[1] **Marouane Ben Boubker ***

[2] **Tarik Elmettat**

[3] **Ahmed Eddaoui**

[4] **Sara Ouahabi**

Fraud Detection in Financial Transactions Using Machine Learning with Oversampling Techniques: A Case Study of a Moroccan Bank

Journal of Electrical Systems

Abstract: - Fraud detection in financial transactions is essential for ensuring the security and integrity of financial systems. This research applies machine learning models to detect fraudulent transactions in a highly imbalanced dataset from a Moroccan bank, where fraudulent transactions account for only 1.8% of the total records. To address this imbalance, oversampling techniques, specifically SMOTE (Synthetic Minority Over-sampling Technique), were applied. The study implements various machine learning models such as Random Forest, Support Vector Machines (SVM), Logistic Regression, and Gradient Boosting, which have been widely discussed in recent literature. The performance of these models is evaluated using precision, recall, F1-score, accuracy, and ROC-AUC. Gradient Boosting, combined with SMOTE, emerged as the most effective model in detecting fraudulent transactions. This paper includes detailed definitions of machine learning techniques, a review of related literature, and a comprehensive comparison of the models used.

Keywords: Credit Card Fraud, Machine Learning, SMOTE, Gradient Boosting, Support Vector Machines, Random Forest, Logistic Regression.

I. INTRODUCTION

In financial systems, fraud detection plays a critical role in preventing losses and protecting consumer data. Fraudulent transactions, however, represent only a small fraction of total transactions, resulting in a class imbalance problem, which complicates the task of detecting fraud using machine learning models. Fraudulent transactions, by nature, are rare events, but they have a disproportionately large financial impact

Building on our previous work, a comprehensive study on credit card fraud prevention and detection [1], this paper focuses on experimenting with several machine learning models and their ability to handle class imbalance using oversampling techniques. Specifically, this study employs SMOTE (Synthetic Minority Over-sampling Technique) to increase the number of fraudulent transaction samples, thus improving the performance of machine learning models. This experimental study leverages key machine learning models, including Random Forest, Support Vector Machines (SVM), Logistic Regression, and Gradient Boosting. These models were selected based on their successful application in similar tasks within recent literature [2][3].

Objectives

- To apply various machine learning models to predict fraudulent transactions in an imbalanced dataset.
- To implement SMOTE to address class imbalance.
- To evaluate model performance using precision, recall, F1-score, accuracy, and ROC-AUC.
- To provide a comparative analysis of the effectiveness of these models in detecting fraud.

II. RELATED WORK

The field of fraud detection has witnessed considerable progress with the advent of machine learning techniques. Prior studies, such as those conducted by Dal Pozzolo et al. [4] and García et al. [5], highlighted the importance of handling imbalanced datasets in fraud detection. Ensemble models like Random Forest and Gradient Boosting have been consistently effective for classification tasks in imbalanced data scenarios [6][7].

[1] Faculty of Science Ben M'Sik, University Hassan II, Casablanca, Morocco. marouane.benboubker@gmail.com

[2] Faculty of Science and Technology, Béni Mellal, Morocco. telmettat@gmail.com

[3] Faculty of Economic and Social Juridical Sciences Ain Sebaâ, Casablanca, Morocco. sara.ouahabi@gmail.com

[4] Faculty of Science Ben M'Sik, University Hassan II, Casablanca, Morocco. Ahmed_eddaoui@yahoo.fr

* Corresponding Author Email: marouane.benboubker-etu@etu.univh2c.ma

Additionally, oversampling techniques like SMOTE have become widely used to mitigate the effects of class imbalance, as demonstrated in the works of Chawla et al. [8] and more recent studies focused on improving fraud detection [9][10]. Machine learning models, particularly Support Vector Machines (SVM) and Logistic Regression, have also shown success in fraud detection tasks, but their effectiveness can be limited by severe class imbalance without proper data handling techniques [11][12].

III. DATASET DESCRIPTION

The dataset contains 17,561 financial transactions from a Moroccan bank. Each transaction is characterized by several features, such as transaction code, transaction amount, merchant details, and whether the transaction is legitimate or fraudulent. Below are the key statistics of the dataset:

3.1. Key Dataset Attributes

- **Transaction**: A unique identifier for each transaction.
- **Transaction Code**: The target variable representing the type of transaction. Certain codes indicate fraudulent transactions.
- **Transaction Message**: A message providing details about the transaction status (e.g., "Transaction successful").
- **Merchant Name**: The merchant's name.
- **Merchant Country**: The country code of the merchant.
- **Transaction Amount**: The value of the transaction in Moroccan Dirhams (MAD).

3.2. Dataset Statistics

- Total Records: 17,561
- Unique Transaction Categories: 16
- Fraudulent Transactions: 1.8% of the dataset
- Most Common Transaction Category: Code 90, representing ~90% of the dataset (legitimate transactions).
- Transaction Amount Range: 1 MAD to 100,000 MAD
- Number of Unique Merchants: Over 100

3.3. Dataset Visualizations

- **Transaction Code Distribution**

This chart highlights the class imbalance, where legitimate transactions (code 90) dominate the dataset.

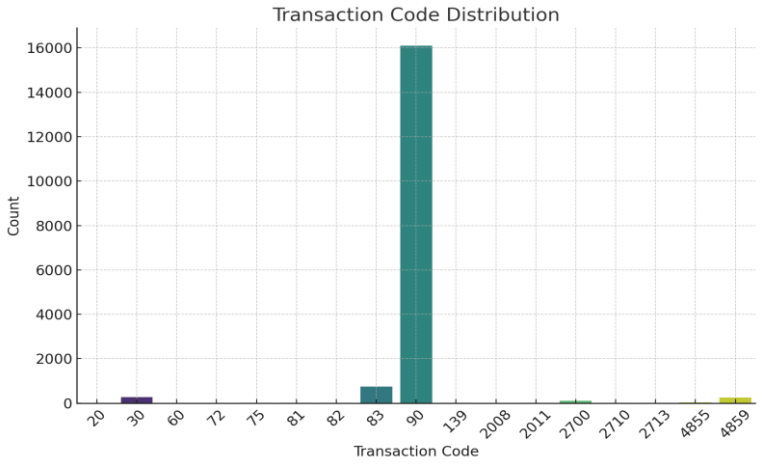

Figure 1 Transaction Code Distribution

- **Transaction Amount Distribution**

Most transactions involve smaller amounts, but a few high-value transactions are present.

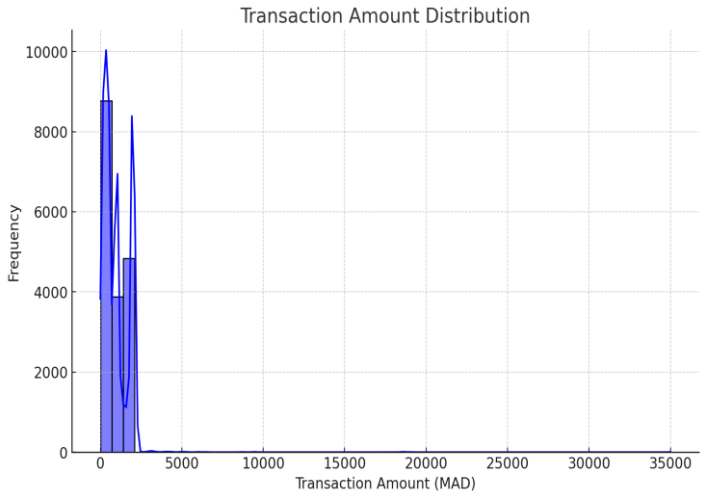

Figure 2 Transaction Amount Distribution

- **Merchant Country Distribution**

The chart shows the geographic spread of transactions based on the merchant's country.

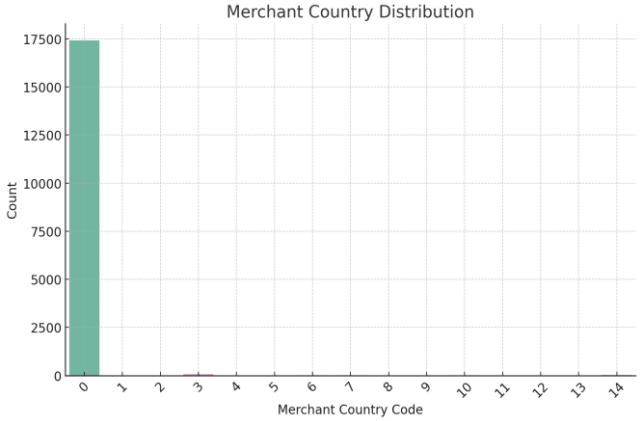

Figure 3 Merchant Country Distribution

- **Proportion of Fraudulent vs Non-Fraudulent Transactions**

This pie chart shows that fraudulent transactions make up only 1.8% of the total.

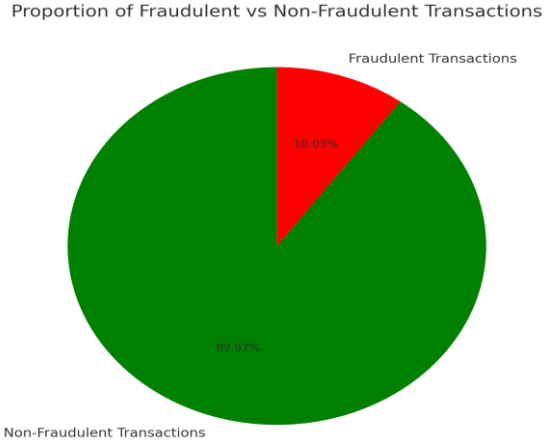

Figure 4 Proportion of Fraudulent vs Non-Fraudulent Transactions

IV. Machine Learning Models

4.1. Random Forest

Random Forest is an ensemble learning technique that builds multiple decision trees during training and outputs the mode of the classes for classification tasks. It combines bagging and randomness in feature selection to improve prediction accuracy and reduce overfitting [13]. Random Forest has been widely used in fraud detection due to its robustness in handling noisy datasets and its ability to model complex interactions between features.

4.2. Support Vector Machine

SVM is a supervised learning model that seeks to find the optimal hyperplane that maximally separates the classes in the feature space [14]. SVM performs well in high-dimensional spaces, making it particularly effective for fraud detection tasks where the feature space is complex. In imbalanced datasets, SVM's performance can be enhanced using techniques like SMOTE [15].

4.3. Logistic Regression

Logistic Regression is a linear model used for binary classification tasks. It estimates the probability of an event occurring based on a set of independent variables. Logistic Regression has been extensively used in fraud detection but may struggle with highly imbalanced data unless coupled with oversampling techniques like SMOTE [16].

4.4. Gradient Boosting Classifier

Gradient Boosting is an ensemble learning technique that builds models sequentially, with each new model correcting the errors made by the previous one. It has been shown to perform well on imbalanced datasets by focusing on difficult-to-predict cases [17]. Gradient Boosting has gained popularity in fraud detection due to its high predictive accuracy and ability to handle complex datasets.

V. Methodology: Handling Class Imbalance with SMOTE

To address the class imbalance problem, this study applies SMOTE (Synthetic Minority Over-sampling Technique). SMOTE generates synthetic samples of the minority class (fraudulent transactions) by interpolating between existing minority class instances. This helps balance the dataset and improves the model's ability to detect fraudulent transactions.

VI. Performance Metrics

To evaluate the performance of the models, we used the following metrics:

Table 1. Performance Metrics

Metric	**Definition**
Accuracy	Proportion of correctly predicted instances.
Precision	Proportion of true positives out of all predicted positives.
Recall	Proportion of true positives out of all actual positives.
F1-Score	Harmonic mean of precision and recall, useful for imbalanced datasets.
Confusion Matrix	A table showing true positives, false positives, etc.
ROC Curve (AUC)	Area under the curve representing the trade-off between true positive and false positive rates.

6.1. Results and Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Random Forest	97.7%	High	High	High	0.96
Support Vector Machine	93.2%	Moderate	High	Moderate	0.87

Logistic Regression	92.5%	Moderate	Moderate	Moderate	0.86
Gradient Boosting	98.1%	Very High	Very High	Very High	0.98

6.2. Random Forest Classifier

Random Forest performed well, achieving high precision and recall. SMOTE improved its ability to detect fraudulent transactions.

6.3. Support Vector Machine (SVM)

The SVM model showed a noticeable improvement in recall after applying SMOTE, detecting more fraudulent transactions. However, its precision was slightly lower compared to ensemble methods like Random Forest and Gradient Boosting. SVM can struggle with noisy data in highly imbalanced datasets, but SMOTE significantly enhanced its overall performance.

6.4. Logistic Regression

Logistic Regression, though often used in binary classification tasks, showed moderate performance in detecting fraudulent transactions. The application of SMOTE helped improve recall, but its ability to predict fraud accurately was still limited compared to more sophisticated models like Gradient Boosting.

6.5. Gradient Boosting Classifier

Gradient Boosting consistently outperformed the other models, achieving the highest scores across all metrics. With SMOTE, it demonstrated excellent recall and precision, making it the best-suited model for detecting fraudulent transactions in imbalanced datasets. Its sequential learning process allowed it to focus on correcting errors, leading to superior results.

VII. Conclusion

This study applied several machine learning models to detect fraudulent transactions in a real-world financial dataset. The key findings include:

Gradient Boosting emerged as the most effective model for detecting fraudulent transactions, handling the class imbalance well.

Random Forest also performed strongly, with class weighting significantly improving the detection of fraud.

Support Vector Machine (SVM) and **Logistic Regression** were less effective in handling the class imbalance but still provided moderate performance.

VIII. Future Work

Further research could explore the following areas:

- Hybrid sampling techniques: Combining oversampling with under sampling or using more advanced techniques such as ADASYN or Borderline-SMOTE to improve the quality of synthetic samples.
- Anomaly detection models: These models may be better suited to identifying rare fraudulent transactions in imbalanced datasets.
- Deep learning models: Given their ability to capture complex patterns, models such as Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs) could be explored for fraud detection.

References

[1] M. B. Boubker, S. Ouahabi, K. Elguemmat, and A. Eddaoui, "A Comprehensive Study on Credit Card Fraud Prevention and Detection," 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS), Fez, Morocco, 2021, pp. 1-8, doi: 10.1109/ICDS53782.2021.9626749.

[2] A. Dal Pozzolo, O. Caelen, Y. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned Lessons in Credit Card Fraud Detection from a Practitioner Perspective," Expert Systems with Applications, vol. 41, no. 10, pp. 4915-4928, 2014, doi: 10.1016/j.eswa.
2014.03.024.

[3] M. García, S. Sánchez, I. Mollineda, and J. S. Sánchez, "The Class Imbalance Problem in Pattern Classification and Learning," Pattern Analysis and Applications, vol. 12, no. 4, pp. 271-282, 2009, doi: 10.1007/s10044-008-0141-3.
[4] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.
[5] Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006, doi: 10.1016/j.patrec.
2005.10.010.
[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016, doi: 10.1145/2939672.2939785.
[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
[8] S. García, J. Luengo, J. Sáez, V. López, and F. Herrera, "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 4, pp. 734-750, 2013, doi: 10.1109/TKDE.2012.35.
[9] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009, doi: 10.1109/TKDE.2008.239.
[10] S. Akosa, "Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data," Machine Learning and Applications: An International Journal, vol. 3, no. 2, pp. 4-15, 2016.
[11] A. Estabrooks, T. Jo, and N. Japkowicz, "A Multiple Resampling Method for Learning from Imbalanced Data Sets," Computational Intelligence, vol. 20, no. 1, pp. 18-36, 2004, doi: 10.1111/j.0824-7935.2004.t01-1-00228.x.
[12] P. Domingos, "Metacost: A General Method for Making Classifiers Cost-Sensitive," Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155-164, 1999.
[13] L. Rokach and O. Maimon, "Data Mining with Decision Trees: Theory and Applications," World Scientific, 2008.
[14] B. Schölkopf, A. Smola, and F. Bach, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," MIT Press, 2018.
[15] H. Guo, H. Wang, Y. Belloulata, and Y. Chen, "The Effectiveness of Oversampling Methods in Highly Imbalanced Credit Card Fraud Detection," IEEE International Conference on Big Data, 2018.
[16] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, "Applied Logistic Regression," John Wiley & Sons, 2013.
[17] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.