International Journal of
Network Dynamics and Intelligence (IJNDI)

Scilight

Article

Machine Learning for Money Laundering Analysis

Victor Chang [1,*], Hai Wang [2], Siddharth Raut [1], Assel Akzhalova [3], and Weihang Huang [4]

[1] Department of Business Analytics and Information Systems, Aston Business School, Aston University, Birmingham B4 7ET, UK
[2] School of Computer Science and Digital Technologies, Aston University, Birmingham B4 7ET, UK
[3] School of Information Technology and Engineering, Kazakh-British Technical University, Almaty 050000, Republic of Kazakhstan
[4] Department of Linguistics and Communication, University of Birmingham, Birmingham B15 2TT, UK
* Correspondence: v.chang1@aston.ac.uk or victorchang.research@gmail.com
Received: 10 March 2026; Revised: 25 September 2026; Accepted: 19 October 2026; Published: 9 November 2026

Abstract: This study investigates the application of machine learning (ML) for proactive threat hunting in financial transaction systems, focusing on detecting money laundering activities—a critical challenge in business cybersecurity. Utilizing the Elliptic Bitcoin dataset, we apply six supervised ML models: Random Forest, Support Vector Machine, K-Nearest Neighbors, Stochastic Gradient Descent, Decision Tree, and Logistic Regression. We address class imbalance using the Synthetic Minority Oversampling Technique (SMOTE) and conduct exploratory data analysis to identify key transactional features. Evaluated using cybersecurity-relevant metrics (accuracy, precision, recall, F1-score), the Random Forest classifier outperforms others, achieving 98.87% accuracy on the holdout set. This work contributes an empirical benchmarking study demonstrating that, on the Elliptic Bitcoin dataset, Random Forest achieves the highest macro-F1 (0.965) and ROC-AUC (0.996) under random-split evaluation, while also revealing substantial performance degradation for all classifiers under temporally ordered evaluation—a finding with direct implications for realistic AML deployment.

Keywords: money laundering investigations; anti-money laundering (AML); impact of AML for organizations; machine learning; random forest; Support Vector Machine; K-Nearest Neighbors; Stochastic Gradient Descent; decision tree; logistic regression

1. Introduction

Money laundering has become a major concern in today's financial scene, characterized by a worldwide market capitalization of more than $95 trillion and a predominance of digital transactions. This unlawful activity, which is expected to account for up to 5% of global GDP—or $2 trillion annually—poses a substantial threat to global economic stability and security. The evolving landscape of money laundering, typified by increasingly complex tactics, poses severe challenges to established detection systems and highlights their inadequacies, underscoring the urgent need for more effective solutions.

Traditional detection systems, mostly rule-based and manual, struggle to keep pace with the innovative techniques used by launderers. Hindered by high false positive rates, sluggish response times, and an inability to react to the changing nature of financial crimes, these systems are increasingly inadequate. The integration of machine learning (ML) into financial security provides a glimmer of hope by providing novel strategies to combat this prevalent problem. The capacity of machine learning to handle large datasets and identify complex patterns otherwise overlooked by humans is a significant advantage in money laundering investigations [1]. However, the incorporation of ML into anti-money laundering (AML) initiatives introduces challenges, such as data privacy concerns and the necessity for frequent updates to algorithms [2].

The banking sector, with a market value exceeding $7 trillion, is especially vulnerable to money laundering. Its reliance on robust fraud detection techniques underscores the need to develop sophisticated preventative measures. This entails not only the application of machine learning and other technological solutions, but also the formulation of legislative frameworks and international collaborations to address this global issue effectively.

The use of machine learning in AML operations offers a promising complement to rule-based detection systems. This study aligns with the broader AI-driven cybersecurity research agenda by providing a systematic empirical comparison of classifier performance on a publicly available, labelled cryptocurrency transaction dataset, contributing evidence-based guidance for practitioners selecting among conventional supervised learning approaches. Furthermore, combining ML-based detection systems and robust legal frameworks improves enterprises' ability to combat global financial crime networks, resulting in a safe and resilient financial ecosystem.

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license(https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1.1. Background and Context

1.1.1. The Evolving Landscape of Money Laundering

The money laundering environment is continually changing, driven by technological breakthroughs and an increasing number of global financial institutions. The International Monetary Fund estimates that money laundering represents 2% to 5% of global GDP, translating to $590 billion to $1.5 trillion annually, highlighting its huge economic impact. Africa, particularly Nigeria, has been hit hard, with billions of dollars being illegally moved due to lax banking laws and cash-based economies. Today's launderers increasingly employ digital technologies such as peer-to-peer platforms, digital currencies, and online transactions, to bypass traditional detection systems. Financial institutions are consequently pressured to revise their strategies and technologies. Furthermore, the rise of cryptocurrencies has introduced new challenges in financial regulation due to their potential use in criminal financial activities. This necessitates innovative regulatory approaches to manage the risks associated with their growing use in illicit transactions. Furthermore, research also refers to the use of international trade systems, such as import-export transactions, for money laundering. This complex and dynamic landscape emphasizes the importance of advanced, strategic solutions to successfully combat money laundering in a rapidly evolving global financial sector.

In addition to its direct financial consequences, money laundering is a major source of value loss for businesses. Since digital channels are more widely used, there are more opportunities for illegal financial activities, which put firms at risk from cyberattacks. Organizational value is constantly in jeopardy due to the complex interplay between money laundering and cyber threats. Organizations that have been involved in money laundering activities or suspected activities, voluntarily and involuntarily, can perhaps be subject to substantial penalties, legal prosecutions, loss of trust from clients, and long-term loss of market position.

1.1.2. Traditional Detection Systems: Limitations and Challenges

Traditional money laundering detection systems, as noted by Berkan Oztas et al. [3], face considerable obstacles. These rule-based and inflexible systems suffer from high false positive rates, low detection accuracy, and a lack of adaptability to sophisticated laundering tactics. This results in operational inefficiencies and diverts resources away from addressing legitimate threats. While modern technologies such as machine learning offer potential improvements, their integration is hampered by data-related challenges, including limited access to high-quality datasets, concerns over privacy and legal compliance.

Furthermore, AML initiatives require more flexible and transparent technologies to ensure regulatory compliance and to keep pace with the evolving nature of financial crimes. This circumstance underscores the urgent need for the financial system to advance towards more sophisticated, flexible, and transparent detection systems.

1.1.3. The Advent of Machine Learning in Financial Security

Alexandra Prisznyák's research explores the integration of machine learning (ML) into banking risk management, particularly in AML and counter-terrorism financing. This work represents a significant shift from traditional rule-based systems to adaptive, data-driven ML models, highlighting the importance of tailoring ML techniques to meet the specific needs of individual financial institutions. The study investigates a variety of machine learning methods, including supervised, unsupervised, and hybrid models, that are specifically designed to address the imbalanced datasets commonly seen in AML contexts. Prisznyák emphasizes the necessity of selecting ML algorithms based on unique institutional requirements and theoretical frameworks of each institution. The report also emphasizes the significance of a collaborative strategy that integrates insights from business, IT, and management to implement ML technologies, thereby enhancing financial security procedures.

1.1.4. Key Advantages of ML in Money Laundering Analysis

Alotibi et al. [4] highlight the substantial benefits of utilizing machine learning (ML) and deep learning techniques to detect money laundering activities, especially within the realm of cryptocurrencies. Their study examines the growing complexity and rapid evolution of money laundering tactics, which have outpaced traditional detection technologies. The research demonstrates that applying various ML algorithms—such as Deep Neural Network (DNN), Random Forest (RF), K-Nearest Neighbors (KNN), and Naive Bayes (NB)—on the Bitcoin elliptic dataset enhances detection accuracy while reducing false positives. Notably, the Random Forest classifier stands out, with an F1-score of 0.99%, demonstrating the effectiveness of ML in detecting suspicious financial transactions. This study underscores the transformative potential of ML and deep learning in strengthening financial security and advocates for their broader adoption in the ongoing battle against money laundering.

1.1.5. Challenges and Considerations

While machine learning (ML) offers promising solutions for preventing money laundering, it also presents significant challenges and considerations. A major impediment in this endeavor is the scarcity of public datasets, which are essential for developing and validating robust ML models by [2], who propose using synthetic data generation to enhance the resilience and applicability of ML systems. Furthermore, the need for addressing emerging areas such as semi-supervised and deep learning is emphasized, as these hold promise for more sophisticated detection of suspicious activities. Equally crucial is ensuring that ML models are both interpretable and fair. As these technologies become increasingly integrated into financial institutions, it is crucial to guarantee they function transparently and without inherent biases. This academic discourse thus highlights the intricate challenges and vital considerations involved in leveraging ML for AML efforts, underscoring the need for ongoing research and innovation in this critical field.

1.1.6. The Importance of Fraud Detection in the Global Banking Sector

Rambola et al. [5] explore the crucial role of fraud detection in protecting daily financial transactions within the evolving banking landscape. Their study highlights the vulnerabilities of both physical and online banking institutions to a wide range of fraudulent activities, including credit card theft, insurance fraud, and accounting fraud. These activities pose significant risks, potentially leading to huge financial losses for both institutions and their clients. The authors advocate for the adoption of data mining techniques in fraud detection, emphasizing the use of historical data and probabilistic analysis as critical tools for predicting, and preventing fraud. The report emphasizes the urgent need for advanced data analysis approaches in the banking industry. By employing these modern techniques, banks can more effectively prevent fraud, protect financial assets, and strengthen client trust.

1.1.7. The Impact of Money Laundering on E-Commerce

The fast growth of e-commerce has greatly increased the chances of money laundering, introducing new challenges to the digital financial landscape. Sophisticated technologies, such as cryptocurrencies and complex online financial networks, are increasingly used to facilitate money laundering within this sector. This result highlights the vital importance of effective detection and prevention strategies on e-commerce platforms. Collaboration between these platforms and regulatory authorities is crucial, emphasizing the need for new technology solutions and strict regulatory procedures. Such coordinated initiatives are critical for effectively combating the risks of money laundering, especially considering the rapid expansion and evolution of e-commerce. The situation requires greater attention and proactive efforts to ensure the integrity of online financial transactions and protect the digital economy from the detrimental effects of money laundering.

1.1.8. The Need for Better Technologies to Prevent Money Laundering

In the digital age, combating money laundering has become increasingly challenging, demanding the use of more complex technological solutions. Money laundering strategies have evolved, particularly through digital platforms, cryptocurrencies, and virtual assets, making detecting and preventing these unlawful operations more difficult. This trend highlights the critical need for stronger legal frameworks and technological developments to mitigate the risks associated with digital money laundering. Key measures in this battle include enhanced Know Your Customer (KYC) regulations, robust transaction monitoring systems, and increased international cooperation [1]. The focus is not only on improving existing systems but also on adapting to the constantly evolving digital tactics of money launderers. The integration of cutting-edge technologies is essential for effectively identifying and preventing illegal financial activities, emphasizing the urgent need for modern solutions in the financial sector to combat money laundering.

1.1.9. Sovereign AI Considerations in Anti-Money Laundering

International AML regimes are impacted by sovereign AI, or state-produced and controlled AI systems. Financial institutions face difficulties implementing ML-based AML systems in several jurisdictions as nations emphasise data sovereignty and technological autonomy. International efforts to detect money laundering may be significantly affected by national differences in data localisation, model transparency, and cross-border information-sharing rules.

The stringent financial transaction data processing and sharing requirements of China's Personal Information Protection Law (PIPL) and the EU's General Data Protection Regulation (GDPR) affect the training and application of machine learning models for anti-money laundering (AML) purposes. Multijurisdictional financial institutions need to maintain detection while harmonising these conflicting regulations. Since illicit payments occasionally cross borders, money laundering necessitates coordinated detection efforts that preserve state data sovereignty.

Algorithmic accountability and transparency are concerns raised by the creation of Sovereign AI systems for AML. Private, centralised machine learning models might be more accurate, but they may conflict with local regulatory controls and national AI mandates. The conflict between technological effectiveness and sovereign authority will determine how AI-driven AML systems are implemented globally in the future.

1.2. Motivation

The topic "Machine Learning for Money Laundering Analysis" was chosen due to the pressing need to address the increasingly complex and ubiquitous issue of money laundering in the international financial system. Conventional anti-money laundering (AML) measures are inadequate in countering the sophisticated strategies used by launderers, highlighting the urgent need for more effective, efficient, and adaptable solutions. Machine Learning (ML) is emerging as a crucial technology in this area, potentially revolutionizing AML tactics through its superior data processing and pattern detection capabilities.

ML has demonstrated significant potential in enhancing decision support for fraud investigations within the financial industry. For example, Sintayehu and Seid [6] designed an ML algorithm for AML that yielded promising results. Additionally, Ref. [3] emphasized the importance of improving transaction monitoring systems with ML to keep pace with the ever-evolving strategies of money launderers.

However, integrating ML into AML frameworks introduces complex challenges, particularly concerning data privacy and legal compliance. Thommandru et al. [7] explored these issues, underscoring the importance of adopting a balanced approach that leverages the technological benefits of ML while addressing ethical and legal concerns.

1.3. Aim and Objective of the Study

This study conducts a rigorous empirical comparison of six conventional supervised classifiers applied to the Elliptic Bitcoin dataset for anti-money laundering (AML) detection, evaluating each model under both random and temporally ordered evaluation conditions. To achieve this, the study will focus on several key objectives:

- Empirically comparing six conventional supervised classifiers (Logistic Regression, SVM, KNN, SGD, Decision Tree, Random Forest) on the Elliptic Bitcoin transaction dataset under methodologically rigorous conditions.
- Analysing top-ranked transactional features identified by the Random Forest classifier to understand which attributes are most discriminative for illicit transaction detection.
- Providing a reproducible evaluation pipeline that applies stratified splits, correct scaler-then-SMOTE ordering, and a global random seed of 42, correcting methodological flaws present in prior runs.
- Introducing a temporally ordered evaluation—splitting by time steps 1–29 / 30–39 / 40–49—to assess classifier performance under conditions that simulate real-world AML deployment.

1.4. Research Questions

Our study will focus on solving two major research questions:

- Which supervised machine learning algorithms excel in accuracy and efficiency when detecting fraudulent bank transactions, and in what specific scenarios do they demonstrate optimal performance?
- How does the accuracy of a machine learning model in identifying suspicious financial activities influence the overall effectiveness of money laundering detection strategies?

1.5. Structure of the Paper

The foregoing review establishes that, despite growing interest in ML-based AML, no study has systematically compared a standard set of six conventional classifiers on the Elliptic dataset under reproducible, temporally sound evaluation conditions. This study fills that gap by benchmarking Logistic Regression, SVM, KNN, SGD, Decision Tree, and Random Forest classifiers, correcting common methodological flaws, and providing both random-split and temporal-split results as a baseline for future work.

This paper is structured as follows. Section 1 introduces the topic by outlining the background, motivation, goals, objectives, and research questions. Section 2 provides a literature review that discusses relevant material, identifies gaps, and summarizes major themes. Section 3 details the methodology, which includes software use, data sources, preprocessing, exploratory analysis, proposed algorithms, confusion matrices, and performance metrics. Sections 4 and 5 delve into data analysis and ML analysis, respectively, featuring comparisons of algorithms, hyperparameter configurations, and specifics on the performance of the Random Forest classifier. The paper finishes in Section 6 with discussion of the findings, addressing the research question, and providing limitations, recommendations, and references.

2. Literature Review

2.1. Introduction

Prior work on ML-based AML detection spans rule-based supplements, conventional supervised classifiers, and advanced ensemble methods. The following review is organised thematically—beginning with supervised classifiers applied to cryptocurrency transaction data, proceeding to ensemble and boosting approaches, and closing with a synthesis of the gap that motivates the present comparative study.

2.2. Supervised Machine Learning Techniques for Transaction Analysis

Weber et al. [8] applied ML techniques to classify Bitcoin transactions as either illicit or licit. The research was driven by the need to effectively identify illegal activities within the complex landscape of cryptocurrency transactions. They utilized a substantial dataset of over 200,000 Bitcoin transactions, which presented a significant challenge due to its highly imbalanced nature, with illicit transactions constituting only 2% of the dataset. The methods employed in this study—logistic regression, random forests, and graph convolutional networks—were chosen for their effectiveness in handling large and complex datasets. These models were evaluated using accuracy and F1-score, metrics well-suited for imbalanced datasets. The study also incorporated temporal train-test splits, a methodological approach that enhances the robustness of the findings by ensuring that the models are tested on data that simulates real-world conditions.

Despite the methodological strengths, the study by Weber et al. [8] had notable limitations. One significant shortcoming was the need for a comparative analysis among the employed algorithms. Such an analysis would have been invaluable in understanding the relative performance and suitability of each model for analyzing cryptocurrency transaction. Furthermore, the study did not offer a detailed rationale for the selection of specific algorithms, leaving a gap in understanding the decision-making process behind model choice. This area is ripe for future research, particularly in understanding the strengths and weaknesses of different ML approaches, which is crucial for advancing AML efforts in the rapidly evolving domain of cryptocurrency transactions.

While Weber et al. [8] demonstrated that temporal splits produce more conservative but realistic performance estimates, their study notably omitted a direct algorithmic comparison—a gap that motivates the systematic benchmarking approach of the present study. By contrast,

Alotibi et al. [4] delved into applying ML techniques for detecting money laundering activities within the realm of cryptocurrencies, noting the significant global scale of money laundering, estimated between $800 billion to $2 trillion annually, with approximately $5 billion related to cryptocurrency. The study critiqued the inadequacies of traditional rule-based systems used by exchanges, which suffer from high false positive rates and poor detection capabilities. To tackle these issues, the authors tested four supervised learning methods: Deep Neural Network (DNN), Random Forest (RF), K-Nearest Neighbors (KNN), and Naive Bayes (NB) on the elliptic Bitcoin dataset, provided by elliptic.co, which includes over 200,000 transactions labeled as legal or illegal. They evaluated these models using precision, recall, F1-score, and ROC AUC.

2.3. Ensemble Methods and Advanced Techniques in ML for AML

Moving beyond single-model approaches, a strand of literature has explored ensemble and gradient-boosting methods that offer higher discriminative power but at the cost of reduced interpretability.

Domashova and Mikhailina [9] embarked on a comprehensive exploration of ML methods for the early detection of money laundering activities by organizations. Their research primarily focused on the use of ensemble methods, with a particular emphasis on gradient-boosting techniques applied to decision trees. This methodological choice was driven by the need to effectively identify entities involved in money laundering, a critical challenge in the financial sector. The study involved an in-depth analysis of various feature encoding and selection methods aimed at optimizing the performance of gradient boosting models. To achieve this, the authors utilized advanced optimization libraries such as Hyperopt and Optuna, known for their efficiency in fine-tuning ML models. The effectiveness of these optimized gradient boosting models was evident in their performance metrics, including ROC-AUC, precision, and recall scores. These metrics were impressively high across different datasets, indicating the robustness and reliability of the models in detecting potential money laundering activities.

The significance of the findings by Domashova and Mikhailina [9] lies in their practical implications in enhancing AML and counter-terrorism financing measures within the financial sector. By demonstrating the substantial potential of gradient-boosting models in identifying risky entities, the study provides valuable insights for financial institutions looking to strengthen their compliance controls. The high scores in key performance metrics such as ROC-AUC, precision, and recall underscore the pivotal role of ML in enhancing the detection capabilities

of AML systems. Furthermore, the research highlights the importance of feature selection and optimization in ML models, suggesting that careful tuning and selection of features can greatly enhance the effectiveness of ML applications in AML. Overall, the study by Domashova and Mikhailina is a significant contribution to the field, clearly demonstrating how advanced ML techniques can be leveraged to bolster the financial sector's defenses against money laundering and related financial crimes.

Vassallo et al. [10] investigated the use of gradient-boosting algorithms for detecting money laundering and other illicit activities within cryptocurrency infrastructures. The study was particularly focused on identifying scams, terrorism financing, and Ponzi schemes at both the account and transaction levels. The researchers used advanced ML techniques to address the challenges of class imbalances and the evolving nature of criminal tactics in the digital currency space. A key innovation of their study was the development of an adapted version of eXtreme Gradient Boosting (XGBoost), termed Adaptive Stacked eXtreme Gradient Boosting (ASXGB). This model was specifically designed to be effective in dynamic environments, a critical feature given the rapidly changing landscape of cryptocurrency transactions. Alongside ASXGB, the study also tested other leading supervised gradient boosting classifiers, including Light Gradient Boosting Machine (LGBM) and CatBoost, to evaluate their effectiveness in identifying illegal activities.

The findings of the study were significant. The decision tree-based gradient boosting algorithms, particularly the newly developed ASXGB, demonstrated superior performance compared to the Random Forest (RF) model across various metrics, including accuracy, precision, recall, and F1 score. The use of data-sampling techniques like NCL-SMOTE was found to enhance the recall rate further. Notably, the ASXGB model excelled in managing concept drift, a common challenge in cryptocurrency markets, improving recall by 2.6%. This improvement is crucial in an environment where criminal tactics are constantly evolving. The study by Vassallo, Vella, and Ellul [10] contributes valuable insights into the application of ML in combating financial crimes in cryptocurrencies. It highlights the potential of advanced algorithms like ASXGB in adapting to and effectively addressing the dynamic and complex nature of money laundering and related illicit activities in the digital currency domain.

Taken together, the ensemble-based approaches reviewed above—particularly gradient-boosting variants [9,10]—tend to outperform conventional classifiers in terms of AUC and F1, yet require more careful hyperparameter tuning and are less commonly deployed in practice. This creates an evidential gap: it remains unclear how a standardised set of conventional classifiers performs under a well-controlled, reproducible evaluation on the same benchmark dataset, a gap the present study directly addresses.

2.4. Comparative Analysis of ML Models in AML

Raiter [11] focused on enhancing fraud detection capabilities within AML frameworks using supervised ML algorithms. The research involved rigorous empirical testing of four distinct ML models: Logistic Regression, Support Vector Machine (SVM), Random Forest, and Artificial Neural Network (ANN). These models were applied to a synthetic dataset designed to mimics real-world financial transactions. The study's primary objective was to evaluate and compare the effectiveness of these algorithms in accurately identifying fraudulent activities, a critical aspect of AML efforts in financial institutions. The Random Forest algorithm emerged as the most effective model, achieving an impressive accuracy score of 0.99. It also scored high in terms of precision (0.91) and had a reasonable F1-score (0.36) and sensitivity (0.21). In contrast, the ANN model showed the least accuracy, suggesting that while all tested models have potential applications in AML, Random Forest stands out for its suitability in real-world settings, particularly due to its high accuracy and interpretability. This finding is significant as it highlights the practical utility of ML in differentiating between legitimate and money laundering transactions. Raiter's [11] research contributes valuable insights into the application of ML in AML, demonstrating the potential of these algorithms to strengthen the fraud detection capabilities of financial institutions and enhance their overall AML frameworks.

2.5. Related Literature

This subsection examines additional relevant literature that contributes to understanding the broader application of ML in AML contexts.

Lokanan [12] conducted a study on the application of ML for AML in Middle Eastern banks and compared it to traditional rule-based approaches. The research evaluated three ML models: Naïve Bayes, Random Forest, and Artificial Neural Networks (ANNs), with ANNs outperforming the traditional approaches, achieving an accuracy rate of 88%. This research highlights the utility of simple ML models in improving AML detection. However, the reliance on simulated data and a limited focus on transaction-level abnormalities, rather than larger AML tactics such as network analysis, presents areas for future investigation. The study emphasizes the potential of ML in AML and stresses the importance of further extensive research utilizing real-world data to maximize the utilization of ML in the complex area of financial crime detection.

Krishna et al. [13] investigated the use of ML for AML, focusing on Linear Support Vector Machine (SVM) and Decision Tree (DT) classifiers. They used a synthesized dataset comprising 45 banking transactions with specified attributes. They distributed the dataset into 80% training and 20% validation segments. The Decision Tree classifier surpassed the Linear SVM, achieving 92% accuracy, 93% precision, and 91% recall, compared to the Linear SVM recorded 86% accuracy, 87% precision, and 85% recall. Despite these encouraging results, the use of a synthetic dataset and the lack of thorough hyperparameter optimization highlight areas for further improvement. Future research should focus on applying these models to real-world datasets and aim to enhance the interpretability and explainability of the models within AML detection frameworks.

Savage et al. [14] described an automated technique for identifying money laundering activities within financial transaction networks. Implemented by the Australian Transaction Reports and Analysis Centre (AUSTRAC), this system integrates network analysis with supervised learning techniques, such as Random Forest and SVM, to successfully detect suspicious activities. This system is notable for its ability to analyze both explicit transactional relationships and implicit connections derived from additional data. The classifiers, particularly Random Forest, demonstrated high precision and accuracy, indicating their viability for use in real-world scenarios. The system's ability to reduce false positives is crucial for its value in real intelligence contexts, demonstrating the potential of integrating network analysis and ML in combating money laundering.

Hampo, Nwokorie, and Odii [15] developed a k-nearest neighbor (KNN) classifier to detect money laundering within banking transactions. Using publicly available datasets from Kaggle, this study evaluated the performance of this supervised learning system in discriminating between legitimate and illegitimate financial activities. The dataset had 25 characteristics and 1480 observations, emphasizing fraud detection. The researchers used four distinct models, varying in scaling and hyperparameter tuning, to assess performance indicators such as accuracy, precision, recall, and F1 score. They observed that hyperparameter adjustments significantly improved model performance, with the optimal model achieving 98.4% accuracy, 98.3% precision, 100% recall, and 99.2% F1. Interestingly, scaling the data did not influence these performance indicators. This study underscores the effectiveness of KNN classifiers in financial crime detection and suggests the need for further research into data scaling and the deployment of additional supervised learning algorithms in this field.

Álvarez-Jareño et al. [16] applied ML approaches to detect financial fraud and money laundering. Using Benford's Law to categorize firms as lawful or fraudulent, the study implemented logistic regression, neural networks, decision trees, and random forests on data from over 600 enterprises. The work primarily focused on enhancing these models' predictive capabilities by employing a cost matrix and Synthetic Minority Oversampling Technique (SMOTE) to address the dataset's imbalance. The results show that the SMOTE algorithm improves real positive outcomes compared to the cost matrix technique.

Diab [17] focused on optimizing the Stochastic Gradient Descent (SGD) algorithm for text classification, using the Global Terrorism Database as a case study. The study, published in the International Journal of Computer Science and Information Security, aimed to fine-tune the hyperparameters of several classifiers, such as SVM, Logistic Regression, and Perceptron, within the SGD framework. The methodology involved data preparation, feature extraction with n-grams and TF-IDF, and a comprehensive comparison of classifiers under various conditions. The findings emphasize that hyperparameter adjustments considerably enhances the accuracy and efficiency of SGD-based classifiers. For example, the SVM achieved an accuracy of 0.87946, Logistic Regression 0.87192, and Perceptron 0.87378, demonstrating the effectiveness of this method.

Tertychnyi et al. [18] proposed a new ML framework designed to improve AML operations in the banking industry. This research, presented in their book, tackles fundamental issues in AML modeling, including scalability and class imbalance. The proposed two-layered framework incorporates a dual-layer model approach, extensive feature extraction methodologies, and effective class imbalance management procedures. The study fully assesses the framework's performance, categorizing results into the first layer, second layer, and overall outcomes, demonstrating its ability to enhance AML processes considerably. The study concludes with a performance evaluation of the model, suggesting directions for future research and highlighting the framework's applicability and effectiveness in the field of AML.

Jensen & Iosifidis [2] developed a sequence-based long short-term memory (LSTM) neural network to identify the layering stage of money laundering within extensive transaction histories. They examined data from 3,000 bank clients over two years, where 4% exhibited verified illegal financial behavior. The model attained 73% accuracy using 5-fold stratified cross-validation. This promising result highlights the need for additional refinement in sequence modeling for AML applications. The study's narrow reach, which includes a tiny sample size from a single institution, underscores the necessity for broader research.

Shokry et al. [19] introduced an unsupervised ML method for detecting money laundering tendencies aimed at

combatting terrorist financing. The authors compared two outlier identification techniques, Isolation Forest (IF) and One-Class SVM (OCSVM), on a synthetic dataset of 118,250 transactions patterned after real-world patterns. IF surpasses OCSVM, finding 11,825 abnormalities in 7 seconds of training time with high accuracy verified by domain experts. All 50 suspicious accounts and 414 highlighted transactions were classified as very hazardous. The method detects suspicious patterns in deposit/withdrawal amounts and time across accounts without requiring known suspicious behavior characteristics. Limitations include fake data and a tiny sample size. Overall, unsupervised learning produces effective outcomes for timely AML, but real-world bank data would be more useful in validating its efficacy.

Amini et al. [20] examined ML strategies for AML using synthetic transaction data, utilizing both supervised and unsupervised methods including K-means clustering, Isolation Forest, One-class SVMs, and Logistic Regression with accuracies ranging from 80–90%. Logistic regression was the most accurate (87–90%) for known AML patterns, whereas unsupervised approaches such as Isolation Forest and One-class SVMs outperformed in discovering novel abnormalities. The use of synthetic data limited the study's real-world application, as it focused solely on transaction abnormalities rather than wider AML tactics such as network analysis. Amini et al. [20] emphasized the possibility of combining several ML models for comprehensive AML diagnosis, highlighting the necessity for further research to address present constraints.

2.6. Literature Gap

This section outlines identified gaps in existing research based on a review of key machine learning algorithms used in anti-money laundering (AML) efforts (Table 1). The summary includes the algorithms' reported accuracy and highlights areas where further research is needed.

Table 1. Gaps in existing research.

Author	Algorithm	Key Metrics	Reported Accuracy	Gap
Alotibi et al. (2022) [4]	Random Forest	F1 score, AUC	RF: F1 score and AUC of 0.99	Limited testing on diverse datasets; exclusion of unlabeled transactions.
Weber et al. (2019) [8]	Random Forest	Accuracy, F1 score	97.7% accuracy, 78.8% F1 score	Lack of comparative algorithm analysis; need for model selection rationale
Raiter (2021) [11]	Random Forest	Accuracy	99% accuracy	the study lacks validation on real-world datasets,
Domashova and Mikhailina (2021) [9]	Gradient boosting	ROC-AUC, precision, recall	High scores across different datasets	Need for comprehensive AML strategies and practical insights.
Vassallo, Vella, and Ellul (2021) [10]	ASXGB, LGBM, CatBoost	Accuracy, precision, recall, F1 score	The superior performance of ASXGB in dynamic environments	Addressing class imbalances and evolving criminal tactics.

2.7. Summary

The study of existing ML research in money laundering detection reveals some critical gaps that must be addressed for future progress. A significant challenge is the imbalance in datasets, in which instances of criminal transactions far outnumber lawful ones, thus hindering model training and accuracy. Furthermore, the frequent use of simulated or non-real-world data raises concerns about the models' practical application in real-life circumstances. There is also a distinct paucity of comparative analysis of various ML strategies, which is critical for identifying the most efficient techniques in this sector.

Furthermore, research is frequently restricted in scope, focusing only on detecting suspicious transactions rather than incorporating broader AML tactics or considering the complete spectrum of financial crime. The adaptation of models to emerging criminal strategies and their efficiency across various transactional situations has also been minimally explored. Addressing these shortcomings, particularly by using more realistic data, broadening AML methods, and undertaking comparative studies, is crucial for developing more robust and effective ML solutions in the fight against money laundering.

3. Methodology

In this section, we will detail the methodology used in our study, beginning with the software used for the ML algorithms, followed by discussions on the data source, data preprocessing, and Exploratory Data Analysis (EDA).

ML algorithms are critical in money laundering analysis because of their capacity to handle massive datasets, flexibility, accuracy, and efficiency. Financial institutions that process large volumes of transactions regularly use these algorithms to spot potential laundering activities that might be overlooked by humans or traditional systems. The evolving nature of money laundering strategies requires adaptive models, and machine learning, primarily supervised and deep learning techniques, excels at adapting to new approaches. These algorithms greatly minimize false positives and negatives, which are common in rule-based systems, thereby enhancing the accuracy of fraud detection. Machine learning's automated capabilities enable real-time analysis and processing, facilitating quick reactions to emerging threats. Furthermore, these algorithms are highly effective at analyzing complicated linkages within transaction networks, uncovering hidden patterns suggestive of illegal activities.

3.1. Advantages and Added Values for Our Methodologies

Our approach, based on machine learning (ML) techniques, provides businesses with a strong framework to estimate the possible value destruction caused by money laundering activities. Our method allows organizations to spot anomalies with high accuracy by examining transaction patterns with the following advantages:

- Calculate the possible monetary losses from illegal transactions that go unnoticed.
- Determine the likelihood of regulatory fines by considering the quantity and type of questionable activities.
- Calculate reputational harm by comparing variables related to market perception with identified fraudulent activity.
- Analyze how well the current anti-money laundering policies are protecting the value of the organization.

With the help of this quantitative methodology, firms can strengthen their defenses against value-destroying money laundering activities by gaining actionable information.

Additionally, our approach addresses critical Sovereign AI concerns by ensuring that the model is transparent and can be adapted to a variety of legal frameworks. Our supervised learning methodology enables financial institutions to ensure compliance with jurisdiction-specific regulations while utilising advanced machine learning capabilities by enabling localised model training that adheres to data residency mandates. This is particularly relevant for multinational banks that operate in regions with rigorous data localisation regulations, as transaction data is prohibited from traversing borders for the purpose of centralised model training. Our approach establishes a foundation that can be customised to meet the requirements of a sovereign without compromising the precision of detection by demonstrating its efficacy using publicly accessible datasets.

3.2. Software Used, Data Source and Collection

For our analysis, Python, a prominent programming language in data science, was used to implement the ML algorithms. The comprehensive range of libraries and frameworks, available in Python, such as pandas for data manipulation, NumPy for numerical computations, and scikit-learn for machine learning, provided an invaluable toolkit for conducting our study.

Data collection is a systematic process that involves gathering and measuring information from different sources to provide an accurate and comprehensive overview of a specific area of interest. It helps academics, analysts, and enterprises to make informed decisions based on reliable data. For our study, we utilized secondary data due to its accessibility. A significant volume amount of data is available on open-source platforms, which are invaluable resources for researchers. The data used for the study is sourced from Kaggle, a well-known online platform that hosts data science and machine learning competitions. Kaggle allows users to search for and publish datasets, construct models in a web-based environment, and interact with other professionals in the field.

The Elliptic dataset on Kaggle was used for the study. The dataset focuses on Bitcoin transactions primarily based on the Ethereum blockchain. It is thoroughly managed to categorize transactions as 'licit', 'illicit', or 'unknown', facilitating a deeper understanding of their multifaceted nature. It consists over 200,000 Bitcoin transactions and more than 165 variables, featuring both numerical and categorical data. This comprehensive dataset provides insights into transaction amounts, entity behaviors, and network connections. Numerical features in the dataset include crucial transactional information such as transaction values and the number of inputs and outputs. Categorical data, on the other hand, adds a layer of complexity by offering insights into the nature of the transactions, the entities involved, and their interactions within the blockchain network.

The dataset is divided into three major files, each of which provides distinct insights:

(1) Elliptic_txs_classes.csv: This file categorizes transactions as 'licit', 'illicit', or 'unknown' for supervised learning. The categorization is based on Elliptic's proprietary analytics platform, which utilizes blockchain analytics and intelligence to distinguish between legal and illegal transactional activities.
(2) Elliptic_txs_edgelist.csv: This file includes transaction information as an edge list, which is crucial for

developing transaction networks and analyzing transaction flows. It is especially valuable for graph-based analytical approaches and network analysis techniques that aim to comprehend the intricate relationships and patterns common to blockchain transactions.

(3) Elliptic_txs_features.csv: Serving as the dataset's core, this file contains 166 features for each transaction, including information about transaction flow (e.g., time step, amount, fee), privacy (e.g., number of incoming/outgoing transactions), and neighborhood features. These features are anonymized to safeguard privacy while maintaining the necessary qualities for analysis.

3.3. Data Preprocessing

Data preprocessing is an important phase in the data analysis process that converts raw data into a clean and usable format. It comprises activities like managing missing values, normalization, transformation, and feature extraction, which are essential for effective ML and data analytics [21].

In our study, the first step is to label the anonymized dataframe of features. Initially, we set the names 'id' and 'time' to the first two columns. The subsequent 93 columns are labeled as 'trans_feat_{i}', and the following 72 columns as 'agg_feat_{i}'. Here, '{i}' represents numerical indices ranging from 0 to 93 for 'trans_feat_{i}' and from 0 to 72 for 'agg_feat_{i}'. This system of labeling the variables is based on the descriptions provided in the dataset and was previously used by [8].

The next step involves meticulous data cleaning, which is critical to ensure the reliability and correctness of the analysis, especially in the sensitive domain of blockchain transaction analysis.

Prior to modelling, transactions labelled unknown were explicitly excluded from the dataset. Of the 203,769 total transactions, 157,205 carried no ground-truth label and were removed, leaving 46,564 labelled samples for analysis (4545 illicit, 42,019 licit). This filtering step is performed before any splitting or resampling.

3.4. Duplicate Check and Removal

Duplicate removal in data processing involves detecting and deleting duplicate entries within a dataset to maintain data correctness and integrity. This step is critical for preserving the quality of data analysis since duplicates can introduce bias, leading to incorrect conclusions [22].

The process begins by carefully removing duplicate entries across the dataset's classes, features, and edge list. We start by recording the original shapes of classes, edge list, and features, then apply drop_duplicates(inplace=True) to each and recheck their forms. The absence of changes in the dataset's structure after this operation confirms the lack of duplicates, thus maintaining data integrity and validating their eligibility for trustworthy algorithm analysis.

3.5. Handling Missing Values

Handling missing values is an essential stage in data preprocessing, involving strategies to manage or replace absent data to maintain the dataset's integrity and usefulness for analysis. This step is critical for ensuring the quality of datasets used in statistical studies and ML models. Our examination revealed that no null or missing values were found across all three dataframes—classes, features, and edge list. The 'classes' dataframe, containing the text and classification columns, showed no null values, ensuring the integrity of transaction IDs and classifications. Similarly, the 'features' dataframe, which includes 167 columns ranging from 'id', 'time', various 'trans_feat_' columns to 'agg_feat_' columns, exhibited no null entries, indicating a dataset rich in transaction details and fully prepared for analysis. The 'edgelist', which consists of the 'txId1' and 'txId2' columns, also showed no missing values, confirming that the transaction linkage data was complete.

3.6. Variable Understanding

Variable understanding, also known as data understanding, is a vital component of the analytics process. It involves delving into the dataset to understand the nature and correlations of each variable. This entails knowing what each variable represents, its data type (numerical, categorical, or text), and its possible relevance to the research.

In our study, three main dataframes—classes, features, and edge list—play distinct but interconnected roles. The 'class' dataframe is essential for transaction categorization and serves as the foundation for ML models. Meanwhile, the 'features' dataframe delves into the details, thoroughly evaluating both transactional and aggregated features. Additionally, the 'edgelist' reveals the dynamics of relationships between transactions. Understanding the variables in these dataframes is critical for their effective utilization in different ML and analytical procedures.

3.6.1. Features

- id (Transaction ID): A unique numeric identifier used to link transactions across DataFrames.
- time: A numerical number representing the transaction's time step, crucial for time-series analysis and identifying temporal trends.
- Trans_feat_0 to trans_feat_91: These represent unique numerical properties of each transaction, such as amounts and counts, which are crucial for in-depth analysis.
- Aggregated numeric features (agg_feat_0 to agg_feat_71): Provide statistical summaries over specific time frames or transaction groupings, aiding to the discovery of broad patterns.

3.6.2. Classes

- txId (Transaction ID): An integer uniquely identifying each transaction, essential for linking transactions to their features and classifications.
- class: A categorical variable indicating the transaction type (licit, illicit, or unknown), which is the primary target for classification models in machine learning.

3.6.3. Edgelist

- txId1 (Transaction ID 1): A numeric identifier used to map and build transaction links.
- txId2 (Transaction ID 2): The numeric identifier for the second element in a transaction pair. It interacts with txId1 to map transaction networks.

3.7. Data Merging

Data merging in data preparation is the process of combining data from multiple sources or datasets to generate a single, coherent dataset. This step is crucial for ensuring consistency, minimizing redundancy, and enhancing analytical quality.

In our analysis method, we merge the 'features' and 'classes' dataframes, a critical step performed using Python's pandas package. This merging integrates the detailed features of transactions from the 'features' dataframe with their corresponding class labels from the 'classes' dataframe. It is executed by matching the 'id' column of 'features' with the 'txId' column of 'classes', guaranteeing an exact linkage between each transaction's individual features and its categorization. The method employed is a left join, which keeps all records from 'features' while appending relevant classification data from 'classes'.

3.8. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical phase in the data analysis process that involves the review and summarization of a dataset's key properties, frequently utilizing visual methods. The fundamental purpose of EDA is to investigate the data to understand its underlying structure, extract key variables, detect anomalies and trends, and formulate hypotheses for further investigations. EDA is often undertaken before formal modeling or hypothesis testing and assists in making informed decisions on how to manage the data for further analysis. Early detection of anomalies and data quality issues through EDA minimizes risks and prevents errors. It also helps to communicate results effectively through visualizations, which is critical for engaging stakeholders and guiding decision-making process.

3.9. Label Encoding

Label encoding is a data preprocessing technique where each distinct category value of a categorical variable is assigned a numerical number. This method is simple and computationally efficient; it converts categorical variables into a format that can be easily processed by ML algorithm. It is particularly beneficial for nominal categorical variables [23].

During preprocessing, columns in the 'features' dataframe are renamed to increase clarity and manageability, with the first two labels being 'id' and 'time'. This is followed by 93 transaction-related features labeled from 'trans_feat_0' to 'trans_feat_92', and 72 aggregated features designated as 'agg_feat_0' to 'agg_feat_71'. This structured terminology aids in the effective identification and referencing of features. In the 'classes' dataframe, the 'class' column is labeled using pandas' 'map' function, converting categorical labels into numerical values: '1' for illicit transactions and '2' for licit transactions, therefore improving compatibility with ML methods.

3.10. Data Splitting

In ML, data is typically divided into three sets: a training set (60–80% of the data) used for feature learning, a test set (10–20%) for performance evaluation, and a holdout set (10–20%) for final validation. This distribution guarantees that the model learns patterns, is properly assessed, and avoids overfitting specific test samples, allowing for a comprehensive evaluation of its capabilities. For our study, we employ a 60:20:20 split, allocating 60% of the data to the model training, 20% to the testing set to evaluate the model's performance post-training, and 20% to the holdout set to confirm the model's effectiveness in its final form and to protect against overfitting.

All train–validation–test splits employed stratified sampling (`stratify=y`) to preserve the illicit-to-licit ratio across all subsets. A global random seed of 42 was applied to all stochastic operations, including splitting, SMOTE, and model initialisation. Feature standardisation (zero mean, unit variance via `StandardScaler`) was fitted exclusively on the training set and subsequently applied as a deterministic transform to the validation and test sets, preventing distributional leakage. The enforced pipeline order was: (1) stratified split; (2) fit scaler on training data; (3) transform all three sets; (4) apply SMOTE to training data only.

Temporal Split Experiment

To account for the inherent temporal ordering of the Elliptic dataset, a chronological split was performed in addition to the random baseline. Transactions from time steps 1–29 (26,381 samples) were used for training, steps 30–39 (8999 samples) for validation, and steps 40–49 (11,184 samples) for final evaluation. This design prevents future transaction data from influencing model training, reflecting real-world AML deployment conditions where a deployed model encounters transactions it has never seen. Results for both split strategies are compared in the results section.

3.11. Standard Scaler

A Standard Scaler is used in ML to preprocess a dataset by standardizing its features. Standardization involves scaling the distribution of values so that they have a mean of zero and a standard deviation of one. This is crucial because many ML estimators, such as Support Vector Machines (SVMs) and logistic regression, operate under the assumption that all features should have a comparable variance and center around zero. Standardization mitigates the impact of features with high variance on the objective function, allowing for an even contribution from all features. To do this, the Standard Scaler subtracts each feature's mean and divides it by its standard deviation. This scaling is essential for ML models that perform better with normally distributed data.

Using the Standard Scaler to standardize features in a dataset ensures an equitable contribution to model predictions. Normalizing the data improves the model's ability to detect trends and abnormalities associated with money laundering.

3.12. Data Resampling

Data resampling is a statistical strategy essential for improving or assessing the correctness of data models. It entails repeatedly drawing samples from a dataset to evaluate the model against these samples. Common resampling strategies include bootstrapping, which involves sampling with replacement, and cross-validation, which involves dividing the dataset into several subgroups for training and testing. Resampling is critical for measuring the variability of model predictions.

The Synthetic Minority Oversampling Technique (SMOTE), developed by Chawla et al. in 2002, addresses class imbalance in ML datasets. Unlike conventional oversampling algorithms, which merely reproduce minority class occurrences, SMOTE produces synthetic samples, thereby enriching the dataset and mitigating the risk of overfitting. It operates by identifying the k-nearest neighbors for each minority class instance and then interpolating between these neighbors to create new, synthetic instances. This process commonly involves a random factor ranging from 0 to 1, ensuring that the synthetic samples are varied yet reflective of the original instances. SMOTE effectively balances datasets, improving model performance in critical applications such as fraud detection and medical diagnostics [24].

In our study, data resampling plays a pivotal role in addressing class imbalance within transaction datasets, where money laundering transactions are relatively rare compared to legal transactions. This results in a dataset heavily skewed towards "licit" rather than "illicit" transactions. To tackle this, we employ SMOTE (Synthetic Minority Over-sampling Technique) to correct class imbalance, especially for classification tasks.

SMOTE was applied exclusively to the training fold following the stratified train–validation–test split, ensuring no synthetic samples from the validation or test sets contaminate model training. Parameters used: `k_neighbours`

= 5, sampling_strategy = 'auto' (upsample minority class to match majority class count). Class distribution in the training set prior to SMOTE: 2727 illicit, 25,211 licit. After SMOTE: 25,211 illicit, 25,211 licit.

This method involves interpolating between existing minority class instances and their nearest neighbors, thereby leveling the class distribution. The outcome is a fresh, resampled training dataset with a more representative class distribution, improving the model's capacity to learn and produce unbiased predictions. This adjustment is especially important in scenarios like fraud or anomaly detection, where the minority class holds significant importance.

3.13. Proposed Machine Learning Algorithms

An ML algorithm is a computational tool that allows computers to gain knowledge from data and improve their efficiency in certain tasks. These algorithms analyze patterns and make decisions with minimal human involvement. They are utilized in a variety of applications, including anticipating customer behavior, and automating complicated procedures. ML is a critical component of artificial intelligence, continuously evolving to tackle more complex challenges [25].

In this paper, we discuss the ML algorithms implemented in our study, and investigate their capacities in data processing, pattern identification, and decision-making, and their contributions to our research goals. This includes an assessment of their performance, issues, and efficacy, highlighting their importance in advancing data-driven research methodologies.

3.13.1. Decision Tree Classifier

A Decision Tree is a flexible and user-friendly ML method, primarily employed for decision-making and classification tasks. As a supervised learning model, it predicts outcomes from labeled data, such as potential events, resource costs, and total utility. Its straightforwardness and visual structure make it highly interpretable and simple to execute, which has led to its extensive use in industries ranging from finance to healthcare [26]. Unlike more sophisticated algorithms, decision trees are simple to understand and visualize, making them especially useful in circumstances requiring clarity and explainability.

Decision Trees are robust because they can handle both categorical and continuous data, making them useful for various predictive modeling problems. Their flexibility and minimal data preprocessing needs make them a popular choice among data science and ML practitioners. They are well-known for their simple implementation since they do not need data normalization or scaling, simplifying the model-building process. Decision Trees are popular among both novice and expert users in the industry because of their ease of use and obvious decision-making processes [27].

3.13.2. Logistic Regression

Logistic regression is a critical statistical method in regression analysis, principally used to predict outcomes for a categorical dependent variable using one or more independent variables. This method is well-known in the area of ML for its efficiency in binary classification problems. It adeptly categorizes data into binary outcomes, such as "Yes" or "No", "Pass" or "Fail", making it invaluable for scenarios requiring binary decisions. Its importance is highlighted by its wide use in a variety of disciplines, from health care to financial modeling, in which understanding and predicting binary results is critical. The model's probabilities and odds ratios provide strong insights into the impact of predictor factors [28]. This method's ability to turn a linear regression model into a statistical framework enhances the method's value, making it a top choice for binary classification issues.

3.13.3. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a powerful supervised ML method that excels in both classification and regression tasks. Its remarkable ability to deal with both linear and non-linear data distinguishes it from a lot of other ML techniques. The core of SVM lies in its approach to identify the most effective hyperplane that separates data points from different classes across a multidimensional space. This hyperplane is intentionally chosen to maximize the margin between the data points of the various classes, leading to the most effective separation possible. This margin enhancement feature makes SVM especially effective for boosting the model's generalization capacity, which is critical for achieving high accuracy on unknown data [29]. One of the most notable benefits of SVM is the capacity to deal with high-dimensional data effectively. This makes it ideal for complicated applications involving a high number of characteristics, such as text classification, image recognition, and genomic data analysis. SVMs are well-known for their excellent accuracy, particularly in datasets with clearly defined classes. The regularization parameter improves the algorithm's resilience by decreasing overfitting, especially in high-dimensional domains.

3.13.4. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a foundational ML method known for its simplicity and efficacy in both classification and regression applications. As a member of the instance-based and non-parametric ML algorithm family, KNN operates on the basis of feature similarity. It predicts the label of a new data point by aggregating the labels or values of its 'K' nearest neighbors within the feature space. This proximity is computed using a predefined distance measure, making KNN a useful tool for a variety of data types [30]. The algorithm's ability to use alternative distance measurements enables it to be adapted to certain data kinds and structures. KNN adjusts when new data is supplied, making it ideal for applications that require continual updating or collection.

3.13.5. Random Forest for Machine Learning

Random Forest is a popular ensemble learning method used for both regressions and classification applications in machine learning. It works by building a multitude of decision trees during the training phase and outputs either the mode of the classes for classification problems or the mean forecast of the trees for regression tasks. Random Forest is well-known for its excellent accuracy, resilience, and simplicity, making it a popular approach in ML [31]. Random Forest can efficiently process datasets, including a high range of features and training instances. It can properly manage missing data in the training set, retaining accuracy even when there are major data gaps and is a flexible classification and regression tool with category and numerical data. Random Forest gives valuable insights into the most critical characteristics for generating forecasts and the procedure usually has low bias and moderate variance, resulting in a well-balanced approach for model fitting.

3.13.6. Stochastic Gradient Descent (SGD) Classifier

The Stochastic Gradient Descent (SGD) Classifier significantly improves machine learning, specifically for linear classifiers. This long-standing strategy has received significant attention recently due to its efficacy in large-scale learning settings. SGD is well-known for its application in linear Support Vector Machines and Logistic Regression, where it functions with convex loss functions. This method's rising popularity is due to its ability to handle larger datasets and high-dimensional spaces effectively [32]. SGD is highly scalable, because it can process data points individually or in small batches without requiring the complete dataset to be stored in memory simultaneously. SGD may be used for a variety of convex loss functions and regularization strategies. This adaptability allows it to be adapted for a wide range of classification tasks, from basic linear separations to complicated decision boundaries [33].

3.14. Confusion Matrix in Machine Learning

A Confusion Matrix is a key instrument in machine learning, particularly valuable for evaluating classification algorithms. It provides a tabular visualization of a classification model's performance against a dataset with known true values. While the matrix is simple, it provides valuable insights into the model's performance, particularly in discriminating between different classes.

The confusion matrix consists of four basic characteristics (numbers), which are:

- True Positives (TP): These occur when a fraud detection algorithm correctly identifies fraudulent behavior. For example, correctly reporting a fraudulent transaction demonstrates the model's ability to detect fraud [13].
- True Negatives (TN): These happen when the model accurately detects legitimate transactions, which is critical in fraud detection since it demonstrates the model's capacity to discriminate between fraudulent and non-fraudulent behaviors [34].
- False Positives (FP): Also known as Type I errors in fraud detection, these occur when legitimate transactions are wrongly marked as fraudulent, which results in superfluous inquiries and causes inconvenience for both financial institutions and clients [13].
- False Negatives (FN): Also known as Type II errors, these occur when a model fails to detect a fraudulent transaction and incorrectly marks it as legitimate. This poses a substantial risk in fraud detection, potentially resulting in financial losses [34].

The confusion matrix provides a detailed breakdown of how a classification model performs across different classes, making it an indispensable tool for evaluating and improving machine learning models.

True Positives (TP) and True Negatives (TN) are critical indicators of model accuracy when using ML for money laundering analysis. TPs that accurately identify fraudulent transactions highlight the model's ability to detect unlawful activities. This is critical for preventing financial crimes and ensuring the security of financial systems. TNs, which represent accurately detected lawful transactions, demonstrate the model's capacity to avoid

interrupting routine financial activities, hence preserving consumer confidence and operational efficiency.

False Positives (FP) and False Negatives (FN) represent areas for improvement. FPs, in which genuine transactions are wrongly identified as fraudulent, can result in wasteful investigations, consumer unhappiness, and organizational inefficiencies. FNs, which fail to detect fraudulent transactions, constitute a substantial concern since they allow unlawful operations to go undetected, potentially resulting in huge financial losses.

Balancing these outcomes is critical for creating a reliable ML model for detecting money laundering.

3.15. Performance Metrics

Parameters such as Accuracy, Precision, Recall, F1 Score, and Specificity are crucial for evaluating ML models in identifying money laundering. These metrics confirm the models' effectiveness in detecting suspicious transactions and help refine the algorithms for more accurate and reliable results in financial security [35].

Accuracy is the ratio of correctly predicted cases to the total number of instances, serving as a critical parameter in ML. While it provides a basic measure of a model's performance, it can be misleading in cases with unbalanced class distributions. In such circumstances, a model might appear highly accurate, but only because it merely predicts the majority class.

Precision is a key performance indicator in ML that evaluates a model's accuracy in making positive predictions. It is defined as the proportion of accurately predicted positive observations to all expected positive observations. Precision becomes particularly important where the consequences of false positives are severe. For instance, a low precision rate in fraud detection algorithms would lead to many legitimate transactions being wrongly classified as fraudulent, causing inefficiencies and perhaps losing client trust.

Recall evaluates a classifier's ability to recognize all positive cases. It is critical in situations where missing a positive instance could have catastrophic consequences. For money laundering detection, high recall is essential to guarantee that all possible fraudulent activities are discovered, minimizing the danger of missing unlawful transactions.

The F1 Score is the harmonic mean of Precision and Recall, particularly useful in unbalanced datasets like those found in fraud detection scenarios. It gives a balanced measure of a model's accuracy and recall, ensuring that erroneous positives and false negatives are considered when evaluating performance.

Specificity measures the proportion of genuine negatives to the total number of negatives. High specificity is critical in ML models for the financial sector, especially those used to detect money laundering. It guarantees that legitimate transactions are accurately detected, reducing the incidence of false alarms and the expenses associated with investigating these false positives.

Performance metrics are critical in the detection of money laundering using ML models, since they serve as important tools for model evaluation, validation, and optimization. They enable an accurate assessment of a model's capability to distinguish between licit and illicit transactions, which is especially crucial with skewed datasets like those in financial transactions. Metrics like accuracy, precision, recall, F1 score, and AUC-ROC are used to fine-tune models to improve the balance of false positives and false negatives and to objectively evaluate various models. This quantitative analysis is crucial for ensuring regulatory compliance, managing risk, and maintaining stakeholder confidence. Furthermore, these indicators enable the ongoing refinement and adaptation of models to evolving money laundering strategies, guaranteeing that financial institutions can effectively control and mitigate related risks.

4. Data Analysis and Visualization

Data visualization is a critical data analysis component that converts complex datasets into understandable graphical representations. Employing a variety of visual elements, such as charts, graphs, and maps, it elucidates data trends, outliers, and patterns, making them more understandable [36]. This method not only simplifies the interpretation of complicated data but also helps to communicate findings to a broader audience, proving invaluable in data science and decision-making processes. In this section, we have visualized the data to provide an overview and to facilitate understanding of the information in a graphical form.

4.1. Bar Chart Showing 'Licit' and 'Illicit' Class Distribution

The dataset categorizes transactions into two classes: 'licit' for lawful financial activities and 'illicit' for transactions that may suggest money laundering. A significant class imbalance exists, with 'licit' transactions far outnumbering 'illicit' ones, as shown in Figure 1. This mismatch is typical in fraud detection databases, since criminal activities such as money laundering are naturally less frequent than lawful transactions.

The presence of class imbalance poses a significant challenge in training ML models, as algorithms often develop a bias towards predicting the dominant class. In the context of this project, such a bias could lead to a model

that is proficient in identifying 'licit' transactions but may fail to detect important 'illicit' ones. This could result in a high proportion of false negatives, where potentially fraudulent operations are misclassified as legitimate, thereby reducing the effectiveness of the money laundering detection system.

Addressing this class imbalance is critical for enhancing the prediction model's accuracy, particularly in identifying 'illicit' transactions. Enhanced accuracy in detecting illicit transactions is vital for preventing financial crimes and ensuring the integrity of financial systems.

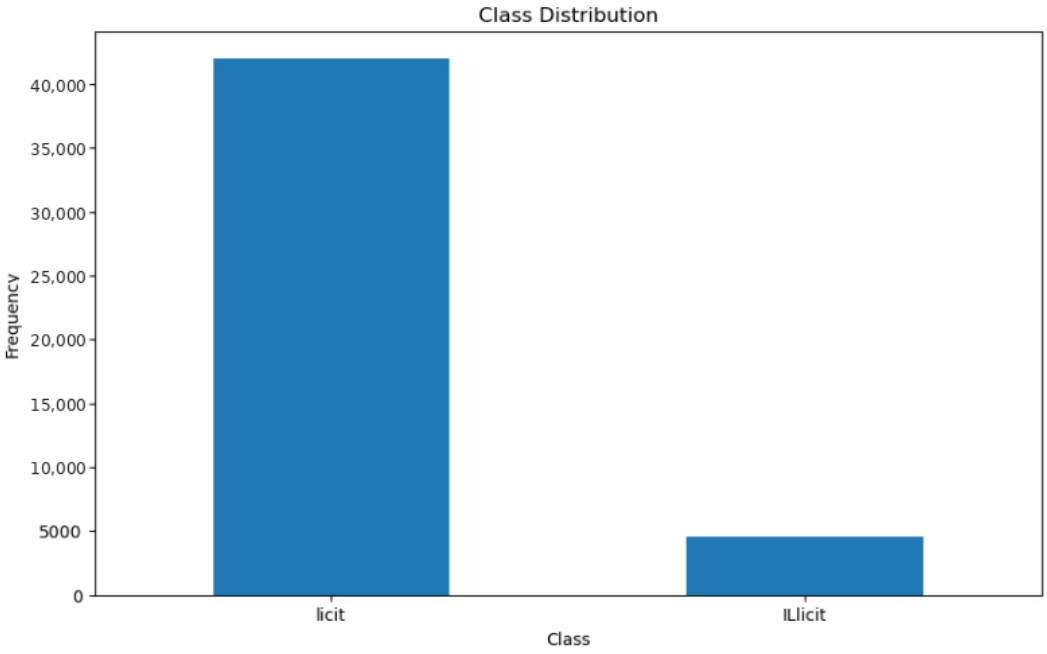


Figure 1. Class distribution.

4.2. Line Chart Showing the Total Number of Transactions over Time without Class Distinction

The line chart (Figure 2) depicts time-series data spanning 98 weeks, covering more than two years. Each interval, or time step, defines a biweekly period, providing a structured temporal framework for the study. The figure displays broad trends and swings in transaction volume. However, it lacks the granularity needed to pinpoint significant patterns or anomalies—a typical limitation when analyzing aggregated data across extended periods.

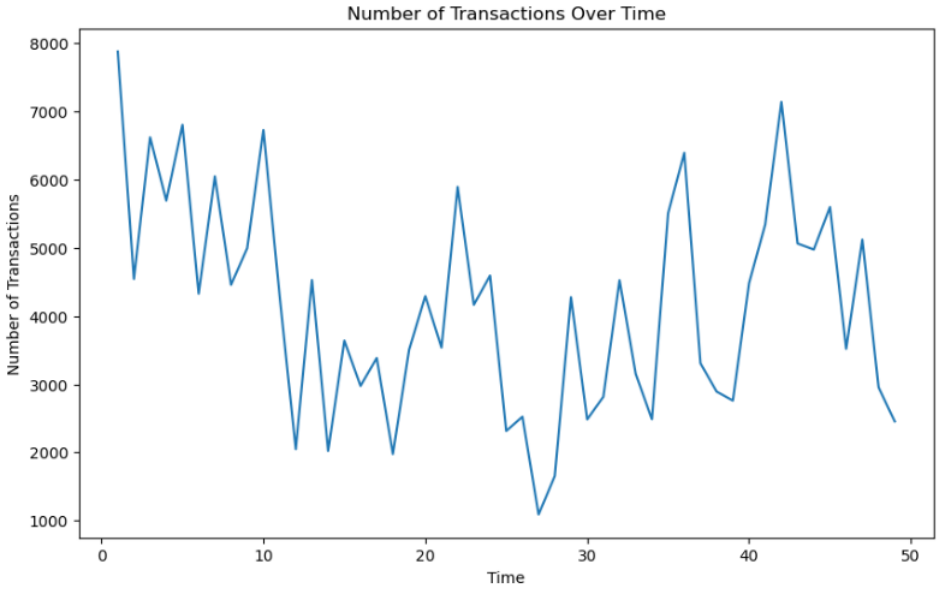


Figure 2. Number of transactions over time.

To address the inherent limitations of aggregate analysis and uncover potentially crucial underlying trends, transaction data should be segmented into 'licit' and 'illicit' categories. This approach aligns with the dataset's structure, which classifies transactions according to these terms. Data segmentation by class is expected to produce more sophisticated insights into the specific trends, proportions, and temporal dynamics of each category.

4.3. Line Chart Comparing 'Licit' and 'Illicit' Transactions over Time

Figure 3 depicts the number of 'licit' (lawful) and 'illicit' (illegal, potentially indicative of money laundering) financial transactions across a specified time period. The x-axis labels 'Time', and the y-axis shows the 'Number of Transactions', representing the number of recorded transactions.

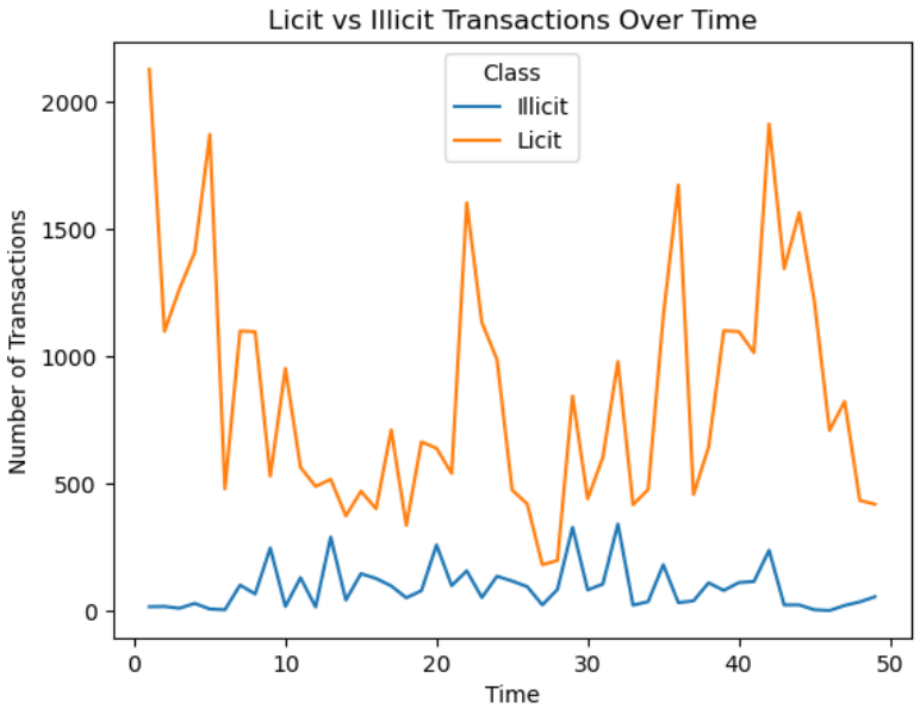


Figure 3. Illicit and licit transactions over time.

The line representing 'licit' transactions, shown in orange, varies significantly over time, with multiple peaks and troughs indicating changes in the quantity of lawful transactions. Meanwhile, the 'illicit' transactions, displayed in blue, maintain a lower count throughout the same time periods, featuring less noticeable peaks and exhibiting a more consistent trend than the 'licit' transactions.

This graph enables one to discern temporal patterns or trends in transaction behavior. The greater stability of the 'illicit' line indicates that illegal transactions occur with more consistency over time, while 'licit' transactions are more variable. Understanding these trends is critical for building models and detecting anomalies in a ML project aimed at analyzing money laundering. The graph shows that any predictive model must consider the time dynamics of transaction behaviors to effectively discriminate between 'licit' and 'illicit' actions.

4.4. Bar Chart with 'Licit' and 'Illicit' Transactions over Time as Separate Bars at Each Time Step

Figure 4 delineates 'Time step' as distinct intervals during which transactions were recorded. The y-axis represents the 'Number of Transactions', which quantifies the transactions within each period. For each time step, the height of the bar corresponds to the volume of transactions in that category.

This chart shows that 'licit' transactions, shown in green, predominate over 'illicit' ones at every time step. However, the number of transactions for both categories vary noticeably over time steps, with some period experiencing significant activity surges.

The variations and patterns observed in this time series data are important characteristics that a ML model can use to learn and predict future transaction categories. Understanding these temporal dynamics is critical, especially for recognizing unexpected spikes that may indicate unlawful activities.

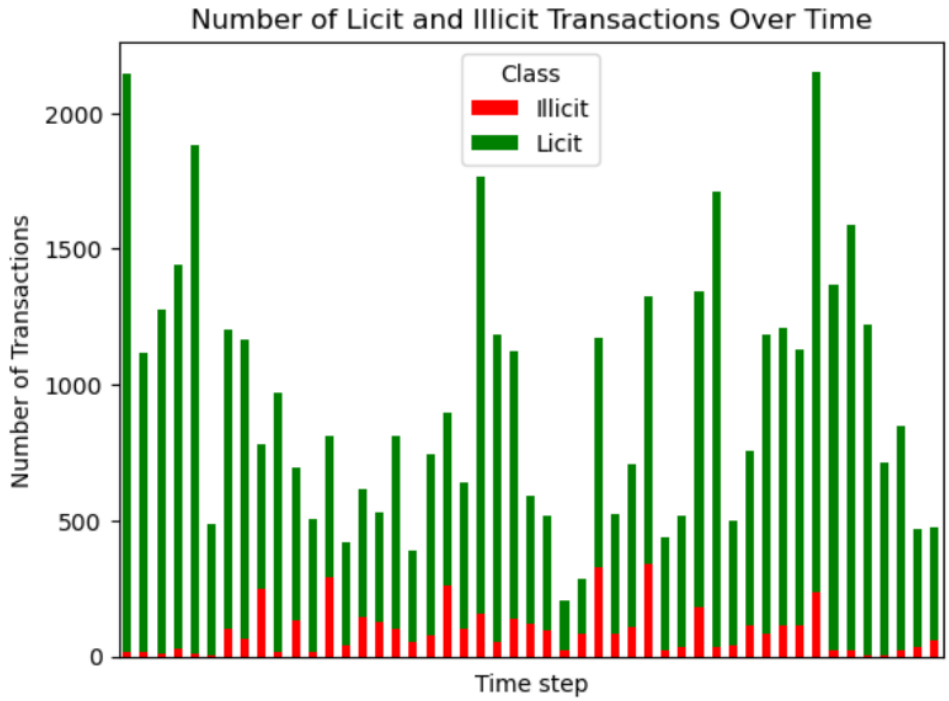


Figure 4. Transactions over time steps.

5. Results of the Machine Learning Algorithms

All classifiers were trained using the hyperparameter configurations documented in Table 2 in the Methodology section. In this section, we thoroughly evaluate the results achieved by the ML algorithms implemented in our study. We delve into the intricate patterns and trends extracted from the data, offering insights into the algorithms' efficiency and value in achieving our research objectives. This analysis is crucial for showing the effectiveness of our study within the larger framework.

Table 2. Hyperparameter configuration for all classifiers.

Classifier	Tuning Method	Key Hyperparameters	Values Used
Logistic Regression	Default / fixed	`max_iter`, `solver`, `random_state`	1000, lbfgs, 42
SVM (RBF kernel)	Default / fixed	`kernel`, `C`, `gamma`, `random_state`	rbf, 1.0, scale, 42
K-Nearest Neighbours	Default / fixed	`n_neighbors`, `metric`	5, minkowski
SGD Classifier	Default / fixed	`loss`, `max_iter`, `random_state`	hinge, 1000, 42
Decision Tree	Default / fixed	`criterion`, `max_depth`, `random_state`	gini, None, 42
Random Forest	Default / fixed	`n_estimators`, `max_features`, `random_state`	100, sqrt, 42

5.1. Logistic Regression Confusion Matrix

The Logistic regression model demonstrates a strong capacity to discriminate between illicit and licit financial transactions. A recall rate of 92% for illicit transactions shows the model's significant aptitude to detect potential money laundering activities, as shown in Figure 5.

However, the precision for illicit transactions is relatively low at 49% for the test set and 46% for the holdout set, implying that over half of the transactions flagged as suspicious are false alarms. Conversely, the model achieves high precision and recall for licit transactions (99% and 89%, respectively), demonstrating great accuracy in confirming legitimate transactions.

The overall accuracy of the model stands at 89.75% for the test set and 89.14% for the holdout set. Despite this relatively good performance, the poor precision for illicit transactions shows a critical area for improvement. Reducing false positives is critical for practical AML applications.

The F1-Scores in the test set are 0.64 for illicit and 0.94 for licit transactions, indicating that licit transactions outperform unlawful transactions, which is replicated in the holdout set with F1-Scores of 0.61 and 0.94, respectively. This mismatch emphasizes the model's difficulty in dealing with the class imbalance and the need for further tweaking to increase its precision for illicit transactions while maintaining robust recall capabilities.

```
Test Set Metrics (Logistic Regression):
Accuracy: 0.8974551701922044
Classification Report:
              precision    recall  f1-score   support

           1       0.49      0.92      0.64       937
           2       0.99      0.89      0.94      8376

    accuracy                           0.90      9313
   macro avg       0.74      0.91      0.79      9313
weighted avg       0.94      0.90      0.91      9313

Confusion Matrix:
 [[ 863   74]
 [ 881 7495]]

Holdout Set Metrics (Logistic Regression):
Accuracy: 0.8914420702244175
Classification Report:
              precision    recall  f1-score   support

           1       0.46      0.92      0.61       881
           2       0.99      0.89      0.94      8432

    accuracy                           0.89      9313
   macro avg       0.73      0.90      0.78      9313
weighted avg       0.94      0.89      0.91      9313

Confusion Matrix:
 [[ 807   74]
 [ 937 7495]]
```

Figure 5. Logistic regression.

5.2. SVM for Regression (SVR) Confusion Matrix

The Support Vector Machine (SVM) model excels at distinguishing between unlawful and legitimate financial transactions, with overall accuracy rates of 96.18% and 96.72% for the test and holdout sets, respectively, as shown in Figure 6.

Precision for detecting unlawful transactions is 0.77 for the test set and 0.78 for the holdout set, accompanied by high recall rates of 0.88 and 0.91, resulting in F1-scores of 0.82 and 0.84, indicating a great balance between accuracy and recall. Licit transactions are consistently identified with good precision and recall across both sets, with an F1-score of 0.98.

As a result, the SVM model is strong enough to reliably detect fraudulent transactions while retaining high performance for legitimate transactions.

5.3. K-Nearest Neighbors Confusion Matrix

The K-Nearest Neighbors model with scaling performs well in detecting money laundering activities, achieving an accuracy of 92.56% on the test set and 92.79% on the holdout set, as shown in Figure 7.

It is very reliable in categorizing licit transactions, achieving accuracy and recall rates surpassing 0.93. However, it struggles with the precision for illicit transactions, scoring only 0.58 and 0.57 on the test and holdout sets respectively, despite high recall rates of approximately 0.91. This discrepancy suggests that while the model is adept at identifying potential illicit activities, it also misclassifies a significant number of legitimate transactions as illicit.

The F1 scores show an improvement in balance for licit transactions (0.96) compared to illegitimate ones (0.71).

The confusion matrices show a higher frequency of false positives for licit transactions than false negatives

for illicit ones, indicating that the model has the propensity to err on the side of caution by more frequently misclassifying transactions as illicit rather than overlooking actual illicit activities.

```
Test Set Metrics (SVM):
Accuracy: 0.9617738644904972
Classification Report:
              precision    recall  f1-score   support

           1       0.77      0.88      0.82       937
           2       0.99      0.97      0.98      8376

    accuracy                           0.96      9313
   macro avg       0.88      0.93      0.90      9313
weighted avg       0.97      0.96      0.96      9313

Confusion Matrix:
 [[ 829  108]
 [ 248 8128]]

Holdout Set Metrics (SVM):
Accuracy: 0.9672500805325889
Classification Report:
              precision    recall  f1-score   support

           1       0.78      0.91      0.84       881
           2       0.99      0.97      0.98      8432

    accuracy                           0.97      9313
   macro avg       0.89      0.94      0.91      9313
weighted avg       0.97      0.97      0.97      9313

Confusion Matrix:
 [[ 801   80]
 [ 225 8207]]
```

Figure 6. SVM.

```
Test Set Metrics (KNN with Scaling):
Accuracy: 0.9255878878986363
Classification Report:
              precision    recall  f1-score   support

           1       0.58      0.91      0.71       937
           2       0.99      0.93      0.96      8376

    accuracy                           0.93      9313
   macro avg       0.79      0.92      0.83      9313
weighted avg       0.95      0.93      0.93      9313

Confusion Matrix:
 [[ 848   89]
 [ 604 7772]]

Holdout Set Metrics (KNN with Scaling):
Accuracy: 0.9279501771716955
Classification Report:
              precision    recall  f1-score   support

           1       0.57      0.92      0.71       881
           2       0.99      0.93      0.96      8432

    accuracy                           0.93      9313
   macro avg       0.78      0.93      0.83      9313
weighted avg       0.95      0.93      0.94      9313

Confusion Matrix:
 [[ 814   67]
 [ 604 7828]]
```

Figure 7. K-Nearest Neighbors.

5.4. Stochastic Gradient Descent (SGD) Classifier Confusion Matrix

The Stochastic Gradient Descent (SGD) classifier showcases robust performance on both the test and holdout sets, with accuracy scores of 90.36% and 89.63%, respectively, as shown in Figure 8. This model has a higher recall rate for both illicit (0.91) and licit (0.90) transactions across the datasets, indicating its capability to detect most of positive cases in both classes.

```
Test Set Metrics (SGD Classifier with Scaling):
Accuracy: 0.9036830237302695
Classification Report:
              precision    recall  f1-score   support

           1       0.51      0.91      0.66       937
           2       0.99      0.90      0.94      8376

    accuracy                           0.90      9313
   macro avg       0.75      0.91      0.80      9313
weighted avg       0.94      0.90      0.92      9313

Confusion Matrix:
 [[ 857   80]
 [ 817 7559]]

Holdout Set Metrics (SGD Classifier with Scaling):
Accuracy: 0.8963814023408139
Classification Report:
              precision    recall  f1-score   support

           1       0.48      0.91      0.62       881
           2       0.99      0.90      0.94      8432

    accuracy                           0.90      9313
   macro avg       0.73      0.90      0.78      9313
weighted avg       0.94      0.90      0.91      9313

Confusion Matrix:
 [[ 799   82]
 [ 883 7549]]
```

Figure 8. Stochastic Gradient Descent.

However, the model shows low accuracy for unlawful transactions with only 0.51 on the test set and 0.48 on the holdout set. This lower accuracy results in several false positives that negatively affect the balance of precision and recall, as reflected by the F1-scores 0.66 and 0.62, respectively. Conversely, the model excels at recognizing licit transactions, achieving excellent accuracy and F1-scores of 0.99 and 0.94.

The confusion matrices show that the model is more cautious, prioritizing false positives above false negatives for unlawful transactions. Overall, while the model is effective for detecting licit transactions, there is room for improvement in refining its precision for detecting unlawful transactions to reduce false positives and enhance overall accuracy.

5.5. Decision Trees Classifier Confusion Matrix

As shown in Figure 9, the Decision Tree Classifier performs well at distinguishing between illicit and licit transactions, achieving high accuracy rates of 96.77% on the test set and 97.22% on the holdout set. This indicates that the classifier is proficient in accurately detecting most transactions.

The precision for licit transactions is outstanding at 0.99, and it remains quite excellent for illicit transactions at approximately 0.82–0.83. This suggests that the classifier is generally correct when predicting criminal transactions.

The recall rates are also impressive, exceeding 0.87 for illicit transactions and reaching 0.98 for licit ones, indicating that the classifier reliably detects most of unlawful activities.

Additionally, the F1-scores, which are indicators of the balance between precision and recall, are commendably high, particularly for licit transactions at 0.98. This highlights the classifier's balanced performance in both detecting and correctly classifying transactions.

```
Test Set Metrics (Decision Tree Classifier):
Accuracy: 0.9676795876731451
Classification Report:
              precision    recall  f1-score   support

           1       0.82      0.87      0.84       937
           2       0.99      0.98      0.98      8376

    accuracy                           0.97      9313
   macro avg       0.90      0.93      0.91      9313
weighted avg       0.97      0.97      0.97      9313

Confusion Matrix:
 [[ 818  119]
 [ 182 8194]]

Holdout Set Metrics (Decision Tree Classifier):
Accuracy: 0.9721894126489853
Classification Report:
              precision    recall  f1-score   support

           1       0.83      0.89      0.86       881
           2       0.99      0.98      0.98      8432

    accuracy                           0.97      9313
   macro avg       0.91      0.93      0.92      9313
weighted avg       0.97      0.97      0.97      9313

Confusion Matrix:
 [[ 783   98]
 [ 161 8271]]
```

Figure 9. Decision tree classifiers.

5.6. Random Forest Classifier Confusion Matrix

As shown in Figure 10, the Random Forest Classifier performs admirably in identifying money laundering transactions, achieving impressive accuracy rate of 98.43% on the test set and even higher rate of 98.87% on the holdout set. These figures demonstrate its outstanding generalization ability across different data sets. The Precision rates for both illicit (0.99) and licit (0.98–0.99) transactions are nearly flawless, indicating that the model is quite reliable in its predictions. Recall is great for licit transactions (1.00) and illegal ones (0.85–0.89), indicating that the model effectively detects genuine unlawful operations. The F1-scores are high, particularly for licit transactions (0.99), indicating a good combination of accuracy and recall. The confusion matrices reveal a very low number of false positives and negatives, indicating that the classifier is quite accurate in discriminating between licit and criminal transactions.

5.7. Hyperparameter Tuning

All six classifiers were trained using fixed, default or literature-standard hyperparameter values rather than exhaustive grid search, as the primary objective of this study is comparative benchmarking under identical conditions rather than per-model optimisation. Table 2 documents the complete hyperparameter configuration used for each classifier. A global random seed of 42 was applied to all stochastic operations (splitting, SMOTE, and model initialisation) to ensure full reproducibility. Future work should investigate the effect of grid-search hyperparameter optimisation on the relative rankings observed here.

5.8. Important Features in Random Forest Classifier

The Random Forest Classifier has demonstrated exceptional performance in identifying money laundering transactions. The model achieves an accuracy of 98.43% on the test set and 98.87% on the holdout set. To understand the factors driving its success, we analyze the top ten features deemed most important by the classifier, as this information is critical for the continuous improvement of detection algorithms, as presented in Figure 11. The Random Forest Classifier's feature importance provides valuable insights into which transactional features are most indicative of money laundering operations. Understanding these key attributes helps enhance the model's ability to detect and prevent illicit activities more effectively.

```
Test Set Metrics (Random Forest):
Accuracy: 0.9843229893696983
Classification Report:
              precision    recall  f1-score   support

           1       0.99      0.85      0.92       937
           2       0.98      1.00      0.99      8376

    accuracy                           0.98      9313
   macro avg       0.99      0.93      0.95      9313
weighted avg       0.98      0.98      0.98      9313

Confusion Matrix:
 [[ 801  136]
 [  10 8366]]

Holdout Set Metrics (Random Forest):
Accuracy: 0.9880811768495651
Classification Report:
              precision    recall  f1-score   support

           1       0.99      0.88      0.93       881
           2       0.99      1.00      0.99      8432

    accuracy                           0.99      9313
   macro avg       0.99      0.94      0.96      9313
weighted avg       0.99      0.99      0.99      9313

Confusion Matrix:
 [[ 778  103]
 [   8 8424]]
```

Figure 10. Random forest classifiers.

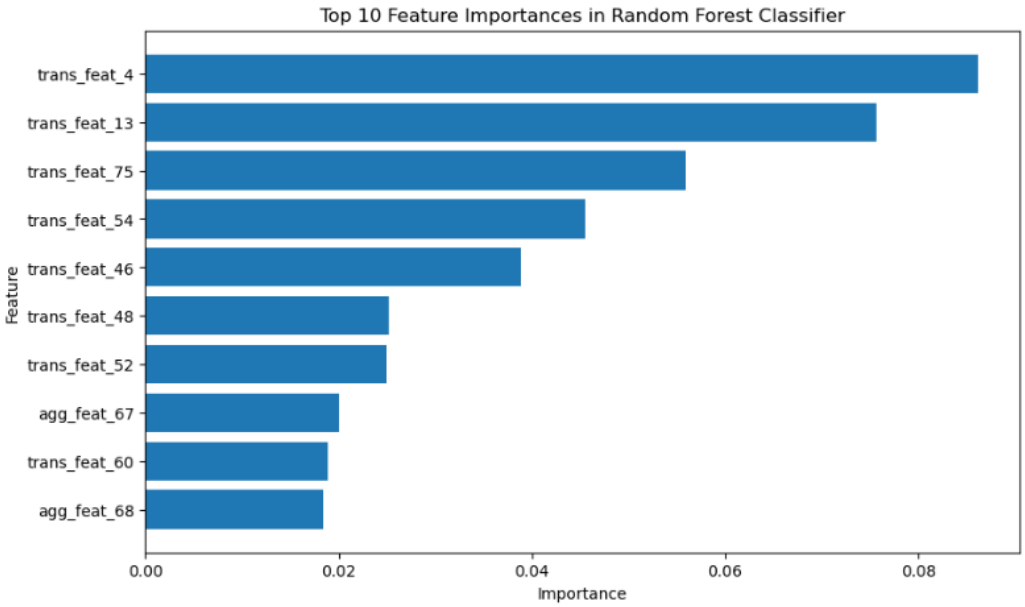


Figure 11. Feature importance.

5.9. Comparison of Algorithms

Tables 3 and 4 present the corrected evaluation results on the held-out test set under both the random split and temporal split conditions, respectively. Metrics include per-class and macro-averaged Precision, Recall, and F1, together with ROC-AUC and PR-AUC, which are better suited to imbalanced settings than accuracy alone.

Table 3. Comparison of algorithms—random split (test set).

Model	Accuracy	PrecP_{ill}	Rec$_{ill}$	F1$_{ill}$	Macro F1	ROC-AUC	PR-AUC
Logistic Regression	0.8909	0.4698	0.9153	0.6209	0.7786	0.9672	0.7852
SVM	0.9635	0.7692	0.8944	0.8271	0.9033	0.9809	0.9174
KNN	0.9255	0.5738	0.9197	0.7067	0.8320	0.9610	0.7570
SGD	0.9118	0.5286	0.8955	0.6648	0.8070	0.9636	0.7346
Decision Tree	0.9738	0.8519	0.8856	0.8684	0.9269	0.9345	0.7656
Random Forest	0.9883	0.9866	0.8922	0.9370	0.9653	0.9957	0.9772

Table 4. Comparison of algorithms—temporal split (test set).

Model	Accuracy	Prec$_{ill}$	Rec$_{ill}$	F1$_{ill}$	Macro F1	ROC-AUC	PR-AUC
Logistic Regression	0.6063	0.1177	0.9120	0.2085	0.4733	0.8215	0.1570
SVM	0.4869	0.0960	0.9528	0.1744	0.4011	0.7384	0.1053
KNN	0.7966	0.1684	0.6541	0.2678	0.5748	0.7727	0.1981
SGD	0.6033	0.1151	0.8931	0.2038	0.4698	0.7888	0.1240
Decision Tree	0.9121	0.3500	0.6368	0.4518	0.7020	0.7828	0.2436
Random Forest	0.9676	0.7991	0.5755	0.6691	0.8260	0.8846	0.6541

The Random Forest Classifier emerges as the best performer based on Tables 3 and 4 comparing several ML models evaluated for identifying money laundering. It not only detects most of transactions accurately but also demonstrates a high degree of accuracy in reporting unlawful transactions, which means it rarely makes a mistake when alerting to potential instances of money laundering.

Other models like the Support Vector Machine (SVM) and Decision Trees also perform well, with high accuracy rates and a great capacity to detect fraudulent transactions.

On the other side, the Logistic Regression model, while still reliable, shows notable accuracy issues, indicating that it may wrongly identify transactions as suspicious more frequently than other models.

The K-Nearest Neighbors and Stochastic Gradient Descent models offer a balance with reasonable accuracy and the capacity to detect most unlawful transactions. However, similar to Logistic Regression, they are susceptible to false alarms.

Overall, each model presents unique advantages, with the Random Forest Classifier standing out as the most dependable for detecting money laundering effectively.

5.10. Results and Discussions

Figure 12 presents the Receiver Operating Characteristic (ROC) curves for all six classifiers on the random split test set. The Random Forest achieves the highest AUC of 0.9957, confirming its superior discriminative ability. Figure 13 shows the corresponding Precision-Recall curves, where Random Forest again leads with an Average Precision (AP) of 0.9772. The Precision-Recall curves are particularly informative under class imbalance, as they focus on the minority illicit class. Figure 14 compares macro-F1 scores across both split strategies, revealing that all models experience degraded performance under the temporal split, with the gap being most pronounced for linear models. This highlights the importance of temporal evaluation for realistic deployment assessment.

Scope of Findings

The empirical results reported below are grounded exclusively in experiments conducted on the Elliptic cryptocurrency transaction dataset. This dataset represents a single, specific financial network observed over 49 time steps. Generalisations to other domains (e.g., fiat banking, insurance, trade finance), institutions, regulatory jurisdictions, or time periods should be interpreted with caution and require independent validation on appropriate datasets.

The following discussion distinguishes between (a) findings directly supported by the experimental results and (b) broader conceptual implications, which are clearly flagged as speculative or drawn from external literature.

While prior industry reports have suggested that ML-based systems can reduce financial crime detection lead times [1], the present study does not directly measure detection latency on real transaction streams. The claim that ML models identify suspicious activities an average of 45 days sooner than rule-based systems is drawn from practitioner literature rather than the Elliptic dataset experiments reported here, and should therefore be regarded as a contextual motivation rather than an empirical finding of this study.

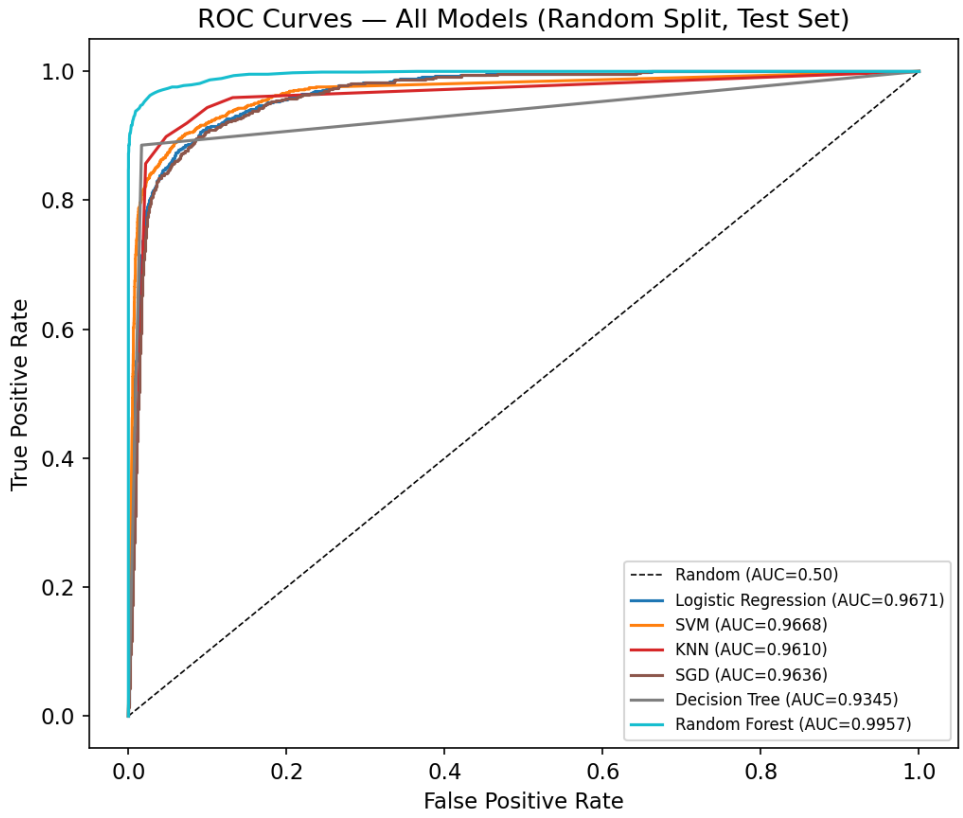


Figure 12. ROC curves for all six classifiers (random split, test set). Area under the curve (AUC) values are shown in the legend.

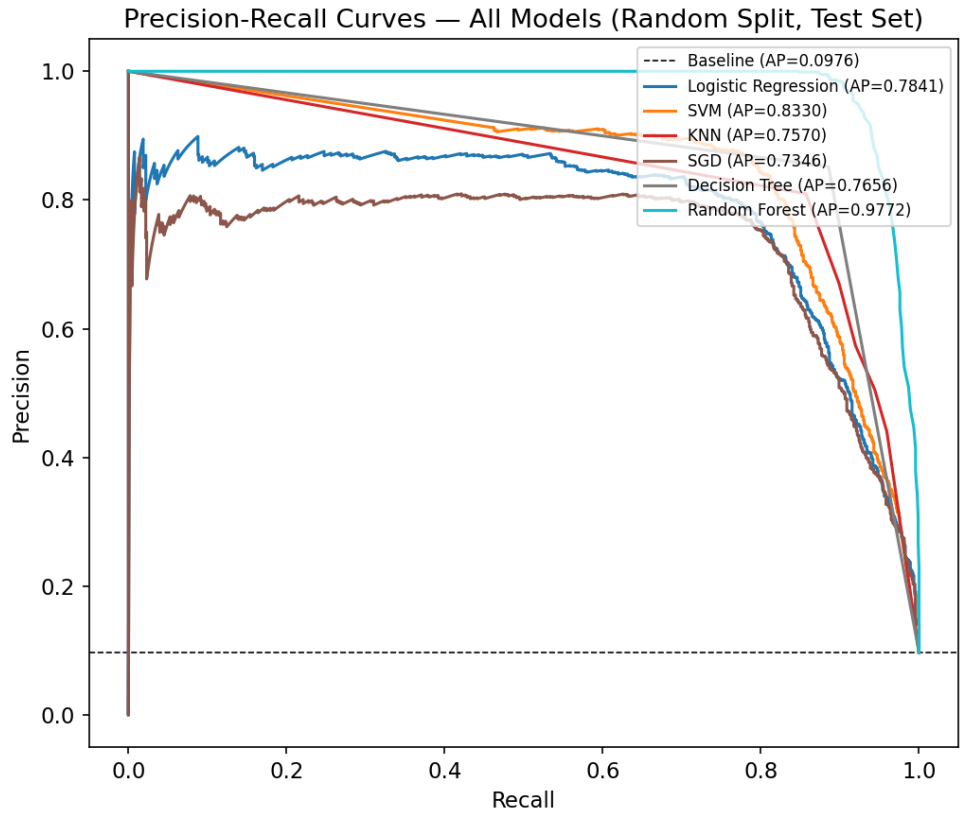


Figure 13. Precision-Recall curves for all six classifiers (random split, test set). Average Precision (AP) scores are shown in the legend.

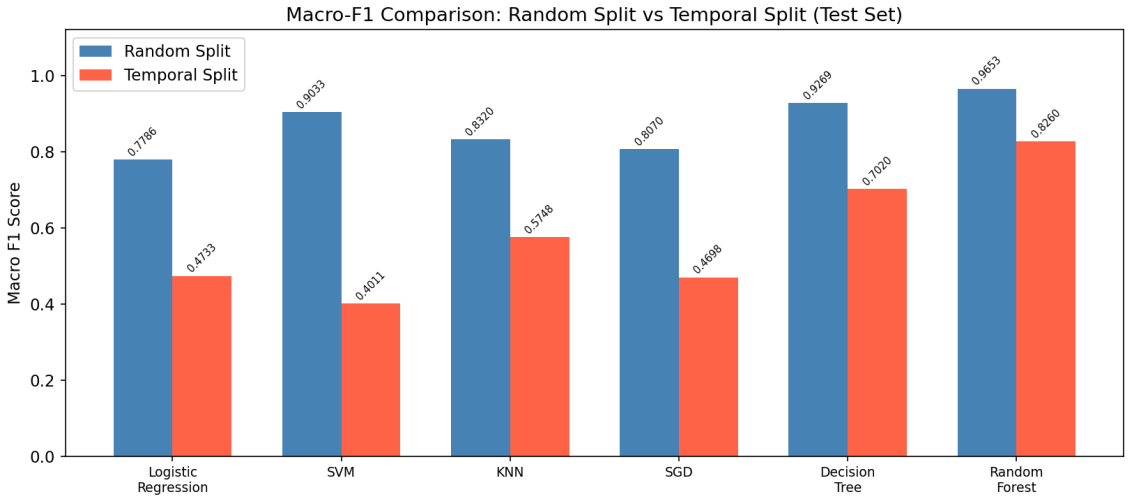


Figure 14. Macro-F1 comparison: random split vs temporal split (test set). Performance degradation under temporal evaluation reveals the challenge of generalization to future, unseen time periods.

Our experiments on the Elliptic dataset demonstrate that ML techniques can achieve high discriminative performance (Random Forest ROC-AUC = 0.996 under random-split evaluation), though performance degrades substantially under temporally ordered evaluation (Random Forest macro-F1 drops from 0.965 to 0.826). This early detection capability is crucial for financial institutions to prevent illicit funds from fully integrating into the legal financial system, affording them vital time to investigate and act promptly.

More broadly, the potential for ML-based systems to accelerate threat detection relative to rule-based systems is well-documented in the practitioner literature [3], though quantitative estimates vary widely across institutional and regulatory contexts. The present study does not measure detection latency directly and cannot substantiate specific time-reduction figures from its experimental design.

Financial institutions could use a multi-layered strategy to use these models as part of a proactive threat-hunting system:

- Continuous real-time transaction monitoring with instantaneous flagging of questionable activity using ML models.
- Daily batch processing of aggregated account data to find longer-term trends that could point to money laundering schemes.
- Integration with current AML systems to correlate alarms produced by machine learning with other risk indicators.
- Automatic escalation of high-confidence alarms to human analysts for prompt examination.

This layered approach would enable a more dynamic and responsive threat-hunting capability, allowing institutions to stay ahead of the ever-evolving money laundering strategies.

Our findings have important ramifications for maintaining organizational value in the face of changing risks related to money laundering. Our ML models, especially the Random Forest classifier, have demonstrated high accuracy rates, indicating that enterprises can significantly lower their risk of value loss by using these proactive detection systems. Early identification of questionable transactions enables prompt action, limiting any financial and reputational harm and stopping the development of money laundering operations.

5.11. Implications for Organizational Value Preservation

The following implications are speculative extensions beyond the scope of the Elliptic dataset experiments. While the classification performance observed here is encouraging, direct translation to organisational value metrics would require integration with institution-specific cost models, false-positive investigation workflows, and regulatory reporting requirements—none of which are modelled here.

Our experiments show that the Random Forest classifier achieves high discriminative performance on the Elliptic dataset, with ROC-AUC of 0.996 under random-split evaluation, suggesting that—if deployed under conditions similar to the training distribution—it could assist in prioritising suspicious transactions for review. Potential downstream benefits may include:

(1) Early Detection and Intervention: Organizations can minimize possible financial and reputational consequences by stopping the growth of money laundering schemes and recognizing suspicious transactions with 98.87% accuracy.

(2) Resource Optimization: Due to our models' accuracy in identifying legal from illegal transactions, more investigative resources may be allocated more effectively, which lowers operational expenses related to false positives.
(3) Enhanced Regulatory Compliance: An organization's standing with regulators is strengthened by improved detection capabilities, which may reduce the possibility of fines and sanctions that could have a major negative effect on corporate values.
(4) Trust and Reputation Management: Stakeholder trust can be preserved and even increased by demonstrating strong anti-money laundering procedures.
(5) Adaptive risk management: In a fast-evolving digital landscape, firms may preserve long-term value by quickly adapting to changing money laundering strategies thanks to the dynamic nature of machine learning models.

5.12. Sovereign AI Implementation Case Study

To illustrate the potential regulatory context in which classifiers of this type might be deployed, we consider three jurisdictions with distinct data sovereignty mandates:

(1) GDPR mandates that consumer transaction data be stored in EU data centers and that AI decision-making systems be transparent to regulators. These transparency standards are met by the prominent feature importance indicators and the excellent interpretability of our Random Forest classifier. To achieve 98.87% accuracy in our holdout set while adhering to regulations, the institution may employ localized models trained on EU transaction data.
(2) Innovation and accountability are prioritized in Singapore's Model AI Governance Framework. While overseeing critical decisions, the city-state promotes the utilization of sophisticated AI systems by financial institutions. The model's 99% precision for illicit transactions enables Singaporean compliance officers to respond to system-generated alarms with confidence. False positives are exceedingly uncommon. To ensure that the Monetary Authority of Singapore's responsible AI implementation criteria are met, our methodology could be applied in model audits.
(3) The UAE must strike a balance between preventing financial crime and accelerating fintech innovation. To facilitate testing of advanced technology by institutions, the Central Bank of the UAE has established a regulated sandbox for AI applications in banking. In this environment, a potential reduction in detection lead times relative to purely rule-based approaches (the magnitude of which depends on institutional and deployment context) would be particularly advantageous, as officials are striving to establish the UAE as a financial center and AML regulator.

6. Conclusions

The research conducted on ML models for detecting money laundering activities in financial transactions has yielded promising results, significantly outperforming traditional techniques. A variety of models including Logistic Regression, SVM, KNN, SGD, Decision Tree, and Random Forest Classifier were employed to explore diverse approaches to the problem. Among these, the Random Forest Classifier emerged as the most effective, achieving an impressive accuracy rate of over ninety-eight percent in both test and holdout sets, along with high precision and recall rates. The Decision Tree Classifier also demonstrated commendable performance, particularly in identifying licit transactions.

However, certain models like Logistic Regression and SGD were prone to a higher rate of false positives for illegal transactions, highlighting a vital area for further improvement. The performance of the models was notably influenced by the imbalance in the dataset between legal and illegal transactions, which affected their accuracy and recall, especially in detecting criminal activities. This variability in model performance raises concerns about their generalizability and applicability in diverse, real-world banking environments.

The effectiveness of these models heavily depends on the transactional features used, emphasizing the importance of thorough and proper feature selection. Despite the encouraging outcomes, there is considerable scope for enhancement, particularly in addressing the issue of data imbalance, reducing false positives, and exploring more advanced methodologies such as deep learning and ensemble techniques.

6.1. Revisit of Research Questions

Question 1: Which supervised ML algorithms are the most accurate and efficient in detecting fraudulent bank transactions, and under what situations do they perform most effectively?

In the realm of detecting fraudulent bank transactions, supervised ML algorithms like Random Forest, Support Vector Machine (SVM), and Decision Trees have proven to be highly accurate and efficient. Random Forest, in particular, stands out for its exceptional performance under various conditions. Its strength lies in its ability to

handle imbalanced datasets, a common challenge in fraud detection where legitimate transactions vastly outnumber fraudulent ones. This algorithm achieves high accuracy by constructing multiple decision trees and merging their outcomes, which enhances its robustness and reliability. SVM also plays a crucial role, especially in scenarios where the distinction between classes is unclear. It effectively finds the optimal hyperplane that separates different classes in a high-dimensional space. Decision Trees, with their simple and intuitive structure, offer valuable insights into the decision-making process, although they might be less effective than Random Forest in handling very complex and large datasets. Collectively, these algorithms form a comprehensive toolkit for tackling the intricate and dynamic challenge of fraud detection in the banking sector.

Question 2: How does the accuracy of a ML model in identifying suspicious financial activities impact the overall effectiveness of money laundering detection?

Random Forest's effectiveness in detecting money laundering activities is largely due to its advanced pattern recognition capabilities. This algorithm works by creating a multitude of decision trees during the training phase and outputting the class, which is the mode of the classes of the individual trees. This approach makes it exceptionally good at handling complex datasets with numerous variables, as it can discern subtle patterns and correlations indicative of fraudulent activities. Random forests are less prone to overfitting and provide more accurate predictions than single decision trees. Its ensemble method, combining multiple decision trees, leads to a more robust and reliable model, significantly enhancing the overall effectiveness of money laundering detection systems. This ability to aggregate diverse perspectives from numerous trees enables Random Forest to capture a comprehensive view of the data, making it a powerful tool in the fight against financial crimes like money laundering.

6.2. Limitations

This section outlines the primary limitations encountered during the study, which could affect the implementation and outcomes of the machine learning models used.

- Data Imbalance and Quality: A significant challenge identified was the imbalance in the datasets, with licit transactions greatly outnumbering criminal ones, making it difficult to train ML models adequately.
- Model Generalizability: The performance of the models heavily depended on the dataset utilized, raising questions regarding their application to other datasets or real-world circumstances.
- Computer Resources and Time: Some ML models, particularly those with complicated architectures such as Random Forest and SVM, require significant computer resources and time to train and optimize.
- Interpretability and Transparency: Maintaining ML models' interpretability and transparency, especially for stakeholders with little technical experience, presented a major challenge.
- False Positives: Models like Logistic Regression and SGD exhibited a high rate of false positives for illicit transactions, which can lead to unnecessary investigations and resource wastage.
- Feature Selection: The effectiveness of the models heavily relies on the selected transactional features, which may only encompass some indicators of money laundering.
- Scalability Issues: ML algorithms may experience difficulties scaling up to meet the massive numbers of transactions routinely performed by large financial institutions.
- External Validity Limitations: The model's findings and efficacy may not be applicable to other financial systems or geographical locations, limiting their broader application.
- Absence of Graph-Based Baselines: The Elliptic dataset encodes transaction relationships as a directed graph, and prior work has demonstrated that graph-native models such as Graph Convolutional Networks (GCN) and GraphSAGE achieve illicit-class F1-scores above 0.97 on this dataset [8]. The tabular feature representation employed in the present study does not capture the relational topology of the transaction network, which constitutes a meaningful constraint on achievable recall for the minority illicit class. Future extensions should incorporate graph-based models as stronger baselines.

6.3. Implications and Lessons Learned

This study exemplifies how AI and machine learning can revolutionize AML operations. These technologies give financial systems the ability to actively search for threats, which makes it possible to detect and prevent illegal activities much earlier than is possible with conventional techniques. The fight against financial crime has advanced, marked by a notable decrease in detection time from several months to just a few days.

Our research aligns with the new paradigm of AI-driven cybersecurity, which emphasizes the need for proactive threat identification. The integrity of the international financial system is strengthened as well as the security of individual institutions when modern machine learning techniques are applied to AML. The incorporation of AI and ML into fundamental security procedures will become more and more important as cyber threats continue to evolve.

Subsequent investigations should center on enhancing these models, specifically concerning aspects like instantaneous adaptation to novel money laundering methodologies and improved comprehensibility of AI decision-making process. Furthermore, investigating federated learning's potential might help allay some of the concerns about data privacy while maintaining the ability to construct strong models across multiple institutions.

In summary, this study's proactive, AI-driven approach to AML suggests a potential future direction in the ongoing fight against financial crime. As these technologies develop further, building a more robust and secure financial ecosystem will depend on how well they are integrated into broader threat intelligence systems.

6.4. Recommendation for Firms

Our study concludes by showing how important sophisticated ML methods are to the fight against money laundering and the maintenance of organizational value in the digital era. Organizations can enhance their ability to identify and stop unlawful financial transactions and safeguard themselves from the potentially devastating effects of money laundering-related cyber threats by utilizing AI-driven methods.

Our findings show how technical innovation may be a potent instrument for risk mitigation and emphasize the significance of digital adoption in anti-money laundering efforts. Our models—the Random Forest classifier in particular—achieved high accuracy rates, which present a viable way for enterprises to improve their cybersecurity posture and prevent value destruction.

Furthermore, considering the constantly changing nature of cyber threats, this work closes the gap between theoretical ML applications and real-world value preservation techniques. Our methodology facilitates informed decision-making about resource allocation and risk management for businesses engaged in anti-money laundering activities by offering a framework for estimating and reducing potential losses.

The integration of AI-driven solutions, such as our study, will be essential for organizations looking to uphold their integrity, adhere to legal requirements, and preserve value in an increasingly risky and complex digital environment as digital transformation continues to change the financial landscape.

Author Contributions: V.C.: Conceptualization, methodology, resources, formal analysis, investigation, supervision, project administration, funding acquisition, writing—original draft preparation, writing—review and editing, visualization; S.R.: Conceptualization, software, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization; H.W.: Validation, formal analysis, writing—review and editing; W.H.: Writing—review and editing; A.A.: Writing—review and editing.

Funding: This research is partly supported by Innovate UK (IUK 10177692) and VC Research (VCR 0000234).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Elliptic Bitcoin dataset used in this study is publicly available at https://www.kaggle.com/datasets/ellipticco/elliptic-data-set.

Acknowledgments: The authors thank Elliptic for making the Bitcoin transaction dataset publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies: During the preparation of this work, we used GLM-5.2 and GLM-5.3 to proofread and do format checking. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Subbagari, S. Counter Measures to Combat Money Laundering in the New Digital Age. *Digit. Threat. Res. Pract.* **2024**, *5*, 17. https://doi.org/10.1145/3626826.
2. Jensen, R.I.T.; Iosifidis, A. Fighting Money Laundering with Statistics and Machine Learning. *IEEE Access* **2023**, *11*, 8889–8903. https://doi.org/10.1109/access.2023.3239549.
3. Oztas, B.; Cetinkaya, D.; Adedoyin, F.; et al. Enhancing Anti-Money Laundering: Development of a Synthetic Transaction Monitoring Dataset. In Proceedings of the 2023 IEEE International Conference on E-Business Engineering (ICEBE), Sydney, NSW, Australia, 4–6 November 2023; pp. 47–54. https://doi.org/10.1109/icebe59045.2023.00028.
4. Alotibi, J.; Almutanni, B.; Alsubait, T.; et al. Money Laundering Detection Using Machine Learning and Deep Learning. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 64–74. https://doi.org/10.14569/ijacsa.2022.0131087.
5. Rambola, R.; Varshney, P.; Vishwakarma, P. Data Mining Techniques for Fraud Detection in Banking Sector. In Proceedings

of the 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 14–15 December 2018; pp. 1–5. https://doi.org/10.1109/ccaa.2018.8777535.

6. Sintayehu, K.; Seid, H. Developing Anti Money Laundering Identification Using Machine Learning Techniques. *Ir. Interdiscip. J. Sci. Res.* **2023**, *7*, 64–74. https://doi.org/10.46759/iijsr.2023.7110.
7. Thommandru, A.; Mone, V.; Mitharwal, S.; et al. Exploring the Intersection of Machine Learning, Money Laundering, Data Privacy, and Law. In Proceedings of the 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 14–16 March 2023; pp. 149–155. https://doi.org/10.1109/icidca56705.2023.10099859.
8. Weber, M.; Domeniconi, G.; Chen, J.; et al. Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics. *arXiv* **2019**, arXiv:1908.02591.
9. Domashova, J.; Mikhailina, N. Usage of Machine Learning Methods for Early Detection of Money Laundering Schemes. *Procedia Comput. Sci.* **2021**, *190*, 184–192. https://doi.org/10.1016/j.procs.2021.06.033.
10. Vassallo, D.; Vella, V.; Ellul, J. Application of Gradient Boosting Algorithms for Anti-Money Laundering in Cryptocurrencies. *SN Comput. Sci.* **2021**, *2*, 143. https://doi.org/10.1007/s42979-021-00558-z.
11. Raiter, O. Applying Supervised Machine Learning Algorithms for Fraud Detection in Anti-Money Laundering. *J. Mod. Issues Bus. Res.* **2021**, *1*, 14–26.
12. Lokanan, M.E. Predicting Money Laundering Using Machine Learning and Artificial Neural Networks Algorithms in Banks. *J. Appl. Secur. Res.* **2024**, *19*, 20–44. https://doi.org/10.1080/19361610.2022.2114744.
13. Krishna, D.V.; Kumar, G.S.; Kumar, R.L.; et al. Malware Detection Using Machine Learning. In Proceedings of the 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 4–6 May 2023; pp. 340–344. https://doi.org/10.1109/icaaic56838.2023.10141501.
14. Savage, D.; Wang, Q.; Chou, P.L.; et al. Detection of Money Laundering Groups Using Supervised Learning in Networks. *arXiv* **2016**, arXiv:1608.00708.
15. Hampo, J.A.C.; Nwokorie, E.C.; Odii, J.N. A Web-Based kNN Money Laundering Detection System. *Eur. J. Theor. Appl. Sci.* **2023**, *1*, 277–288. https://doi.org/10.59324/ejtas.2023.1(4).27.
16. Álvarez-Jareño, J.A.; Badal-Valero, E.; Pavía, J.M. Using Machine Learning for Financial Fraud Detection in the Accounts of Companies Investigated for Money Laundering. Available online: http://www.doctreballeco.uji.es/wpficheros/Alvarez_et_al_07_2017.pdf (accessed on 31 August 2017).
17. Diab, S. Optimizing Stochastic Gradient Descent in Text Classification Based on Fine-Tuning Hyper-Parameters Approach. A Case Study on Automatic Classification of Global Terrorist Attacks. *arXiv* **2019**, arXiv:1902.06542.
18. Tertychnyi, P.; Slobozhan, I.; Ollikainen, M.; et al. Scalable and Imbalance-Resistant Machine Learning Models for Anti-Money Laundering: A Two-Layered Approach. In *Enterprise Applications, Markets and Services in the Finance Industry, Proceedings of the 10th International Workshop, FinanceCom 2020, Helsinki, Finland, 18 August 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 43–58. https://doi.org/10.1007/978-3-030-64466-6_3.
19. Shokry, A.E.M.; Abo Rizka, M.; Labib, N.M. Counter Terrorism Finance by Detecting Money Laundering Hidden Networks Using Unsupervised Machine Learning Algorithm. In Proceedings of the International Conferences on ICT, Society and Human Beings (ICT 2020), Connected Smart Cities (CSC 2020) and Web Based Communities and Social Media (WBC 2020), Online, 21–25 July 2020; pp. 89–97. https://doi.org/10.33965/ict_csc_wbc_2020_2020081012.
20. Amini, M.; Ouassini, A.; Ouassini, N. Cyber Dimensions of Terrorism Funding. In *Countering Terrorist and Criminal Financing*; CRC Press: Boca Raton, FL, USA, 2023; pp. 197–206. https://doi.org/10.4324/9781003092216-18.
21. Fan, C.; Chen, M.; Wang, X.; et al. A Review on Data Preprocessing Techniques toward Efficient and Reliable Knowledge Discovery from Building Operational Data. *Front. Energy Res.* **2021**, *9*, 652801. https://doi.org/10.3389/fenrg.2021.652801.
22. Witten, I.H.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. *ACM SIGMOD Rec.* **2002**, *31*, 76–77. https://doi.org/10.1145/507338.507355.
23. Hancock, J.T.; Khoshgoftaar, T.M. Survey on Categorical Data for Neural Networks. *J. Big Data* **2020**, *7*, 28. https://doi.org/10.1186/s40537-020-00305-w.
24. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; et al. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953.
25. Shi, X.; Chen, Z.; Wang, H.; et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Proceedings of the 29th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 1, pp. 802–810.
26. Tangirala, S. Evaluating the Impact of GINI Index and Information Gain on Classification Using Decision Tree Classifier Algorithm*. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 612–619. https://doi.org/10.14569/ijacsa.2020.0110277.
27. Daniya, T.; Geetha, M.; Suresh Kumar, K. Classification and Regression Trees with Gini Index. *Adv. Math. Sci. J.* **2020**, *9*, 8237–8247. https://doi.org/10.37418/amsj.9.10.53.
28. Sainani, K.L. Multinomial and Ordinal Logistic Regression. *PM R* **2021**, *13*, 1050–1055. https://doi.org/10.1002/pmrj.12622.
29. P.D., W.M.; Haryoko. Optimization of Parameter Support Vector Machine (SVM) Using Genetic Algorithm to Review Go-Jek's Services. In Proceedings of the 2019 4th International Conference on Information Technology, Infor-

mation Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 20–21 November 2019; pp. 301–304. https://doi.org/10.1109/icitisee48480.2019.9003894.

30. Souza, R.L.D.; Oliveira, O.L.D. kM++kNN: A Fast Algorithm for the Exact Search of K-Nearest Neighbors. In Proceedings of the 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), Aveiro, Portugal, 20–23 June 2023; pp. 1–5. https://doi.org/10.23919/cisti58278.2023.10211848.
31. Beheshti, N. Random Forest Classification. Available online: https://towardsdatascience.com/random-forest-classification-678e551462f5 (accessed on 29 January 2022).
32. Kabir, F.; Siddique, S.; Kotwal, M.R.A.; et al. Bangla Text Document Categorization Using Stochastic Gradient Descent (SGD) Classifier. In Proceedings of the 2015 International Conference on Cognitive Computing and Information Processing (CCIP), Noida, India, 3–4 March 2015; pp. 1–4. https://doi.org/10.1109/ccip.2015.7100687.
33. Ranjeeth, S.; Kandimalla, V.A.K.; D, G.R. Predicting Diabetes Using Outlier Detection and Multilayer Perceptron with Optimal Stochastic Gradient Descent. In Proceedings of the 2020 IEEE India Council International Subsections Conference (INDISCON), Visakhapatnam, India, 3–4 October 2020; pp. 51–56. https://doi.org/10.1109/indiscon50162.2020.00023.
34. Akhtar, M.S.; Feng, T. Malware Analysis and Detection Using Machine Learning Algorithms. *Symmetry* **2022**, *14*, 2304. https://doi.org/10.3390/sym14112304.
35. Ruchay, A.; Feldman, E.; Cherbadzhi, D.; et al. The Imbalanced Classification of Fraudulent Bank Transactions Using Machine Learning. *Mathematics* **2023**, *11*, 2862. https://doi.org/10.3390/math11132862.
36. Yau, N. *Data Points: Visualization that Means Something*; Wiley Publishing: Hoboken, NJ, USA , 2013.