A Proposed Paradigm Using Data Mining to Minimize Online Money Laundering

Shimaa Ouf [1*], Meram Ashraf [2], Mohamed Roushdy [3]
[1,2] Information Systems Department, Faculty of Commerce and Business Administration, Helwan University, Helwan, Egypt
[3] Faculty of Computers & Information Technology, Future University in Egypt New Cairo, Cairo, Egypt
E-mail: shimaaouf@commerce.helwan.edu.eg, Meram.ashraf21@commerce.helwan.edu.eg, Mohamed.Roushdy@fue.edu.eg
[*]Corresponding author

Keywords: online money laundering, data mining, clustering, classification, association

Received: April 21, 2024

Since the global financial crisis (GFC), banks have been compromised by various risks. One of the significant risks is online money laundering. It is the third-largest business in the world after currency exchange and the automotive industry. As technology has advanced, the methods of online money laundering have become more evasive. Banks' traditional methods cannot deal with online money laundering. The absence of contemporary anti-money laundering techniques has led to the rise of this criminal activity. As a result, the existing systems need effective technology to accommodate the development of online money laundering. Data mining is the technology that applies mathematical, statistical, and machine-learning techniques to extract patterns like malicious behavior of money launderers and gain information about whether the online transaction belongs to money laundering. Many research papers apply different data mining techniques to predict online money laundering. By analyzing these research papers, we found that authors focused on applying one or two data mining techniques to predict online money laundering but ignored combining many techniques like classification, clustering, and association to improve the accuracy of prediction and overcome the limitations of each of them. Therefore, this paper proposes and implements a paradigm (APPD-OML) based on classification, clustering, and association techniques to improve the accuracy of predicting and detecting online money laundering. The result of testing the proposed paradigm illustrates that the prediction and detection of online money laundering based on applying data mining techniques like classification, clustering, and association achieve a strong accuracy of 94%, f-measure of 95%, and AUC 95% which means that the proposed paradigm outperforms each technique used separately in predicting and detecting online money laundering and outperformed the other research that used data mining in this field.

Povzetek: V članku je predstavljena nova paradigma APPD-OML, ki združuje tehnike klasifikacije, združevanja in asociacije za izboljšanje napovedovanja in zaznavanja spletnega pranja denarja..

1 Introduction

The advancement of technology, along with the COVID-19 pandemic, has prompted businesses and the public to use digital technologies even more intensively. Online services, online payments, and the use of electronic payment systems have significantly increased [1].

Despite the many advantages of online transactions, including time savings, reduced sales queues, a decrease in the need for cash and checks, ease of management, 24-hour accessibility, etc. Online transactions remain a double-edged weapon because they are a fertile environment for malicious attackers and criminals to do what they want without any tracking or monitoring [2].

Many criminals use online transactions as a way of laundering their money [3]. Money laundering is the process of giving fraudulent (dirty) money a legitimate look. It is the process of changing or converting assets that came from a criminal source to assist the criminal who is conducting the crime [4].

Money laundering has a destructive economic effect and relationship to the financing of terrorism. It weakens the stability and integrity of financial institutions as well as national economic stability in a way that distorts global capital flows and deters foreign direct investment [4]. Due to the digital era, many criminals use the internet to turn monetary gains from illegal activities into "clean" funds, which is known as online money laundering [3].

Online money laundering is a sophisticated practice that is carried out via a variety of cyber media, such as e-commerce, e-banking, online gaming, online gambling, electronic money, and other cyber means [5].

Therefore, banks' existing approaches are insufficient for predicting and detecting such activity. As a result, the requirement for automated prediction and detection technology, such as data mining, became essential [3].

Data mining is a method for locating hidden information in datasets. It is the process of extracting and

identifying potential and valuable knowledge stored in huge amounts of data by utilizing statistical techniques, mathematics, artificial intelligence, and machine learning [6]. In recent years, researchers have focused a lot of attention on the role that data mining plays in predicting and detecting money laundering in general and online money laundering in particular. Many methods are investigated, proposed, and studied for providing anti-money laundering solutions [7]. For such reasons, the main goal of this paper is to present a paradigm based on data mining techniques to predict and detect online money laundering.

The rest of this paper is divided as follows: Section 2 presents the research methodology and related work related to the topic. In Section 3, the proposed paradigm is presented. Section 4 demonstrates the experiments and results. In Section 5, the evaluation is discussed. In Section 6, the discussion is presented and finally, Section 7 presents the conclusion and future research.

2 Research methodology & literature review

2.1 Research methodology

The goal of this paper is to minimize online money laundering by employing different data mining techniques that can predict and detect such criminal activity and assist bank experts in making the right decisions regarding clients who engage in such activity. All of which will strengthen the viability of financial institutions and boost national economies. Therefore, a search was conducted through the most popular databases, Science Direct, Springer, and IEEE, to find articles that were relevant to the paper's goal. Only articles from academic journals released between 2018 and 2024 have been considered.

The search strategy was performed based on the main terms of the paper's goal to ensure that the material employed was related to it and served as a solid basis for our paper. Thus, the search terms that were utilized were "anti-money laundering and data mining", "data mining and online money laundering," "clustering and anti-money laundering," "classification and anti-money laundering," and "association and anti-money laundering." As illustrated in Figure 1.

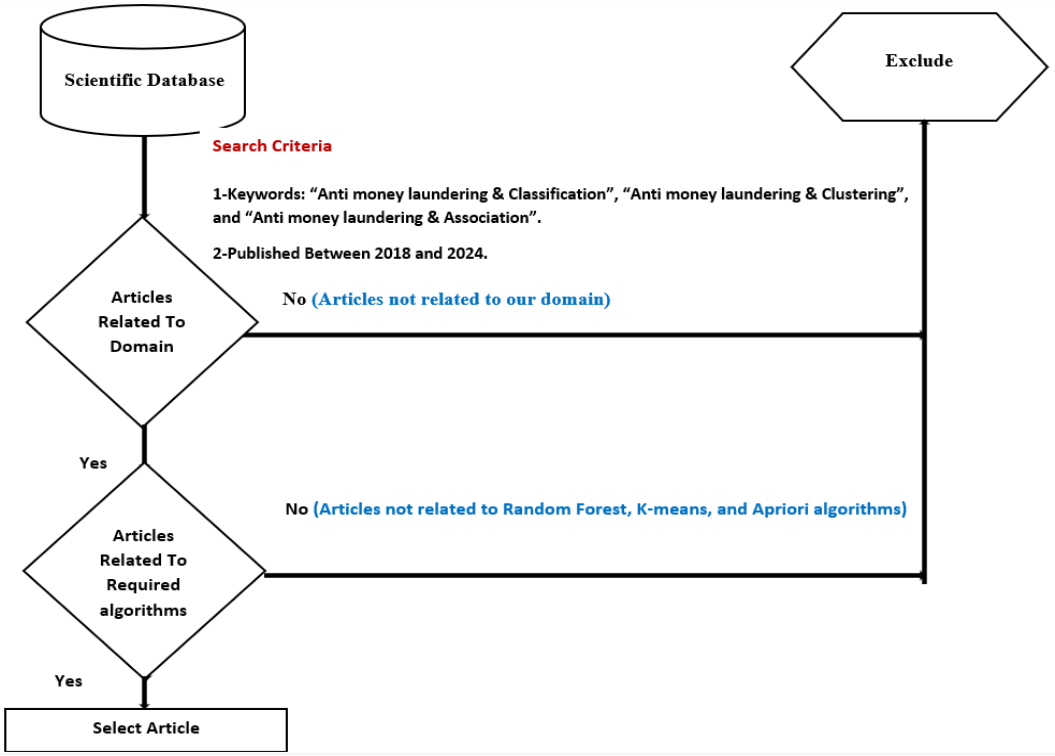

Figure 1: Research methodology.

The search yielded 123 articles. The search results are displayed according to the year of publication in Figure 2.

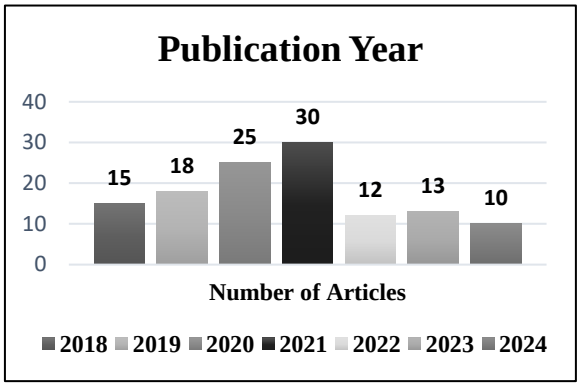

Figure 2: Classification of papers by year of publication.

2.2 Literature review

2.2.1 Online money laundering

The numerous obstacles that banks face today have an impact on their ability to compete, how they distribute their resources, and how they usually develop their strategic objectives. Online money laundering is one of these difficulties. A plethora of scholarly literature exists regarding online money laundering and the several approaches that mitigate this type of illicit activity.

[8] Provided a comprehensive review of data mining algorithms and methods applied to identify suspicious transactions. And introduced methods for anti-money laundering like link analysis, behavioral modeling, risk scoring, anomaly detection, and geographic capability. Also discussed were the essential steps of data preparation, data transformation, and data mining techniques.

[9] Discussed how data mining methods like association, clustering, and classification may be used to recognize and avoid fraud in the banking industry. The findings showed that a wide range of algorithms may be applied to fraud detection. And each of which has benefits as well as drawbacks.

[10] Demonstrated how link analysis may be applied to detecting suspicious bank transactions. And proposed a framework that includes four phases: task analysis, system design, implementation, and testing for detecting money laundering.

[4] Provided an overview of online money laundering, identifying the various anti-money laundering methods and their efficacy, the impact of online money laundering on other industries and the economy, the new avenues for online money laundering, and the detection of money laundering. The results showed that most of the research concentrated on detecting instances of online money laundering through the application of technology such as machine learning and data mining.

[11] Investigated how data mining techniques enable banks to predict and detect online money laundering. The findings indicated that due to the unavailability of high-quality datasets regarding online money laundering, there is limited scope for using supervised learning and ignorance of the key roles of unsupervised learning in detecting online money laundering.

[1] Provided a comprehensive analysis of the most recent studies on financial fraud detection, spanning the years 2009 to 2019, and categorized according to the forms of fraud and data mining techniques used. A sample of 75 pertinent papers that fell into four primary categories—bank fraud, insurance fraud, financial statement fraud, and cryptocurrency fraud—was produced because of the review. The results showed that many data mining techniques were applied in different financial applications to detect fraud.

[12] Identified data mining techniques, such as clustering, classification, and association, for financial fraud detection (FFD) and anti-money laundering (AML). And highlight how they help anticipate and identify online money laundering.

[13] Explained the data mining phases—data extraction, data cleaning and integration, data transformation, pattern discovery, pattern analysis, and information presentation—that were used to predict and detect online money laundering. Additionally, it covered numerous methods and techniques that can be applied at every phase.

[14] Proposed a system for detecting online payment fraud that makes use of data mining algorithms including random forests, logistic regression, and decision trees. The finding shows that the suggested system helps large organizations analyze the vast volume of transaction data to look for anomalies or potential fraud.

[15] Proposed a classification-based automated secure computing system for fraud detection. According to the results, the system helped process and analyze transaction data securely, look for anomalies, and notify users of potentially fraudulent activity.

2.2.2 Data mining techniques used to minimize money laundering

Analyzing and mining clients' online transactions is a crucial stage for predicting and detecting online transactions related to online money laundering that requires reliability and time. As a result, a significant number of studies were done on the best ways to produce accurate results using various data mining techniques.

[16] Illustrated and applied the supervised-based classification algorithms that predict and detect fraudulent online transactions, like naïve Bayes, logistic regression, and decision tree classifiers. The results showed that decision trees outperformed the other classifiers with an accuracy of 92%.

[17] Applied support vector machines to build a model that classifies the usual and unusual behavior that clients exhibit in online bank transactions. The findings demonstrate the effectiveness of the method in predicting banking fraud. Additionally, it demonstrates that it achieves better results when the splitting ratio is 80:20.

[18] Proposed a semi-supervised hybrid model to identify suspicious financial transactions and fraud that is based on association rule mining and clustering methods. The normal behavior patterns of the clients are extracted using the fuzzy clustering approach. Transactions that are

abnormal and do not fit into any of these clusters will be regarded as high-risk. Combining the outcomes of association rules derived from the apriori and clustering patterns allowed for the ultimate comprehension of a transaction. The results show that more frauds are detected faster when both rule-based and clustering-based components are included.

[19] Employed naïve Bayes, logistic regression, support vector machines, and k-nearest neighbors to estimate the likelihood of fraud in online banking transactions. The findings indicated that the support vector machine outperformed the other algorithms with an accuracy of 91%.

[20] Outlined the various clustering approaches, such as partitioned clustering and hierarchical clustering, that can be used to determine if a transaction is related to money laundering or not. It demonstrated how several algorithms for clustering, like sequential clustering, simultaneous clustering, divisive clustering (top-down), and agglomerative (bottom-up) clustering, can identify instances of money laundering. The results show that because simultaneous clustering is more informative, has fewer parameters, is adaptable, and can successfully integrate row and column information, it is a good fit for high-level data analysis of suspected money laundering operations.

[21] Applied frequent pattern mining to detect money laundering, the results showed that, despite the time it requires, it is very important and should not be neglected.

[22] Provided a comprehensive review of studies on the application of data mining techniques in the battle against money laundering. The results of this review demonstrated that academics were more focused on identifying suspicious transactions and that data mining techniques such as clustering and classification were applied more widely to identify money laundering.

[23] Analyzed the studies conducted on the issue of fraud detection focused on discovering indications of money laundering by applying data mining techniques like clustering and classification. The findings showed how useful algorithms for data mining are for identifying anomalous behavior and patterns of money laundering.

[24] Provided an overview of the application of classification methods like logistic regression, random forests, and support vector machines to identify online money laundering under two sampling techniques: under- and over-sampling. The effectiveness of the various approaches was evaluated using real transaction data that was supplied by a financial institution in the United States. The results showed that Random Forest outperformed the other methods.

[24] Utilized the naive Bayes classifier to predict money laundering in Bangladesh. The outcomes show that naive Bayes is a straightforward supervised learning algorithm that can classify large data sets with excellent accuracy and mobility. It is also fast, accurate, and dependable.

[26] Discussed the benefits of several supervised and unsupervised methods for anticipating and identifying online money laundering. Additionally, it demonstrated the various classification techniques and how they are used to predict suspicious transactions, including logistic regression, decision trees, random forests, and support vector machines. It also discusses how the apriori algorithm works to detect occurrences of online money laundering.

[27] Outlined the function of k-means clustering in spotting dubious clients. The findings show that it was successful in identifying clients in bank fraud situations like fishing, vishing, and skimming.

[28] Proposed an anti-money laundering system that uses the Apriori algorithm to determine the laundered money's traversal path. The findings demonstrate that the Apriori algorithm was successful in identifying odd patterns that might point to the possibility of fraud by utilizing location scanning and user behavior.

[29] Utilized historical data along with logistic regression to estimate the likelihood of money laundering. The customer's level of risk is considered the dependent variable (high or low), and the contracted product variables (legal entity, origin, economic activity, seniority) are considered the predictors. The findings indicate that logistic regression achieved an accuracy of 89%.

[30] Applied two classification algorithms, RF and SVM, to detect and predict money laundering. The results show that the RF classifier outperforms the SVM.

[31] Utilized the gradient boosting method to classify online transactions into one of the following classes: legitimate transactions that follow the law, transactions that are marked as suspicious by a built-in alert system, and possible money laundering cases that are reported to the authorities. The results show that gradient boosting achieved an AUC of 90%.

[32] Applied decision tree and support vector machine classifiers to forecast online money laundering activity. The Python programming language has been used for the analysis. The outcome demonstrated that in terms of accuracy, recall, and precision, the decision tree performed better than the support vector machine.

[33] Applied four classification algorithms—KNN, RF, LR, and SVM—to detect and predict whether the transaction was fraudulent or not. The results show that the RF classifier outperforms the other classifiers with an AUC of 93.92%.

[7] Introduced a prediction model based on naive Bayes, logistic regression, and random forest classifiers to determine whether money laundering is likely to occur in banks. The data were derived from a simulation of actual transactions at Middle Eastern banks that were suspected of money laundering. The results indicate that the random forest classifier was found to be the best at predicting transactions related to money laundering in the bank sector.

[34] Proposed a model that uses the gradient boosting algorithm to detect and predict the existence of abnormal behavior in online transactions in the financial sector. The results indicate that the model achieved an accuracy of 93% in predicting abnormal behavior.

[35] Proposed a model that classifies online transactions as fraud or not based on DT and KNN classification methods. The results indicate that KNN outperformed DT with an accuracy of 90%.

[36] Proposed a model based on the KNN algorithm to combat online money laundering. The results indicate that the model achieved an accuracy of 93% in detecting money laundering.

[37] Demonstrated how well-supervised machine learning algorithms work for spotting money laundering activity in financial institutions. K-nearest neighbors, gradient boosting, and random forest were applied. The accuracy, precision, recall, fi-measure, and area under the curve of each classifier were evaluated to test the performance of each classifier. The findings demonstrate that the Random Forest classifier produced the best result with 93% accuracy.

Table 1 summarizes researchers' work in the domain of money laundering and illustrates the different data mining techniques used, dataset characteristics, performance metrics, and key findings.

Table 1: A summary of researchers' work in the field of money laundering.

Authors	publication year	Techniques used	Dataset characteristics	Performance metrics	Key findings
[16]	2018	Classification	A dummy dataset was used to evaluate the integrity of the data for online fraud detection.	Accuracy "92%"	A decision tree algorithm is more effective in predicting fraudulent online transactions than naive Bayes and logistic regression.
[17]	2018	Classification	The German and Australian databases of credit card datasets were used.	Precision "80%"	A support vector machine is effective in predicting usual and unusual behavior in online bank transactions under a splitting ratio of 80:20.
[18]	2018	Clustering Association	An Iranian bank dataset that includes card data from February 2015 to January 2016 was used. Each transaction has 12 raw attributes which are transaction id, time, account number, card number, transaction type, entry mode, amount, merchant code, merchant group, gender, age, and bank.	AUC "85%"	More frauds are detected faster when both rule-based and clustering-based components are included.
[19]	2019	Classification	The fraud transactions log file and the all-transactions log file were the two data sources that were combined to form the dataset. All instances of online credit card fraud are contained in the fraud transactions log file, and all transactions recorded by the relevant bank within a given time frame are contained in all transactions log files.	Accuracy "91%"	A support vector machine algorithm is more effective in predicting fraudulent online transactions than naive Bayes, logistic regression, and k-nearest neighbors.
[20]	2019	Clustering	-	-	Compare different types of clustering and their role in identifying money laundering.
[21]	2019	Association	A bank transaction dataset is used to identify money laundering transactions.	Accuracy "91%"	Graph matching algorithm is faster than frequent mining
[22]	2019	Classification Clustering	-	-	Provide literature about the role of classification and clustering in predicting and detecting money laundering.

[23]	2019	Classification Clustering	-	-	Provide literature about the role of classification and clustering in predicting anomalous behavior
[24]	2019	Classification	A dataset containing roughly a hundred explanatory variables was used, based on rich information on customer/account characteristics, recent transactions, and alert management outcomes if applicable.	AUC "91%"	A random forest algorithm is more effective in predicting fraudulent online transactions than support vector machine, and logistic regression.
[25]	2020	Classification	A bank dataset that contains attributes like "job category", "net income", "open mode", "VT", "NT", "VCT", "NCT", and "Risk class" was used.	Accuracy "78%"	A naive Bayes algorithm is fast and dependable in classifying money laundering transactions.
[27]	2020	Classification Association	-	-	Provide literature about the role of classification and association in predicting and detecting money laundering.
[28]	2020	Clustering	-	-	Identify the role of the k-means algorithm in fraud detection in banks.
[29]	2020	Association	A credit card dataset was used.	Accuracy "88%"	Identify odd patterns that might point to the possibility of fraud by utilizing location scanning and user behavior.
[30]	2020	Classification	The customer's level of risk dataset that contains attributes of the dependent variable (high or low), legal entity, origin, economic activity, and seniority was used.	Accuracy "89%"	Identify the role of the logistic regression algorithm in fraud detection in banks.
[31]	2020	Classification	"Synthetic Financial" Datasets for Fraud Detection were used. The synthetic dataset consists of several features, including the type of transaction: CASH_IN, CASH_OUT, DEBIT, PAYMENT, and TRANSFER.	Accuracy "93%"	A random forest algorithm is more effective in predicting fraudulent online transactions than a support vector machine.
[32]	2021	Classification	An alerted transactions dataset spanning from 1 April 2014 to 31 December 2016 was used.	AUC "90%"	Identify the role of gradient boosting algorithm in predicting money laundering.
[33]	2022	Classification	A dataset downloaded from the UCI Machine Learning repository was used. This dataset contains a total of 699 instances and each instance consists of 11 attributes.	Accuracy "92%" Precision "94%" Recall "94%"	A decision tree algorithm is more effective in predicting fraudulent online transactions than a support vector machine.
[7]	2022	Classification	A bank's clients' online transaction dataset that includes 10,000 records of legitimate and fraudulent	AUC "93.92%"	A random forest algorithm is more effective in predicting fraudulent online transactions than support vector machines,

			online transactions made by clients between January 1 and December 31, 2020, was used.		K-nearest neighbors, and logistic regression.
[34]	2023	Classification	A simulated dataset of money-laundering activities in Middle Eastern banks based on a real dataset was used.	Accuracy "77%"	A random forest algorithm is more effective in predicting money laundering transactions than naïve Bayes and logistic regression.
[35]	2023	Classification	A credit card dataset was used.	Accuracy "93%"	Identify the role of gradient boosting algorithm in predicting money laundering.
[36]	2023	Classification	A banking fraud transactions dataset was used.	Accuracy "90%"	A K-nearest neighbor is more effective in predicting money laundering transactions than a decision tree.
[26]	2023	Classification	A money laundering transactions dataset was used.	Accuracy "54%"	A Part is more effective in predicting money laundering transactions than ZerO, and OneR.
[37]	2023	Classification	The Saudi General Organization for Social Insurance dataset was used.	Accuracy "90%"	A random forest algorithm is more effective in predicting money laundering transactions than gradient boosting and K-nearest neighbors.
[14]	2024	Classification	A fraud transactions dataset was used.	Accuracy "93%"	A random forest algorithm is more effective in predicting money laundering transactions than logistic regression.
[15]	2024	Classification	A credit card transactions dataset was used.	Accuracy "93%"	Proposed classification-secure fraud detection system that analyzed data securely.

According to the related research, the majority of online money laundering prediction and detection tactics use many techniques to assist banks in identifying this illicit activity, but none of the studies provide all of them, as Table 2 illustrates.

Table 2 demonstrates, most researchers used two or more data mining approaches, but no one used them all to identify, minimize, and overcome the limitations of each technique. Therefore, the main goal of this paper is to propose and implement a paradigm based on data mining techniques like classification, clustering, and association to predict and detect criminals who engage in online money laundering and mitigate the detrimental effects of online money laundering on the country's economy. In addition, as illustrated in Table 2, the three algorithms that proved their power in predicting suspicious transactions are random forest for classification, k-means for clustering, and apriori algorithm for association. Therefore, the proposed paradigm will be based on these three algorithms as illustrated in the next section.

3 A proposed paradigm (appd-oml) for predicting and detecting online money laundering

To effectively predict and detect online money laundering, the anti-money laundering methods must be updated efficiently to address online money laundering, which increases daily. Therefore, a paradigm (APPD-OML) that is based on data mining is proposed, for restructuring the process of anti-money laundering. Using APPD-OML, the information provided enables banks to easily predict and detect suspicious transactions that are associated with online money laundering.

Table 2: A summary of the data mining techniques employed by some researchers.

Author	Classification	Clustering	Association
[16]	✓	-	-
[17]	✓	-	-
[18]	-	✓	✓
[19]	✓	-	-
[20]	-	✓	-
[21]	-	-	✓
[22]	✓	✓	-
[23]	✓	✓	-
[24]	✓	-	-
[25]	✓	-	-
[27]	✓	-	✓
[28]	-	✓	-
[29]	-	-	✓
[30]	✓	-	-
[31]	✓	-	-
[32]	✓	-	-
[33]	✓	-	-
[7]	✓	-	-
[34]	✓	-	-
[35]	✓	-	-
[36]	✓	-	-
[26]	✓	-	-
[37]	✓	-	-
[14]	✓	-	-
[15]	✓	-	-
The proposed paradigm	✓	✓	✓

Therefore, the architecture of the proposed paradigm involves three layers. The first layer is the data pre-processing layer; the second layer is the mining layer; and the third layer is the report generation layer. Figure 3 presents the layers of the proposed paradigm that will assist bank experts in recognizing instances of online money laundering that take place far from bank surveillance.

3.1 Layer 1 (data pre-processing layer)

Data pre-processing is the most important layer that prepares online transactions for further analysis. Online transactions are unstructured data that may contain missing values, duplicates, and irrelevant noise that are not suitable to be used in the mining layer. As a result, the data pre-processing layer is responsible for cleaning, integrating, converting, and minimizing data to prepare it for the mining layer. Data preprocessing is made up of three important phases: the data cleaning phase, the data transformation phase, and the data reduction phase. Each of these phases has a significant role in preparing online transactions to be accurately analyzed in the mining layer.

3.1.1 Data cleaning phase

This phase oversees identifying and resolving any data anomalies, including duplicates, and missing values. The two methods that are utilized to find such anomalies are finding records that are identical in all aspects and finding empty or Nan values. Anomalies are found and fixed using a variety of techniques. Two methods are used to fix missing values: imputation and deletion. The imputation approach is used to fill a categorical variable with the dummy variable "missing" when its percentage is greater than 50%; the mode is used to fill it when its percentage is less than 50%; and the deletion approach is used to remove transactions for which specific information was missing.

Conversely, duplicates are corrected by retaining the original and deleting all others. At the end of this phase, the data becomes cleansed, free of anomalies, and prepared for the data transformation phase.

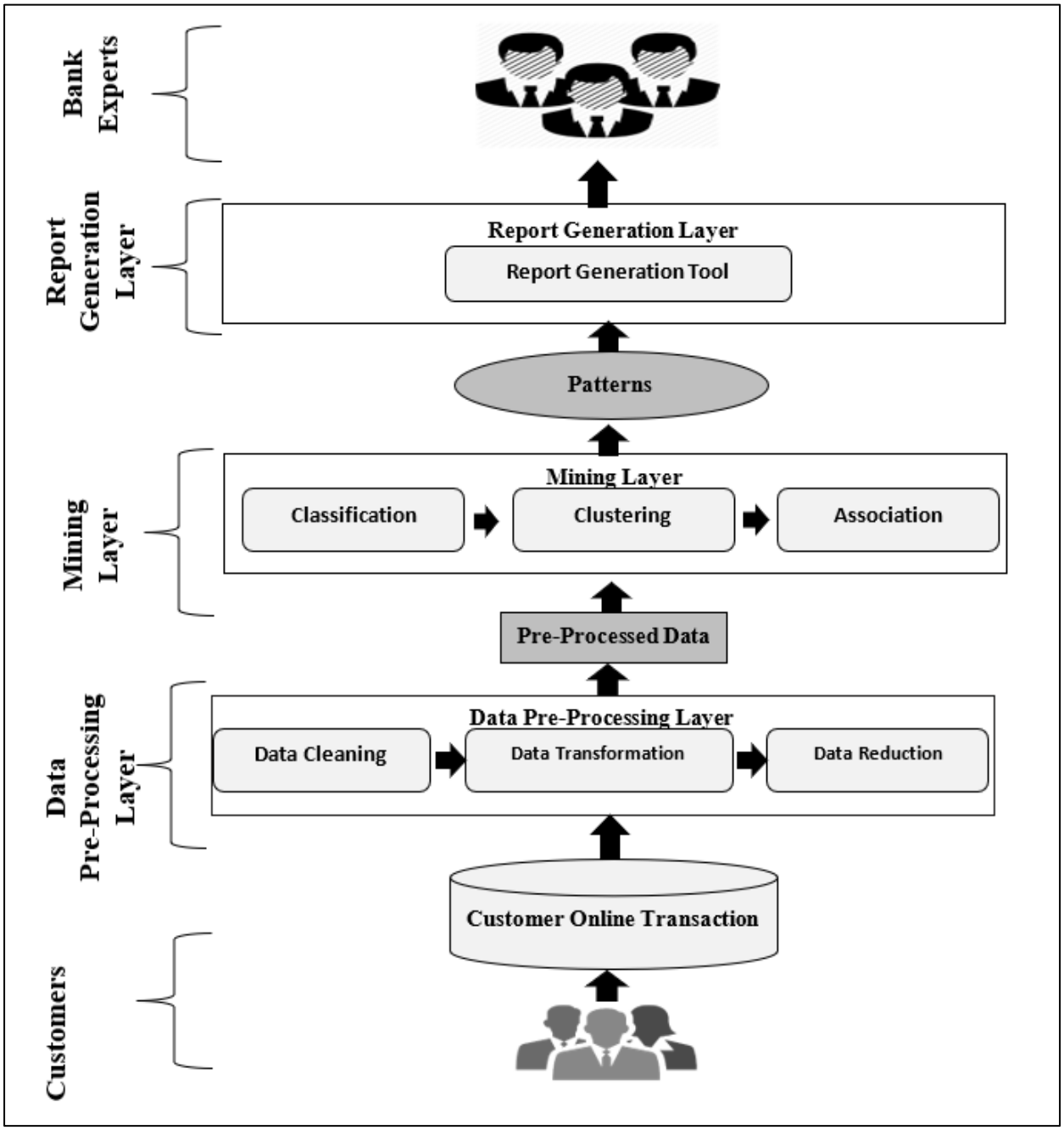

Figure 3: The proposed (APPD-OML) paradigm.

3.1.2 Data transformation phase

This phase is responsible for transforming the data into appropriate forms and generating new features that are suitable for the mining layer. It includes two stages: the first stage features construction, and the second stage is discretization.

- **The feature construction stage:** Is responsible for generating new features from the existing data that can identify clients' online behavior, the timestamps decomposition approach is employed to construct features such as, "trans_date", "trans_day", "trans_month", and "trans_hour" features derived from the "trans_date_trans_time" field. The "age" feature is derived from the "dob" field by using the subtraction method. The "distance" feature which contains the distance between the customer and the merchant derived from the "lat", "long", "merchant_lat", and "merchant_long" fields by using the calculation approach. From the "cc_num", "unix_time", and "amt" fields, the "history_30" feature is derived that contains the last 30-day spending of each credit card by applying the combination approach. From the "history_30" and "amt" fields, the "interaction_30" feature is derived by applying a combination, which contains the ratio of the current purchase price to the amount spent in the last 30 days.

- **The discretization stage:** this is responsible for dividing the range of values into intervals and placing each data point in the proper bin, the width binning approach in Equation 1 [38] is used.

$$W = \frac{\max - \min}{\text{Number of bins}} \quad (1)$$

Where
W represents the width binning that has an equal width.
max is the maximum value.
min is the minimum value.
Number of bins represents the total number of bins.

At the end of the data transformation phase, all the features that demonstrate client online behavior were created and became ready for the data reduction phase.

3.1.3 Data reduction phase

This phase is responsible for eliminating unnecessary data that didn't provide any information in the mining layer while preserving important data. This is done to improve the efficiency of data analysis and avoid overfitting the model. Furthermore, this phase is also responsible for converting data to a format that can suit data mining techniques. As a result, two approaches were used in this phase: feature selection and feature extraction.

- **Feature selection:** This is responsible for preserving only important features that are used in the mining layer and eliminating other features. In online transactions, many features exist that don't add any value for predicting whether the transaction is related to online money laundering or not. In addition, some features can be identified from another feature, such as a "city" that can be identified from the client's "lat" and "long" fields. So, the existence of such features can consume a lot of processing time and doesn't add value. As a result, the feature selection approach is used to select the most significant features that can be used in the predictive model. The correlation method in Equation 2 [39] is used to measure the correlation between features and target features. Only features with a high correlation will be kept, and the others will be eliminated.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]} \quad (2)$$

Where

r represents the Pearson correlation coefficient.

n is the number of values or elements.

$\sum x$ is the sum of the 1st values list.

$\sum y$ is the sum of the 2nd values list.

$\sum xy$ is the sum of the product of 1st and 2nd values.

$\sum x^2$ is the sum of squares of 1st values.
$\sum y^2$ is the sum of squares of 2nd values.

- **Feature extraction:** This is responsible for converting features such as "categories" and "clients ages" to numerical values to be used in the mining layer. The ordinal encoding method in Equation 3 [40] is employed for this task. In ordinal encoding, each distinct category is given an integer value [41]. After completing this layer, the online transactions become ready to be used in the mining layer.

$$X(m) => N(m) \quad (3)$$

Where
X represents the categorical value of feature m.
N represents the numerical value of feature m.

3.2 Layer 2 (mining layer)

This layer is responsible for applying the different data mining techniques to the preprocessed online transactions obtained from the data pre-processing layer. Three data mining techniques, classification, clustering, and association were employed to build the prediction model that helps bank experts in the prediction and detection of online money laundering-related transactions.

3.2.1 The role of the classification technique in prediction

This phase is responsible for building the classifier that will classify online transactions whether it is related to online money laundering or not, based on the correlation between the different features that exist in the data and the target feature. In this phase, the classifier is constructed using the random forest classifier in Equation 4 [42], and 80% of the online transactions are used to train it to produce accurate results.

$$G = 1 - \sum_{i=1}^{C} (p_i)^2 \quad (4)$$

Where
G represents the Gini index.
C is the total number of classes.
p(i) is the probability of picking the data point with the class *i*.

3.2.2 The role of the clustering technique in prediction

This phase is responsible for grouping online transactions based on the amount since there is a strong association between the amount and the fraud status. The clusters of the amount generated will be used as one of the inputs for the association technique. In this phase, the K-means algorithm in Equation 5 [43] is used to create the clusters.

(5)

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\left|\left|x_i - v_j\right|\right|)^2$$

Where
J(V) represents an objective function.
$||x_i - v_j||$ *is* the Euclidean distance between x_i and v_j.
c_i represents the number of data points in i^{th} cluster.
c represents the number of cluster centers.

3.2.3 The role of the association technique in prediction

This phase is responsible for discovering the relationship between features in the same transaction using the association technique. As the classification technique still has a percentage of error. It indicates that certain online money laundering-related transactions may have been incorrectly categorized, which poses a serious risk to the financial industry, and to the economy. As a result, the association rule was used to analyze the client's transaction and determine whether the transaction was fraudulent based on some features that are strongly correlated with the target variable, such as" categories", "amount", and "clients' age". In this phase, the apriori algorithm in Equations 6, 7, 8, and 9 [44] is used to create the rules.

$$\text{Support} = \frac{\text{Frequency}(x,y)}{N} \quad (6)$$

$$\text{Confidence} = \frac{\text{Frequency}(x,y)}{\text{Frequency}(x)} \quad (7)$$

$$\text{Lift} = \frac{\text{Support}}{\text{Support}(x) * \text{Support}(y)} \quad (8)$$

$$\text{Conviction} = \frac{1 - \text{Support}(y)}{1 - \text{Confidence}(x \geq y)} \quad (9)$$

Where
N represents the total number of transactions.
x represents rule antecedent.
y represents rule consequent.

After completing this layer, the necessary patterns and data that demonstrate whether the transaction is related to online money laundering or not become ready to be used in the report generation layer.

3.3 Layer 3 (report generation layer)

This layer is responsible for taking the patterns generated in the mining layer and generating reports for banks' experts. Two types of reports are generated, including informational reports that contain information about online money laundering transactions. And analytical reports that contain qualitative and quantitative information used in decision-making. By the completion of the proposed paradigm (APPD-OML), the prediction and detection processes used to identify money laundering have been modified to incorporate the ability to identify online money laundering. This fulfills the paper's goal of enhancing the prediction and detection of online money laundering, helping banks promptly identify such criminal activity, and helping governmental authorities take the appropriate action against them.

4 Experiments and results

To validate the performance of the proposed (APPD-OML) paradigm, the clients' data collected must be prepared to help banks predict and detect online money laundering and provide experts with the appropriate information that helps them make the right decision regarding these suspicious transactions.

4.1 Data acquisition

In this paper, a bank's clients' online transaction dataset [33] that includes records of legitimate and fraudulent online transactions made by clients between January 1 and December 31, 2020, was used. These data were labeled completely and are available for open research. Table 3 illustrates dataset features.

Table 3: Features description.

Data Fields	Description
trans_date_trans_time	Contains the time and date of the transaction.
cc_num	Contains the number of credit cards.
Merchant	Contain names of merchants.
Category	Contains the domain in which transactions were performed.
Amt	Contains the amount involved in the transaction
First	Contains the first name of the client who initiated the transaction.
Last	Contains the last name of the client who initiated the transaction.
Gender	Contains the gender of the client who initiated the transaction.
Street	Contains the street of the client who initiated the transaction.
City	Contains the city of the client who initiated the transaction.
State	Contains the state of the client who initiated the transaction.
Zip	Contains the zip code of the client who initiated the transaction.
Lat	Contains the latitude of the client who initiated the transaction.
Long	Contains the longitude of the client who initiated the transaction.
City_pop	Contains the number of the city's population

Job	Contains the job of the client who initiated the transaction.
Dob	Contains the birthdate of the client who initiated the transaction.
trans_num	Contains the transaction number.
unix_time	Contains the time delay between the previous and current transaction
merch_lat	Contains the latitude of the merchant where the transaction ended.
merch_long	Contains the longitude of the merchant where the transaction ended.
is_fraud	Contains whether the client is fraudulent or not.0 means he is not fraudulent and 1 is fraudulent.

As shown in Table 4, there are no missing values in the dataset.

Table 4: Missing values and their percentage.

Column name	Percentage of missing values
trans_date_trans_time	0%
cc_num	0%
Merchant	0%
Category	0%
Amt	0%
First	0%
Last	0%
Gender	0%
Street	0%
City	0%
state	0%
Zip	0%
Lat	0%
Long	0%
City_pop	0%
Job	0%
Dob	0%
trans_num	0%
unix_time	0%
merch_lat	0%
merch_long	0%
is_fraud	0%

As previously mentioned, the (APPD-OML) has 3 layers. The data preprocessing layer is the first layer executed after data acquisition and is the most complex and time-consuming. Python, a high-level, object-oriented, and interpreter-based multipurpose programming language that gives its programmers access to a wealth of resources [45], was used to implement the proposed paradigm.

4.2 Data pre-processing layer (layer 1)

This layer is responsible for the preprocessing of the data to obtain appropriate mining in layer 2. As a result, the goal of this layer is to get the input data from Table 3 ready for the following analysis. In this layer, the following phases were applied:

4.2.1 Data cleaning

Detecting and fixing anomalies in data is one of the major difficulties in data preprocessing. Anomalies such as missing values and duplicates can lead to incorrect decisions and unreliable analysis [46]. Therefore, this phase is responsible for finding and fixing such anomalies.

- **Detecting and handling missing values**

Missing values occur when no value is recorded for a particular field in an observation [47]. The percentage of missing values in each field is shown in Table 4.

- **Detecting and handling duplicates**

Duplicates refer to records that contain identical data [48]. As shown in Table 5, there are no duplicates in the data.

Table 5. Duplicate Values in Each Record.

No. of record	Duplicate values
0	False
1	False
2	False
3	False
4	False
…..	…..
9995	False
9996	False
9997	False
9998	False
9999	False

Upon completion of this phase, it was verified that there were no duplicates or missing values in the data. It's prepared to undergo the next phase, which is data transformation.

4.2.2 Data transformation

This phase is responsible for transforming the data into appropriate forms and constructing new features that are suitable for the data mining analysis that will be applied in the mining layer.

- **Feature construction**

This stage is responsible for deriving data from the raw data to transform it into information that can be used in the prediction model, the timestamps decomposition approach, the subtraction method, the calculation approach, and the combination approach are employed to

construct the following features that were used in the prediction and detection of online money laundering:

- From the "trans_date_trans_time" field, four important features, "trans_date", "trans_day", "trans_month", and "trans_hour" were derived.
- From the "dob" field, the "age" feature was derived to understand the relationship between age and whether the client is fraudulent or not.
- From the "lat", "long", "merchant_lat", and "merchant_long" fields, the "distance" feature was calculated, which contains the distance between the client and merchant.
- From the "cc_num", "unix_time", and "amt" fields, the "history_30" feature was derived that contains the last 30-day spending of each credit card.

From the "history_30" and "amt" fields, the "interaction_30" feature was derived, which contains the ratio of the current purchase price to the amount spent in the last 30 days. Table 6 illustrates the output of this stage.

Table 6. The results of the feature construction stage.

category	amt	gender	is_ fraud	trans_dat e	trans_da y	trans_ month	trans_hour	age	distance	history _30	interaction _30
misc_net	4.97	F	1	01/01/20 19	Tuesday	Januar y	0	33	0.87283	3	0
grocery_ pos	107. 23	F	0	21/02/20 19	Tuesday	Februa ry	1	45	0.27231	0	10
entertain ment	220. 11	M	0	01/01/20 19	Tuesday	Januar y	0	53	0.97584 5	1	0
gas_trans port	45	M	1	01/01/20 19	Tuesday	Januar y	0	56	0.91980 2	0	0
misc_pos	41.9 6	M	0	13/05/20 19	Sunday	May	1	39	0.86850 5	0	0

- **Discretization**

This stage is responsible for splitting continuous data into discrete intervals to format the data for further data mining analysis. In this paper, the width binning method in Equation 1 was applied to group "cust_age" into six intervals which are '10-20', '20-30', '30-40', '40-50', '50-60', and '60 and above'. Table 7 illustrates the results.

Table 7: The results of binning the customers' ages.

category	amt	gender	is_ fraud	trans_date	trans _day	trans_ month	trans_hour	cust _ag e	distance	history _30	interactio n_30
misc_net	4.97	F	1	01/01/2019	Tues day	January	0	30- 40	0.87283	3	0
grocery_ pos	107. 23	F	0	21/02/2019	Tues day	Februar y	1	40- 50	0.27231	0	10
entertain ment	220. 11	M	0	01/01/2019	Tues day	January	0	50- 60	0.97584 5	1	0
gas_trans port	45	M	1	01/01/2019	Tues day	January	0	50- 60	0.91980 2	0	0
misc_pos	41.9 6	M	0	13/05/2019	Sund ay	May	1	30- 40	0.86850 5	0	0

After completing the data transformation phase, the unstructured raw data was transformed into the structured format needed for the following phases. Additionally, the raw data was utilized for constructing the key features that the prediction paradigm will employ to achieve the paper's goal.

4.2.3 Data reduction

This phase is responsible for shrinking the size of the dataset while maintaining the key information. This is done to increase the effectiveness of the data mining processes and prevent the model from overfitting.

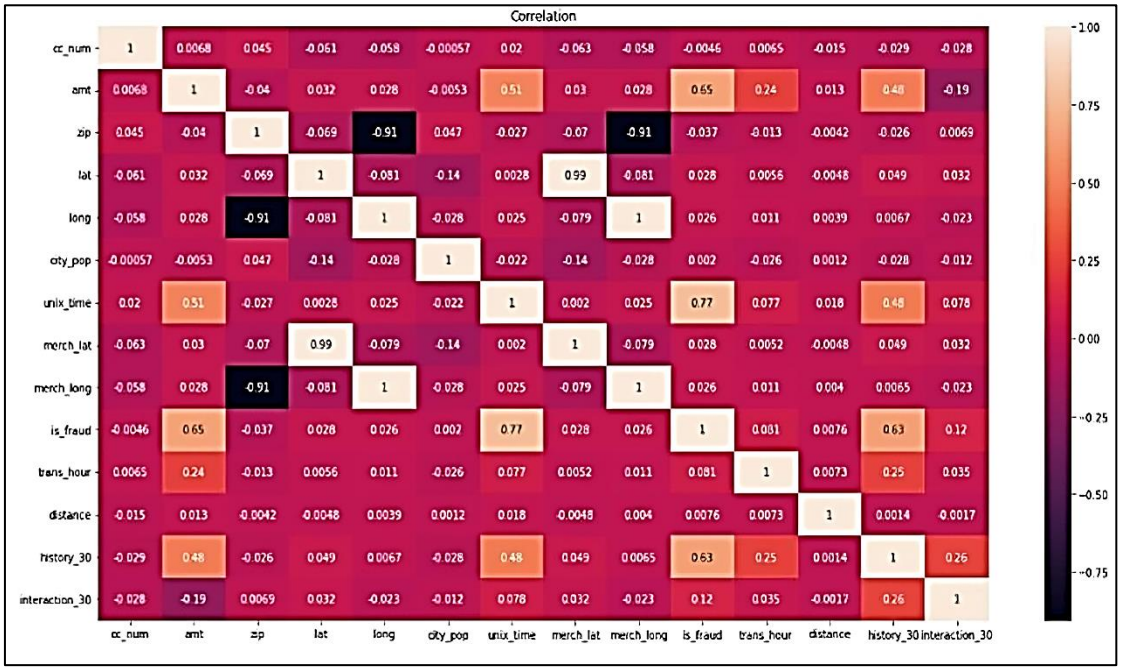

Figure 4: The correlation heatmap.

- **Feature selection**

This stage is responsible for selecting the most relevant subset of features out of the original features to be used as inputs for the mining layer. And since the utilized dataset was supervised, the filtering technique that is based on Pearson correlation in Equation 2 was used to filter the features and measure their correlation to the target feature "is_fraud". Figure 4 presents a correlation heatmap that illustrates the correlation between the variable.

As illustrated in Figure 4, there is a multicollinearity issue that arises when two features have a correlation of more than 0.7 that must be eliminated. Furthermore, features that consumed a significant processing time and did not provide any new values to the prediction were removed. Upon the completion of this stage, as indicated in Table 8, only the features utilized for prediction remained.

Table 8: The predictive features.

category	amt	city_ pop	is_fraud	Trans _date	Trans _day	Trans _month	Trans _hour	cust_age	distance	History _30	Interactio n _30
misc_net	4.97	3495	1	01/01/2019	Tuesday	January	0	30-40	0.87283	3	0
grocery_p os	107.23	149	0	21/02/2019	Tuesday	February	1	40-50	0.27231	0	10
entertain ment	220.11	4154	0	01/01/2019	Tuesday	January	0	50-60	0.975845	1	0
gas_trans port	45	1939	1	01/01/2019	Tuesday	January	0	50-60	0.919802	0	0
misc_pos	41.96	99	0	13/05/2019	Sunday	May	1	30-40	0.868505	0	0

- **Feature extraction**

This stage is responsible for transforming categorical data into numerical features. The ordinal encoding method in Equation 3 was used in this stage to encode features such as "category" and "cust_age".

At the end of the data preprocessing layer, the data became ready for the different data mining techniques that would be applied in the mining layer.

4.3 The mining layer (layer 2)

In this layer, different data mining techniques were applied to the preprocessed data to accomplish the paper's

goal. The classification, clustering, and association techniques were used in this layer to predict and detect online money laundering.

4.3.1 Classification technique

This phase is responsible for classifying online transactions to determine if a transaction is fraudulent or not. Or to put it another way, whether it is related to online money laundering or not. In this layer, the following stages were applied:

- **Data splitting**

In this stage, an "80:20" spitting ratio was set for the classification model.

- **Build and train the model**

In this stage, the Random Forest (RF) algorithm in Equation 4 was used and was trained with 80% training data.

- **Test the model**

In this stage, the classification model was tested. The results are shown in Table 9.

Table 9: Results of the classification model.

Algorithm	Performance Measures			
	Accuracy	*Precision*	*Recall*	*F1*
RF	93%	92%	91%	91%

The results demonstrated that RF classifies online transactions with an accuracy of 93%, and an F1 measure of 91%. However, about 7% of transactions may be categorized as legitimate but include fraud. For this reason, the clustering and association techniques were applied to predict suspicious transactions based on specific features highly correlated with fraud.

4.3.2 Clustering technique

In this phase, the k-means algorithm in Equation 5 was utilized to cluster transactions according to amount since there is a high correlation between the amount and status of fraud, as illustrated in Figure 4. The elbow method was used to determine the appropriate number of clusters. As shown in Figure 5.

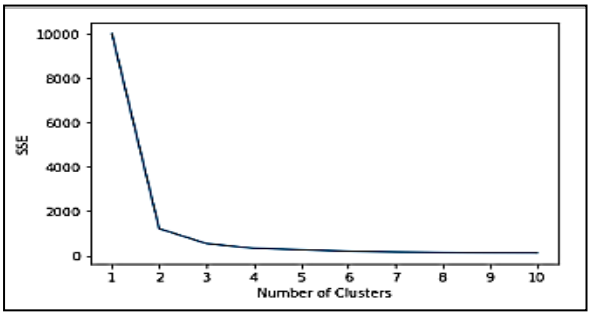

Figure 5: The results of the elbow method.

As illustrated in Figure 5, the appropriate number of clusters was two. Thus, two clusters, the "high amount" and the "low amount" were generated.

4.3.3 Association technique

In this phase, the Apriori algorithm in Equation 6 was utilized to analyze the customer's transaction and determine whether it was fraudulent based on "category", "amount", and "cust_age" features that have a high correlation to the "is_fraud" field as demonstrated in Figure 4.

A sample of the rules that were generated is illustrated in Table 10 and only rules with high confidence (>0.3) were chosen to predict whether the transaction was risky or not. The confidence of each rule represents its accuracy.

Table 10: A Sample of the generated association rules.

antecedents	consequents	antecedent support	consequent support	support	confidence	Lift	leverage	convictio n
(misc_net,30-40,low_amount)	(not_fraud)	0.137856	0.245077	0.074398	0.539683	2.202098	0.040613	1.640006
(grocery_pos,40-50,low_amount)	(not_fraud)	0.245077	0.137856	0.074398	0.303571	2.202098	0.040613	1.237951
(shopping_pos,50-60,low_amount)	(not_fraud)	0.157549	0.245077	0.131291	0.833333	3.400298	0.092679	4.52954
(grocery_pos,20-30,low_amount)	(fraud)	0.245077	0.157549	0.131291	0.535714	3.400298	0.092679	1.814509
(shopping_pos,60 - Above,high_amount)	(fraud)	0.102845	0.245077	0.070022	0.680851	2.778116	0.044813	2.365427

4.4 The report generation layer (layer 3)

In this layer, two different kinds of reports were generated: analytical and informative, which provide information for banks' experts to make the right decisions regarding the different transactions that are found to be related to online money laundering.

After completing the proposed paradigm's layers, the paper's goal of predicting and detecting online money laundering that exists through online transactions such as

buying and selling of goods was met, which helps banks' experts make the right decision in different situations.

5 Evaluation

To demonstrate that the (APPD-OML) paradigm outperforms the other approaches that use data mining techniques to predict and detect online money laundering various experiments were conducted that highlight the effectiveness of prediction and detection based on classification, clustering, and association. And which are made based on applying those three data mining techniques.

5.1 Experimental data

A bank's clients' online transaction dataset [33] that contains records of both legitimate and fraudulent online transactions performed by clients between January 1, 2020, and December 31, 2020, was used. These data were labeled completely and are available for open research.

5.2 The experiments methodology

The evaluation process depends on two different experiments. The first experiment tests the ability of each technique against the paradigm (APPD-OML) to achieve highly accurate results regarding the prediction and detection of online money laundering. In the second experiment, the effectiveness of the (APPD-OML) in predicting and detecting online money laundering is tested against other approaches employing the same dataset.

5.2.1 Experiment 1: (Evaluate the effectiveness of the APPD-OML paradigm against each technique)

This experiment demonstrates that APPD-OML outperforms each technique in generating highly accurate results regarding the prediction and detection of online money laundering.

To test the paradigm, in this experiment, classification, clustering, and association are applied to the data, and then the proposed paradigm is applied.

The evaluation process compares the effectiveness of the APPD-OML against each technique in providing accurate results. The APPD-OML's effectiveness and accuracy were measured in terms of accuracy in Equation 10 [49], F measure in Equation 11 [49], and AUC in Equation 12 [50]. Table 10 shows the results of Experiment 1.

- **Accuracy**

Accuracy is one of the most widely used metrics for measuring performance which is calculated as the ratio of samples that are correctly predicted to all samples

$$\text{Accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} \qquad (10)$$

Where

True positive (TP): This is the number of predictions when the classifier properly identifies the positive class as positive.

True negative (TN): This is the number of predictions in which the classifier properly identified the negative class as negative.

False positives (FP): This is the proportion of predictions in which a classifier predicts a negative class as a positive one.

False negative: This is the proportion of predictions in which the classifier misinterprets the positive class as the negative class. [49]

- **F1-Score**

F1- score also known as the F-measure. It denotes the harmonic mean of recall and precision. Precision is the ratio of accurately positive samples to the total number of positive predicted samples, and recall shows the proportion of positively identified positive samples to all positive samples [49]

$$\mathbf{F1\text{-}Score} = \mathbf{2} * \frac{\mathbf{Precision} * \mathbf{Recall}}{\mathbf{Precision+Recall}} \qquad (11)$$

Where

Precision equals TP/(TP+FP).

Recall equals TP/(TP+FN).

- **AUC**

AUC is an acronym for the Area Under the Receiver Operating Characteristic Curve and is also referred to as ROC AUC. The score it produces ranges from 0.5 to 1, with 1 signifying the best outcome and 0.5 meaning the model performs as well as a chance [50].

$$\mathbf{AUC} = \int \text{recall}\, d(\text{FPR} \qquad (12)$$

Where

FPR refers to a False positive rate and it is equal to FP/(FP+TN).

Table 11: Experiment 1 results.

Technique	Evaluation measures (%)		
	Accuracy	*F1 measure*	*AUC score*
Classification	93	91	93
Clustering and Association	91	86	86
The Proposed Paradigm (APPD-OML)	94	95	95

According to Table 11, the findings of testing the APPD-OML indicate that the prediction based on applying the three techniques outperformed each technique with an accuracy of 94%, F1 95%, and AUC 95%. This means that the proposed paradigm outperforms each technique. In addition, it achieves a high coverage

and accuracy power that is crucial for any model used in the financial industry.

5.2.2 Experiment 2: (Compare the effectiveness of the APPD-OML paradigm against the [33] approach)

The second experiment aims to compare the results of APPD-OML against [33]. The results of the proposed paradigm are compared with the ones obtained from the [33] approach and the AUC score in Equation 12 is used to evaluate the performance. Table 12 shows the results of Experiment 2.

Table 12: Experiment 2 results.

Models	Algorithms	Accuracy
[16]	DT	92%
[19]	SVM	91%
[21]	GRAPH MATCHING	91%
[25]	NB	78%
[30]	APRIORI	88%
[29]	LR	89%
[31]	RF	93%
[33]	DT	92%
[34]	RF	77%
[36]	KNN	90%
[26]	PART	54%
[37]	RF	90%
The Proposed Paradigm (APPD-OML)	RF K-MEANS APRIORI	94%

As shown in Table 12, the proposed APPD-OML outperformed the [33] approach that utilized only the classification technique in predicting fraudulent transactions.

Based on the outcomes of the two experiments, it has been proven that the APPD-OML succeeded in meeting the paper's goal of predicting and detecting online money laundering. Additionally, it decreases the effort that banks exert to monitor online transactions.

6 Discussion

In the previous sections, the impacts of various prediction and detection techniques on the online bank transaction dataset were presented. The findings show that the accuracy of prediction differs concerning the various techniques applied. As shown in Table 13.

Table 13: Comparison between different models used in prediction and detection of OML.

Criteria	The Proposed Paradigm (APPD-OML)	[22]
AUC score	95%	93.92%

In Table 13, the proposed paradigm outperformed the other models in terms of accuracy of prediction and detection, due to the application of classification, clustering, and association. In addition, it demonstrated that some algorithms are more powerful in predicting malicious activities than others.

The research results were analyzed as presented. The statistical data produced by the analysis was used to describe the findings. The results as reported are reviewed in this part with the findings of other relevant studies. The main objective of this paper is to increase the accuracy of prediction and detection of online money laundering activities. The results revealed that applying classification, clustering, and association techniques to data increases the accuracy of prediction and detection and overcomes the limitations of each technique in the banking sector.

7 Conclusion and future work

This paper provides a paradigm that predicts and detects potential instances of online money laundering in any nation. The proposed paradigm combines three data mining techniques to efficiently predict such criminal activity and promptly deliver the appropriate results to government authorities to take the right action against those criminals. The results of testing the APPD-OML demonstrate that it outperforms each data mining technique implemented individually in terms of accuracy and coverage; in addition, it assists banks' experts in monitoring and analyzing online transactions more efficiently. There are numerous ways to conduct further research. The following list includes a few of these directions:

- Our proposed paradigm can be extendable for minimizing online money laundering by using AI techniques such as artificial neural networks. The artificial neural network is a subset of machine learning models that are constructed using connectionism's discoveries of the principles of neuronal organization in the biological neural networks that make up animal brains, which helps in recognizing patterns in massive amounts of data across various formats, which makes it ideal for identifying suspicious transactions and risks. Previous studies ignored many artificial intelligence techniques despite their advantages. As a result, our future work was to integrate artificial neural networks into our paradigm to enable banks to effectively uncover many online transactions with various formats.
- Constructing a tool that can speed up the processing of data in the data preprocessing phase.

References

[1] Al-Hashedi, K. G. and P. J. C. S. R. Magalingam (2021). "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019." 40: 100402. https://doi.org/10.1016/j.cosrev.2021.100402

[2] Pour, M. S., et al. (2023). "A Comprehensive Survey of Recent Internet Measurement Techniques for Cyber Security." 103123. https://doi.org/10.1016/j.cose.2023.103123

[3] Wronka, C. J. J. o. M. L. C. (2022). "Cyberlaundering": the change of money laundering in the digital age." 25(2): 330-344. https://doi.org/10.1108/jmlc-04-2021-0035

[4] Tiwari, M., et al. (2020). "A review of money laundering literature: the state of research in key areas." 32(2): 271-303. https://doi.org/10.1108/par-06-2019-0065

[5] Anwar, M. J. E. A. J. o. M. R. (2023). "The Urgency of Reforming Regulations for Money Laundering in the Digital Era." 2(7): 2895-2906. https://doi.org/10.55927/eajmr.v2i7.5009

[6] Sain, P. and S. Puri (2018). Detection of money laundering accounts using data mining techniques, Mar. https://www.researchgate.net/profile/Shalini-Puri-3/publication/323834276_Detection_of_money_laundering_accounts_using_data_mining_techniques/links/5cf45d4e92851c4dd0240b3a/Detection-of-money-laundering-accounts-using-data-mining-techniques.pdf

[7] Lokanan, M. E. J. J. o. A. S. R. (2022). "Predicting money laundering using machine learning and artificial neural networks algorithms in banks." 1- 25. Lokanan, M. E. J. J. o. M. L. C. (2019). "Data mining for statistical analysis of money laundering transactions." 22(4): 753-763. https://doi.org/10.21203/rs.3.rs-2161095/v1

[8] Chen, Z., et al. (2018). "Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review." 57: 245-285. https://doi.org/10.1007/s10115-017-1144-z

[9] Rambola, R., et al. (2018). Data mining techniques for fraud detection in banking sector. 2018 4th International Conference on Computing Communication and Automation (ICCCA), IEEE. https://doi.org/10.1109/ccaa.2018.8777535

[10] Singh, K. and P. J. I. J. o. A. I. S. Best (2019). "Anti-money laundering: Using data visualization to identify suspicious activity." 34: 100418. https://doi.org/10.1016/j.accinf.2019.06.001

[11] Canhoto, A. I. J. J. o. b. r. (2021). "Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective." 131: 441-452. https://doi.org/10.1016/j.jbusres.2020.10.012

[12] Goecks, L. S., et al. (2022). "Anti-money laundering and financial fraud detection: A systematic literature review." 29(2): 71-85. https://doi.org/10.1002/isaf.1509

[13] Shrivastava, A., et al. (2023). "Literature Review on Tools & Applications of Data Mining." https://doi.org/10.26438/ijcse/v11i4.4654

[14] Tawde, S. D., et al. (2024). Online Payment Fraud Detection for Big Data. International Conference on Distributed Computing and Intelligent Technology, Springer. https://doi.org/10.1007/978-3-031-50583-6_22

[15] Singh, K., et al. (2024). "Automated Secure Computing for Fraud Detection in Financial Transactions." 177-189. https://doi.org/10.1002/9781394213948.ch9

[16] Yee, O. S., et al. (2018). "Credit card fraud detection using machine learning as data mining technique." 10(1-4): 23-27. https://jtec.utem.edu.my/jtec/article/view/3571

[17] Gyamfi, N. K. and J.-D. Abdulai (2018). Bank fraud detection using support vector machine. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE. https://doi.org/10.1109/iemcon.2018.8614994

[18] Kargari, M. and A. Eshghi (2018). A model based on clustering and association rules for detection of fraud in banking transactions. Proceedings of the 4th World Congress on Electrical Engineering and Computer Systems and Sciences EECSS, vol. MVML. https://doi.org/10.11159/mvml18.104

[19] Thennakoon, A., et al. (2019). Real-time credit card fraud detection using machine learning. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE. https://doi.org/10.1109/confluence.2019.8776942

[20] Lokanan, M. E. J. J. o. M. L. C. (2019). "Data mining for statistical analysis of money laundering transactions." 22(4): 753-763. https://doi.org/10.1108/jmlc-03-2019-0024

[21] Dias, L. F. C. and F. S. Parreiras (2019). Comparing data mining techniques for antimoney laundering. Proceedings of the XV Brazilian Symposium on Information Systems. https://doi.org/10.1145/3330204.3330283

[22] Sobreira Leite, G., et al. (2019). "Application of technological solutions in the fight against money laundering—A systematic literature review." 9(22): 4800. https://doi.org/10.3390/app9224800

[23] Chuparkoski, D. (2019). DATA MINING TECHNIQUES FOR ANTI MONEY LAUNDERING. Proceedings of the International scientific and practical conference "Bulgaria of regions". https://science.uard.bg/index.php/regions/article/view/602

[24] Zhang, Y. and P. J. C. E. Trubey (2019). "Machine learning and sampling scheme: An empirical study of money laundering detection." 54: 1043-1063. https://doi.org/10.1007/s10614-018-9864-z

[25] Islam, M. A., et al. (2020). "EVALUATION OF MONEY LAUNDERING RISK OF BANK ACCOUNTS USING NAIVE BAYES

CLASSIFICATION." 15(5): 3481-3493. https://jestec.taylors.edu.my/Vol%2015%20issue%205%20October%202020/15_5_43.pdf

[26] Ahluwalia, A., et al. (2023). Money Laundering Fraudulent Prediction Using Classifiers. 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE. https://doi.org/10.1109/esci56872.2023.10099770

[27] Ghosal, A., et al. (2020). "A short review on different clustering techniques and their applications." 69-83. https://doi.org/10.1007/978-981-13-7403-6_9

[28] Bagle, A. A., et al. (2020). "Anti-Money Laundering System to Detect Suspicious Account." https://d1wqtxts1xzle7.cloudfront.net/62033068/IRJET-V7I111620200208-19382-jra341-libre.pdf?1581157745=&response-content-disposition=inline%3B+filename%3DIRJET_Anti_Money_Laundering_System_to_De.pdf&Expires=1721241633&Signature=G8Ryg1itdocFq9CaeEG1THOzYiw9Vmfp6KnAFXJM3rr8nuXJWaAyJP89Qmha5q4DG6meajYSAv7uClDkZHraC9W7h2moWH3NFQevqz-xn2Yx3IKE57tmA9c0~w63R-dG5MHYgO3HlP6pat2a51o0SrmHuyytagA9Qb-BplIT0vf~kTJR8s5M5NEpEXlNmro2M3Q-7M-8ds9QRUYenxgjdoa4VcxlomPEeszWs~zCPwKlOdvtFTw6LAP45DVzrvascDPvJmK33h0Mn6aBKUnLcSNjWWnVQVcb~u6kG3e-KoFGymxXlVicXG~0t9psDALauvDUZHdL1QCfPrKZUlU1iA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

[29] Martínez-Sánchez, J. F., et al. (2020). "Money laundering control in Mexico: a risk management approach through regression trees (data mining)." 23(2): 427-439. https://doi.org/10.1108/jmlc-05-2022-0061

[30] Vu, H. J. C. s. M. "Machine learning in money laundering detection." 59. https://www.researchgate.net/profile/M-Afzal-Upal/publication/341277472_Proceedings_of_the_Third_Annual_Great_Lakes_Data_Science_Symposium/links/5eb8055d4585152169c14c52/Proceedings-of-the-Third-Annual-Great-Lakes-Data-Science-Symposium.pdf#page=67

[31] Jullum, M., et al. (2020). "Detecting money laundering transactions with machine learning." 23(1): 173-186. https://doi.org/10.1108/jmlc-07-2019-0055

[32] Kumar, V. J. N. C. and Applications (2021). "Evaluation of computationally intelligent techniques for breast cancer diagnosis." 33(8): 3195-3208. https://doi.org/10.1007/s00521-020-05204-y

[33] He, Y. J. H. i. S., Engineering and Technology (2022). "Machine Learning Methods for Credit Card Fraud Detection." 23: 106-110. https://doi.org/10.54097/hset.v23i.3204

[34] Bakhtiari, S., et al. (2023). "Credit card fraud detection using ensemble data mining methods." 1-19. https://doi.org/10.1007/s11042-023-14698-2

[35] Esmail, F. S., et al. (2023). "Review of Loan Fraud Detection Process in the Banking Sector Using Data Mining Techniques." 14(2): 229-239. https://doi.org/10.32985/ijeces.14.2.12

[36] Hampo, J. A., et al. (2023). "A Web-Based kNN Money Laundering Detection System." 1(4): 277-288. https://doi.org/10.59324/ejtas.2023.1(4).27

[37] Alsuwailem, A. A. S., et al. (2023). "Performance of different machine learning algorithms in detecting financial fraud." 62(4): 1631-1667. https://doi.org/10.1007/s10614-022-10314-x

[38] Saraswat, P., et al. (2022). "Data pre-processing techniques in data mining: A Review." 10(1): 122-125. https://doi.org/10.55524/ijircst.2022.10.1.22

[39] Fan, C., et al. (2021). "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data." 9: 652801. https://doi.org/10.3389/fenrg.2021.652801

[40] Farahnak-Ghazani, F. and M. S. Baghshah (2016). Multi-label classification with featureaware implicit encoding and generalized crossentropy loss. 2016 24th Iranian conference on electrical engineering (ICEE), IEEE. https://doi.org/10.1109/iraniancee.2016.7585772

[41] Alexandropoulos, S.-A. N., et al. (2019). "Data preprocessing in predictive data mining." 34: e1. https://doi.org/10.1017/s026988891800036x

[42] Gupta, B., et al. (2017). "Analysis of various decision tree algorithms for classification in data mining." 163(8): 15-19. https://doi.org/10.5120/ijca2017913660

[43] Deng, Y., et al. (2020). "A study on e-commerce customer segmentation management based on improved K-means algorithm." 18(4): 497-510. https://doi.org/10.1007/s10257-018-0381-3

[44] Edastama, P., et al. (2021). "Implementation of data mining on glasses sales using the apriori algorithm." 1(2): 159-172. https://doi.org/10.34306/ijcitsm.v1i2.46

[45] Python, W. J. P. R. f. W. (2021). "Python." 24. Rambola, R., et al. (2018). Data mining techniques for fraud detection in banking sector. 2018 4th International Conference on Computing Communication and Automation (ICCCA), IEEE. https://doi.org/10.1109/ccaa.2018.8777535

[46] Ilyas, I. F. and X. Chu (2019). Data cleaning, Morgan & Claypool. https://doi.org/10.1145/3310205

[47] Yadav, N. and N. J. I. J. o. F. S. E. Badal (2020). "Data preprocessing based on missing value and

discretisation." 1(2-3): 193-214. https://doi.org/10.1504/ijfse.2020.10032758

[48] Nauman, F. and M. Herschel (2022). An introduction to duplicate detection, Springer Nature. https://doi.org/10.1007/978-3-031-01835-0_4

[49] Tharwat, A. J. A. C. and Informatics (2020). "Classification assessment methods." https://doi.org/10.1016/j.aci.2018.08.003

[50] Yang, S., et al. (2017). "The receiver operating characteristic (ROC) curve." 5(19): 34-36. https://doi.org/10.12746/swrccc.v5i19.391