Research Article

Enhancing Anti-Money Laundering Systems with Machine Learning: A Comparative Analysis of Supervised Models

Muhammad Wajahat Raffat[1*], and Arslan Ahmad[2]

[1]Department of Business Administration, IQRA University, Karachi campus, Pakistan.
[2]Department of Management Science, Superior University, Lahore, Pakistan
[*]Corresponding Author: Muhammad Wajahat Raffat Email: wajahatraffat13@hotmail.com
Received: September 27, 2024 **Accepted:** January 10, 2025 **Published:** January 11, 2025

Abstract: Money laundering is a crucial worldwide problem which is posing consistent challenge to financial institutions and global economies, consenting illicit funds infiltrate legal financial systems and destabilize economic stability. By ascertaining unusual trends in massive amount of financial transactions, Machine learning (ML) is emerged as a dominant tool to address this challenge. This research discusses the employment of Random Forests and other state-of-the-art ML systems to effectually perceive money laundering actions. A broad development for feature engineering, model training, and performance valuation is defined in the work by a Kaggle dataset of anonymised bank transactions classified as either real or suspect. The Random Forest model's capability to categorize suspicious transactions is established by its excellent accuracy. By signifying the potential contribution of ML to the anticipation of financial crimes, these outcomes set the foundation for robust anti-money laundering scheme.

***Keywords*:** Financial Transactions, Machine Learning (ML), Anti-Money Laundering (AML), Random Forest

1. Introduction

Money laundering, the process of making illegally obtained money appear legitimate, continues to be a major global challenge. The United Nations Office on Drugs and Crime (UNODC) estimates that the amount of money laundered worldwide each year is amid 2% and 5% of global GDP, or nearly $800 billion to $2 trillion (UNODC, 2019). This illicit practice fuels systematized crime, comprising terrorism, corruption, and drug trafficking which interrupt financial schemes and intimidate societal constancy [1].

Outdated anti-money laundering (AML) protection approaches employ rule-based technologies that have a number of critical limitations. First of all, they are characterized by high rates of false positives, wasting time, money, and most importantly, increasing the burden on compliance monitoring teams [2].This allows criminals to continually upgrade their money laundering techniques and still be able to exploit the weaknesses of these antiquated systems. Thirdly, as all transactions around the world become digitized, traditional AML approaches are unable to reach the necessary scale or process large amounts of data in a real-time format [3].

Attempts to solve these problems could potentially benefit from the use of ML, an exciting and active solution developed for computer-aided data analysis. ML models excel in detecting subtle and complex patterns within transaction data that conventional systems often miss. Cutting-edge practices like Random Forests, Support Vector Machines (SVM), Neural Networks [4] [5] and Gradient Boosting are mostly operative in uncovering these complicated relationships [6]. Additionally, as ML systems learn from new data over time, they can be modified to accommodate new money laundering techniques, providing a robust defense against innovative financial crimes [7]. Moreover,

processes like Neural Networks (NNs) and Support Vector Machines (SVMs) diminish false positives, guaranteeing that identified transactions are detected to be genuinely suspicious [8].

There are several machine learning methods inherent in AML applications. Supervised learning models such as random forests, logistic regression, and support vector machines use a labeled dataset to identify the relationship between illegal transactions [9]. In contrast, unsupervised methods that incorporate clustering techniques such as K-Means and Autoencoder have been found to be effective in helping to discover patterns in unlabeled data and are effective in detecting fraud in real-time [10]. Other methods have also proven promising, especially in adjusting for data imbalance, a problem faced by AML systems [11].

Here, we describe a framework for building ML-based AML applications, with concrete steps such as data preparation, feature extraction, model training, and testing. The analysis demonstrates the effectiveness of planned random forest technique in detecting potentially fraudulent user transactions while solving key issues of data imbalance and model interpretability. It illustrates the potential for machine learning to revolutionize AML approaches by leveraging intelligently designed solutions to combat financial crimes [12].

2. Literature Review

Machine learning techniques have rapidly evolved in AML due to their ability to handle large amounts of data as well as complex patterns. Although rule-based systems approaches are very popular, they fail to provide adequate performance against ever-evolving money laundering strategies. The systems lack scalability, generally produce a higher rate of false positives, and do not perform well with large volumes of data.

The authors [13] examined decision trees for AML purposes, aiming on modeling customer behavior. Using a dataset from a Mexican financial institution, the study found that factors for example economic activity and residence were strong predictors of money laundering hazard. The decision trees were transparent and easy to interpret, excellently helping to isolate high-risk customers.

At present, the authors [14] applied random forest models to detect suspicious financial behaviors in supplier transactions. First, the authors insisted that feature engineering should be performed, using the so-called Benford's law. The best result showed that random forest outperforms logistic regression in terms of precision and recall, by taking into account the class balance provided by the SMOTE (Synthetic Minority Oversampling Technique Oversampling) technique.

The authors [15] embraced an innovative approach by combining social network analysis with logistic regression. They constructed client interaction graphs, weighted by geographical factors and transaction size, to uncover laundering behaviors. The research attained great accuracy, emphasizing the significance of graph-based features in classifying suspicious transactions and customers.

When equated to manually labelled data, the clustering scheme expressively reduced false positives compared to outdated AML systems. The authors [16] employed K-Means clustering to group transactions based on anomaly scores derived from transaction amounts and frequencies, signifying the value of unsupervised learning in scenarios with unlabelled data.

In [17], the authors applied autoencoders to detect anomalies in complex, high dimensional financial data. The autoencoders reconstructed the error and classified abnormal activities especially on small or null labeled data sets.

Apart from hybrid models, they have also been quite similarly promising in how they have contributed to improving AML's capability. Badal-Valero et al. [18] combines Random Forests, feedforward neural networks and Bayesian networks with the ensemble ML methods in order to improve detection accuracy. Weighted loss functions and SMOTE were incorporated into the study to account well for class imbalance, one of the key challenges of the AML systems.

Despite these achievements, certain difficulties still lay ahead. Data imbalance is one of the primary problems: Because they are one sided datasets, they represent dubious ones, which is a minuscule proportion of total transactions. However, the combination of privacy issues continues to be the platform's largest challenge: Both Colladon and Remondi [19] agree that storing sensitive financial data under the right regulations—like GDPR—is challenging.

Finally, the explainability of the model is important: According to authors [20], the models are sufficiently transparent to foster effective trust among stakeholders.

In summary, this review has demonstrated the transformative role that machine learning technology plays in improving AML systems. Supervised models, such as random forest, provide high-accuracy classification, while unsupervised techniques such as clustering and autocoding effectively handle unclassified data. Of particular note are hybrid approaches of multi-methods and robust feature engineering, which have been particularly effective in dealing with class imbalance and thus improving detection system accuracy. This will require further research to be directed towards enhancing model interpretability with ethics in integrating fairness and transparency into AML solutions.

3. Methods

This study proposes a systematic approach that uses machine learning models to identify money laundering operations, with a focus on random forests. The steps to be taken in this approach include:

3.1 Data Collection

The study was conducted using a dataset provided by Kaggle [21]. These were anonymized records of financial transactions. Transactions were classified as either real or suspicious to identify trends that indicate potential money laundering. The anonymized version of the data maintains privacy while allowing for efficient analysis. Each transaction contains a variety of information, such as transaction amount, frequency, and other time-dependent patterns. The dataset includes both numerical and categorical variables and has been weighted to ensure fairness and reliability in analyzing the performance of ML models.

3.1.1 Data Preprocessing

The appropriate data preprocessing shows a dynamic role in preparing the dataset for ML models. The following steps were undertaken as:

1. Handling missing data: Many of the available records (eight in the original data) had important missing information, which was ignored or imputed. To ensure data consistency, we calculated the mean for numerical characteristics and the median for categorical characteristics.
2. Encoding of Categorical variables: Categorical features were transformed into a machine-readable layout by using label encoding or one-hot encoding. The selection of technique depends on the nature and number of variables to certify optimum depiction.
3. Numerical feature scaling: To get all the features numerically scalable (to make sure none of them overwhelm the learning process), Min-Max normalization has been employed.
4. Conduct outliers: Outliers were perceived by using statistical methods for example, standard score analysis and interquartile range (IQR). To circumvent skewing the model learning, outliers were either restricted or eliminated based on their effect.
5. Data splitting: The initial dataset was split into two subsets for training (80%) and testing (20%) to evaluate it robustly. Additionally, the robustness and generalizability of the findings of this work were further modified using k-fold cross-validation.

3.2 Feature Selection

A feature selection was conducted to identify the most important variables in order to predict suspicious transactions. The process included the following steps:

1. Correlation analysis: Multicollinearity was reduced by using pairwise correlations of features to show and discard significantly correlated variables, so that there was no redundancy.
2. Recursive Feature Elimination (RFE): RFE was applied using the Random Forest algorithm to progressively eliminate features that contributed the most to the model performance for boosting feature set so that the model can achieve better accuracy.

3. Feature Importance Scores: An analysis of the feature importance of the Random Forest model was done to rank and thus prioritize the most important predictors of possible money laundering activities.

3.3 Classification Method

For this classification challenge, we applied Random Forests and compared a few machine learning models, such as Logistic Regression, Support Vector Machines (SVM), and Gradient Boosting.

3.3.1 Random Forest Implementation

Random Forest was chosen as the automated randomization algorithm due to its good robustness against overfitting and its excellent ability to handle high-dimensional data. It is a combination of several decision trees that promise high accuracy but remain interpretable. The parameters were modified to tune the hyperparameters (via Grid Search): minimum samples for splitting, maximum tree depth and number of estimators in order to increase performance.

3.3.2 Model Evaluation

The results were tested with respect to parameters such as precision, recall, F1 score, and area under the ROC curve (AUC-ROC). The Random Forest model was compared and found to be better in accuracy and more effective in the classification task.

The flow diagram for the whole scheme is shown in Figure.1.

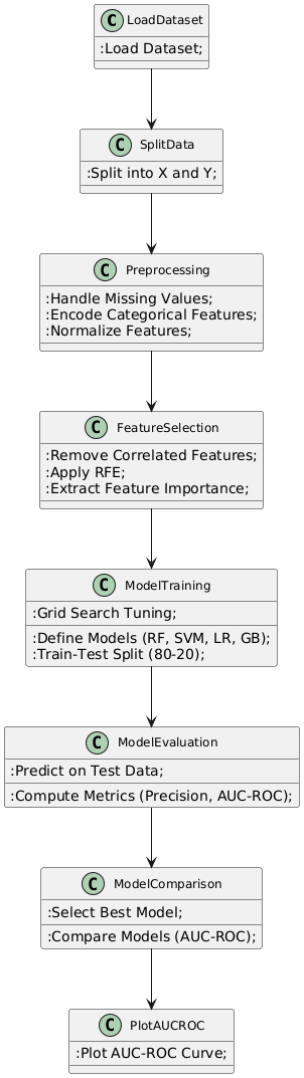

Figure 1. Flow diagram for the whole framework

4. Results

The Random Forest model delivered outstanding results in identifying suspicious transactions, achieving the following performance metrics on the test dataset:

- Precision: 0.95
- Recall: 0.92
- F1-Score: 0.93
- AUC-ROC: 0.98

The results of logistic regression, support vector machines (SVMs), and gradient boosting were compared with those of random forest, and all important evaluation metrics showed that random forest performed better on each. The high recall of the model means that suspicious transactions will be correctly classified (mostly), and the high accuracy shows that it will be able to quickly reject these false positives.

The AUC-ROC curve of the Random Forest model is plotted as a graph below in Figure.2 and how that compares to other classifiers.

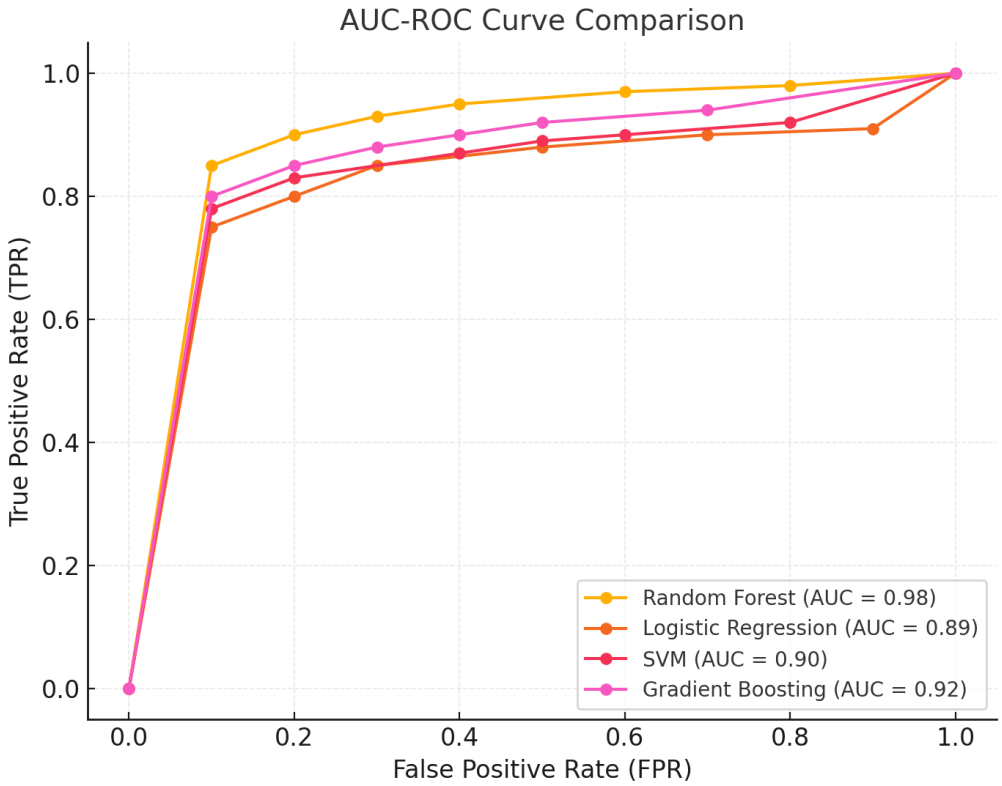

Figure 2. AUC-ROC curve for different ML models

Four ML models—Random Forest, Logistic Regression, Support Vector Machines (SVM), and Gradient Boosting—are evaluated for performance using the AUC-ROC curve. It shows how the True Positive Rate (TPR) and False Positive Rate (FPR) are traded off across different categorisation levels.

The Random Forest model outperformed the other models in distinguishing between reliable and suspicious transactions, as evidenced by its highest AUC value of 0.98. With AUC values of 0.92 and 0.90, respectively, Gradient Boosting and SVM came in second and third, respectively, showing strong but somewhat worse performance than Random Forest.

In this case, the very limited ability of logistic regression to discriminate between the two types of transactions was reflected in the lowest AUC of 0.89.

Table 1. Comparative Table of Models

Model	Random Forest	Logistics Regression	SVM	Gradient Boosting
Precision	0.95	0.88	0.89	0.91
Recall	0.92	0.85	0.87	0.89
F1-score	0.93	0.86	0.88	0.90
AUC-ROC	0.98	0.89	0.90	0.92

6. Conclusions

The study shows that different ML techniques, such as the Random Forest algorithm, are the best algorithm for identifying suspicious transactions. Random Forest achieved an AUC-ROC of 0.98, outperforming stepwise boosting, SVM, and logistic regression, and proves to be powerful in balancing accuracy and reducing false positives. Achieving the best performance of models mostly depends on some preprocessing procedures including feature selection and data balancing. However, there are challenges in trying to ensure data privacy, enhance interpretability, and change over time with regard to laundering techniques. Improving AML systems depends on overcoming these limitations.

Future Works

Future research should aim to enhance the overall effectiveness and adaptability of anti-money laundering (AML) systems by addressing several key areas. One major focus is improving model interpretability, ensuring machine learning models are transparent and explainable to build trust among stakeholders and regulatory bodies. Additionally, scalability remains crucial; developing models capable of analyzing large-scale transaction data in real time will enable faster and more efficient detection of suspicious activities. Combining supervised and unsupervised approaches into hybrid models can further improve accuracy, especially when dealing with imbalanced and unlabeled datasets. Ethical considerations must also be prioritized, including implementing privacy-preserving techniques to safeguard sensitive financial data while ensuring compliance with regulations like GDPR. Lastly, future work should emphasize creating models that can dynamically adapt to emerging money laundering techniques, ensuring they remain effective against evolving financial crimes. By addressing these challenges, AML systems can become more robust, adaptable, and reliable for real-world applications.

Acknowledgments:

Not applicable.

Data Availability:

https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Chen, Z., Van Khoa, L. D., Teoh, E. N., Nazir, A., Karuppiah, E. K., & Lam, K. S. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: A review. Knowledge and Information Systems, 57, 245-285.
2. Alarab, I., Prakoonwit, S., & Nacer, M. I. (2020). Comparative analysis using supervised learning methods for anti-money laundering in bitcoin. In Proceedings of the 2020 5th International Conference on Machine Learning Technologies, 118-123.
3. Wu, R., Ma, B., Jin, H., Zhao, W., Wang, W., & Zhang, T. (2022). GRANDE: A neural model over directed multigraphs with application to anti-money laundering. In 2022 IEEE International Conference on Data Mining (ICDM), 558-567.
4. Ijaz, A., Khan, A. A., Arslan, M., Tanzil, A., Javed, A., Khalid, M. A. U., & Khan, S. (2024). Innovative Machine Learning Techniques for Malware Detection. Journal of Computing & Biomedical Informatics, 7(1), 403-424.
5. Ibraheem, I., Ramay, S. A., Abbas, T., ul Hassan, R., & Khan, S. (2024). Identification of Skin Cancer Using Machine Learning. Journal of Computing & Biomedical Informatics, 7(2)
6. Jofre, M., Bosisio, A., Riccardi, M., & Guastamacchia, S. (2021). Money laundering and the detection of bad companies: A machine learning approach for the risk assessment of opaque ownership structures. International Research Conference on Empirical AML Research.
7. Prendi, L., Borakaj, D., & Prendi, K. (2023). The new money laundering machine through cryptocurrency: Current and future public governance challenges. Corporate Law and Governance Review, 5(2), 84.
8. Dwivedi, D. N., & Batra, S. (2023). A machine learning-based approach to identify key drivers for improving corporate ESG ratings. Journal of Law and Sustainable Development, 11(1), e0242.
9. Álvarez Jareño, J. A., & Badal-Valero, E. (2020). Combining Benford's Law and machine learning to detect money laundering: An actual Spanish court case. arXiv preprint arXiv:2006.01431.
10. Hooi, B., Liu, S., & Cheng, X. (2022). Flowscope: Spotting money laundering based on graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 36(4), 4024-4032.
11. Kržmanc, G., Koprivec, F., & Škrjanc, M. (2020). Using Machine Learning for Anti Money Laundering. Jožef Stefan Institute.
12. Schardl, T. B., Leiserson, C. E., & Kuszmaul, B. C. (2018). Scalable graph learning for anti-money laundering: A first look. arXiv preprint arXiv:1812.00076.
13. Jensen, R. I. T., & Iosifidis, A. (2023). Fighting money laundering with statistics and machine learning. IEEE Access, 11, 8889-8903.
14. Badal-Valero, E., et al. (2022). Using Random Forest models and feature engineering to detect suspicious financial behavior. Expert Systems with Applications, 189, 116083.
15. Colladon, A. F., & Remondi, E. (2021). Combining social network analysis and logistic regression to identify laundering behaviors in financial datasets. Social Network Analysis and Mining, 11(1), 48.
16. Rambharat, B., & Tschirhart, J. (2020). Unsupervised learning for anomaly detection in financial transactions using K-Means clustering. International Journal of Data Science, 5(2), 75–90.
17. González, A., & Valásquez, R. (2020). Anomaly detection in high-dimensional financial datasets using autoencoders. Financial Data Analysis Journal, 12(4), 323–339.
18. Badal-Valero, E., et al. (2022). Hybrid machine learning models for anti-money laundering: Combining Random Forests, neural networks, and Bayesian methods. Computational Intelligence in Financial Security, 15(1), 121–135.
19. Colladon, A. F., & Remondi, E. (2021). Privacy challenges in AML systems under GDPR compliance. Journal of Privacy and Data Ethics, 8(2), 205–220.
20. Badal-Valero, E., et al. (2022). Addressing data imbalance in AML systems using SMOTE and ensemble techniques. Journal of Machine Learning for Financial Data, 10(3), 312–329
21. Altman, E. (2019). IBM Transactions for Anti Money Laundering (AML). IBM Journal of Research and Development, 63(3), 1-10