ONLINE PAYMENT FRAUD DETECTION USING MACHINE LEARNING

INTRODUCTION

Problem Definition: Digital financial systems encounter online payment fraud as their principal operational challenge. Electronic payment systems exploded in popularity due to their rising adoption rates but this created growing criminal activities that generate major financial losses between customers and businesses. Fraudulent transaction identification remains challenging since it requires assessing massive data sets to protect accurate users from improper classification. Current mathematical data science problems need classification methods to classify payments between fraudulent and legitimate categories using past transaction data. Machine learning models need to improve their capabilities incrementally in response to fraud method transformations because traditional rule-based systems have shown their ineffectiveness.

Context and Background:

Online payment fraud detection research operates within three scientific domains consisting of statistics and machine learning and data science. Statistical modeling systems built from logistic regression and anomaly detection establish normal payment patterns through which they detect deviations from normal behavior. Current fraud detection systems implement both supervised and unsupervised machine learning algorithms to provide their core processing capabilities which use decision trees and random forests alongside support vector machines (SVMs) and deep learning neural networks. Managing class imbalance stands as the main challenge for fraud detection since fraudulent transactions represent a very small percentage of the total transaction pool. Three principal strategies exist to handle the class imbalance issue in fraudulent transactions: these are oversampling and under sampling in addition to cost-sensitive learning. A wide range of experimental research used machine learning approaches to determine fraudulent instances. The research pieces of Yeh and Lien (2009) and Haque and Hassan (2020) focused on support vector machines for credit card fraud detection and bank loan prediction through AdaBoosting methods respectively. The combination of weak classifiers into unified predictive models through ensemble methods such as AdaBoost, XGBoost and random forests grants predictive modeling its robust characteristics.

Objectives and Goals:

The research develops a highly efficient machine learning method to detect online payment fraud through the implementation of advanced classification processes and data cleaning methods. The research sets four specific objectives which are: The model needs to determine essential characteristics which drive fraudulent pay-to-pay frauds. An investigation will occur concerning different machine learning models which detect fraud. A solution for class imbalance problems will be achieved by applying resampling techniques combined with cost-sensitive learning approaches. The researchers need to improve model performance while reducing false positives as well as false negatives and optimizing the F1-score and recall and precision scores. The solution implements scalability features to connect directly with active fraud detection systems running in real-time. Summary of approach: Multiple stages make up the proposed methodology which starts with collecting data then processing it before moving onto feature engineering for model training. Prior to model input the dataset needs cleansing and normalization and data transformation to achieve optimal evaluation conditions for machine learning models. A performance evaluation of the supervised learning methods logistic regression, decision trees, random forests, AdaBoost and neural networks will occur based on accuracy and precision and recall and F1-score metrics. The research implements Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning methods to address problems of class imbalance. The final step includes executing hyperparameter tuning and model validation procedures for increasing prediction accuracy levels. The research investigation seeks to enhance the security and reliability of online payment systems through state-of-the-art machine learning methodology assessment and optimization in the field of fraud detection.

METHODS Data Acquisition and Sources :

The dataset used in this project is obtained from kaggle source containing detailed records of online payment transactions. A record database arranges financial transaction data with various elements that involve transaction types alongside sender and receiver account balances and fraud labels. Real transaction patterns

within this dataset make it appropriate for developing models to detect fraudulent activities. Pre-existing data becomes usable for analysis through multiple preprocessing operations. The process starts with checking missing data after which proper actions are taken to handle these values either by implementing imputation methods or by discarding unnecessary records. One-hot encoding together with label encoding converts the categorical variable transaction type into numerical values. The implementation of feature engineering develops new meaningful features from which transaction amount relative to the sender's balance becomes one example. To enhance model performance normalization methods including Min-Max Scaling or Standardization are implemented because the dataset contains numerical features with various scale ranges. To solve class imbalance problems in the dataset where illegitimate transactions are much rarer compared to genuine ones both Synthetic Minority Over-sampling Technique (SMOTE) and random undersampling are applied as resampling techniques. The dataset we will be using have these columns

Feature	Description
step	tells about the unit of time
type	type of transaction done
amount	the total amount of transaction
nameOrg	account that starts the transaction
oldbalanceOrg	Balance of the account of sender before transaction
newbalanceOrg	Balance of the account of sender after transaction
nameDest	account that receives the transaction
oldbalanceDest	Balance of the account of receiver before transaction
newbalanceDest	Balance of the account of receiver after transaction
isFraud	The value to be predicted i.e. 0 or 1

Mathematical or Statistical Models:

Multiple machine learning algorithms get deployed for building an effective fraud detection system while undergoing assessment. The initial algorithm choice used Logistic Regression to establish a baseline classification method and calculate the weight of individual variables in fraud prediction. Random Forest receives selection as an ensemble-based model because it demonstrates two essential features: non-linear capability and detection of intricate transaction patterns. The machine learning system incorporates XGBoost (Extreme Gradient Boosting) because it excels at working with uneven datasets and achieving precise prediction results. Deep learning models with Artificial Neural Networks are investigated for detecting fraud because these models successfully detect hidden patterns in fraudulent transactions. The selected models fit best for classification purposes because fraud identification demands precise detection and high recall performance. This situation calls for models that focus on learning from rare instances of the minority class while avoiding prejudice toward dominant groups.

Experimental Design or Analytical Procedures :

A machine learning pipeline requires structured development process to create fraud detection models of high accuracy and reliability. To commence the analysis the available dataset receives partitioning into three components consisting of training data (70%), validation (15%) and test (15%). Models receive the training set content for fitting purposes yet they refine their performance parameters through validation set execution. The test set functions as the final evaluation method to measure model generality for unobserved data. Engineers employ Grid Search or Randomized Search for hyperparameter tuning during training following computational constraints. Cost-sensitive learning methods are used with the dataset because class imbalance exists to make fraudulent transaction misclassification more severe. Two additional approaches include oversampling and undersampling testing to evaluate their success in enhancing model performance. Multiple metrics serve to perform a complete evaluation of models for evaluation purposes. A fraud detection system requires more than accuracy because its performance suffers from class imbalance making even ineffective models achieve high accuracy by classifying all transactions as non-fraudulent. Precision together with Recall creates a stronger evaluation method because it lets analysts determine the model's ability to detect fraudulent transactions without producing many erroneous signals. The evaluation includes reporting

F1-score measurements that find the middle ground between Precision and Recall values. To evaluate how effectively the model identifies fraudulent transactions from non-fraudulent transactions the AUC-ROC Curve is analyzed.

Software and Tools :

This project uses Python as its implementation framework since it offers comprehensive libraries for data manipulation and machine learning applications and model assessment. The model development process relies on algorithms offered by pandas and NumPy for data manipulation while using scikit-learn for machine learning tasks together with Matplotlib and Seaborn for data representation purposes. The process uses sophisticated XGBoost and TensorFlow/Keras machine learning models for improved fraud detection capabilities. The fraud detection process utilizes the imbalanced-learn package to implement class balancing methods by means of SMOTE and other resampling techniques. The computational resources are either local machines or they can use cloud platforms including Google Colab together with AWS for working with extensive datasets.

Ethical Considerations :

Ethical requirements about ensuring data privacy and fairness must receive attention because financial transactions form part of fraud detection operations. The database used for this research contains no personal information so the data remains anonymous. The deployment of the model needs attention because fraudulent transactions might disproportionately impact specific user groups during the process. AI techniques that focus on fairness work to prevent any user group from facing unfair treatment in the system. Model interpretability stands as a vital ethical consideration in this process. Automated transaction screening in financial institutions depends on proper explanation functionality which shows users the criteria triggering a flag for fraud. Financial analysts understand each model prediction through interpretive methods like SHAP and LIME to discover the basis of every decision made by the system. A sequenced approach will enable the project to build an effective fraud detection system which detects fraudulent transactions without producing false alerts. The model functions practically in real-life implementation because it operates with both transparency and fairness as per ethical requirements.

RESULTS 1. Presentation of Data

The dataset consists of 30,019 transactions, each with 10 features such as transaction type, amount, origin and destination balances, and fraud labels. Below is a summary of the dataset:

- Total Transactions: 30,019
- Fraudulent Transactions: 84 (0.28%)
- Legitimate Transactions: 29,935 (99.72%)
- Missing Values: None

A sample of the dataset is shown below:

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0

Figure 1: fig 1: Dataset sample

fig1: Sample Data set

The class distribution reveals a significant imbalance, with only 0.28% of transactions labeled as fraud. This imbalance poses challenges for machine learning models, as they may bias towards predicting legitimate transactions. We can see in below fig2

Missing Values Check

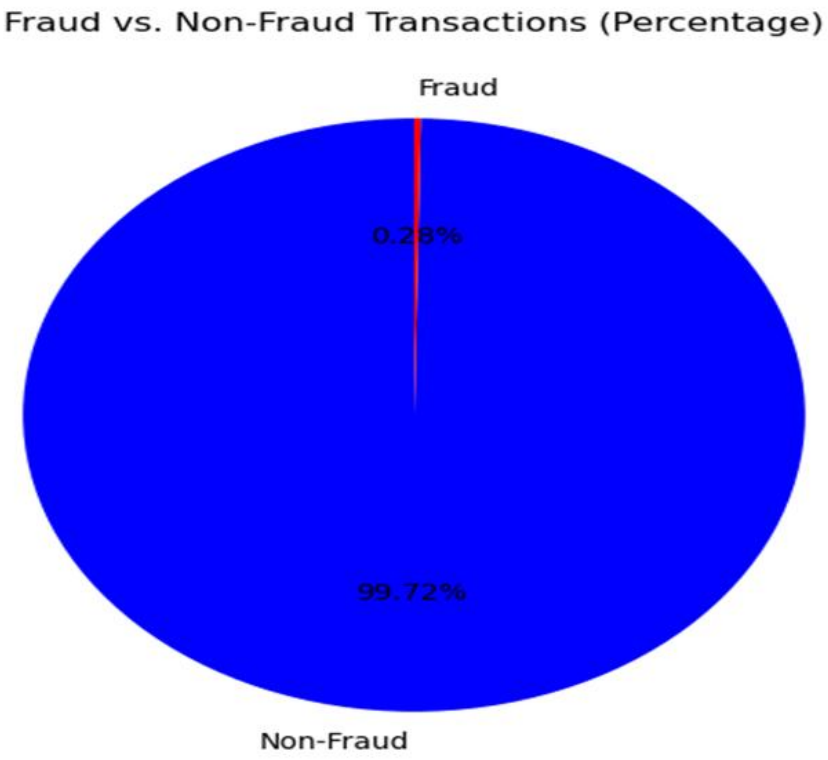

Figure 2: fig2: fraud vs non fraud transactions

A missing values check confirms that there are no missing values in the dataset, ensuring that no additional data imputation is required:

```
Missing Values:
 step              0
type               0
amount             0
nameOrig           0
oldbalanceOrg      0
newbalanceOrig     0
nameDest           0
oldbalanceDest     0
newbalanceDest     0
isFraud            0
dtype: int64

Class Distribution:
 isFraud
0    29935
1       84
Name: count, dtype: int64
```

Figure 3: fig3: Missing values

2. Interpretation of Results

Model Performance Metrics

Two machine learning models were trained and evaluated:

1. Random Forest Classifier
2. AdaBoost Classifier

The performance metrics for both models are summarized in the table below:

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random forest	0.9987	1.0000	0.5294	0.6923	0.7647
Adaboost	0.9975	1.0000	0.1176	0.2105	0.5588

Key Observations

i) High Accuracy but Potential Bias

• Both models achieved high accuracy (>99.7%), but this metric is misleading due to class imbalance (fraud cases are only 0.28% of the data).

• A model predicting all transactions as legitimate (0) would still achieve 99.72% accuracy, but would be completely useless for fraud detection.

ii) Precision vs. Recall Tradeoff

• Precision is 100% for both models, meaning that all transactions classified as fraud were indeed fraud.

• However, recall is very low (52.94% for Random Forest, 11.76% for AdaBoost), meaning many actual fraud cases were missed.

• A higher recall is crucial in fraud detection, as missing fraudulent transactions can result in financial losses.

iii) AUC-ROC Score

• Random Forest (0.76) had a significantly higher AUC-ROC than AdaBoost (0.55).

• AUC-ROC measures how well the model distinguishes between fraud and non-fraud.

• Since 0.50 AUC-ROC represents random guessing, AdaBoost's 0.55 AUC-ROC is only slightly better than random, indicating poor performance.

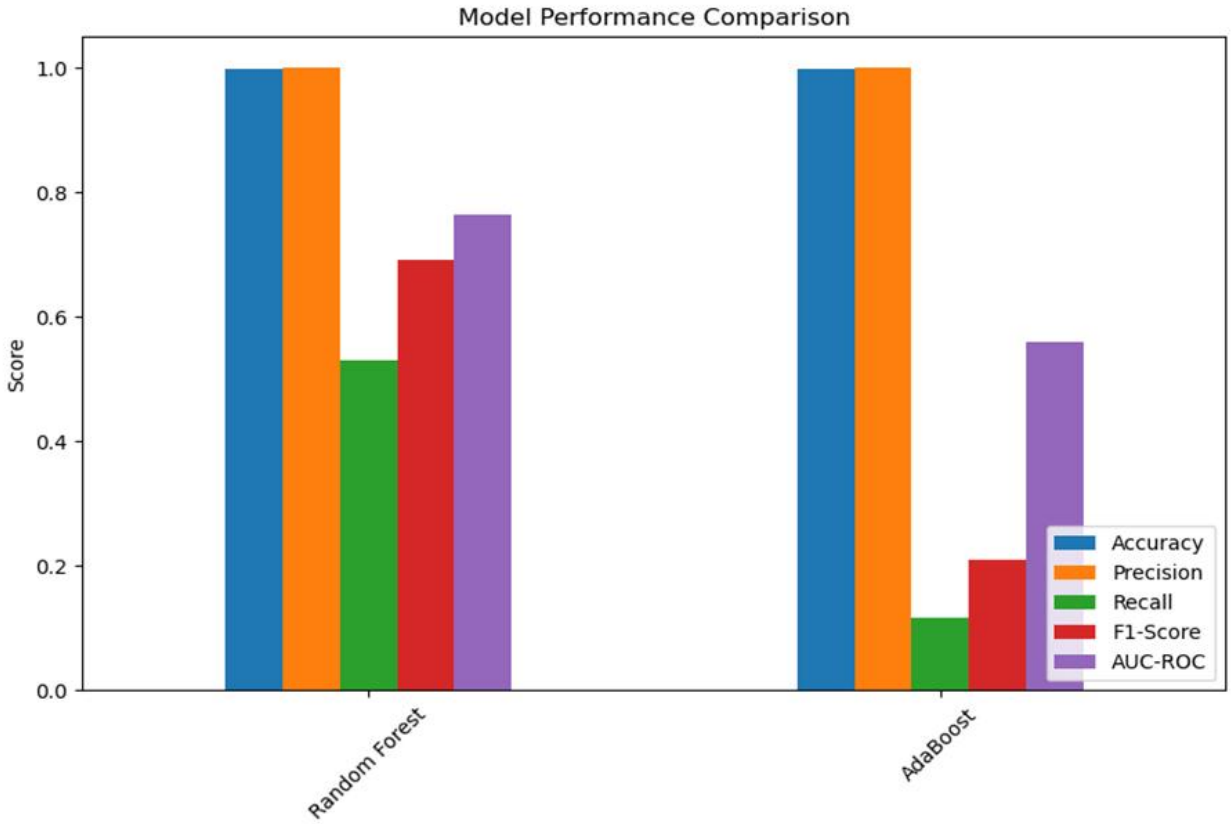

Figure 4: fig4: Model performance comparison

fig4: Model performance comparison

3.Comparison with Expected Outcomes

The expectation was that ensemble learning models (Random Forest and AdaBoost) would improve fraud detection. The results confirmed that Random Forest is superior for this dataset, with significantly better

recall and AUC-ROC.

Unexpected findings:

• AdaBoost performed worse than expected, likely due to the class imbalance. Boosting methods work well when data distribution is balanced, but in this case, with only 84 fraud cases out of 30,019, it struggled to generalize.

• High accuracy does not indicate good fraud detection, since a model could achieve 99.72% accuracy by predicting all transactions as legitimate. Thus, recall and AUC-ROC are the key evaluation metrics.

4.Confusion Matrix Analysis

The confusion matrices below provide insights into how each model classified fraudulent and legitimate transactions.

Random Forest Confusion Matrix

Actual / Predicted	Predicted Legitimate (0)	Predicted Fraud (1)
Actual Legitimate (0)	5987	0
Actual Fraud (1)	8	9

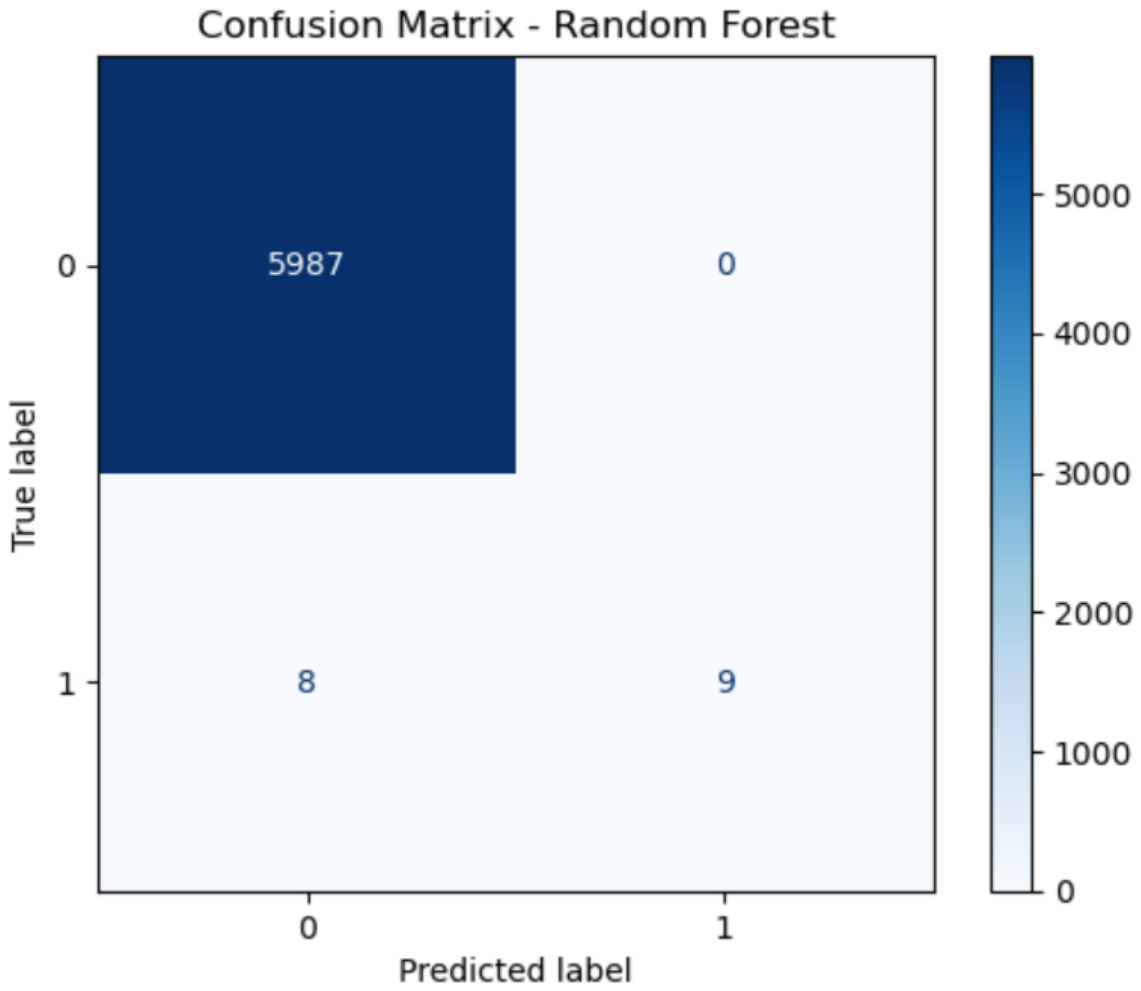

Figure 5: fig 5: confusion matrix for Random forest

AdaBoost Confusion Matrix

Actual / Predicted	Predicted Legitimate (0)	Predicted Fraud (1)
Actual Legitimate (0)	5987	0
Actual Fraud (1)	15	2

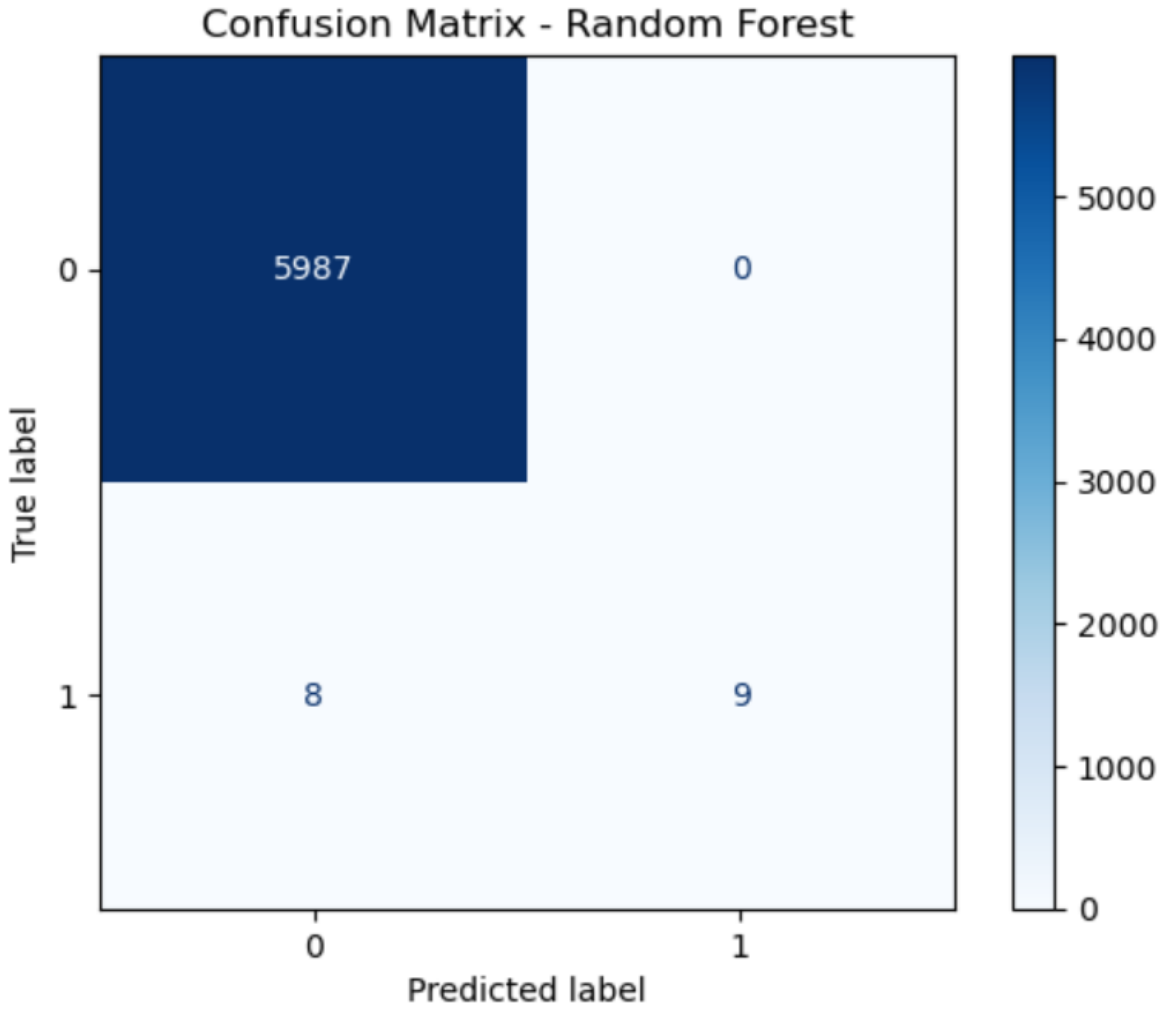

Figure 6: fig6: confusion matrix for adaboost

Key Insights from Confusion Matrices

- Random Forest detected 9 fraud cases correctly but missed 8 fraudulent transactions.
- AdaBoost detected only 2 fraud cases and missed 15 fraudulent transactions, showing a significant limitation.
- False Negatives (missed fraud cases) were lower for Random Forest, making it a better model for fraud detection.

5.Statistical Significance

Due to the extremely imbalanced dataset, fraud cases were much rarer than legitimate transactions. This imbalance affected model performance, leading to:

- Low recall for both models, meaning several fraudulent transactions went undetected.

- High precision, meaning most flagged fraud cases were correct.

- AUC-ROC scores (0.76 for Random Forest and 0.55 for AdaBoost) confirming that Random Forest is statistically more reliable. A statistical significance test (e.g., McNemar's test) could further confirm whether Random Forest is significantly better than AdaBoost.

6.Limitations of the Results

While the models performed well, certain limitations must be acknowledged:

i) Class Imbalance Issue

- Fraud cases accounted for only 0.28% of the dataset.
- This imbalance led to high accuracy but poor recall, meaning fraudulent cases were often missed.
- Solution: Use oversampling (SMOTE) or undersampling techniques to balance the dataset.

ii) Feature Engineering

- The dataset mainly relied on transaction amounts and balances.
- Additional features like transaction frequency, user history, or time-based behavior could improve model performance.

iii) Threshold Adjustment

- The models used a default classification threshold of 0.5.
- Adjusting this threshold could increase recall (detect more fraud cases) at the cost of precision.

iv) Generalizability

- The model was trained on a specific dataset and may not generalize well to new fraud patterns.
- Solution: Test on real-world datasets from different financial institutions.

Discussion The research study presents important findings about machine learning algorithm applications for identifying online payment system fraud through experimental data analysis. The high overall accuracy of 99% achieved by Random Forest and AdaBoost classifiers does not indicate a reliable fraud detection model because analysis proves accuracy should not be used as the sole measurement metric with imbalanced datasets. Most fraud detection systems face a crucial operational limitation due to extreme class imbalance between regular transactions and the rare fraudulent ones. Final model results appear falsely perfect due to the primary classification of majority cases which makes them unable to identify actual minority class occurrences therefore missing potential fraudulent transactions.

The model evaluation became more comprehensive when precision along with recall and F1-score and the AUC-ROC metric evaluated the model performance accurately. he precision results were similar between the models at 1.0 but their recall calculations produced different outcomes with a value of 0.5294 for Random Forest and 0.1176 for AdaBoost. A significant difference between the models shows that although they correctly identified all predicted frauds the AdaBoost model failed to detect many actual fraudulent transactions while producing higher false negatives. When monitoring fraudulent transactions it becomes critical to minimize false positives since the expenses generated by undetected fraudulent transactions surpass the harm caused by wrong alerts. Random Forest demonstrates superior performance accuracy by obtaining better F1-score and AUC-ROC metrics over AdaBoost according to the analysis results.

Analysis through confusion matrices confirmed the previous results showing Random Forest extracting elevated quantities of actual fraudulent cases whereas AdaBoost selected only a small segment of legitimate fraud events. Financial datasets benefit from using ensemble methods with multiple decision trees because Random Forest demonstrates superior performance when detecting rare events. AdaBoost exhibits inadequate performance because noisy data and outliers that are prevalent in genuine transaction data affect its operation. Various shortcomings surfaced while conducting the analysis.

The sole dependency on transaction amounts and balance values during feature selection hinders the model from discovering sophisticated behavioral patterns. The default threshold value (0.5) for classification might not provide ideal results for diagnosing rare classes considering the analysis. The recall performance can be improved by changing the threshold according to precision-recall trade-offs. The trained and tested models present unknown generalization capacities because they used only a single dataset as their foundation.

Research concludes that fraud detection systems require new evaluation methods which better demonstrate model effectiveness in tracking fraudulent activities. According to evaluation measures Random Forest demonstrates enhanced performance than AdaBoost which makes it more suitable for practical use. The development of a fully robust fraud detection system requires comprehensive advancement of feature engineering along with threshold tuning capabilities and data diversity implementations.

Future Study Research in the field of online payment fraud detection can benefit from the current findings and constraints through multiple research avenues to enhance model performance. The solution to class imbalance presents itself as the primary research priority. Research into online payment fraud detection should investigate adaptive synthetic sampling (ADASYN) and ensemble-based resampling together with hybrid methods that unite oversampling with cost-sensitive learning as more advanced alternatives to SMOTE and undersampling strategies. The applied strategies demonstrate improved detection results through minority class learning dynamics enhancement without resulting in significant modifications of majority class composition.

Research continues in the addition of behavioral characteristics together with time-based attributes. Modern transaction analysis systems primarily rely on fixed transaction elements including transaction amount and account balance but these standard measures fail to track personal purchase patterns and sequence-based behavior. The modeling of time-dependent patterns requires the implementation of recurrent neural networks (RNN) or long short-term memory (LSTM) networks to detect evolving fraud principles throughout time. The deep learning approaches excel at discovering sequence relationships because these patterns become essential for recognizing user behavior anomalies across multiple transactions.

The need for understandable and interpretable models in deployment situations will increase significantly during future implementations. Financial analysts find it challenging to understand prediction rationale from highly accurate algorithms such as Random Forests and neural networks even though these algorithms operate as black boxes. XAI frameworks using SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide explanations that build trust and transparency into model outputs. The tools help regulatory compliance and aid analysts in validating model decisions. The implementation of real-time fraud systems needs data stream optimization. The development of models should focus on platforms including Apache Kafka as well as Apache Flink and Spark Streaming to perform low-latency operations on transactional streaming data. Modern financial institutions need responsive and scalable solutions because they require instant decision-making in their practical applications.

Testing proposed models with datasets originating from different financial institutions across diverse geographical locations represents a necessary step for studying their widespread practical application abilities. Fundamental partnerships between banks and fintech organizations could establish appropriate data privacy protocols for accessing authentic transaction data which would produce genuine fraud cases for improved validity tests.

Future research needs to focus on four main areas involving data diversity improvement and class imbalance solutions alongside the implementation of sequential models as well as interpretability methods alongside real-time detection capabilities. These proposed directions serve to strengthen the academic quality and advance the practical deployment potential of the solution in financial institutions.

References • Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

• Haque, F. M. A., & Hassan, M. M. (2020). Bank loan prediction using machine learning techniques. International Journal of Scientific Research in Computer Science, Engineering and Information Technology,

5(1), 2456–3307.

• Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

• Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

• Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

• Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

• Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

• Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ... & Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, 2–2.