STM JOURNALS

Current Trends in
Information Technology

https://journals.stmjournals.com/ctit

ISSN: 2249-4707 (Online)
ISSN: 2348-7895 (Print)
Volume 14, Issue 2, 2024
May–August
DOI (Journal): 10.37591/CTIT

Review CTIT

Money Laundering Transaction with Machine Learning

R. Lingeswari[1,*], S. Brindha[2]

Abstract

This study discusses the use of machine learning algorithms to discover firms that are prone to money laundering. The purpose of this research is to develop, describe, and test a machine learning model for determining which bank transactions should be physically scrutinized for money laundering activities. To train a supervised machine learning model, three categories of historical data are required: legitimate "normal" transactions, transactions flagged as suspicious by the bank's internal alert system, and instances of suspected money laundering reported to authorities. The model is trained to use various data sources, including background details of both the sender and receiver, past behaviors, and transaction records, to predict the likelihood that a new transaction warrants reporting. Our newly designed procedure outperforms the bank's current methodology in terms of a fair assessment of performance. This research represents one of the scarce published anti-money laundering (AML) models for identifying suspicious transactions that has been tested on a dataset of realistic size. In addition, the paper provides a new performance indicator that was created particularly to compare the proposed solution against the bank's existing anti-money laundering system. The greatest quality prediction results are obtained when a gradient boosting method is used instead of decision trees. The speed at which the optimal set might be produced was determined using the Hyperopt and Optuna Python package, and the quality of hyper parameter selection was studied. The model makes it possible to compile a list of the most crucial indicators for the early detection of money laundering and terrorist financing (ML/TF) organizations, and also provides appropriate suggestions for enhancing the compliance control process. Monetary laundering is a result of corruption, illegal activities, and organized crime, all of which have a social effect and have embroiled numerous groups, both directly and indirectly, in the laundering of illicit cash via a number of methods. This paper suggests employing a machine learning method to identify potentially suspicious activities among non-banking correspondents. A non-banking correspondent is a kind of financial agent that executes financial transactions on behalf of particular banking customers. Our results show that certain ways are more suited to this circumstance than others, and that these methodologies make it simpler to spot irregularities and suspicious transactions in this kind of financial intermediary.

***Author for Correspondence**
R. Lingeswari
E-mail: lavalingu@gmail.com

[1]Research Scholar, Department of Computer Science, St. Peter's Institute of Higher Education & Research, Chennai, Tamil Nadu, India
[2]Associate Professor, Department of Computer Applications, St. Peter's Institute of Higher Education & Research, Chennai, Tamil Nadu, India

Received Date: February 29, 2024
Accepted Date: April 25, 2024
Published Date: July 03, 2024

Citation: R. Lingeswari, S. Brindha. Money Laundering Transaction with Machine Learning. Current Trends in Information Technology. 2024; 14(2): 1–15p.

Keywords: Anti-money laundering (AML), machine learning, terrorist financing, Non-bank correspondents, money laundering, transactions

INTRODUCTION

Because of the nature of a bank's business activities, it is conceivable for it to be involved in money laundering schemes. There is no bank that is fully immune to the possibility of becoming a money laundering accomplice, and noncompliance with anti-money laundering regulations is the most prevalent cause of penalties, loss of corporate image, and, eventually, termination of a credit

organization's license. Banks are progressively strengthening the criteria for their operations and procedures, as well as implementing new technology and risk-reduction approaches, in order to mitigate the risks connected with the illicit use of financial instruments. This holds especially true in the realm of money laundering, where stringent security measures are paramount. For example, the process of automating anti-money laundering and counter-terrorist financing processes [1] is now undertaking a number of tasks [2, 3]. The large amount of data to be reviewed is the work's distinguishing feature, and only the use of high-tech techniques allows fraud to be tracked down and stopped with pinpoint accuracy. The number of dubious transactions has been progressively dropping in recent years because of targeted anti-money laundering and counter-terrorist funding initiatives.

The Bank of Russia periodically conducts analysis on this matter. During the initial 6 months of 2020, the count of suspicious transactions in the banking sector decreased by 1.5 times compared to the corresponding period in 2019.

A possible explanation for this phenomenon is the rise in the number of supervisory and control procedures, which frequently result in the closure of compliant businesses, resulting in significant revenue losses for credit institutions; additionally, the fact that money laundering has occurred is discovered after the customer has completed the transaction, i.e., after the illegal conduct has occurred. Identifying frequent money laundering and terrorist financing signals of suspicious organizations when opening a current account is important since it helps you to spot a problem early and take the required steps to address it. As a consequence, identifying firms that are vulnerable to money laundering at an early stage is critical. In this study, machine learning algorithms are being used to predict the risk of money laundering schemes happening during the process of opening a company's current account. The goal is to enhance the overall compliance control process. The study's practical relevance lies in the production of a list of the most important indicators for identifying money laundering and terrorist financing firms, as well as the formulation of relevant recommendations for improving the compliance control system. As a consequence of the outcomes of such modeling, AML/CFT measures will be maintained at a high level, preventing criminals from exploiting deficient control measures in credit organizations. Banks will be able to detect money laundering organizations early on and report pertinent information to the Federal Financial Monitoring Service for inquiry as soon as possible [3].

According to the World Bank's Universal Financial Access Initiative, there are 1.7 billion people worldwide without access to a basic transaction account. Non-bank correspondents (NBC) serve as the initial step in the financial inclusion and poverty alleviation process, as stated by the World Bank. They have the potential to be more accessible than traditional financial service providers, and they have the ability to boost financial service provider trust via typical adult activities. As a consequence, non-bank correspondents are non-financial firms that are part of the everyday economy and whose principal activity involves currency exchange, as a strategy to give a new option for financial inclusion and to incorporate customers into the formal financial system. Customers in various developing countries profit from this strategy, but due to a lack of anti-money laundering procedures, non-bank correspondents are unable to notice suspicious transactions or implement an anti-money laundering system. Non-banking correspondents have expanded their operations in countries and regions such as China, India, Bangladesh, Pakistan, Indonesia, Turkey, the Democratic Republic of the Congo (DRC), South Africa (South Africa), Brazil (Brazil), Colombia (Colombia), Ecuador (Mexico), and others, where certain conditions, such as economic informality, make it easier for criminal organizations to launder money through traditional or new schemes. As a result, this type of correspondent is particularly susceptible to criminal organizations and illicit activities [4, 5].

This problem has piqued the interest of academics from a wide range of fields. However, in the framework of anti-money laundering methods, the non-banking correspondent is a fresh phenomenon to explore, and data science approach inspires us to design an effective plan. Our suggestion is to use algorithmic analysis and visualization to make it simpler to spot questionable behavior in non-banking

connections. Our contribution involves identifying abnormal transactions within a non-banking correspondent using real-world data analysis and machine learning techniques, alongside empirical rules and insights on money laundering gleaned from previous research conducted by other scholars. We use basic machine learning and visualization methodologies in this research to identify suspicious transactions in non-bank correspondents, an industry that is growing in popularity in developing countries such as India and China [6]. We used unsupervised machine learning approaches since we could not use supervised machine learning algorithms because the test data did not provide exit labels. The true size of money laundering transactions is unknown and vague, perhaps due to financial institutions' lack of desire and tools to examine the extent of money laundering activities in their accounts. Money laundering was estimated to account for between 0.05 and 0.1% of all transactions handled via the Society for Worldwide Interbank Financial Telecommunications system. The entire amount of money laundered via the financial system is comparable to roughly 2.7% of worldwide gross domestic product, or US$1.6 trillion in 2009, according to a meta-analysis undertaken by the United Nations Office of Drugs and Crime (2011). According to the World Bank, the estimated value of money laundering amounts to $2.85 trillion. Financial fraud of this size presents a serious threat to society and economies all around the world, thus it is vital to detect as many fraudulent transactions as possible [7, 8]. As a consequence, the present research is focused on developing a system that can discriminate between a small number of suspicious/fraudulent transactions and a large number of legitimate transactions. Without knowing which data points are connected to money laundering and which are not, the algorithms try to detect patterns in data. The algorithms aim to identify patterns that differentiate between money laundering and legitimate operations by utilizing labeled data, where the outcome (whether it involves money laundering or not) is known.

Supervised learning is often favored over unsupervised learning when data with known outcomes/labels is available. This is significant in the context of AML because, unlike other types of financial fraud, a financial institution seldom finds whether a money laundering suspect is genuinely responsible for a crime. However, by mimicking "suspicious" behavior rather than actual money laundering, this issue may be avoided. We devise a method to distinguish between legitimate transactions and those exhibiting signs of potential money laundering utilizing supervised machine learning. Rather of relying on dubious accounts or groups of so-called partners [9], we built our model based on the transactions [10].

Most supervised anti-money laundering techniques operate under the assumption that experts flag suspicious activities, while legitimate activities are drawn from a pool of typical customers, this is based on the understanding that the probability of a random (or normal) activity being deemed suspicious is extremely low. We broaden the scope of lawful transactions to cover transactions from both random customers and AML warnings that did not result in an AML report, as well as third-party vendor transactions. As we shall see, developing a thorough prediction model necessitates the inclusion of both types of legitimate transactions. Although data sources and modeling frameworks in the AML literature are relatively restricted, we condense background customer data, transaction details and history, as well as any past instances of dubious conduct (semi-questionable), into a predefined set of clearly defined explanatory variables (data presented in matrix format) [11]. We train a prediction model utilizing state-of-the-art supervised machine learning methods in combination with proper model tweaks in order to learn a binary output (suspect or real).

Previous AML research has focused mostly on small or simulated data sets based on real data. We put our methodologies to the test in a real-world AML scenario in Norway, with a large and genuine data set. As a consequence, in a real-life circumstance, the performance outcomes as expected. To our knowledge, no previous AML study of this magnitude has been published. We refer to each transaction as having a sending party and a receiving party throughout this work, which are both defined as follows: Account administrators are the people or companies in control of the sending and receiving accounts. We also go through the data and how we organized it into relevant explanatory elements [1].

MATERIAL AND THEORETICAL BASES OF RESEARCH

There are two basic methods for recognizing suspicious businesses. The first is based on commercial principles. Expert criteria are poor in identifying new legalization techniques, which is a key downside, in addition to a large amount of false positives. The alternative method employs mathematical techniques and machine learning methodologies. Because the emergence of machine learning methods has enabled the automation of decision-making processes, the study will be built on the basis of models. The purpose is to identify organizations that are vulnerable to money laundering and terrorist funding as soon as possible, such as when accounts are opened. The stages of the research are represented in Figure 1.

The most effective model and a collection of the most significant components for the identification of malignant lymphoma and tumor progression are developed using quality assessments. Typical ways for carrying out the procedures mentioned above were explored prior to the study. There are various ways for converting category data into numerical characteristics, each with its own set of benefits and drawbacks. Figure 2 depicts coding methods for categorical variables. Within this category, there are three subgroups: basic encoders, level-based encoders, and target variable knowledge-based encoders [12, 13].

The distinctiveness of the objective and data validates the selection of a particular algorithm for the task. Working with data that had categorical variables requires the use of cross-validation techniques such as: None, single validation, and double validation are examples of validation procedures.

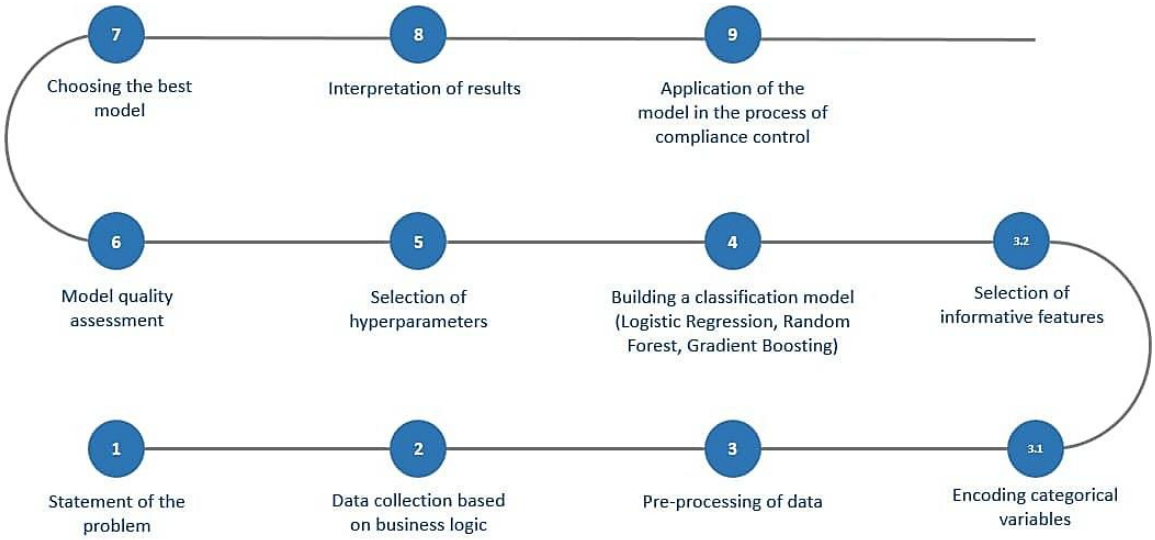


Figure 1. Scheme of the Research.

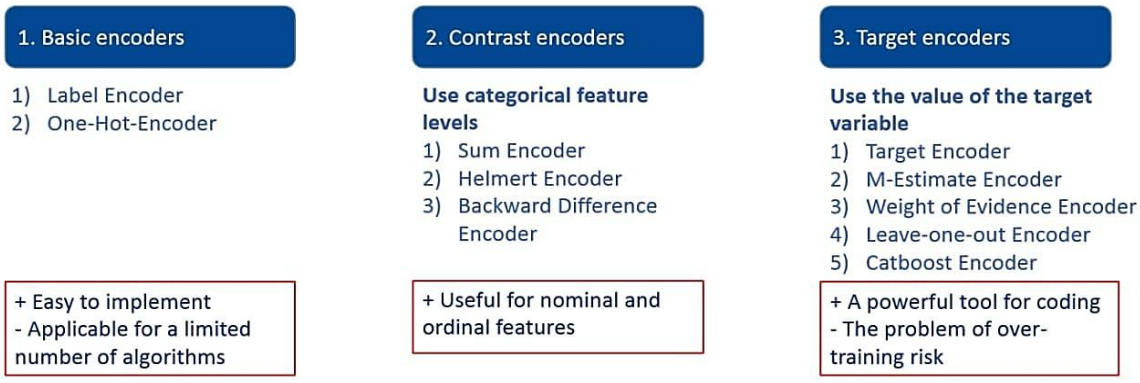


Figure 2. Analysis of the coding methods of category variables.

Cross-validation is often missed, resulting in an over fitted model. To evaluate an organization's propensity for ML/TF, methods such as logistic regression and ensemble algorithms were used to address the classification problem: boosting, bagging, random forest, and stacking were all used to assess an organization's propensity for ML/TF. The usage of a variety of algorithms allows for high-quality predictions.

Machine learning algorithms often contain a large number of hyper parameters that must all be set before the method can be utilized. The primary objective of parameter selection is to find an equilibrium between the model's complexity and its capacity to generalize across a wider range of data. Several optimization approaches are studied in this research work, including grid search, random search, and SMBO (Sequential Model-Based Optimization) [9].

Data Collection

We have been given anonymized data on a big class of flagged transactions across a broad variety of time periods thanks to a partnership with the bank. We also know which transactions resulted in the filing of a complaint and which transactions resulted in the submission of a report. We also have a totally random sample of ordinary transactions from the same time period that were not exposed to an alert or otherwise scrutinized, which might be used to augment these alert-based transactions. When these transactions are made public, background variables for both the sending and receiving parties become available. Apart from that, we have 2 months' worth of transaction data for any accounts that the parties have access to. It is conceivable that many warnings are linked to the same suspicious behavior. Currently, throughout the case-building process, warnings that are connected to the same party are grouped into a single case.

In order to generate a data set with the highest level of transparency possible and to eliminate the underlying reliance between the simulated transactions, we had to filter out a number of transactions from our research set. We made assured that there was only one alerted transaction connected with a given case in our data set by ensuring that only one alerted transaction linked with a specific case was present. Furthermore, 2 days is the smallest duration between two transactions that originate from or end with the same party. In this scenario, the specifications were selected as a trade-off to minimize dependence on observations while still retaining a sufficient amount of data. The non-bank correspondent collected transaction receipts daily, in chronological order, at the end of each day. Furthermore, the digital data was downloaded and kept on a monthly basis based on the date. To initiate the investigation and merge the datasets, the proprietor of the non-bank correspondent supplied all of this information.

This study's data was obtained through a non-banking correspondent who gave the information. The non-banking correspondent receives and transmits transaction data through the financial institution's point-of-sale machine, which is linked to the financial institution's information system. After each transaction is completed, two receipts are generated: one is sent to the transaction holder, and the other is utilized as a backup for the transaction information retained by the non-bank correspondent. The receipt stores the date and time of the purchase, the transaction type, and the transaction number for each transaction. This data may be linked to various products offered by the financial institution, such as debit cards, credit cards, personal accounts, business accounts, or loan accounts, depending on the financial institution.

The receipts data had to be organized into a dataset so that it could be analyzed and visualized, and then analytical and machine learning approaches could be utilized. The financial institution operates an information system that enables each non-bank correspondent to access transaction details, excluding the product number associated with the transaction. For instance, deposits made to current or savings accounts do not include the product number, whereas deposits made to business accounts or other financial instruments do, and vice versa. However, it was an oversight that the product number was not

included in debit card withdrawals. On the other hand, such facts were found on the receipt. In order to receive the whole set of data in digital format and make full use of the database, it was necessary to include some receipt data in the digital file. We did this by experimenting with the Google Vision API and the OCR service (Optical Character Recognition), as well as building computer programs to retrieve missing data from a receipt image. Despite the fact that the tool recognized the text, due to numerous physical properties of receipts, such as the thermal printing process, it lacked the accuracy to record the whole receipt's contents. We resorted to manually inserting the receipt information into the database in light of these disclosures. To ensure the correctness of the procedure, we examined 100 randomly selected transactions from the database and cross-referenced them with receipts, resulting in a 100% match.

Data Analysis

The database has numeric and category columns, as well as a date and time column for each transaction. The purpose of using this structure is to uncover unusual transactions that might be linked to money laundering in any of the 64 crimes mentioned in Colombia's criminal code. Despite the fact that transaction amounts in a non-bank correspondent are limited and low when compared to transactions at banking institution branches, we noticed some distinct patterns, such as transactions that were similar in nature, transactions that occurred on specific days and at specific times, and transactions that occurred on a regular basis. Withdrawals and savings account deposits account for the great majority of transactions (Table 1), and 72% of transactions are closed values, which is notable considering that closed values are seldom utilized for debt or service payments (Table 1). We make an attempt to approach these transactions in an innovative manner.

Transactions are subject to valuation limitations, which vary depending on the kind of transaction. As a consequence, the maximum withdrawal amount is ten Colombian million pesos, the maximum transfer amount is six Colombian million pesos, and the maximum amount for other transactions is three Colombian million pesos. As a consequence, we observed that neglecting to account for this discrepancy leads to a variable disproportion problem. To address this, we divided the raw database into three portions, each of which took into account the three transaction constraints. Withdrawals (Withdrawal), transfers (Transfer), and other transactions (Other Transactions) are all included in the first dataset (Various). Based on the value, we were able to discover the asymmetric behaviour of distributions for the withdrawal database as well as other databases, as well as how some closed values greatly exceed the mean. When it comes to categorical variables, we determine the features count for each variable, with the savings account product being the most essential, and the deposit in this product being the most relevant transaction type for the various datasets. The categorical variable from the withdrawal dataset, on the other hand, has a dense distribution with a broad range of values that fall outside of the distribution's normal ranges. We used the box plot in particular to see whether a distribution was skewed and if any possibly unusual observations (outliers) were included in the withdrawal dataset. In order to make the database stronger overall, we removed some variables that did not provide important information and included more variables that were derived from the pertinent variables. We subdivided the time variable into various components for withdrawal scenarios, including month, day of the month, day of the week, hour, and time intervals as shown in Figures 3 and 4.

Table 1. Relative frequency by transaction type.

Type	Code	Frequency
Withdrawal	Withdrawal	0.4286
Savings account deposit	Deposit savings	0.3734
Collection	Collection	0.1111
Current account deposit	Current deposit	0.0643
Credit card collection	Credit card pass	0.0096
Transfer	Transfer	0.0066
Other collections	Purse	0.0064

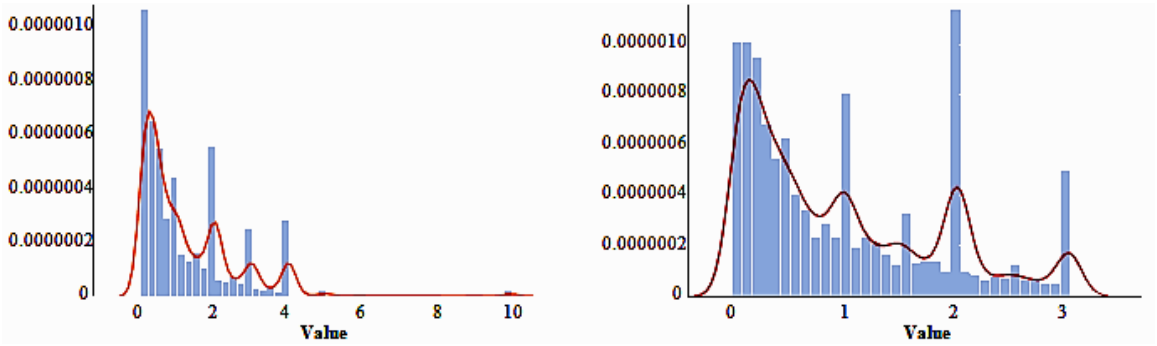


Figure 3. Value Histogram.

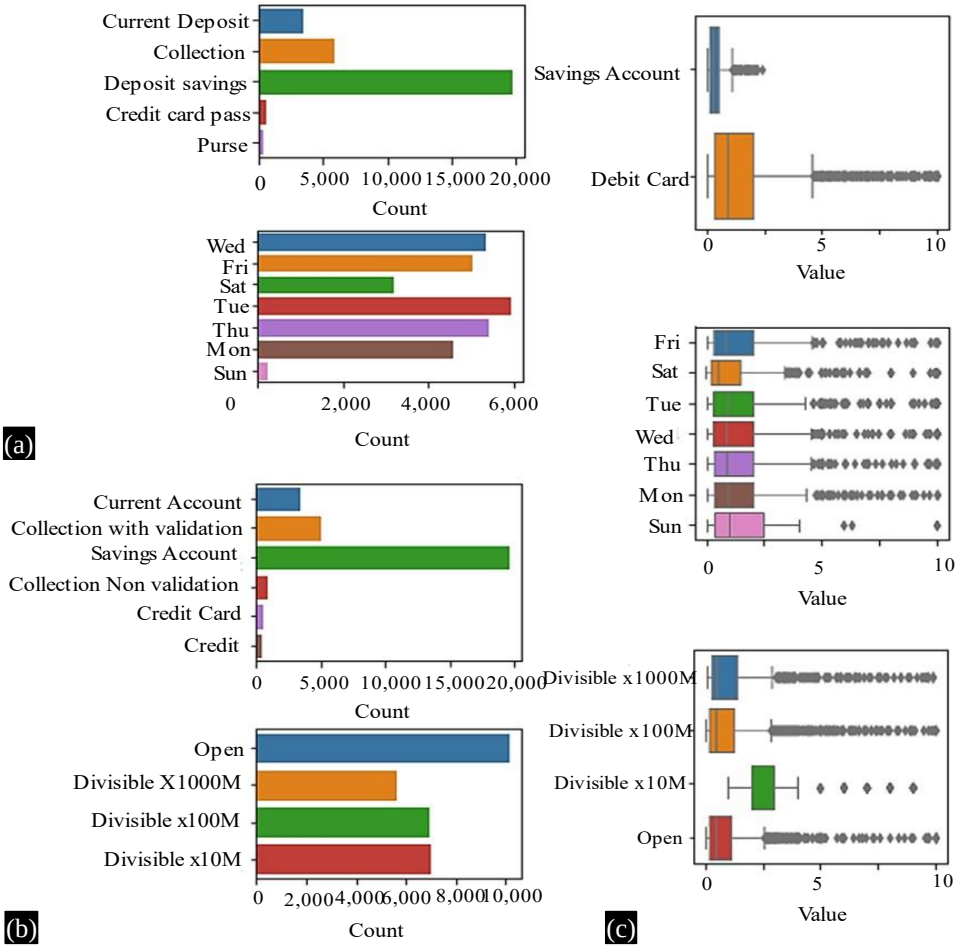


Figure 4. Categorical variables, (a) Withdrawal data set, (b) Various data set.

a. Transaction type in various dataset.
b. Product type in various dataset.
c. Boxplot(Box and whisker plot) in WITHDRAWAL dataset.

This allowed us to extract additional insights from the dataset, such as daily frequency, among other examples. In summary, the dataset originally consisted of 64,000 records, but after preprocessing and cleaning, it was reduced to 52,512 records and nine variables. We divided this raw data into three distinct datasets, as previously indicated.

Isolation Forest

It is an unsupervised learning approach that uses decision trees to locate abnormalities in a dataset. A statistical anomaly is an occurrence or observation that differs significantly from prior ones,

increasing the likelihood that it was generated by a different mean than the rest of the data. Typically, algorithms create the profile of typical data before isolating data that is out of the ordinary; however, Isolation Forest uses the opposite approach to uncover anomalies, isolating unusual data as soon as it is found. As a consequence, anomaly detection has evolved into a two-stage process. Initially, the method constructs isolation trees from subsets of the training set, indicating that trees are created by repeatedly partitioning the training set until instances are isolated or the trees reach a specified height limit l, roughly equal to the average tree height specified by the sub-sampling size. Thus, the algorithm structure is:

Input: X- input data, *T*- number of trees, ψ- sub-sampling size
Output: a set of *T*i Trees

1 *Initialize* Forest:
2 setheight limit l= ceiling(log2ψ);
3 for$i=1$*to*do
4 X^{l}←sample(X,ψ)
5 Forest←ForestUiTree(X^{l},0,l)
6 End
7 returnForest

Algorithm 1. iForest(X,T,ψ)

In the second stage, test instances are processed through a sequence of isolation trees to obtain an anomaly score for each instance. This score is calculated based on the predicted route length for each test case in the data set. As a consequence, this technique is more specialized than earlier algorithms and uses less computer resources. The dispersion diagram is presented after executing this approach on the WITHDRAWAL dataset. This graph depicts which transactions are unusual and which are not. Since a consequence, we learn that the dollar amount of a transaction is not the only factor that determines data attributes, as there are both anomalous and usual points within the value range.

The support vector machine (SVM) is a linear classifier based on the notion of margin maximisation. The support vector machine defines this algorithm's methodology (SVM). The SVM creates an appropriately split hyperplane that separates the data into two groups while conducting a classification task. This hyperplane can be used for both classification and regression tasks, and it needs to be given a labelled training data set that looks like this: x1,...xn, where n is the number of observations and is a number collection. In this case, the model was trained using data from just one kind of information, that is, positive information. To put it another way, this algorithm tries to categorise one kind of thing and distinguish it from the other possible types of objects. Prior to performing this task, the model needs to be trained to identify and classify this object as an outlier. It also infers the standard class features and predicts which data is unique based on these characteristics as shown in Figure 5.

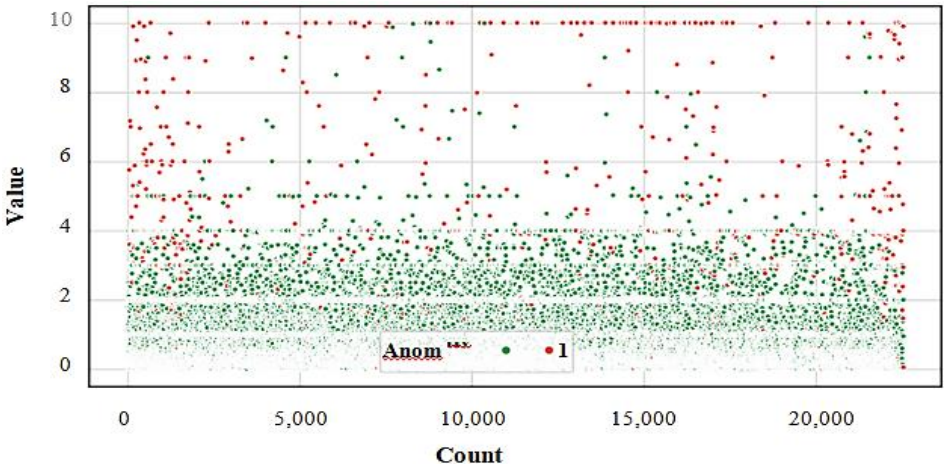


Figure 5. iForest dispersion diagram for Withdrawal dataset.

We apply this method to the training and testing datasets for the Withdrawal dataset, respectively, to assess the detection performance of the classifier utilizing the One-Class SVM algorithm. Because there are abnormal and typical points within the range, just as there were in the prior technique, the transaction value is not the sole component determining data characteristics, and this approach is used to both training and testing data.

Complementary Analysis

We looked at the top 10 transactions reported as unusual by each approach using two separate methodologies. In both cases, we noticed that values labeled as closed had the highest rates. We also noticed that numbers do not necessarily represent the highest or lowest values in each dataset since the algorithms used to determine the transaction value took into account additional elements that were not previously taken into account as shown in Figure 6. When compared to the other methods, the One-class SVM approach demonstrates a more consistent categorization of anomalous and normal values than the Isolation Forest methodology [14].

Furthermore, we used the K-prototype approach to do statistical analysis. This method, which is similar to k-means clustering, clusters objects with mixed, quantitative, and category attributes together. When things are clustered together, it is because they have the most comparable properties. Objects are grouped together against k-prototypes rather than k-means in this scenario [15]. Numeric data is grouped using k-means clustering based on Euclidean distance, while categorical data is clustered using k-modes. In terms of performance, the k-prototypes technique is equivalent to the k-means algorithm when applied to numeric data. We were able to demonstrate how numeric and categorical variables interact with one another throughout the clustering process, as well as uncover additional issues that may be utilized to augment the anomaly research, using this method.

RESULTS

The difficulty in identifying an organization that is likely to participate in money laundering or terrorist financing is exacerbated by a lack of knowledge about its transactional behavior. The data for the study was gathered from law firms and individual businesses who had been involved in money laundering in the preceding 2 years. We need to develop a model that can predict the possibility of money laundering activities taking place in the absence of correct transaction information. The modeling technique is implemented using the Python 3.6 programming language and the Pilchard development environment [15]. To cope with big data, the Hadoop and Apache Spark platforms, which allow for distributed processing of enormous data sets, were used in the problem-solving process. A comparison of pre-processing approaches, machine learning algorithms for classification problems, and hyper parameter selection methods was carried out in accordance with the plan as shown in Figure 7.

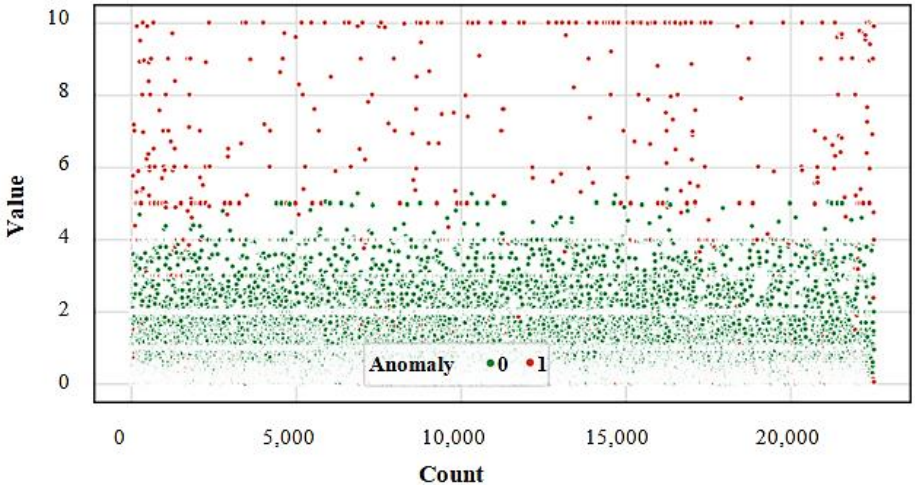


Figure 6. One-Class SVM dispersion diagram for withdrawal dataset.

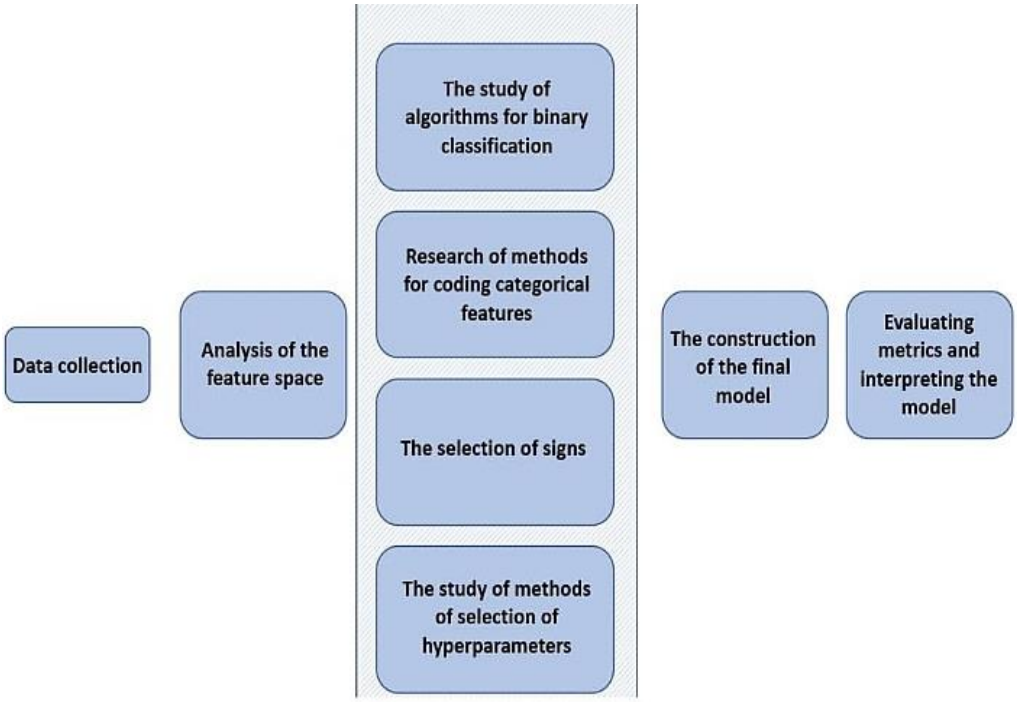


Figure 7. Extended study scheme.

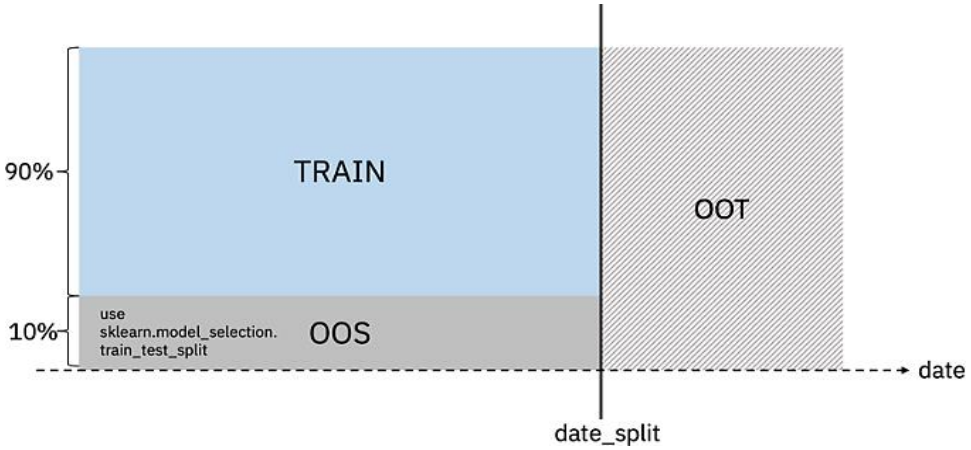
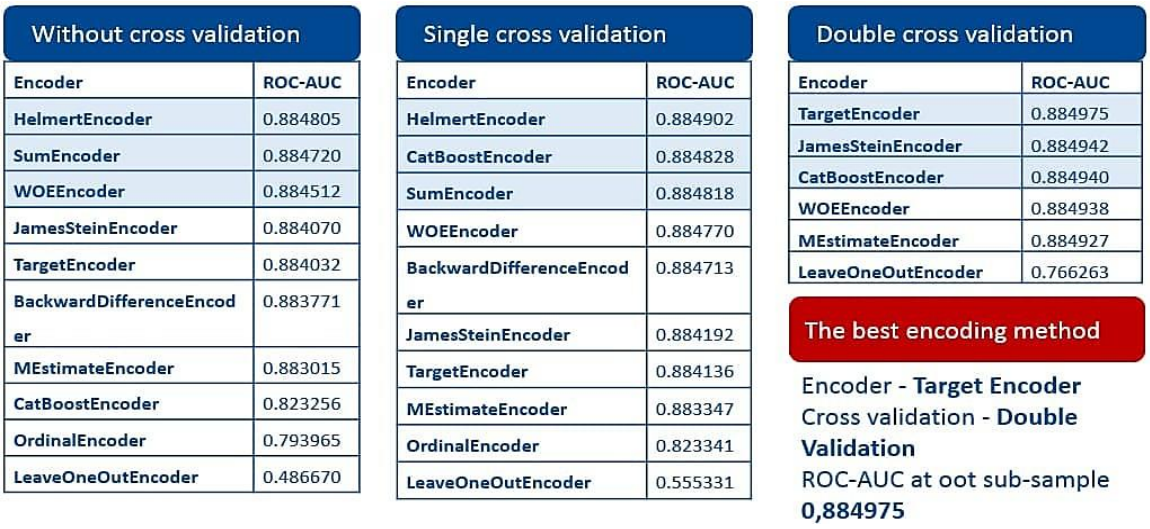


Figure 8. Splitting a dataset into sub samples.

The following feature space was built in order to identify organisations that are vulnerable to ML/TF: There are 42 quantitative and qualitative features offered, divided into four categories: general information, accounting data, data on public procurement participation, and legal information. The original data set was separated into three categories: We can use out-of-sample (OOS) and out-of-time (OOT) data to provide a final assessment of the performance of the machine learning model after it has been trained and validated. The original sample is divided into pieces depending on the date of the goal event to form the OOT sample (the fact of ML) as shown in Figure 8.

The OOT subsample was utilized to evaluate the algorithms' overall quality. The researchers devised the techniques employed in the study, which included logistic regression, random forest, and gradient boosting over solver trees. The ROC-AUC measure was used to train each model using k-fold cross-validation and then to evaluate its performance to determine model quality. The gradient boosting model outperformed logistic regression and random forest when it came to predicting the result. The categorical variable processing approaches described above were tested as part of a study on the impact of categorical feature encoding on categorical variable processing performance as shown in Figure 9.



Without cross validation

Encoder	ROC-AUC
HelmertEncoder	0.884805
SumEncoder	0.884720
WOEEncoder	0.884512
JamesSteinEncoder	0.884070
TargetEncoder	0.884032
BackwardDifferenceEncoder	0.883771
MEstimateEncoder	0.883015
CatBoostEncoder	0.823256
OrdinalEncoder	0.793965
LeaveOneOutEncoder	0.486670

Single cross validation

Encoder	ROC-AUC
HelmertEncoder	0.884902
CatBoostEncoder	0.884828
SumEncoder	0.884818
WOEEncoder	0.884770
BackwardDifferenceEncoder	0.884713
JamesSteinEncoder	0.884192
TargetEncoder	0.884136
MEstimateEncoder	0.883347
OrdinalEncoder	0.823341
LeaveOneOutEncoder	0.555331

Double cross validation

Encoder	ROC-AUC
TargetEncoder	0.884975
JamesSteinEncoder	0.884942
CatBoostEncoder	0.884940
WOEEncoder	0.884938
MEstimateEncoder	0.884927
LeaveOneOutEncoder	0.766263

Figure 9. The results of the use of encoders.

Table 2. Shows the final model metrics for train, out-of-sample, and out-of-time sub samples (Final models metrics).

Sub sample	ROC-AUC	Precision	Recall
Train	0,8873	0,7736	0.2452
OOS	0.8832	0,7731	0,2487
OOT	0.8849	0.6730	0,32

DISCUSSION

During the training phase, the final model was trained using the tactics that had showed the most effectiveness in the earlier rounds of the study. The following are the most important criteria in the model: the organization's kind of activity, its age, the amount of permissible capital, the founders' composition, the amount of taxes paid, the amount of profit, and so on as shown in Table 2.

The key characteristics of the final model are:

- Training on five learning sample blocks.
- Categorical features are encoded with Target Encoder with Double Validation.
- Forward feature selection.
- Gradient boosting over decision trees (Light GBM implementation).
- SMBO using the TPE surrogate function in the optima implementation for hyper parameter search.

This approach may be used to guarantee that requirements are satisfied as part of the compliance control process. A regulator, for example, may conduct an assessment of all companies registered in the Russian Federative. The calculated risk value may be presented to banks on a daily basis, and these ideas must be taken into account when dealing with enterprises. Alternatively, the model might be used by the bank to assess the risk of new customers participating in money laundering and terrorism funding. Credit-related businesses may use internal customer information (transaction data) to enhance the quality of their models and hence boost their level of control. Anomaly detection approaches, when combined with prior exploratory analysis, may immediately and consistently discover aberrant transactions in a transactional database, whereas the clustering methodology can find various behaviors in a dataset. Anomalies in bank transactions are not new, but their detection in the case of non-bank correspondents is limited because, in addition to providing a financial service in non-financial establishments, they have the distinct feature of developing in specific socioeconomic, cultural, and geographical environments. In these circumstances, it is vital to adopt guidelines and technological

solutions that would aid in the mitigation of the growing problem of money laundering. Algorithms also allowed for a more in-depth examination of several users' behaviors, which was especially relevant in the context of the use of specific financial instruments. In this case, we noticed that various user activities exceeded the maximum amounts permitted by the Tax Code, as well as an unusual situation: the use of a personal account every 35 h for the duration of the period, which is not normal in personal accounts. By adjusting hyper-parameters and the amount of available validations, we were able to improve the results of the anomaly detection algorithms as shown in Figure 10. However, in order to discover anomalies, we must include algorithm procedures that boost the tool's robustness and give more precise detection results for each cluster as shown in Table 3.

Table 3. Cluster statistical description for withdrawal database.

Cluster 0. Count 2,226.

Measure	Value	Month	Day month	Week day	Time
Min	2,500,000	1	1	0	8
25%	3,500,000	5	10	1	14
50%	4,000,000	7	15	2	15
75%	4,000,000	9	21	3	16
Max	9,999,999	12	31	6	19

Cluster 1. Count 3,768.

Min	1,600,000	1	1	2	7
25%	250,000	3	4	4	10
50%	600.000	5	8	4	11
75%	1,500,000	6	12	5	14
Max	4,390,000	10	19	6	18

Cluster 2. Count 3,203.

Min	3,000	1	11	0	8
25%	250,000	2	20	1	14
50%	630,000	3	25	2	15
75%	1,500,000	5	28	3	16
Max	4,000,000	7	31	5	18

Cluster 3. Count 4,016.

Min	3,500	1	1	0	8
25%	260,000	4	8	0	9
50%	734,000	7	13	1	10
75%	1,900,000	9	19	2	10
max	4,190,000	12	30	3	12

Cluster 4. Count 2,634.

Min	2,000	1	12	2	8
25%	230,000	5	21	3	9
50%	600,000	7	25	4	10
75%	1,500,000	10	28	5	11
Max	4,190,000	12	31	5	16

Cluster 5. Count 3,496.

Min	1,300	6	4	0	11
25%	210,000	8	16	1	14
50%	598,000	10	20	2	15
75%	1,527,500	11	25	4	16
Max	3,590,000	12	31	5	18

Cluster 6. Count 3,166.

Min	3,000	1	1	0	11
25%	240,000	3	4	1	15
50%	650,000	5	8	1	16
75%	1,535,000	7	12	2	16
Max	3,860,000	10	18	4	18

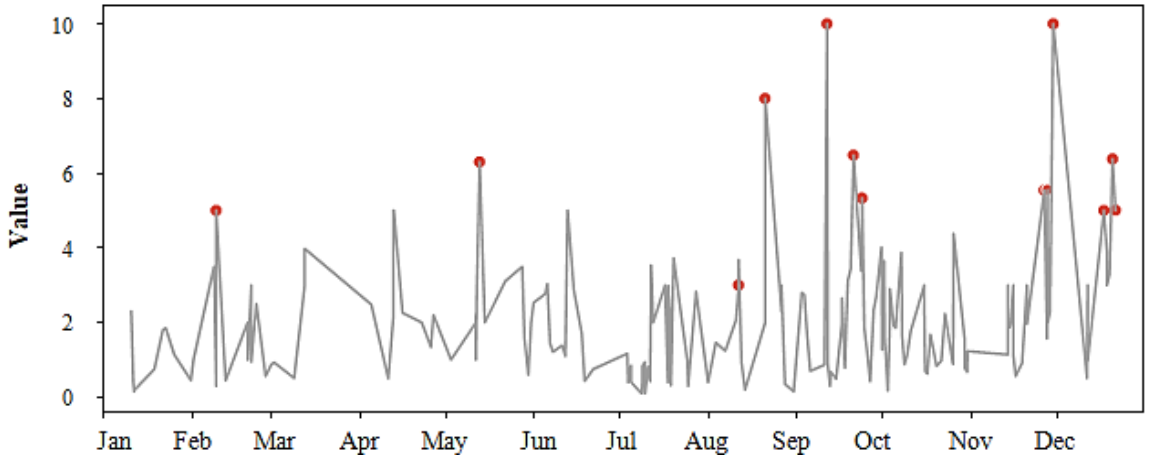


Figure 10. Anomalies example of a specific user.

When tracking the movements of a non-bank correspondent, on the other hand, a new challenge emerges, namely the vulnerability to robbery, which adds to money laundering. They are at danger as a consequence of an increase in the number of activities, making them more appealing targets for both common and organized crime. On a normal operation day, non-banking correspondents may have operations that are five to ten times larger than those of a typical retail outlet such as a store, gas station, or pharmacy, according to many statistical metrics. Non-banking correspondents may also be allowed to do business in the same way that a banking branch does [16–18]. Due to the ongoing evolution of criminal organizations engaged in money laundering and other illicit activities, non-banking correspondents must receive expert guidance from financial institutions regarding cash management, money laundering, and other risks. As a consequence, financial institutions must give competent advice on cash management, money laundering, and other hazards to non-banking correspondents. Such situations, in particular, may have an influence on benefits and result in a decrease in the number of non-banking correspondents.

CONCLUSION

This research looked at existing data pre-processing and machine learning technologies that have been utilized to handle the classification problem with the purpose of identifying organizations that are engaged in ML/TF early on. The data was analyzed using machine learning algorithms, and the initial feature space, consisting of 42 features, was developed and produced. At the level of data pre-processing, the influence of categorical features encoding was explored, and the need of cross-validation for categorical features pre-processing was shown. The usage of greedy direct selection was

used to choose informative attributes. Gradient boosting outperformed decision trees in terms of prediction quality when models were trained using a range of classification strategies.

Credit institutions will be able to identify customers who are at risk of money laundering at an early stage as a result of such modeling, allowing them to maintain a high level of anti-money laundering and counter-terrorist financing (AML/CFT) measures while preventing criminals from exploiting weak measures in individual banks. The goal of this work is to show how machine learning and visualization approaches may be used to spot irregularities in a sequence of transactions involving a non-banking correspondent (non-banking correspondent). This study addresses this issue by offering a fresh outlook on financial transactions within a burgeoning type of financial entity found in various countries, particularly in developing nations like Bangladesh, Brazil, China, Colombia, Ecuador, India, Mexico, Pakistan, and others. It also examines the mechanisms involved in money laundering across diverse socio-economic contexts.

In conclusion, our findings can aid in the development of monitoring policies, the development of preventive actions, and the reduction of money laundering in non-bank correspondents; however, it is critical to have the ongoing support of financial institutions and expert personnel in order to combat a growing situation.

We created and carefully tested a model for prioritizing which transactions should be evaluated further by AML investigators to decide which transactions should be probed further using machine learning. We demonstrated that the frequent practice of ignoring non-reported alerts/cases throughout the model training phase might lead to less-than-ideal results. If you have a bad habit of ignoring unreported warnings or instances, it is possible that this is because these alerts or cases are regarded insignificant in the course of your everyday responsibilities. Nevertheless, legal regulations mandate the monitoring of particular occurrences, and regular transactions are consistently accessible. The ideal data set, according to our results, comprises all data, including transactions associated with reported cases, non-reported alerts/cases, and ordinary transaction data. Hence, our findings should be relatively similar to what one would anticipate in a real-world operational environment. Finally, our method calculates the probability of a transaction requiring disclosure. Instead, the same technology may be used to predict the chance of a customer being reported to authorities with little effort.

REFERENCES

1. Lingeswari R, Brindha S. Identification of Vulnerability During Cross Border Transaction in IoT. J Harbin Eng Univ. 2023 Sep; 44(9). ISSN-1006-7043.
2. Jullum M, Løland A, Huseby RB, Ånonsen G, Lorentzen J. Detecting money laundering transactions with machine learning. J Money Laund Control. 2020 Jan 21; 23(1): 173–86.
3. Brindha S, Ajisha T. An Efficient Ensemble Classifier for Heart Disease Diagnosis and early Prediction. Int J Recent Technol Eng (IJRTE). 2020 Nov; 9(4): 109–114. ISSN: 2277-3878,
4. Rao AA, Kanchana V. Dynamic approach for detection of suspicious transactions in money laundering. Int J Eng Technol. 2018; 7(3): 10–3.
5. Lingeswari R, Brindha S. Validation of Blockchain transaction in banking sector. Humanities and Social Science studies (HASS). 2023 Jan–Jun; 12(1) NO 09. ISSN 2319-829X
6. Bayona-Rodríguez H. Money laundering in rural areas with illicit crops: empirical evidence for Colombia. Crime Law Soc Chang. 2019 Nov; 72(4): 387–417.
7. Ali MZ, Shabbir MN, Liang X, Zhang Y, Hu T. Machine learning-based fault diagnosis for single- and multi-faults in induction motors using measured stator currents and vibration signals. IEEE Trans Ind Appl. 2019 Jan 27; 55(3): 2378–91.
8. Brindha M, Sakthivel NK, Subasree S. Virtual Machine Dynamic Migration Strategy based Intelligent Flow Forecast Technique for Cloud Data Centers. Int J Emerg Technol Innov Res (IJETIR). 2019 Apr; 6(4): 368–76.

9. Colladon AF, Remondi E. Using social network analysis to prevent money laundering. Expert Syst Appl. 2017 Jan 1; 67: 49–58.
10. Lingeswari R, Brindha S. Analysing and classification of Attacks in Financial Transactions using Machine Learning. International Conference on Innovative Technologies and their Applications in Higher Education-Science. 2022 Oct. ISBN No. 978-93-92042-31-7
11. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Advances in neural information processing systems 24. 2011.
12. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016 Aug 13; 785–794.
13. Alexandre C, Balsa J. Client profiling for an anti-money laundering system. arXiv preprint arXiv:1510.00878. 2015 Oct 3.
14. Bholowalia P, Kumar A. EBK-means: A clustering technique based on elbow method and k-means in WSN. Int J Comput Appl. 2014 Jan 1; 105(9): 17–24.
15. Lacoste A, Larochelle H, Laviolette F, Marchand M. Sequential model-based ensemble optimization. arXiv preprint arXiv:1402.0796. 2014 Feb 4.
16. Adankon MM, Cheriet M. Support vector machine. Encyclopedia of biometrics. Boston, MA: Springer; 2009; 1303–8.
17. Bolton RJ, Hand DJ. Statistical fraud detection: A review. Stat Sci. 2002 Aug; 17(3): 235–55.
18. Breiman L. Random Forests. Berkeley, CA: University of California; 2001 Jan. Available from: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

