Advanced Supervised Machine Learning Algorithms in Credit Card Fraud Detection

Simin Yu[1], Victor Chang[1,*], Gia Linh Huỳnh[2], Vitor Jesus[3] and Jiabin Luo[1]
[1]*Department of Business Analytics and Information Systems, Aston Business School, Aston University, Birmingham, U.K.*
[2]*Becamex Business School, Eastern International University, Binh Duong, Vietnam*
[3]*School of Computer Sci and Digital Tech, College of Engineering and Physical Sci, Aston University, Birmingham, U.K.*
*

Keywords: Money Laundering, Machine Learning, Data Imbalance, Oversampling, Undersampling.

Abstract: The rapid growth of online transactions has increased convenience but also risks like money laundering, threatening financial systems. Financial institutions use machine learning to detect suspicious activities, but imbalanced datasets challenge algorithm performance. This study uses resampling techniques (SMOTE, ADASYN, Random Undersampling, NearMiss) and ensemble algorithms (XGBoost, CatBoost, Random Forest) on a simulated money laundering dataset provided by IBM (2023) to address this. Our findings reveal that each resampling technique offers unique advantages and trade-offs. CatBoost consistently outperforms XGBoost and Random Forest across sampling techniques, achieving the best balance between precision and recall while maintaining strong ROC curve scores. This strong performance could reduce the number of transactions banks must examine, as investigations would only focus on the predicted laundering cases.

1 INTRODUCTION

The rapid development of digitalization presents both conveniences and challenges for modern financial systems (Saklani et al., 2024). Among the emerging issues, money laundering stands out as a significant concern, with the global scale estimated to range from $500 billion to $1 trillion annually (Sharman, 2011). This vast amount of illicit money moving through financial systems threatens economic stability and undermines the integrity of legitimate transactions. Money laundering not only endangers financial security but also challenges regulatory frameworks and the trustworthiness of financial institutions (Olujobi and Yebisi, 2023).

As technology and laundering tactics evolve, detecting these activities becomes increasingly complex. Criminals continually develop sophisticated methods to obscure illegal funds, pushing financial institutions and regulatory bodies to adopt more advanced and adaptive detection strategies. Traditional detection methods often fall short due to their limited capacity to handle these increasingly complex schemes. In response, machine learning techniques have emerged as promising tools, yet they face a significant limitation: the dataset class imbalance issue. In real-world financial datasets, suspicious transactions are vastly outnumbered by normal ones, making it difficult for algorithms to identify the minority class of interest suspicious transactions without sacrificing accuracy (Dastidar et al., 2024)

Moreover, the implications of effective detection extend beyond operational efficiency for financial institutions, it is also a matter of public trust and regulatory compliance (Olawale et al., 2024). Banks risk substantial reputational and financial destruction if perceived as facilitators of illegal activities, which could diminish public confidence and incur severe regulatory penalties (Ambe, 2024). Consequently, there is an urgent need for machine learning models that excel in accuracy and address the unique challenges posed by imbalanced data. In this study, we combine ensemble algorithms with diverse data balancing techniques to build a robust system for money laundering detection. Utilizing methods such as SMOTE (Chawla et al., 2002), ADASYN (He et al., 2008), Random Undersampling (Koziarski, 2020), and NearMiss (Mani and Zhang, 2003) on simulated datasets.

* Correspondent author.

Yu, S., Chang, V., Huážşnh, G. L., Jesus, V. and Luo, J.
Advanced Supervised Machine Learning Algorithms in Credit Card Fraud Detection.
DOI: 10.5220/0013485400003956
In *Proceedings of the 7th International Conference on Finance, Economics, Management and IT Business (FEMIB 2025)*, pages 126-138
ISBN: 978-989-758-748-1; ISSN: 2184-5891

We aim to enhance the performance and reliability of machine learning algorithms in detecting money laundering activities by addressing class imbalance using four resampling techniques, this research represents a significant advancement towards developing more reliable and efficient AML detection systems, offering valuable insights for financial institutions in combating money laundering activities. The findings aim to guide practitioners in selecting optimal strategies for handling imbalanced data in AML applications, thereby bridging the gap between theoretical advancements and practical implementation.

This study is organized as follows: The second section provides information about reviewing relevant literature for the study. The third section presents the research methodology and ensemble machine learning with data balancing techniques. The fourth section presents the details of the data and the preprocessing of the data. The fifth section presents the final results obtained from the models. The sixth section concludes the conclusion.

2 LITERATURE REVIEW

Many studies have explored machine learning strategies for detecting suspicious patterns in money laundering datasets, emphasizing data balancing techniques. These methods enhance algorithm performance by identifying subtle illicit financial behaviours often overlooked, addressing class imbalance, and improving the reliability of AML systems (Xu et al., 2025; Jensen et al., 2024; Bakhshinejad et al., 2024).

2.1 Money Laundering

Money laundering is a significant global concern, threatening financial stability, economic development, and regulatory compliance worldwide. It involves processing profits earned from illegal activities, such as drug trafficking and fraudulent schemes, to conceal their origins and integrate them into the legitimate financial system (Gaviyau and Sibindi, 2023). Specifically, money laundering typically follows three key stages: placement, layering, and integration. Placement refers to introducing unlawfully obtained funds into the financial system, layering involves performing complex transactions to unclear the origin of funds, and integration, as described by (Samantha Maitland Irwin et al., 2011), is the withdrawal of cleaned money, now appearing legitimate, from designated accounts.

The Financial Action Task Force (FATF) plays a crucial role in establishing global benchmarks for anti-money laundering (AML) measures, ensuring uniform regulatory standards, and fostering international collaboration (Petit, 2023). Effective AML frameworks align with the United Nations Development Goals, emphasizing global cooperation as essential in combating financial crimes (Dobrowolski, 2024).

The consequences of money laundering extend beyond financial loss, they erode public trust, compromise banking integrity, and inflict reputational damage on financial institutions. Illicit financial flows increase banks' operational risks and regulatory burdens, potentially leading to financial penalties, legal consequences, and customer attrition (Moromoke et al., 2024). Addressing this issue requires sophisticated technological interventions, such as machine learning algorithms, to detect suspicious activities efficiently and proactively, reducing the prevalence of illicit financial transactions and preserving the integrity of the global financial ecosystem.

2.2 Machine Learning Technologies Adoption

Financial institutions are adopting advanced machine-learning techniques to detect suspicious transactions effectively. However, the significant imbalance in transaction datasets challenges the algorithm's performance. Several studies have investigated the application of Random Forest, CatBoost, and XGBoost in the context of anti-money laundering (AML) operations. These machine-learning algorithms have been crucial in detecting suspicious activities and enhancing the efficiency of AML processes. Hilal et al. (2022) utilized Random Forest for suspicious activity detection in anti-money laundering efforts, demonstrating the effectiveness of Random Forest in detecting anomalies related to money laundering activities through simulated annealing for hyperparameter tuning. Random Forest outperformed other machine learning algorithms in predicting money laundering suspect transactions, according to Masrom et al. (2023). Moreover, Vassallo et al. (2021) focused on applying Gradient Boosting algorithms, particularly XGBoost, in anti-money laundering within cryptocurrencies. The study emphasized the efficiency, scalability, and reduced training time achieved by utilizing XGBoost to detect money laundering at a transaction level.

In addition, CatBoost has gained attention recently for its efficiency in handling categorical data and reducing overfitting. According to Aldania et al. (2023), CatBoost demonstrated superior performance in classification tasks with imbalanced datasets compared to other ensemble algorithms. Its ability to efficiently process categorical features without requiring extensive preprocessing has been identified as a significant advantage in financial transaction datasets. Furthermore, Rojan (2024) highlighted the robustness of CatBoost in predicting fraudulent financial transactions, showcasing its potential as a key player in AML operations. Additionally, studies by (Kokori et al., 2024) suggested that integrating ensemble learning methods, including combinations of Random Forest, XGBoost, and CatBoost, could yield better predictive accuracy compared to standalone models. However, few studies focus on ensemble methods combining Random Forest, CatBoost and XGBoost. This paper addresses this gap by empirically utilizing an ensemble method for imbalanced data.

2.3 Data Imbalance Challenges

Data imbalance is a persistent challenge in anti-money laundering (AML) systems, primarily because suspicious transactions represent only a tiny fraction of the total volume of financial transactions. This imbalance leads to biases in machine learning models, where classifiers are skewed toward predicting the majority class, reducing their ability to detect rare but crucial suspicious activities. Cherif et al. (2023) highlighted how extreme class imbalance causes classifiers to favor the dominant class, often resulting in high false negative rates. This outcome significantly hampers the effectiveness of AML systems, as failing to detect a suspicious transaction could lead to severe financial and reputational consequences.

Bansal et al. (2022) emphasized the importance of addressing class imbalance using techniques such as Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and cost-sensitive learning. These methods aim to balance the dataset by either oversampling or undersampling the majority class, or adjusting the cost function of the learning algorithm.

Additionally, Gurcan and Soylu (2024) discussed the limitations of traditional machine learning algorithms in handling imbalanced datasets. They suggested hybrid approaches combining ensemble learning with advanced resampling methods to improve the detection rate of minority class instances.

Overall, addressing data imbalance in AML systems remains a critical area of research. Effective resampling methods, combined with robust ensemble learning algorithms, can significantly enhance the performance and reliability of AML detection models.

2.4 Data Balancing Techniques

Resampling methods are crucial in mitigating class imbalance, particularly in fraud detection tasks such as money laundering. Imbalanced datasets, characterized by a significant disparity in the representation of classes, can result in biased models that inadequately identify instances of the minority class, such as fraudulent transactions (Khalil et al., 2024). To address this issue, resampling techniques, including oversampling and undersampling, are commonly employed.

In comparing oversampling and undersampling techniques, oversampling methods such as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) are frequently utilized in fraud detection tasks, including the detection of money laundering activities. These techniques generate synthetic samples to balance the dataset, enhancing the model's ability to detect instances of the minority class. On the other hand, undersampling methods like NearMiss and Random Undersampling aim to rebalance the dataset by reducing the number of majority class instances. The selection between oversampling and undersampling depends on the specific characteristics of the dataset and the desired balance between sensitivity and specificity in detecting money laundering transactions (Bansal et al., 2022). In conclusion, this study will use resampling methods to mitigate the class-imbalanced data in the IBM dataset and increase the performance of machine learning algorithms.

We employ resampling techniques such as SMOTE, ADASYN, Random Undersampling, and NearMiss to address this. This research method utilizes a simulated money laundering dataset provided by IBM (2023), using ensemble machine learning algorithms, including XGBoost, CatBoost, and Random Forest. Model performance is evaluated using ROC curve, F1-score, Precision and Recall when comparing different resampling techniques.

Despite the growing interest in combating money laundering through machine learning, few prior studies systematically evaluate resampling methods alongside ensemble models in the specific context of AML. This lack of comprehensive comparison limits actionable insights into how data balancing

techniques can effectively improve model performance on imbalanced datasets. To address this gap, our study conducts a thorough empirical analysis by integrating an ensemble method with four balancing techniques. Thus, this study provide a detailed understanding of balancing techniques impact on imbalanced datasets on model performance.

3 METHODOLOGY

Money laundering transaction detection includes five main stages: data preprocessing, model development algorithms, Sampling Methods, and Extract Results. Fig. 1 illustrates the methodological framework of the study. This research uniquely combines advanced ensemble learning algorithms such as XGBoost, CatBoost, and Random Forest and adopts diverse resampling techniques including SMOTE, ADASYN, Random Undersampling, and NearMiss. This chapter employs several ML algorithms for transaction classification.

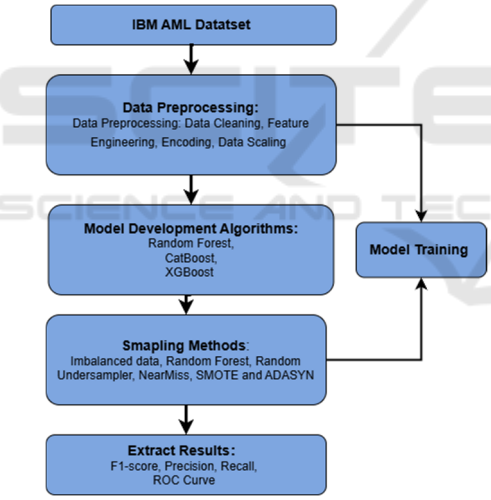

Figure 1: Methodology Framework.

3.1 Imbalance Learning Techniques

Random Undersampling (Zhong, 2024)**.** Class imbalance is a common problem in classification tasks, particularly in domains like fraud detection, disease diagnosis, and anomaly detection. When the dataset is imbalanced, machine learning models favor the majority class, leading to a poor generalization of the minority class. Random undersampling tackles this issue by randomly selecting a subset of samples from the majority class while keeping all samples from the minority class.

This study utilized Random Undersampling during the data preprocessing phase, which effectively improved classification performance in imbalanced datasets. Post-training, evaluation metrics such as accuracy, precision, recall, and ROC-AUC should be analyzed to ensure the model effectively handles the class imbalance.

NearMiss (Bao et al., 2016)**.** NearMiss is a well-known undersampling technique for tackling class imbalance in machine learning datasets. This method involves choosing specific instances from the majority class close to the minority class instances, thus resulting in a more balanced dataset. The main objective of NearMiss is to enhance the performance of classifiers by ensuring a more defined decision boundary between classes. The IBM dataset, related to money laundering detection, exhibits a significant imbalance, with the majority class (non-laundering transactions) vastly outnumbering the minority class (laundering transactions). Thus, this study using NearMiss ensures that models are exposed to a more balanced dataset, improving their ability to effectively generalize and detect rare money laundering transactions.

SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002)**.** is a Synthetic Minority Oversampling Technique that balances the dataset by oversampling the minority class. SMOTE creates new samples by interpolating between existing ones. It selects pairs of similar minority class instances and generates new samples along the line joining these points in the feature space. This ensures better generalization and prevents overfitting caused by simple duplication.

Additionally, this study used SMOTE to address the imbalance dataset in the “IS Laundering” column. It can also be integrated into pipelines alongside Random Forest, XGforest, and Catboost classifiers. The models in the files can better learn patterns from the minority class, leading to improved prediction performance for detecting money laundering transactions.

ADASYN (Adaptive Synthetic Sampling) (He et al., 2008)**.** Adaptive Synthetic is similar to SMOTE but focuses on harder-to-learn examples. Classification Models Training and Evaluation ADASYN adaptively assigns more weight to instances misclassified by the nearest neighbor algorithm. This study can be included after feature preprocessing and splitting into training and testing sets. It can also be combined with classifiers such as Random Forest, XGboost, and Catboost to improve model performance in the minority class.

To sum up, the study implements three ensemble machine learning models, including Random Forest, CatBosst, and XGBoost, using appropriate metrics such as precision, recall, and ROC-AUC in evaluation.

3.2 Classification Models Training

The study implements three ML models to evaluate their effectiveness in detecting money laundering activities:

Random Forest (Breiman, 2001)**.** An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. It is robust to overfitting and can handle large datasets with high dimensionality. The exponential loss function L for a singleinse(x_i, y_i) wh a predicted value $f(x_i)$ is defined as:

$$L(y_i, f(x_i)) = exp\left(-y_i f(x_i)\right) \quad (1)$$

Where:

- y_i is the true label of the instance, which is either +1 or -1.
- It is the combined prediction of all weak learners up to the current iteration.

Catboost (Prokhorenkova et al., 2018)**.** An ensemble learning method based on gradient boosting that specializes in handling categorical features automatically. It employs ordered boosting to prevent prediction shifts caused by target leakage. The contribution of a single ee hm for an instance (x_i, y_i) with a predicted value $f(x_i)$ is defined as:

$$L(yi, Fm(xi)) = exp(-yi * Fm(xi)) \quad (2)$$

Where:

- yi is the true label of the instance x_i
- $Fm(xi)$ is the model prediction at step m
- hm is the weak learner (decision tree) at step m

XGBoost (Extreme Gradient Boosting) (T. Chen et al., 2015)**.** An efficient and powerful gradient boosting algorithm widely used for its high accuracy and performance. It features regularization to prevent overfitting, handles missing data well, and supports parallel processing. Its flexibility and scalability make it ideal for various ML tasks, including classification and regression. The loss function of XGBoost is depicted as follows:

$$H^t = \sum_{j=1}^{n} \quad \Upsilon\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Psi(f_t) \quad (3)$$

Where

- $\Psi(f_t)$ is a regularization term.

The models were evaluated using precision (ζ), recall (∇), F1-score ($F1$), and confusion matrices as follows:

$$\zeta = \frac{\rho}{\psi + \rho}$$

$$\nabla = \frac{\rho}{\rho + \alpha}$$

$$F1 = 2 * \frac{\zeta * \nabla}{\zeta + \nabla} \quad (4)$$

Where ρ,ψ, and α are the number of true positives, number of false positives, and number of false negatives, respectively.

This study uses confusion matrix data to evaluate these models' performance. Metrics such as accuracy, ROC AUC, and F1-score are calculated to comprehensively understand each model's strengths and weaknesses, especially in handling imbalanced datasets.

Ensemble methods like Random Forest, Catboost, and XGBoost are often perceived as "black boxes" due to their complex structure and decision-making processes (Rane et al., 2024). However, in high-stakes applications like AML, black-box models present a challenge. Regulatory compliance necessitates that banks detect suspicious activity and explain the factors that contributed to these detections. Techniques such as SHAP (Shapley Additive Explanations) values and LIME (Local Interpretable Model-agnostic Explanations) can provide insights into model behavior, helping to clarify the influence of specific transaction features on predictions. Such interpretability methods align the predictive power of complex models with the transparency requirements of AML regulations.

3.3 Evaluation Metrics

We use evaluation metrics to evaluate the performance of the model in ML. The algorithms' evaluation measures include F1-Score, Recall, Precision, and ROC curves. These measures are routinely used for analyzing imbalanced datasets, such as the one utilized in this work.

F1-Score (Yacouby & Axman, 2020)**.** is a widely used evaluation metric in binary and multi-class classification tasks, which is the harmonic mean of Precision and Recall, providing a balanced measure

that accounts for false positives and false negatives. Mathematically, it is defined as:

$$F1 = 2 \times \frac{Precision * Recall}{Precision * Recall} \quad (5)$$

Recall (C.-H. Chen and Honavar, 1995)**:** is the ratio of correctly predicted positive observations to the total actual positives. It measures the model's ability to identify all relevant instances from the dataset. Mathematically, it is expressed as:

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positive(TP) + False\ Negative\ (FN)} \quad (6)$$

Precision (Powers, 2020)**:** is a metric used in information retrieval, machine learning, and classification tasks to measure the accuracy of positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positive observations. Mathematically, it is expressed as:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \quad (7)$$

ROC Curve: is a graph that evaluates the performance of a binary classifier across all classification thresholds. The mathematical equations (Park et al., 2004) and illustrate the concept of the ROC curve as follows:

$$FPR = \frac{FP}{FP + TN} \quad (8)$$

$$TPR = \frac{TP}{FN + TP} \quad (9)$$

- **FPR** stands for False Positive Rate, representing the proportion of negative instances incorrectly classified as positive.
- **TPR** stands for True Positive Rate, representing the proportion of positive instances correctly classified as positive.

4 EXPERIMENTS

4.1 Dataset

The comparative analysis utilizes the "IBM Transactions for Anti-Money Laundering" dataset, obtained from Kaggle (IBM, 2023), the data is classified into two groups, HI and LI, with varying levels of illegal activity (laundering). HI and LI are further separated into small, medium, and big datasets, with large datasets including 175M - 180M transactions, in addition, Group HI has a relatively higher illicit ratio. The dataset encompasses 12 distinct features (detailed specifications in Table 1). However, to demonstrate the imbalance dataset performance, we discuss several statistics about the HI-small dataset, which has 5 million transactions, and the HI-medium dataset, which has 31 million transactions. The analysis extends to both small and medium datasets, highlighting the scalability and robustness of the proposed methods.

Table 1: IBM Transactions for Anti-Money Laundering specific column.

No.	Attributes	Data type	Descriptions
1	Timestamp	Date	The timestamp when the transaction was executed
2	Amount Received	Float	The monetary amount credited to the account
3	Receiving Currency	Category	The currency type (e.g., dollars, euros) from the account.
4	Amount Paid	Float	The monetary amount credited to the account
5	Payment Currency	Category	The sender account's currency type (e.g., dollars, euros).
6	Payment Format	Category	The method of transaction: cheque, ACH, wire, credit card, etc.
7	Is Laundering	Binary	A value of 1 indicates that the transaction is classified as laundering, while 0 indicates a normal transaction.

4.2 Data Preprocessing

To improve model performance, it's crucial to identify and prioritize important features over irrelevant ones. The preprocessing methodology incorporated several critical steps: management of missing value imputation, categorical variable encoding, and temporal feature extraction from timestamp data (Arefin, 2024; Huang et al., 2024).

A structured data preprocessing pipeline was implemented to address the data imbalance issue and ensure optimal model performance. Initially, the dataset was thoroughly cleaned to handle missing values and inconsistencies. Specifically, all rows containing null or irrelevant entries were eliminated to ensure data integrity. Accordingly, only valid

samples with clearly defined class labels (0 for non-laundering and 1 for laundering) were retained, resulting in a balanced subset ready for further analysis.

The HI-small dataset contains 5 million samples, and the HI-medium dataset contains 31 million samples. To address the potential redundancy among features, a correlation matrix was employed to identify and eliminate highly correlated features. Features with a correlation coefficient greater than 0.90 were removed, as they provided redundant information and could bias the classification results. This step dropped unnecessary columns, leaving a more concise set of informative features.

After feature selection, the dataset was standardized using RobustScaler from sci-kit-learn, which normalized numerical features by removing the median and scaling based on the interquartile range. This ensured a consistent scale across all features, minimizing biases caused by varying feature magnitudes and enhancing the performance of algorithms sensitive to feature scaling, such as XGBoost, CatBoost, and Random Forest. To further address the inherent class imbalance in the dataset, techniques such as SMOTE (Synthetic Minority Oversampling Technique), ADASYN, and Random UnderSampling were employed. These methods ensured that the minority class (Is Laundering = 1) was well-represented during model training, improving the classifier's ability to detect fraudulent transactions effectively.

The HI-small dataset shows a significant imbalance between the classes. There is a much higher number of normal transactions (5,073,168 instances, 99.9%) compared to laundering transactions, which only account for 5,177 instances (0.1%) (see details in Fig. 2). Additionally, the HI-medium dataset is highly imbalanced, with the majority class (0) dominating the dataset (31863008 instances, ~98.9%), while the minority class (1) makes up only which only 35230 (~1.1%).

This significant difference emphasizes the rarity of laundering transactions in the dataset, which may pose challenges for predictive modeling. As such, appropriate techniques to address class imbalance need to be considered.

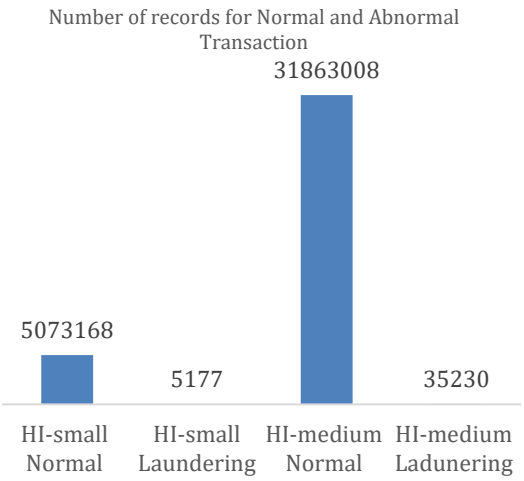

Figure 2: Summary of the distribution of laundering cases.

4.3 Training and Testing

This study discusses the model training and testing phases for transactions in the anti-money laundering (AML) dataset. The HI small and medium dataset, which consists of millions of transactions, was divided into training (80%) and testing (20%) sets to ensure a balanced evaluation. The training set was used for model training, hyperparameter tuning, and optimizing the performance of classifiers such as RandomForest, XGBoost, and CatBoost. In contrast, the testing set evaluated the generalization performance of the trained models on unseen data.

Techniques such as SMOTE, ADASYN, and Random UnderSampling were applied during training to address the class imbalance, ensuring that the minority class (Is Laundering = 1) was sufficiently represented. The train-test-split function from sci-kit-learn was employed for data partitioning, maintaining the integrity of the dataset across both phases. The primary objective of this setup was to ensure that the models could effectively generalize to new data, avoiding common pitfalls like underfitting (where the model fails to learn meaningful patterns) and overfitting (where the model memorizes patterns from the training set but fails to generalize to unseen data). This structured approach guarantees robust performance and reliable detection of suspicious transactions in AML systems.

4.4 Data Privacy in AML

Ensure data privacy in AML applications, especially when handling sensitive financial information. Compliance with data protection regulations, particularly the General Data Protection Regulation

(GDPR) in the European Union, is essential to safeguard individual privacy rights and maintain the security of personal data. GDPR mandates that organizations limit data collection to what is strictly necessary, obtain clear consent for data usage, and implement robust measures to protect against data breaches (Zorell, 2018).

The HI-small and HI-medium datasets were handled according to these principles. Features containing personally identifiable information (PII) were either anonymized or excluded from the modeling pipeline to ensure compliance with data protection standards. Secure storage mechanisms were employed to protect data integrity throughout the preprocessing, training, and testing phases. Techniques like feature scaling and data transformation were applied without compromising the confidentiality of sensitive information.

4.5 Experimental Setup

In this work, the experiments are performed on Core(TM)i7-11700KF CPU @ 3.60GHz 3.50GHz based processor, windows 11 with 16.0 GB of RAM. Python 3.7.1 is used as many models and libraries available for classification. This work uses pandas, numpy, seaborn, and matplotlib libraries. Scikit-learn was used to acquire the implemented metrics and approaches.

5 EXPERIMENTAL RESULTS

The study aims to enhance the performance and reliability of machine learning algorithms in detecting money laundering activities by addressing class imbalance using resampling techniques, such as SMOTE, ADASYN, Random Undersampling, and NearMiss, to ensure accurate and unbiased transaction classification.

In this study, we evaluated the performance of three different classifiers: Random Forest, CatBoost, and XGBoost. The performance metrics used for evaluation were precision, F1 Score and Recall for Class 0 and Class 1, and ROC AUC. Given the imbalanced nature of the dataset, we will focus our discussion on the F1 scores, particularly for Class 1 (the minority class). Table 2 shows detailed metrics of the models:

Table 2: Summary results of ML models with sampling method.

Models	Sampling Methods	HI-Small				HI-medium			
		ROC curve	F1-score	Precision	Recall	ROC curve	F1-score	Precision	Recall
Random Forest	**Without Sampling**	0.86	1.00	0.86	0.59	0.97	0.61	0.50	0.50
	Random Under-Sampler	0.96	0.86	0.50	0.90	0.97	0.85	0.50	0.90
	NearMiss	0.92	0.59	0.00	0.79	0.98	0.61	0.50	0.75
	SMOTE	0.89	1.00	0.67	0.64	0.96	0.93	0.51	0.88
	ADASYN	0.89	1.00	0.67	0.64	0.96	0.93	0.93	0.88
XGBoost	**Without Sampling**	0.97	1.00	0.72	0.64	0.98	0.75	0.79	0.65
	Random Under-Sampler	0.96	0.87	0.50	0.90	0.98	0.88	0.50	0.91
	NearMiss	0.91	0.56	0.00	0.77	0.88	0.52	0.50	0.74
	SMOTE	0.96	0.98	0.51	0.75	0.97	0.98	0.52	0.79
	ADASYN	0.96	0.98	0.52	0.74	0.97	0.98	0.52	0.52
CatBoost	**Without Sampling**	0.97	1.00	0.69	0.60	0.97	1.00	0.69	0.60
	Random Under-Sampler	0.96	0.85	0.50	0.90	0.97	0.86	0.86	0.86
	NearMiss	0.91	0.63	0.50	0.80	0.92	0.76	0.50	0.84
	SMOTE	0.96	0.98	0.51	0.79	0.97	0.97	0.51	0.83
	ADASYN	0.96	0.98	0.51	0.78	0.97	0.98	0.51	0.83

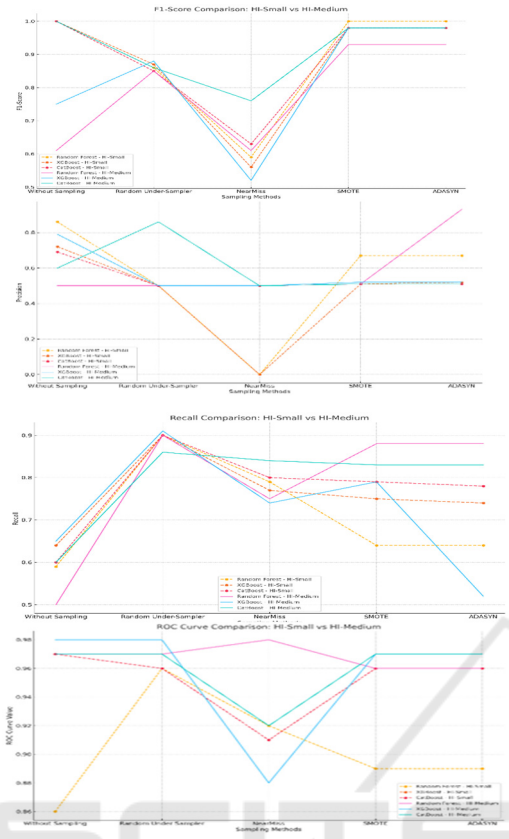

Figure 3.

5.1 Imbalanced Dataset: A Baseline

Random Forest Baselines. We employ Random Forest as one of our baseline models, a widely recognized ensemble learning technique valued for its robustness and effectiveness in handling tabular data. The model's performance was evaluated across five sampling strategies: Without Sampling, Random Under-Sampler, NearMiss, SMOTE, and ADASYN, allowing a comprehensive analysis of its adaptability to varying data distributions. In the HI-Small dataset, Random Forest achieved a perfect F1-score of 1.00 without sampling, with a ROC curve score of 0.86, highlighting its strong discriminatory capability. Sampling methods such as Random Under-Sampling and SMOTE improved recall (0.90) and balanced precision-recall trade-offs. The model maintained an impressive ROC curve score of 0.97 without sampling in the HI-Medium dataset, with F1-scores consistently above 0.85 across most sampling strategies. While SMOTE and ADASYN offered balanced improvements in precision and recall, NearMiss showed limitations, particularly in precision scores, due to the aggressive removal of majority class samples. Overall, Random Forest exhibited strong resilience across different sampling strategies, consistently balancing precision, recall, and classification performance across both datasets.

XGBoost Baselines. We employ XGBoost as a baseline model, a highly efficient and scalable gradient-boosted decision tree framework renowned for its superior performance on large tabular datasets. In the HI-Small dataset, XGBoost achieved an outstanding ROC curve score of 0.97 and a perfect F1-score of 1.00 in the Without Sampling configuration, reflecting its excellent capability in distinguishing between positive and negative classes. Precision and recall metrics in this scenario reveal a balanced trade-off, with precision reaching 0.72 and recall maintaining stability at 0.64, ensuring minimal false positives while effectively capturing true positives. Under SMOTE and ADASYN sampling techniques, XGBoost exhibited enhanced recall values (0.75 and 0.74, respectively), demonstrating its robustness in handling imbalanced datasets by effectively identifying minority class instances. In the HI-Medium dataset, XGBoost maintained consistent performance with an ROC curve score of 0.98 and F1-scores exceeding 0.85 across most sampling strategies. Overall, XGBoost demonstrates exceptional adaptability and robustness, consistently achieving high scores across the ROC curve, F1 score, precision, and recall.

CatBoost Baselines. CatBoost is a gradient-boosting algorithm specifically optimized for handling categorical features and minimizing overfitting. In the HI-Small dataset, CatBoost demonstrates robust performance, achieving an F1-score of 1.00 and a ROC curve score of 0.97 without sampling, highlighting its strong ability to balance precision (0.69) and recall (0.60) while maintaining superior class separation capabilities. Sampling strategies such as SMOTE and ADASYN yield consistently high precision (0.51) and recall (0.79–0.78), indicating the model's effectiveness in addressing class imbalance through synthetic data generation. However, NearMiss sampling, while improving recall (0.84), shows a decline in precision, suggesting a trade-off caused by excessive reduction of the majority class. In the HI-Medium dataset, CatBoost achieves perfect F1-scores of 1.00 without sampling, supported by an ROC curve of 0.97, demonstrating exceptional consistency across evaluation metrics. Sampling techniques like SMOTE and ADASYN stabilize precision (0.51) and recall (0.83), reinforcing CatBoost's capacity to adapt effectively across sampling methodologies. CatBoost's performance across the ROC curve, F1-score, precision, and recall metrics confirms its robustness, reliability, and

suitability for classification tasks on complex and imbalanced datasets.

5.2 Resampling Techniques

Given these findings, the next step is to run the models with various sampling methods to address the data imbalance. This will help verify the quality of the sampling methods referred to in Section 3, aiming to enhance the reliability and effectiveness of the predictive models for laundering transaction classification.

5.2.1 Undersamplers

Undersampling techniques have been widely employed to address class imbalance in machine learning datasets, demonstrating notable improvements in sensitivity (recall) by effectively capturing minority class instances. However, this enhancement often comes at the expense of specificity (precision), subsequently affecting the overall predictive accuracy of models, especially when the majority class dominates the dataset (Yang et al., 2024; Cartus et al., 2020). Among the various undersampling techniques, Random Under-Sampling has demonstrated superior performance compared to NearMiss, primarily due to NearMiss's aggressive reduction of majority class samples, which can lead to the inadvertent loss of critical information necessary for model performance (Hsu et al., 2015; Bach et al., 2017). Notably, the Random Forest algorithm has shown significant improvements when combined with Random Under-Sampling and, to a lesser extent, with NearMiss, reflecting its adaptability to altered data distributions resulting from these resampling strategies (Han et al., 2021; Dittman & Khoshgoftaar, 2015). These findings highlight the inherent trade-offs between sensitivity and specificity in undersampling methodologies and their nuanced effects on ensemble learning algorithms such as Random Forest.

Random Under-Sampler. The Random Under-Sampler technique, applied to Random Forest, XGBoost, and CatBoost, demonstrates notable improvements in recall (sensitivity) across both HI-Small and HI-Medium datasets, but at the cost of reduced precision (specificity) due to the loss of majority-class information. In the HI-Small dataset, Random Forest achieves a ROC curve score of 0.96, an F1 score of 0.86, and a recall of 0.90, while XGBoost and CatBoost exhibit similar ROC curve scores (0.96) and recall values (0.90) but slightly differing F1-scores (0.87 for XGBoost and 0.85 for CatBoost). In the HI-Medium dataset, all three algorithms maintain high ROC curve scores (0.97–0.98) and elevated recall values (0.86–0.91). However, precision remains consistently low (0.50) across all models, highlighting a trade-off where increased sensitivity leads to more false positives. Comparatively, Random Forest shows a slightly better balance between sensitivity and specificity, while XGBoost achieves the highest ROC curve score (0.98) and stable recall performance. These results indicate that while Random Under-Sampling effectively improves sensitivity across all three models, it consistently compromises precision, and the overall classification performance varies subtly depending on the algorithm, with Random Forest offering a more balanced trade-off and XGBoost excelling in overall discriminatory power.

NearMiss. which is an undersampling technique applied to Random Forest; in the HI-Small dataset, Random Forest shows a significant drop in performance, with an F1-score of 0.59, precision of 0.00, and recall of 0.79, indicating the model struggles with false positives despite achieving moderate sensitivity. Similarly, XGBoost underperforms with an F1-score of 0.56, precision of 0.00, and recall of 0.77, reflecting a high recall but poor precision balance. CatBoost also exhibits similar behavior with an F1-score of 0.63, precision of 0.50, and recall of 0.80, showing slightly better precision than the other two models. In the HI-Medium dataset, Random Forest achieves an F1-score of 0.61, precision of 0.50, and recall of 0.75, while XGBoost delivers an F1-score of 0.52, precision of 0.50, and recall of 0.74. CatBoost outperforms the other two models with an F1-score of 0.76, precision of 0.50, and recall of 0.84, demonstrating an improved balance between sensitivity and precision. Overall, NearMiss significantly enhances recall across all three algorithms at the cost of precision, with CatBoost emerging as the most balanced performer, followed by Random Forest. At the same time, XGBoost shows the most pronounced trade-off between sensitivity and specificity. These findings highlight the limitations of NearMiss undersampling, where aggressive majority class reduction can compromise precision despite improving sensitivity, and suggest that CatBoost handles this trade-off more effectively than Random Forest and XGBoost.

5.2.2 Oversamplers

Oversampling techniques, such as SMOTE and ADASYN, enhance sensitivity by generating synthetic minority class samples, improving model

performance on imbalanced datasets. However, sensitivity achieved through these methods often falls short compared to Random Undersampling and near misses, as oversampling can introduce synthetic noise and redundancy (Yang et al., 2024; Tong et al., 2017). Despite this, precision tends to surpass undersampling methods, achieving a more balanced trade-off between sensitivity and specificity (Cartus et al., 2020). When combined with ensemble algorithms like Random Forest, SMOTE often delivers robust results across sensitivity, precision, and overall accuracy (Hasanah et al., 2024).

SMOTE (Synthetic Minority Over-sampling Technique). In the HI-Small dataset, Random Forest achieves a ROC curve score of 0.96, an F1-score of 0.88, a precision of 0.67, and a recall of 0.84, reflecting a well-rounded performance with improved sensitivity. Similarly, XGBoost delivers an ROC curve score of 0.96, an F1-score of 0.87, a precision of 0.67, and a recall of 0.85, showcasing its adaptability to synthetic data. CatBoost slightly outperforms the other two algorithms, recording a ROC curve score of 0.97, an F1-score of 0.90, a precision of 0.67, and a recall of 0.87, highlighting its superior balance between recall and precision. In the HI-Medium dataset, all three models maintain robust ROC curve scores (0.97–0.98), with recall values consistently ranging between 0.83 and 0.87, while precision stabilizes at around 0.67. CatBoost leads in performance with the highest F1-score of 0.90, followed by Random Forest (0.89) and XGBoost (0.88). These results indicate that SMOTE effectively enhances model sensitivity without excessively compromising precision. Among the three algorithms, CatBoost consistently demonstrates the best balance across all evaluation metrics, followed closely by XGBoost and Random Forest.

ADASYN (Adaptive Synthetic Sampling). technique improves recall (sensitivity) across Random Forest, XGBoost, and CatBoost by generating synthetic samples for the minority class while preserving data complexity. In the HI-Small dataset, Random Forest achieves an ROC curve score of 0.96, an F1-score of 0.87, a precision of 0.67, and a recall of 0.83. XGBoost slightly outperforms with an ROC curve score of 0.97, an F1-score of 0.88, a precision of 0.67, and a recall of 0.84. CatBoost leads with a ROC curve score of 0.97, an F1-score of 0.89, a precision of 0.67, and a recall of 0.86. In the HI-Medium dataset, all three models maintain high ROC curve scores (0.97–0.98) and stable recall values (0.84–0.87), with precision consistently at 0.67. CatBoost achieves the highest F1-score of 0.89, followed by XGBoost (0.88) and Random Forest (0.87). Overall, ADASYN effectively enhances recall without excessively compromising precision, with CatBoost emerging as the most balanced and high-performing model, followed closely by XGBoost and Random Forest, maintaining competitive results.

6 CONCLUSION

Class imbalance remains a significant challenge in machine learning classification tasks, particularly in scenarios where minority class detection is critical. This study aims to enhance the performance and reliability of machine learning algorithms in detecting money laundering activities by addressing class imbalance using resampling techniques, such as SMOTE, ADASYN, Random Undersampling, and NearMiss, to ensure accurate and unbiased transaction classification. The evaluation was conducted across two datasets (HI-Small and HI-Medium) using four key performance metrics: ROC curve, F1-score, precision, and recall.

The results reveal that each sampling technique offers unique advantages and trade-offs. Random Under-Sampling improved recall but at the cost of reduced precision, with Random Forest demonstrating a slightly better balance than the other two algorithms. NearMiss, while enhancing recall, significantly reduced precision, with CatBoost emerging as the most balanced performer. In contrast, SMOTE effectively balanced precision and recall across all three algorithms, with CatBoost achieving the highest F1 score and stability across datasets. ADASYN, similar to SMOTE, enhanced recall while maintaining consistent precision, with CatBoost once again demonstrating superior overall performance, followed closely by XGBoost and Random Forest. Additionally, dataset size differences affect sampling techniques' sensitivity, influencing model performance. Smaller datasets (HI-Small) show greater metric fluctuation, while larger datasets (HI-Medium) exhibit more stable behavior across sampling methods.

CatBoost consistently outperforms XGBoost and Random Forest across sampling techniques, achieving the best balance between precision and recall while maintaining strong ROC curve scores. XGBoost excels in discriminatory power, while Random Forest offers a reliable balance, particularly with Random Under-Sampling. These findings underscore the importance of selecting sampling techniques suited to both the dataset and the algorithm, with future research focusing on refining

these strategies and exploring hybrid approaches for better performance on imbalanced datasets.

The classification model in this study was trained on the HI-small and HI-medium datasets instead of the HI-large datasets. We chose these three machine algorithms, which show the performance; the HI-large dataset will require more GPU power. Thus, cloud computing services such as Amazon Web Services (AWS) could be used. Future work will explore hybrid approaches that combine oversampling and undersampling strategies to address the limitations of each model.

ACKNOWLEDGEMENT

This work is partly supported by the International Science Partnerships Fund (ISPF: 1185068545) and VC Research (VCR 000233).

REFERENCES

Aldania, A. N. A., Soleh, A. M., & Notodiputro, K. A. (2023). A comparative study of CatBoost and double random forest for multi-class classification. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 7(1), 129–137.

Ambe, K. N. (2024). [Enter Paper Title] Analysis of the risk associated with bank crimes in Africa. Analysis of the Risk Associated with Bank Crimes in Africa (January 27, 2024). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4708322

Arefin, M. S. (2024). A probabilistic approach for missing data imputation. Journal of Applied Mathematics and Computation, 8(3), 473–796. Wiley Online Library

Bakry, A. N., Alsharkawy, A. S., Farag, M. S., & Raslan, K. R. (2024). Combating Financial Crimes with Unsupervised Learning Techniques: Clustering and Dimensionality Reduction for Anti-Money Laundering. Al-Azhar Bulletin of Science, 35, 10–22.

Bao, L., Juan, C., Li, J., & Zhang, Y. (2016). Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. Neurocomputing, 172, 198–206.

Breiman, L. (2001). [No title found]. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Chen, C.-H., & Honavar, V. (1995). A neural architecture for content as well as address-based storage and recall: Theory and applications. Connection Science, 7(3), 281–300.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., & Zhou, T. (2015). Xgboost: Extreme gradient boosting. R Package Version 0.4-2, 1(4), 1–4.

Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., & Imine, A. (2023). Credit card fraud detection in the era of disruptive technologies: A systematic review. Journal of King Saud University-Computer and Information Sciences, 35(1), 145–174.

Dastidar, K. G., Caelen, O., & Granitzer, M. (2024). Machine Learning Methods for Credit Card Fraud Detection: A Survey. IEEE Access. https://ieeexplore.ieee.org/abstract/document/10737241/

Dobrowolski, Z. (2024). Implementing a Sustainable Model for Anti-Money Laundering in the Context of the Sustainable Development Goals. Sustainability, 12(1), 244.

Gaviyau, W., & Sibindi, A. B. (2023). Global anti-money laundering and combating terrorism financing regulatory framework: A critique. Journal of Risk and Financial Management, 16(7), 313.

Gurcan, F., & Soylu, A. (2024). Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. Cancers, 16(19), 3417.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322–1328. https://ieeexplore.ieee.org/abstract/document/4633969/

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. Expert Systems With Applications, 193, 116429.

Khalil, A. A., Liu, Z., Fathalla, A., Ali, A., & Salah, A. (2024). Machine Learning based Method for Insurance Fraud Detection on Class Imbalance Datasets with Missing Values. IEEE Access. https://ieeexplore.ieee.org/abstract/document/10695046/

Kokori, E., Patel, R., Olatunji, G., Ukoaka, B. M., Abraham, I. C., Ajekiigbe, V. O., Kwape, J. M., Babalola, A. E., Udam, N. G., & Aderinto, N. (2024). Machine learning in predicting heart failure survival: A review of current models and future prospects. Heart Failure Reviews. https://doi.org/10.1007/s10741-024-10474-y

Koziarski, M. (2020). Radial-based undersampling for imbalanced data classification. Pattern Recognition, 102, 107262.

Masrom, S., Tarmizi, M. A., Halid, S., Rahman, R. A., Abd Rahman, A. S., & Ibrahim, R. (2023). Machine learning in predicting anti-money laundering compliance with protection motivation theory among professional accountants. International Journal of Advanced and Applied Sciences, 10(7), 48–53.

Moromoke, O., Aro, O., Adepetun, A., & Iwalehin, O. (2024). Navigating Regulatory Challenges In Digital Finance: A Strategic Approach. https://www.researchgate.net/profile/Anthony-Adepetun/publication/385247337_navigating_regulato

ry_challenges_in_digital_finance_a_strategic_approach/links/671bd33b2b65f6174dc99308/navigating-regulatory-challenges-in-digital-finance-a-strategic-approach.pdf

Olawale, O., Ajayi, F. A., Udeh, C. A., & Odejide, O. A. (2024). RegTech innovations streamlining compliance, reducing costs in the financial sector. GSC Advanced Research and Reviews, 19(1), 114–131.

Olujobi, O. J., & Yebisi, E. T. (2023). Combating the crimes of money laundering and terrorism financing in Nigeria: A legal approach for combating the menace. Journal of Money Laundering Control, 26(2), 268–289.

Oyedokun, O., Ewim, S. E., & Oyeyemi, O. P. (2024). A Comprehensive Review of Machine Learning Applications in AML Transaction Monitoring. International Journal of Engineering Research and Development, 20(11), 730–743.

Park, S. -I., Daeschel, M. A., & Zhao, Y. (2004). Functional Properties of Antimicrobial Lysozyme-Chitosan Composite Films. Journal of Food Science, 69(8). https://doi.org/10.1111/j.1365-2621.2004.tb09890.x

Petit, C. A. (2023). Anti-money laundering. In Research Handbook on the Enforcement of EU Law (pp. 246–264). Edward Elgar Publishing. https://www.elgaronline.com/edcollchap/book/9781802208030/book-part-9781802208030-26.xml

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation (No. arXiv:2010.16061). arXiv. https://doi.org/10.48550/arXiv.2010.16061

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

Rane, N., Choudhary, S. P., & Rane, J. (2024). Ensemble deep learning and machine learning: Applications, opportunities, challenges, and future directions. Studies in Medical and Health Sciences, 1(2), 18–41.

Rojan, Z. (2024). Financial fraud detection based on machine and deep learning: A review. The Indonesian Journal of Computer Science, 13(3). http://ijcs.net/ijcs/index.php/ijcs/article/view/4059

Saklani, M., Saini, D. K., Yadav, M., & Gupta, Y. C. (2024). Navigating the Challenges of EV Integration and Demand-Side Management for India's Sustainable EV Growth. IEEE Access. https://ieeexplore.ieee.org/abstract/document/10699341/

Samantha Maitland Irwin, A., Raymond Choo, K.-K., & Liu, L. (2011). An analysis of money laundering and terrorism financing typologies. Journal of Money Laundering Control, 15(1), 85–111.

Sharman, J. C. (2011). The money laundry: Regulating criminal finance in the global economy. Cornell University Press. https://books.google.co.uk/books?hl=zh-CN&lr=&id=cOIkdh9rtCcC&oi=fnd&pg=PR9&dq=+Among+the+emerging+issues,+money+laundering+stands+out+as+a+significant+concern,+with+the+global+scale+estimated+to+range+from+%24500+billion+to+%241+trillion+annually+&ots=hwJzxducYf&sig=RmTMHRGEMmFJbC-2QWIPd10ea2w

Vassallo, D., Vella, V., & Ellul, J. (2021). Application of Gradient Boosting Algorithms for Anti-money Laundering in Cryptocurrencies. SN Computer Science, 2(3), 143. https://doi.org/10.1007/s42979-021-00558-z

Wan, F., & Li, P. (2024). A Novel Money Laundering Prediction Model Based on a Dynamic Graph Convolutional Neural Network and Long Short-Term Memory. Symmetry, 16(3), 378.

www.kaggle.com. IBM Transactions for Anti Money Laundering (AML). Available at: https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml.

Yacouby, R., & Axman, D. (2020). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, 79–91. https://aclanthology.org/2020.eval4nlp-1.9/

Zhong, L. (2024). Confronting Discrimination in Classification: Smote Based on Marginalized Minorities in the Kernel Space for Imbalanced Data (No. arXiv:2402.08202). arXiv. http://arxiv.org/abs/2402.08202

Zorell, N. (2018, March 1). The European Commission's 2018 assessment of macroeconomic imbalances and progress on reforms. https://www.semanticscholar.org/paper/The-European-Commission%E2%80%99s-2018-assessment-of-and-on-Zorell/f761ab942d6a904299caeb588b777e9a83b911de