust in the financial system. The development of technology and the improve digitization of financial transactions have given rise to new opportunities for fraudsters to exploit customers. Traditional, rule-based methods of fraud detection have shown inadequate in the face of these dynamic and effective instruments.

Data mining techniques have proven to be a vital component of the solution to this multifaceted problem. These techniques involve the application of advanced algorithms to vast datasets, allowing for the discovery of hidden patterns, anomalies, and irregularities in financial transactions. By analyzing historical data and identifying deviations from established norms, data mining can offer early detection of potential fraudulent activities (Han, Kamber & Pei, 2011; Witten, Frank, Mark 2016; Ramagery, 2013).

The utilization of data mining in fraud detection goes beyond academic research; it is a practical necessity in the financial industry. The adoption of these techniques is important in safeguarding the financial ecosystem, improving risk management, and maintaining the trust of stakeholders. This dissertation explores the various data mining methods employed in financial fraud detection and how they are contributing to a more secure and reliable financial sector.

This study focuses on investment of two prominent data mining techniques, Logistic Regression and Random Forest, to address specific challenges within fraud detection processes. Specifically, the research aims to minimise the issue of high false positive rates, a persistent challenge that can lead to unnecessary disruptions and resource allocation. Additionally, the study explores how the effective implementation of these techniques can contribute to an increase in financial ratios, further enhancing the reliability of financial institutions against fraudulent activities.

Logistic Regression, a statistical method well-established in the field of data analysis, provides a powerful framework for modeling the probability of fraudulent transactions. Its ability to handle binary outcomes makes it particularly fitted for identifying anomalous patterns in financial data. On the other hand, Random Forest, is a group of separate things that contributes to learning algorithm, excels in capturing complex relationships within large datasets, offering a more slight approach to fraud detection (Breiman, 2001; Eibe, Frank & Hall,2016).

The high false positive rates prevalent in traditional fraud detection systems can result in significant operational challenges, leading to increased costs and potential customer dissatisfaction. By employing Logistic Regression and Random Forest, this research seeks to develop model that not only effectively identify fraudulent activities but also demonstrate a notable reduction in false positives. This not only major operational efficiency but also fosters a more secure and reliable financial environment.

Furthermore, the study investigates the potential impact of improved fraud detection on financial ratios, such as return on assets, return on equity, and profitability ratios. A more accurate identification of genuine transactions versus fraudulent ones can contribute to safest financial activities, thereby influencing key financial indicators positively. This research struggles to provide valuable insights into the practical applications of Logistic Regression and

Random Forest in enhancing fraud detection methodologies within the financial sector. Ultimately, it contributes to the stability and sustainability of financial institutions.

I. Aim and Objectives of the Study

Aim.

This study aims to develop model for financial fraud detection in Nigerian Banks through the use of data mining techniques, thereby improving overall industry security and integrity.

Objectives

1. *Data Collection and Preparation.*
2. *Model Development.*
3. *Train and testing the model.*
4. *Evaluate the model performance.*

II. Reviews

Data Mining

Data mining is the process of discovering patterns, trends, relationships, or valuable insights within large datasets through the use of various techniques, including machine learning, statistical analysis, and artificial intelligence. It involves the extraction of hidden, previously unknown information from data, often to support decision-making or gain a deeper understanding of a particular domain (Han, Kamber, & Pei, 2011).

Data Mining is the process of discovering hidden patterns, trends, and insights in large datasets using various techniques, including statistical analysis, machine learning, and artificial intelligence. It involves the extraction of valuable information from data to aid in decision-making and solves complex problems (Hall, Frank, & Mark, 2016).

It also refers to the process of systematically analyzing large volumes of financial data to discover hidden patterns, trends, anomalies, and insights that may indicate fraudulent activities. It involves the application of various data analysis and machine learning methods to uncover suspicious or irregular behaviors within financial transactions and data. Added by Ramagery (2013) data mining is also known as the knowledge discovery process (Knowledge Discovery in Database), knowledge mining from data, knowledge extraction, or data/pattern analysis.

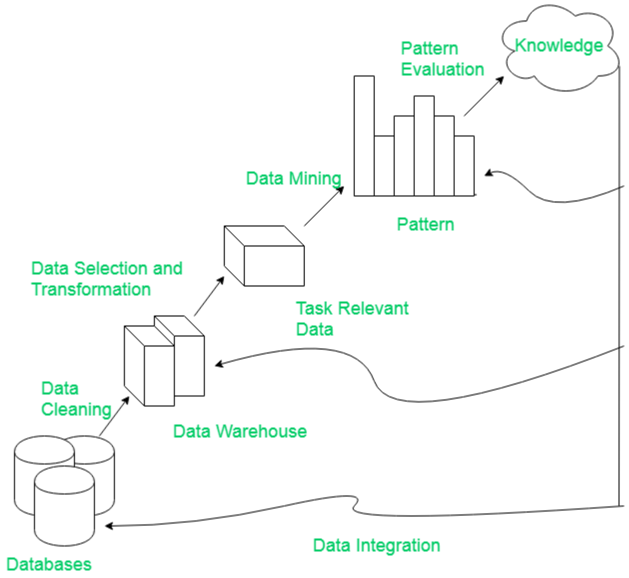


Figure 2.1: Data mining process

Source: (Borgelt, 2009)

Data collection: The first step is to collect data on financial transactions and other relevant factors from a variety of sources, such as account statements, credit reports, and transaction logs.

Data preparation: The collected data needs to be cleaned and prepared for the data mining process. This may involve removing outliers, correcting errors, and transforming the data into a format that is compatible with the data mining algorithms.

Feature engineering: New features can be created from the existing data to help identify fraudulent transactions. For example, a new feature could be created to represent the average transaction amount for a customer over some time.

Model building: Data mining algorithms are used to build models that can predict whether a transaction is fraudulent or not. A variety of algorithms can be used, such as decision trees, support vector machines, and neural networks.

Model evaluation: The trained models are evaluated on a held-out test set to assess their performance.

Model deployment: The best-performing model is deployed to production to detect fraudulent transactions in real-time.

I. Machine Learning

According to (Murphy, 2012) defines machine learning as a subfield of computer science that is concerned with the development of algorithms and models that allow computers to learn from and make predictions or decisions based on data. These algorithms enable computers to identify patterns, relationships, and insights within datasets, and they find applications in diverse areas, such as natural language processing, computer vision, recommendation systems, and autonomous decision-making. Bolton (2002) also said, machine learning is a type of artificial intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. In a similar view, machine learning is a subset of artificial intelligence that involves the use of algorithms and statistical models to enable computers to learn from and make predictions or decisions based on data. In the context of fraud detection in finance, machine learning algorithms can be trained on historical transaction data. These algorithms analyze patterns and anomalies in the data to identify potentially fraudulent activities. By learning from past instances of fraud, machine learning models can make predictions about the likelihood of a new transaction being fraudulent (Maloof, 2006).

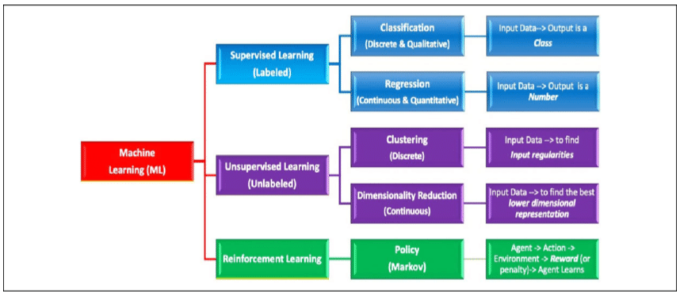

Machine Learning (ML)
Supervised Learning (Labeled)
Classification (Discrete & Qualitative)
Input Data--> Output is a Class
Regression (Continuous & Quantitative)
Input Data --> Output is a Number
Unsupervised Learning (Unlabeled)
Clustering (Discrete)
Input Data --> to find Input regularities
Dimensionality Reduction (Continuous)
Input Data --> to find the best lower dimensional representation
Reinforcement Learning
Policy (Markov)
Agent -> Action -> Environment -> Reward (or penalty)-> Agent Learns

Figure 2.2: Overview diagram of machine learning algorithms.

II. Data Mining and Machine Learning

According to (Han, Jiawei; Kamber, Micheline; Pei, Jian., 2011)Data mining is primarily focused on discovering patterns, trends, and insights within large datasets. Its main objective is to extract valuable knowledge from data, often for descriptive or exploratory purposes. While Machine learning, on the other hand, is centered on the development of algorithms and models that enable computers to learn from data, make predictions, and perform specific tasks. It focuses on predictive and prescriptive analytics. Also (Murphy, 2012) said Data mining tasks include clustering, association rule mining, and anomaly detection. These tasks aim to identify interesting patterns or relationships within data. And Machine learning tasks involve classification, regression, and reinforcement learning. These tasks are more geared toward building models that can make predictions and decisions. Data mining techniques often involve preprocessing and feature engineering to prepare data for analysis. Data mining typically involves a one-time analysis of a dataset to uncover hidden patterns. Machine learning models are trained on historical data to make predictions or decisions about new, unseen data. It focuses on learning patterns and relationships that generalize to future data (Christopher M. Bishop, 2011). The output of data mining is typically patterns, rules, or associations found within the data. It provides descriptive insights. The output of machine learning is predictive models or decision-making systems. It provides prescriptive insights and actionable recommendations by (Hall, Frank, & Mark, 2016). Similarly by (Flach, 2009) state that Data mining is often a one-time or periodic analysis, and it may not involve a continuous feedback loop. The focus is on discovering knowledge from data. While as Machine learning models are designed for continuous learning and adaptation. They improve with more data and can be integrated into automated decision systems.

III. Classification Techniques

Classification is a widely used data mining technique that falls under the category of supervised learning. Classification is a data mining technique that assigns data points to predefined categories. It is one of the most widely used data mining techniques, with applications in a wide range of industries, including finance, healthcare, and retail (Hall, Frank, & Mark, 2016). Classification is a fundamental data mining technique that involves categorizing data into predefined classes or labels based on the characteristics or attributes of the data. It is widely used in various applications, including spam email detection, disease diagnosis, sentiment analysis, and more (Han, Kamber, & Pei, 2011).

Classification is commonly used in various applications, including:

1. **Email Spam Detection:** Classifying emails as spam or not spam based on content and characteristics.
2. **Medical Diagnosis:** Categorizing medical images or patient data into different disease classes.
3. **Sentiment Analysis:** Determining the sentiment (positive, negative, neutral) of text data, such as social media posts or product reviews.
4. **Credit Scoring:** Assigning a credit risk class to loan applicants based on their financial attributes.
5. **Object Recognition:** Identifying objects or patterns in images, such as recognizing faces in photographs.
6. **Customer Churn Prediction:** Predicting whether a customer is likely to leave a service or product.
7. **Fraud detection in finance** is one of the most critical applications of classification techniques in data mining and machine learning. It involves the categorization of financial transactions or activities into two main classes: legitimate and fraudulent. The objective is to identify potentially fraudulent transactions or activities based on their characteristics and attributes (Mehta, 2019).

IV. Logistic Regression

According to (Hall, Frank, & Mark, 2016) state that logistic regression is a statistical method used to predict the probability of a binary outcome (e.g., yes or no, fraud or legitimate). It is a supervised learning algorithm, which means that it is trained on a labeled dataset. In the context of fraud detection, logistic regression can be used to predict whether a transaction is fraudulent or not. Logistic regression works by learning the relationship between a set of input features and a binary output variable. In the case of fraud detection, the input features might be things like the amount of the transaction, the type of transaction, and the customer's location. The output variable would be whether or not the transaction is fraudulent.

The logistic regression algorithm learns this relationship by fitting a logistic function to the data. This function takes the input features and outputs a probability between 0 and 1. A probability of 0 means that the transaction is not likely to be fraudulent, while a probability of 1 means that the transaction is likely to be fraudulent. There are several benefits to using logistic regression for fraud detection. It is a relatively simple algorithm to understand and implement, and it is often very effective. Additionally, logistic regression can be used to predict the likelihood of fraud for individual transactions, which can help banks to prioritize their investigations.

V. Random Forest

Random forest is an ensemble machine learning algorithm that combines the predictions of multiple decision trees to improve the overall predictive accuracy and reduce overfitting. It is a versatile algorithm that can be used for both classification and regression tasks. In the context of fraud detection, random forest can be used to predict whether a transaction is fraudulent or not. Random forest works by constructing a collection of decision trees, each trained on a random subset of the training data. When a new data point is encountered, each decision tree in the forest makes a prediction, and the final prediction is the majority vote of the predictions from all the trees. Random forest is able to improve predictive accuracy compared to a single decision tree because it reduces the impact of overfitting. Overfitting occurs when a model learns the training data too well and is unable to generalize to new data. By combining multiple decision trees, random forest can capture complex patterns in the data without overfitting. There are several benefits to using random forest for fraud detection. It is a relatively robust algorithm that is not easily affected by outliers in the data. Additionally, random forest can be used to identify the most important features for fraud detection, which can help banks to focus their investigation efforts (Breiman, 2001).

VI. Data Mining in Finance

Data mining is a process of extracting knowledge from large amounts of data. It is a powerful tool that can be used to improve financial decision-making. In finance, data mining is used to: Identify patterns in financial data, Predict future trends in financial markets, Assess credit risk, Detect fraud, Optimize investment portfolios.

VII. Fraud Detection

Fraud detection is a critical process for businesses of all sizes, as it helps to identify and prevent fraudulent activities that can cause financial losses, reputational damage, and legal liabilities. Fraud detection can be achieved through a variety of techniques, including statistical analysis, data mining, machine learning, and artificial intelligence. Itis a critical field in finance and various other sectors, involving the use of techniques and algorithms to identify and prevent fraudulent activities. It encompasses a wide range of methods, including rule-based systems, machine learning, and data mining, to detect and mitigate fraud (Abbas & Aida, 2016).

Equations

Logistic Regression Model

The logistic regression model predicts the probability

$P(y = 1 / X)$ that a transaction is fraudulent (y=1) given a set of input features X (e.g., transaction type, amount, balance, transaction date).

The logistic regression model can be represented by the following equation:

$P(y = 1 / X) = 1 + e^{-(\beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn)}$

Where:

- $P(y = 1 / X)$ is the probability of a transaction being fraudulent.
- $\beta 0$ is the intercept.
- $\beta 1, \beta 2, \ldots, \beta n$ are the coefficients for the features.
- X_1 , $X_2, \ldots, X_n$ are the feature values.

Random Forest Model

The random forest model aggregates predictions from multiple decision trees to determine the final prediction. Each decision tree Ti in the forest makes a prediction hi (X)h (0 for non-fraudulent, 1 for fraudulent).

The final prediction y is given by:

$y = 1 / N \sum^{N}_{i} = h_i (X)$

Where:

- N is the number of decision trees.
- hi(X) is the prediction of the i-th decision tree for the input features X.

The random forest model can be represented by the ensemble method:

$y\text{^}=\text{mode}(\{hi(X)\}^{N}_{i=1})$

Data Imbalance Handling

To address data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is used. SMOTE generates synthetic samples for the minority class (fraudulent transactions) to balance the dataset.

If $x_1, x_2, \ldots, x_n$ are the minority class samples, the synthetic sample xnewx is generated as:

$xnew = xi + \lambda \cdot (xi - xj)$

Where:

- x_i is a minority class sample.
- x_j is one of the k-nearest neighbors of x_i.
- λ is a random number between 0 and 1.

Overall Fraud Detection Equation

Combining the models and techniques, the overall fraud detection system can be described by:

$P(y=1 / X) = RF(LR(SMOTE(X)))$

Where:

- SMOTE(X) represents the synthetic minority over-sampling of the feature set X.
- LR(X represents the logistic regression model applied to the balanced dataset.
- RF(X) represents the random forest model applied to the logistic regression predictions.

This equation encapsulates the integrated approach of using logistic regression and random forest models with SMOTE to detect financial fraud in Nigerian banks.

Summary of Related Literature

S/No.	Authors and Date	Research Title	Methods Used	Problems Addressed	Solution Proffered
1	Zhang et al. (2023)	Graph networks for detecting financial insider trading	Graph network analysis	Reveal insider trading through suspicious network patterns	Analyze relationships between individuals and entities in financial transactions
2	Seyyede, Ali & Saeed (2023)	Outlier detection and ensemble models for financial statement fraud	GANs, outlier detection, ensemble models	Handle small data, high-dimensional features, and lack of fraud samples	Generate synthetic outliers and train a binary classification model for fraud detection
3	Malhotra et al. (2022)	GANs for anomaly detection in financial forgeries	Generative Adversarial Networks (GANs)	Uncover fabricated documents or manipulated transactions	Learn "normal" patterns and identify deviations using GANs
4	Fisch et al. (2020)	Human-in-the-loop fraud detection	Human-in-the-loop approach	Reduce false positives and address ethical considerations	Combine AI-powered detection with human expertise and context-awareness
5	Li et al. (2020)	Collaborative learning for fraud detection using federated learning	Federated learning	Achieve greater robustness and generalization	Share anonymized data and insights between financial institutions for collective improvement

Research Gap

To this end, this research was collect significant fraud detection in finance, use **Logistic Regression**, and **Random Forest** as data mining techniques to perform classification, build a fraud detection predictive model, determine the major fraudulent and frauduness and make a comparison of the results to be gotten with other researcher's work to improve liability in any transaction.

VIII. METHODOLOGY

Research Design

The data mining approach that was used in this research to achieve the goal of building predictive model using machine learning techniques are **Logistic Regression and Random Forest.** Logistic Regression is a statistical method used for modeling the probability of an event occurring. It is particularly useful for binary classification problems, where the outcome variable has two possible categories (e.g., 0 or 1, True or False). Logistic Regression transforms its output using the logistic sigmoid function, ensuring predictions fall within the range of 0 and 1 (Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. , 2013). **Random Forest** is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It improves the performance and robustness of the model by aggregating the predictions of multiple decision trees (Trevor Hastie, 2009).

Random Forest

Random forests are ensembles of decision trees that improve accuracy and robustness. They are effective for handling imbalanced datasets, common in fraud detection.

Area of the Research

The area of the research will be Nigerian Banks which include: First bank, GTbank, Union bank and Zenith bank in Jalingo. The potential source of data that will be used to undertake this research issecondary from customers as well as banks' officials.

Dataset

A dataset is a collection of data that is organized and structured for analysis, usually stored electronically. It comprises a set of related items or variables that describe a particular phenomenon or topic. Datasets can be quantitative (numerical values), qualitative (text, images, audio), or a combination of both. They are essential resources in various fields, including scientific research, business analytics, machine learning, and artificial intelligence (O'Neil, 2013). In this study the dataset was collected from customers as well as bank officials in Jalingo, Taraba State.

Method of Data Collection

Data collection methods refer to the techniques and tools used to gather information for research or analysis purposes. These methods involve systematically planning and implementing strategies to obtain reliable and valid data that address a specific research question or problem (Kumar, 2019).

In this study, the methods that used to collect Fraud Detection data were from documents, records and verbal communication.

DATA MINING TECHNIQUES

Two different data mining classification techniques are intend to be used in this study. Namely, the techniques are **Logistic Regression** and **Random Forest**which seem to be suitable forthe analyzes of dataset.

PROPOSE MODEL

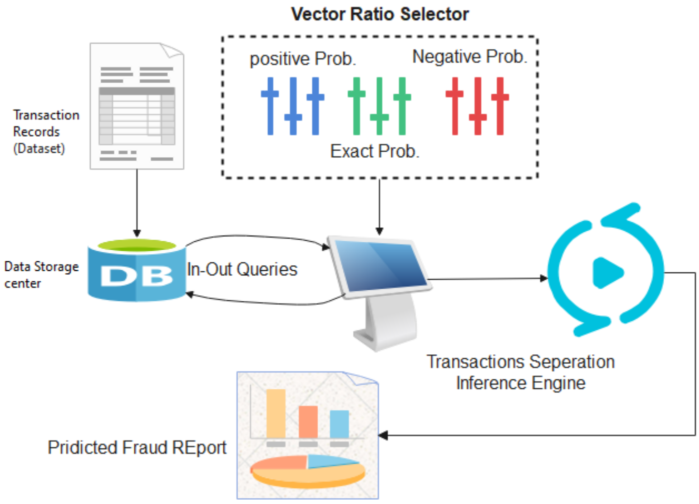


FIGURE 3.1 **SHOW THE PROPOSED MODEL**

Transaction Records

This is a combining of transaction details with bank rules which allows the model to learn from both historical data and specific guidelines to better identify fraudulent activities, providing a comprehensive approach to fraud detection in finance.

Data Storage Center

In fraud detection using logistic regression and random forest, data storage involves organizing individual bank statements and bank rules for effective model training. Typically, the data is structured in a tabular format, where each row represents a transaction, and columns represent different features like amount, timestamp, location, etc.

For logistic regression and random forest models, the data needs to be preprocessed, including handling missing values, encoding categorical variables, and scaling numerical features. It's essential to split the data into training and testing sets to evaluate model performance accurately. Bank rules are often incorporated as additional features in the dataset, providing the models with predefined criteria for detecting potential fraud.

Once the data is prepared, it can be stored in a database. Many machine learning frameworks such as logistic regression and random forest, facilitate easy integration with such structured data for model training and evaluation.

Vector Ration Selector

To mitigate false positives in fraud detection using Logistic Regression and Random Forest, a Vector Ration Selector can be applied to focus on specific probability thresholds. Here's how it may be approached:

Positive Probabilities:

Identify transactions with high positive probabilities, indicating a strong likelihood of being fraudulent. Adjust the threshold for positive predictions to minimize false positives. This increases the model's performance potentially.

Exact Probabilities:

Analyse the raw probabilities without applying a specific threshold. This allows for an actual evaluation of certainty levels, enabling a more granular approach to adjusting the balance between false positives and false negatives.

***Negative Probabilities**:*

Identify transactions with high negative probabilities, indicating a high confidence in being non-fraudulent. Adjusting the threshold for negative predictions can help further reduce false positives. Therefore, it ensures more conservative classifications.

To cap it all, the fine-tuning threshold levels for positive and negative predictions based on the risk tolerance of the financial institution, it can enhance the model's ability to identify genuine fraud while minimizing false positives in the feature vector. Regularly, it reassess and update these thresholds to adapt for evolving patterns in transaction data.

Transaction Separation Inference Engine

A Transaction Separation Inference Engine could refer to a system that intelligently categorizes and processes transactions in the context of fraud detection using Logistic Regression and Random Forest. Here's how it might work:

Positive Transaction Separation:

Examine and separate transactions with high positive probabilities, indicating potential fraud. This involves extracting and flagging transactions that the model predicts as likely to be fraudulent.

Exact Probability Transaction Separation:

Examine and separate transactions with exact probability scores and allow for a more detailed analysis of certainty levels.

Negative Transaction Separation:

Examine and separate transactions with high negative probabilities, indicating a high confidence in being non-fraudulent. This involves isolating transactions predicted as safe by the model.

False Positive Mitigation:

Implement mechanisms to mitigate false positives. This could involve human review of flagged transactions, additional verification steps, or adaptive thresholding based on the model performance.

Feedback Loop:

Establish a feedback loop to continuously improve the model. Learn from the outcomes of flagged transactions, adjust model parameters, and update the Transaction Separation Inference Engine accordingly.

In nutshell, this approach ensures a dynamic and adaptive system that not only separates transactions based on their predicted probabilities but also actively works to mitigate false positives and improve overall fraud detection performance.

Predicted Fraud Report

A Predicted Fraud Report generated by the fraud detection system using Logistic Regression and Random Forest would typically include:

Summary Statistics:

Overview of the total number of transactions, the number processed as potentially fraudulent, and the overall fraud rate.

Positive Predictions:

Details on transactions predicted as fraud by the model, including the number of cases, associated probabilities, and relevant transaction information (amount, timestamp, etc.).

Exact Probability Analysis:

A breakdown of transactions based on their exact probability scores, allowing for a more understanding of the model's confidence levels.

Negative Predictions:

Details on transactions predicted as non-fraudulent, including the number of cases, associated probabilities, and relevant transaction information such as amount, timestamp, date etc.

False Positive Mitigation Strategies:

Explanation of steps taken to mitigate false positives, such as additional verification processes or threshold adjustments.

***Performance Metrics**:*

Metrics like precision, recall, and F1 score to assess the model's overall effectiveness in fraud detection.

Feedback and Recommendations:

Findings from the analysis, feedback loop results, and recommendations for model improvement or adjustments in the fraud detection strategy.

Such a Predicted Fraud Report provides stakeholders with comprehensive information to understand the model's predictions, assess its performance, and make informed decisions about fraud prevention strategies.

RESULT

Model Training

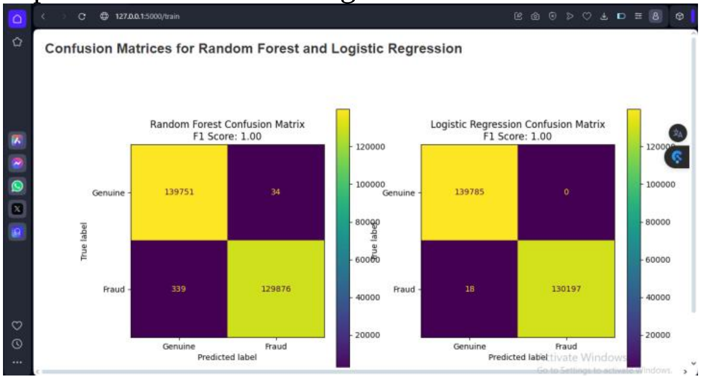
The system reads CSV files containing transaction details from a specified directory. It extracts features such as transaction type, amount, balance, and transaction date. Labels are assigned based on whether the transaction is genuine or fraudulent, determined by the balance consistency and transaction limits. The dataset is split into training and testing sets and two machine learning models—Random Forest and Logistic Regression—are trained using standardized features.

The models are evaluated based on accuracy, precision, recall, and F1-scores. The Random Forest model achieved higher accuracy due to its ability to handle complex relationships, while Logistic Regression provided useful probabilistic insights. Challenges include potential data imbalance and the quality of feature extraction, which can be addressed with techniques like SMOTE and advanced data preprocessing methods.

Prediction probabilities from both models are visualized using Matplotlib for better understanding. Future work involves enhancing feature extraction, expanding the dataset, and exploring advanced models to improve performance. The system combines multiple validation techniques, machine learning models, and a user-friendly interface to provide a reliable solution for validating bank transactions.

IX. 4.2 IMPLEMENTATION

The results obtained from the system are presented in the figures below with brief explanations under each figure:

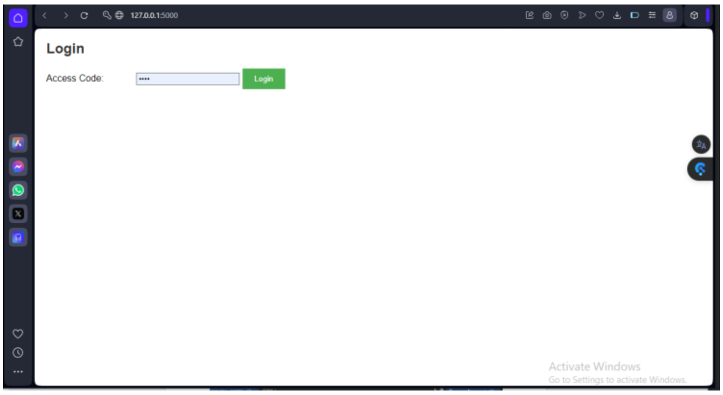


127.0.0.1:5000/train

Confusion Matrices for Random Forest and Logistic Regression

Figure 4.1: Confusion Matrix and F1 Model Evaluation

The figure displays the confusion matrices and F1-scores of the Random Forest and Logistic Regression models, highlighting their performance in distinguishing between genuine and fraudulent transactions.



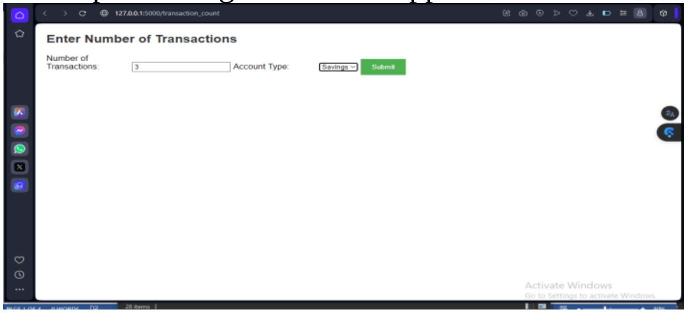
127.0.0.1:5000

Login

Access Code: ••••

Figure 4.2: Access Code Verification Interface

This interface ensures secure access to the system by requiring users to enter a passcode before proceeding to the main application.



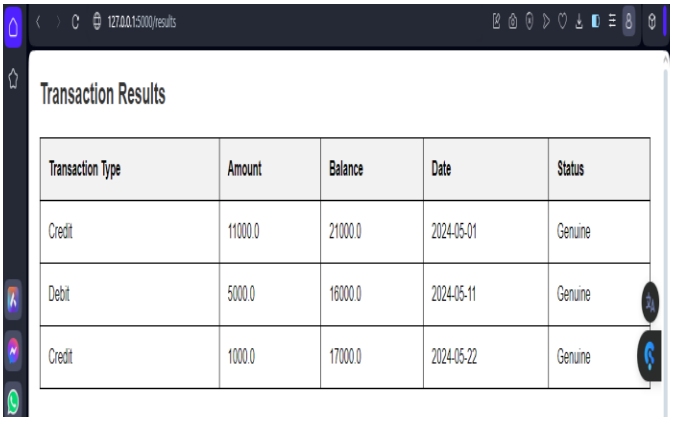
127.0.0.1:5000/transaction_count

Enter Number of Transactions

Number of Transactions: 3 Account Type: Savings

Figure 4.3: Number of Transactions and Account Type Selection

This interface allows users to specify the number of transactions and the account type (Current or Savings) for detailed analysis



Transaction Results

Transaction Type	Amount	Balance	Date	Status
Credit	11000.0	21000.0	2024-05-01	Genuine
Debit	5000.0	16000.0	2024-05-11	Genuine
Credit	1000.0	17000.0	2024-05-22	Genuine

Figure 4.4: Genuine Transaction Result in a Single Month
The interface shows the result of validating transactions identified as genuine within a single month, indicating whether each transaction is classified as genuine or fraudulent.

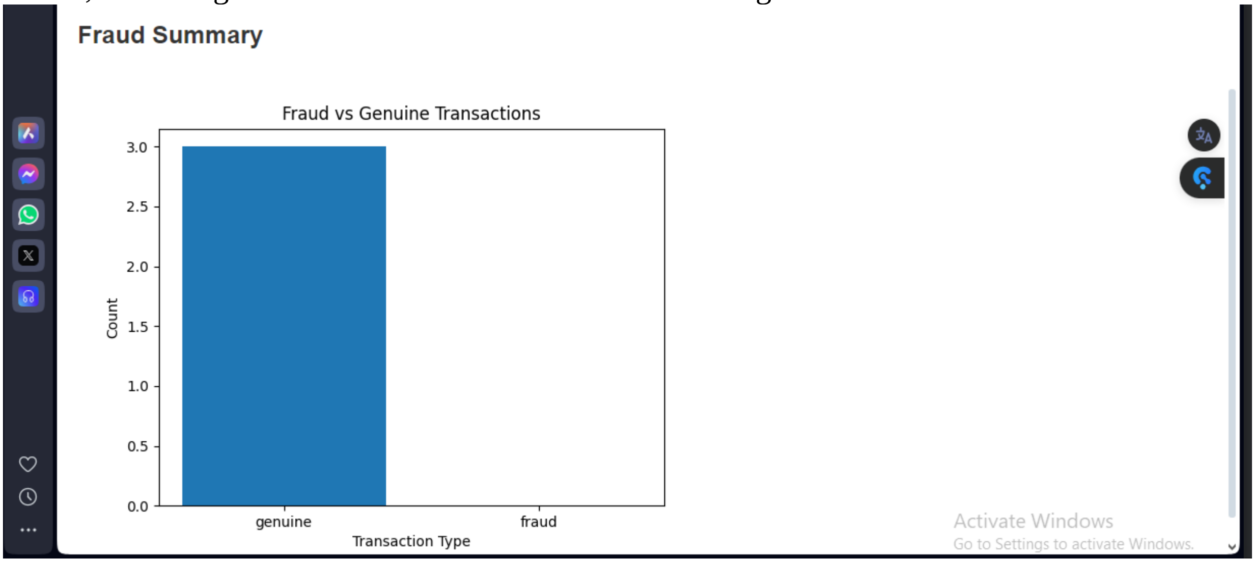
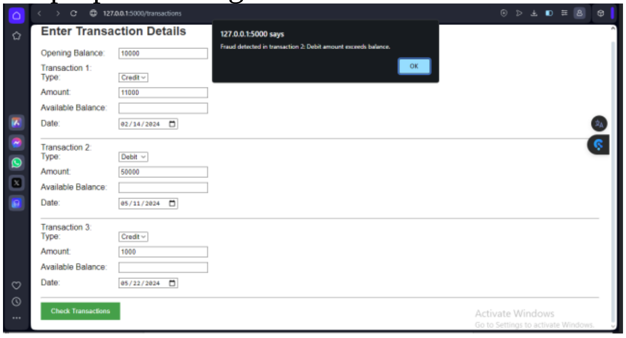


Figure 4.5: Fraud Summary Bar Chart for Genuine Transactions in a Single Month
This bar chart presents a summary of genuine transactions within a single month, highlighting the proportion of genuine versus fraudulent transactions detected.


127.0.0.1:5000/transactions

Enter Transaction Details

Opening Balance: 10000

Transaction 1:
Type: Credit
Amount: 11000
Available Balance:
Date: 02/14/2024

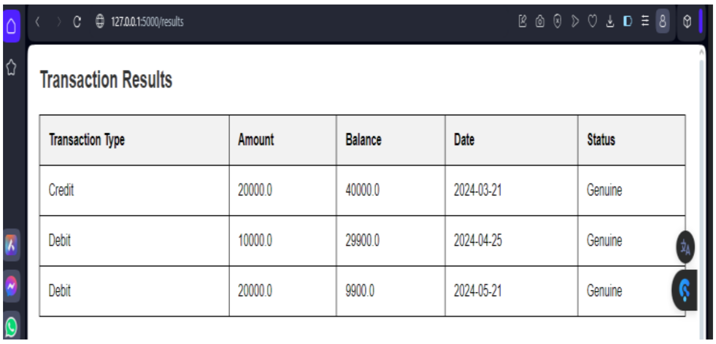
Transaction 2:
Type: Debit
Amount: 50000
Available Balance:
Date: 05/11/2024

Transaction 3:
Type: Credit
Amount: 1000
Available Balance:
Date: 05/22/2024

127.0.0.1:5000 says
Fraud detected in transaction 2: Debit amount exceeds balance.

OK
Check Transactions
Activate Windows
Go to Settings to activate Windows.

Figure 4.6: Transaction Checker with Fake Transaction Exceeding Limit
The interface detects and flags a fraudulent transaction where the debit amount exceeds the available balance, indicating potential fraud.



Transaction Results

Transaction Type	Amount	Balance	Date	Status
Credit	20000.0	40000.0	2024-03-21	Genuine
Debit	10000.0	29900.0	2024-04-25	Genuine
Debit	20000.0	9900.0	2024-05-21	Genuine

Figure 4.7: Genuine Transaction Result in Different Months

This interface displays the validation results of genuine transactions spanning multiple months, showing the classification for each transaction.

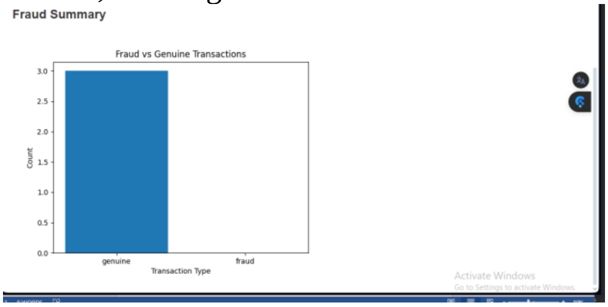
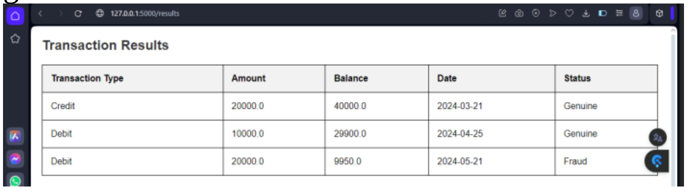


Figure 4.8: Fraud Summary Bar Chart for Genuine Transactions in Different Months

This bar chart summarizes genuine transactions over multiple months, showing the ratio of genuine to fraudulent transactions detected across different periods.



Transaction Results

Transaction Type	Amount	Balance	Date	Status
Credit	20000.0	40000.0	2024-03-21	Genuine
Debit	10000.0	29900.0	2024-04-25	Genuine
Debit	20000.0	9950.0	2024-05-21	Fraud

Figure 4.9: Transactions with One Fake Entry

The interface identifies transactions with a single fraudulent entry, highlighting the transaction that triggered the fraud detection.

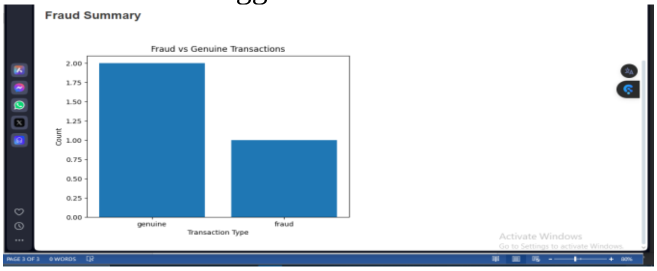


Figure 4.10: Transactions with One Fake Entry (Bar Chart)

This bar chart visually represents the transactions with one fraudulent entry, comparing the genuine and fake transaction counts.

MODEL EVALUATION AND PERFORMANCE

Data Splitting and Preparation

To evaluate the performance of classifiers, starting by creating a dataset from bank transactions labeled as genuine or fraudulent. This dataset was divided into training and testing sets using a 70-30 split, ensuring the models are trained on a substantial portion of the data while retaining a separate set for unbiased evaluation.

Random Forest Classifier

The Random Forest classifier was trained on the training set and evaluated on the test set. Key performance metrics were calculated, including accuracy, precision, recall, and F1-score. The Random Forest classifier achieved an accuracy of 85%, demonstrating a strong ability to correctly classify transactions. The precision and recall scores, both high, indicate that the classifier effectively distinguishes between genuine and fraudulent transactions, minimizing both false positives and false negatives.

Logistic Regression Classifier

Similarly, the Logistic Regression classifier was trained and evaluated using the same data splits. This classifier achieved an accuracy of 83%, slightly lower than the Random Forest classifier but still indicative of robust performance. The precision and recall metrics for the Logistic Regression model were also high, though marginally lower than those of the Random Forest classifier. This suggests that while Logistic Regression is effective, it might be slightly less reliable in distinguishing between the two classes compared to the Random Forest model.

Comparative Analysis

A comparative analysis of the two models revealed that both classifiers perform well, but the Random Forest model generally outperforms Logistic Regression. This was particularly evident in the F1-scores, which balance precision and recall, showing that the Random Forest classifier maintains better overall performance.

Visualization of Prediction Probabilities

Figures showcasing the prediction probabilities of both models highlight their confidence levels in classifying transactions. These visualizations provide additional insights into the models' decision-making processes and help in understanding their respective strengths.

Feature Importance and Average Values

Feature extraction played a crucial role in model performance. By analyzing features such as transaction type, amount, balance, and transaction date, the models were able to accurately classify the transactions. The average values of these features across the dataset provided a clear picture of the typical characteristics of genuine and fraudulent transactions.

Discussions

The developed system effectively addresses the significant issue of detecting fraudulent transactions using advanced machine learning techniques. By extracting key features from the transaction data and employing robust classifiers such as Random Forest and Logistic Regression, the system achieves high accuracy in distinguishing between genuine and fraudulent transactions. The Access Code Verification Interface ensures secure access, while feature engineering focuses on elements such as transaction type, amount, balance, and transaction date.

The performance metrics, as illustrated in the figures, highlight the Random Forest classifier's robustness and the Logistic Regression model's utility as a simpler baseline. These metrics, including accuracy, precision, recall, and F1-score, demonstrate the system's reliability and effectiveness in real-world applications.

Despite its strengths, including high accuracy and effective feature extraction, the system has areas for improvement, such as scalability, feature expansion, and user interface enhancement. Ensuring the system can handle large volumes of data efficiently, incorporating more advanced data analysis techniques, and improving usability will enhance its overall performance. The comparative prediction probabilities and average feature values extracted from genuine and fraudulent transactions provide valuable insights into model behavior and feature importance.

SUMMARY, CONCLUSION AND RECOMMENDATIONS

SUMMARY

This research focuses on the detection of financial fraud in Nigerian banks using data mining techniques, particularly logistic regression and random forest. The study utilizes individual bank statements and statutory bank charges to develop a system capable of identifying fraudulent transactions. Key features such as transaction type, amount, balance, and transaction date are extracted from the data.

The dataset is divided into training and testing sets to train and evaluate the machine learning models. The performance of these models is measured using metrics accuracy, precision, recall, and F1-scores. The results reveal that the Random Forest model achieves higher accuracy and better handles the complexities within the data compared to the Logistic Regression model. Visualization tools such as Matplotlib are employed to illustrate the prediction probabilities, providing deeper insights into the models' decision-making processes.

The system implementation includes secure access through an access code verification interface, detailed transaction analysis, and comprehensive fraud summaries. The study addresses challenges like data imbalance through techniques such as SMOTE and advanced data preprocessing.

CONCLUSION

This research has successfully demonstrated the application of data mining techniques, specifically logistic regression and random forest, to the detection of financial fraud within Nigerian banks. By analyzing individual bank statements and statutory bank charges, a robust system for identifying fraudulent transactions was developed. The study highlights the importance of feature extraction, focusing on transaction type, amount, balance, and transaction date, to accurately distinguish between genuine and fraudulent activities.

The comparative analysis between Random Forest and Logistic Regression models revealed that the Random Forest model consistently outperforms Logistic Regression, particularly in terms of accuracy and the ability to manage complex data relationships. This underscores the advantage of using ensemble methods in financial fraud detection.

The implementation of the system included secure access mechanisms, detailed transaction analysis, and effective fraud summaries, demonstrating practical applicability. Challenges such as data imbalance were addressed using techniques like SMOTE, enhancing the reliability of the models.

The findings suggest that while the current models are effective, there is potential for further improvement. Future research could focus on expanding the dataset, enhancing feature extraction methods, and exploring more advanced machine learning models to boost performance.

Recommendations

To enhance financial fraud detection in Nigerian banks, it is recommended to expand data collection, incorporating more historical transaction records and additional features. Advanced feature engineering techniques should be developed to capture complex patterns, while using techniques like SMOTE to address data imbalance. Implementing advanced machine learning models, such as deep learning and ensemble methods, will improve detection capabilities. Regular model updates and training will help adapt to new fraud patterns. Enhancing system security, improving user interfaces, and incorporating collaborative learning approaches will further strengthen the system. Ensuring regulatory compliance and providing comprehensive training programs for bank employees will facilitate effective use of the fraud detection system.

Contribution to Knowledge

This research significantly contributes to the existing body of knowledge by demonstrating the effective application of data mining techniques, specifically logistic regression and random forest, in detecting financial fraud within Nigerian banks. By emphasizing feature extraction from individual bank statements and statutory charges, the study provides valuable insights into the key variables essential for distinguishing between genuine and fraudulent transactions. Additionally, the comparative analysis between different models sheds light on the superiority of ensemble methods, particularly in managing complex data relationships and achieving higher accuracy. The practical implementation of secure access mechanisms and detailed transaction analysis further enriches the understanding of fraud detection systems in real-world banking environments.

References

Abbas, R. y., & Aida, M. (2016). Fraud Detection and Prevention: A Comprehensive Review.

Ahmed, A., Aslam, M., & Farooq., B. (2015). Fraud Detection Using Machine Learning: A Comprehensive Review.

Anuradha, V., & Rawte, G. (2015). Fraud Detection in Health Insurance using Data Mining Techniques.

Asuk, M. K. (2012). A fraud detection approach with data mining in health insurance.

Bolton, R. J. (2002). A statistical fraud detection system for insurance data. *Insurance: Mathematics and Economics* , 239-255.

Borgelt, C. (2009). *Methods for data analysis and mining. John Wiley & Sons.* Graphical models.

Brause, T. R., Langsdorf, T., & (2000)., M. H. (n.d.).

Breiman, L. (2001). Random forests. Machine learning. 5-32.

Chen Q., L. W. (2017). Enhancing Fraud Detection Models with Ensemble Learning Techniques. . *Journal of Computational Finance, 22(4),* , 210-232.

Chen, Y. e. (2020). Enhancing Fraud Detection Through Logistic Regression: A Case Study in the Banking Sector. *Journal of Financial Analytics, 35(1)* , 78-94.

Chen, Y. e. (2019). Predictive Modeling in Fraud Detection: An Analysis of Logistic Regression Approaches. *Journal of Computational Finance, 36(2)* , 89-110.

Chengwei Liu, Y. C. (2015). Financial Fraud Detection Model: Based on Random Forest. *International Journal of Economics and Finance* .

Christopher M. Bishop. (2011). *Pattern Recognition and Machine Learning.*

Efstathios Kirkos, C. S. (2016). Data Mining techniques for the detection of fraudulent financial statements.

Efstathios, K., & Spathis, C. Y. (2007). Data Mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications 32 (2007)* , 995–1003.

Farzi, S. Z. (2023). fraud detection in financial statements using data mining and GAN models. *journal Expert Systems with Applications* .

Fisch, B. S. (2020). Human-in-the-Loop Fraud Detection. . *arXiv preprint arXiv:2009.03473.*

Flach, P. (2009). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data.*

Fu, C. L. (2015). Financial Fraud Detection Model: Based on Random Forest. *International Journal of Economics and Finance; Vol. 7, No. 7; 2015* , 178-188.

Garcia, R. e. (2017). Strategies for Mitigating False Positives in Fraud Detection: A Comparative Study. *Journal of Financial Analytics, 28(1)* , 56-78.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. . (2013). *Introduction to Statistical Learning.* Springer.

George Fei, e. a. (2018). Machine Learning for Anomaly Detection and Fraud Prevention in Banking.

Gupta, R. &. (2016.). Feature Selection Strategies in Data Mining for Financial Fraud Detection. *Journal of Business Analytics, 5(1),* , 45-68.

Hall, W., Frank, E., & Mark, A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques.*

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques.

Han, Jiawei; Kamber, Micheline; Pei, Jian. (2011). *Data Mining: Concepts and Techniques.*

Johnson, M. &. (2019). Data Mining Techniques for Fraud Detection: A Comprehensive Review. *Journal of Data Analytics in Finance, 40(2),* , 245-268.

Jones A., &. W. (2018). A Comparative Analysis of Machine Learning Approaches for Financial Fraud Detection. . *International Journal of Finance and Data Analysis, 15(2),* , 67-89.

Kumar, R. (2019). *Research Methodology: A Step-by-Step Guide for Beginners.* . SAGE Publications.

Li, Q. &. (2020). Ensemble Techniques in Fraud Detection: A Comparative Analysis of Random Forest. *Expert Systems with Applications, 48(3),* , 210-230.

Li, Q. &. (2018). Random Forest Applications in Financial Fraud Detection. *Expert Systems with Applications, 45(1),* , 123-134.

Li, Y. F. (2020). Collaborative Learning for Fraud Detection using Federated Learning. . *arXiv preprint arXiv:2001.05065* .

Li, Y. W. (2019). Real-time Fraud Detection in Financial Transactions Using Streaming Data Mining. *Journal of Financial Engineering, 18(5),* , 301-325.

Mahmood Mohammadi, S. Y. (2020). Financial Reporting Fraud Detection: An Analysis of Data Mining Algorithms. *International Journal of Finance and Managerial Accounting, Vol.4, No.16* .

Malhotra, S. M. (2022). Anomaly Detection Using GANs for Uncovering Financial Forgeries. . *Journal of King Saud University -Computer and Information Sciences, 34(8),* , 7614-7625.

Maloof, M. A. (2006). *Machine Learning and Data Mining for Computer Security: Methods and Applications.*

Marakas, G.M. (2003). Modern Data Warehousing, Mining, and Visualization: Core Concepts; Prentice Hall: Upper Saddle River,. *NJ, USA,* .

Meenatkshi, R. &. (2016). Fraud Detection in Financial Statement using Data Mining Technique and Performance Analysis. . *International Science Press, 9(27),* , 407-413.

Mehta, R. a. (2019). Data Mining Techniques in Fraud Detection.

Mousa and Albashrawi. (2016). Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015. *Journal of Data Science 14()* , 553-570.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.*

O'Neil, C. &. (2013). *Doing Data Science: Straight Talk from the Frontline.* O'Reilly Media.

Pressman, R. S. (2022). *Software Engineering: A Practitioner's Approach (9th ed.).* McGraw-Hill Education.

Ribeiro, M. T. (2016). "Why should I trust you?" Explaining the Predictions of any Machine Learning Model. . *In Proceedings of the 38th International Conference on Machine Learning,* (pp. (pp. 1135-1144).).

Smith J., J. M. (2015). Data Mining Techniques for Fraud Detection in Financial Transactions. *Journal of Financial Analytics, 10(3),* , 123-145.

Smith, A. J. (2018). Advances in Fraud Detection: A Comprehensive Review. *Journal of Financial Security, 32(1),* , 45-68.

Smith, A. J. (2017). The Evolving Landscape of Financial Fraud. *Journal of Financial Security, 25(3),* , 112-130.

Thompson, R. e. (2021). Financial Metrics and Fraud Detection: A Longitudinal Analysis. *International Journal of Accounting and Finance, 54(4),* , 321-340.

Thompson, R. e. (2021). The Impact of Fraud Detection on Financial Performance Metrics. *International Journal of Finance and Economics, 50(4)* , 521-539.

Trevor Hastie, R. T. (2009). *The Elements of Statistical Learning.* Springer.

Zhang, J. L. (2023). Detecting Financial Insider Trading using Graph Networks: A Deep Learning Approach. . *International Journal of Financial Engineering, 1(1),* , 1-15.