ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 6 No. 1 (2026) 397-408
https://publisher.uthm.edu.my/periodicals/index.php/ekst

Detection of Money Laundering Activities using Supervised Learning Techniques

Christine Yap En Xi[1], Rohayu Mohd Salleh[1]*

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600 Pagoh, Muar, Johor, MALAYSIA.*

*Corresponding Author: rohayu@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2026.06.01.046

Article Info

Received: 4 January 2026
Accepted: 20 January 2026
Available online: 9 July 2026

Keywords

Money Laundering, Descriptive Analysis, Supervised Learning Models, Random Forest, MCC, DOR

Abstract

Money laundering is a financial crime that occurs globally, including in Malaysia. The existing traditional detection methods are not adequate to identify complex money laundering activities. This study aims to identify the laundering and non-laundering transactions based on receiving currency, payment currency and payment format using descriptive analysis, apply the supervised learning models, including Random Forest (RF), Naïve Bayes (NB) and K-Nearest Neighbour (KNN), and compare their performances in the classification of laundering and non-laundering transactions. The dataset used in this study was International Business Machines Corporation (IBM) transaction for Anti-Money Laundering (AML) which was available on the Kaggle website with the class imbalance ratio is 0.00077% of laundering cases. The results indicate that the US dollar has the greatest number of laundering and non-laundering transactions in the receiving and payment currency, while the payment format of Automated Clearing House (ACH) has the greatest number of laundering transactions which is 461 equals to 84.1% and the payment format of reinvestment has the greatest number of non-laundering transactions which is 241702 equals to 34.0%. The RF model was not the most effective model in classifying the laundering status of transactions with low MCC value which is 0.0287 and low DOR value which is 41.88. These findings demonstrate that the type of currency and payment format are critical to differentiate the laundering and non-laundering transactions, showing that the supervised learning approach is not an effective predictive model in detecting money laundering activities.

1. Introduction

Money laundering is an illegal activity that involves the process of converting illegal money obtained from criminal activities, including drug trafficking or terrorism, to become a legal source of funds [1]. This process involved three stages: placement, layering, and integration. Placement refers to the process of moving the 'dirty' money into a legitimate financial system, layering involves multiple layers of transactions to make it difficult to trace the origin of illegal money, and finally, integration, which is the 'clean' money is reentered into the economy as the legal money [2]. The placement stage is related to the transaction amount and payment format, which is the illegal funds are divided into smaller values through different channels. The layering stage is captured through a variety of currencies and payment formats to hide the source of funds.

Money laundering refers to the process of hiding the source of money, which comes from illegal earnings, allowing the criminal to have the money in a legal origin [3]. The estimated amount of money laundered is

between 2% to 5% of global Gross Domestic Product (GDP) annually, or it means between USD 800 billion to USD 2 trillion. There is between 30% and 40% of all money laundering cases worldwide involve drug trafficking, which contributes to the largest source of laundered funds. The top three countries with the highest laundered funds are United States, which achieved USD 800 billion; the Russia, with USD 500 billion; and China, with USD 200 billion [4].

As money laundering spreads worldwide, Malaysia is also threatened by this issue. A case in 2020 is related to a gold investment company named Genneva Malaysia Sdn Bhd (GMSB), which illegally collected RM 7 billion of deposits from the public and laundered RM 4.5 billion [5]. The most popular in Malaysia was 1Malaysia Development Berhad (1MDB), which happened between 2009 and 2015, and involved money laundering of USD 4.5 billion of dollars from bribery activities and embezzlement [6]. In short, the growing threat of money laundering has highlighted the need for financial institutions to implement detection methods for money laundering.

Existing traditional detection methods, including transaction monitoring systems, KYC, and CDD are used to solve the money laundering problem [7]. Due to the existing traditional detection methods are insufficient to detect complex money laundering activities, a study to detect money laundering activities using machine learning is necessary. Machine learning approaches, including supervised learning models are helpful in handling large datasets with labelled data, training with labelled data and identifying complex data patterns through classification to provide a more accurate detection rate.

A study utilised several supervised machine learning algorithms to examine the transaction dataset and found that RF was useful in detecting anti-money laundering with its highest accuracy of 99.0% [8]. Besides, a study found that the ensemble learning method, which combines RF, Extra Trees and Bagging classifier has better performance compared to the RF model in predicting licit and illicit transactions in Bitcoin [9]. Another study identified Logistic Regression (LR), RF, Decision Tree (DT) and XGBoost perform poorly in the imbalanced suspicious and non-suspicious transactions data with Area Under Precision Recall Curve (AUPRC) values are less than 60%, indicating that is challenging to predict the transaction as suspicious [10].

This study aims to identify the laundering and non-laundering transactions based on receiving currency, payment currency and payment format using descriptive analysis, to apply the supervised learning models such as RF, NB and KNN in classifying laundering and non-laundering transactions, to evaluate and compare the performance of the supervised learning models such as RF, NB and KNN in classifying laundering and non-laundering transactions.

2. Methodology

This section discusses the methodology used in this study. Data description, data pre-processing, descriptive analysis, supervised learning techniques and evaluation metrics are clearly explained in detail.

2.1 Data Description

The dataset was obtained from International Business Machines Corporation (IBM) transaction for AML which was available on the Kaggle website [11]. The dataset used is small transactions with a higher illicit ratio (more laundering) starting from 1st September 2022 to 9th September 2022. There are 1048575 observations in the dataset, with 615 transactions labelled as money laundering transactions. Table 1 shows the variables used in this study, such as amount received, receiving currency, amount paid, payment currency, payment format and money laundering. Receiving currency and payment currency are used to identify the potential specific currency involved in money laundering, while payment format is used to determine the frequent methods to make the laundering and non-laundering transactions.

Table 1 List of banking transaction variables

Variables	Description	Type of data
Amount received	Transaction amount	Continuous
Receiving currency	Transaction currency	Nominal
Amount paid	Transaction amount received by bank account of receivers	Continuous
Payment currency	Currency of transaction amount received	Nominal
Payment format	Type of payment used for the transaction	Nominal
Money laundering	Whether the transaction is	Binary

laundering (1) or not (0)

2.2 Data Pre-processing

The data pre-processing involves data cleaning process to identify missing values, remove noisy data, detect outliers and solve inconsistent data [12]. These problems in the data will affect the analysis result. The noisy data is the random errors or variances that exist in a measured variable while inconsistent data is the data that have same input values but it is represented in different class labels [13].

2.2.1 Preliminary Analysis

The step of data preliminary analysis includes data cleaning, which is used to remove missing values, duplicate data and outliers. The Python software was used to check the missing values and duplicate data. In this study, there is no missing values and duplicate data, remaining 1048575 observations in the dataset. Then, the Interquartile Range (IQR) method was used to detect the outliers. There are 162110 outliers, resulting in 710835 observations after removing the outliers, with 710287 observations are non-laundering transactions and 548 observations are laundering transactions. A study utilised outlier removal in an imbalanced credit card transactions dataset because it can affect the model's performance and produce inaccurate predictions. Removing outliers help to stabilise the model by eliminating extreme differences from the features. The IQR method reduces the effect of noisy or unusual data which could cause unreliable or misleading patterns during training phase [14].

2.2.2 Data Splitting

Data splitting is a separation of data into training and validation sets in order to train the model, determine the best model parameters and measure the model's performance [15]. The dataset is randomly divided into an 80:20 ratio, with the training dataset (80%) used for model training and the testing dataset (20%) used to evaluate the model's performance.

In this study, an 80:20 random split may not preserve enough minority samples in the test set. A study showed that the random split does not ensure that the minority class are included in the training dataset, causes imbalance in the training dataset, and not all classes appear equally in the training dataset. The balanced split is recommended to overcome the problems exist in the random split because it can determine the equal number of majority and minority class samples in the training dataset to prevent imbalance classes during the data splitting [16]. Moreover, inaccurate train-test split ratio can affect the evaluation and performance of model which results in overfitting or underfitting, especially in the dataset with imbalanced classes [17].

2.3 Descriptive Analysis

Descriptive statistics is a process of summarising data to identify the relationship between variables in a sample or population. It comprises various types of variables, which include nominal, ordinal, interval and ratio along with measures of frequency, measures of central tendency, measures of dispersion and measures of position [18]. In this study, since payment currency and payment format are nominal variables, the descriptive analysis was utilised to identify the amount of laundering and non-laundering transaction using measures of frequency.

2.4 Supervised Learning Techniques

In this study, the supervised learning techniques such as RF, NB and KNN are used to classify the laundering and non-laundering transactions. The performance of these three supervised learning techniques were evaluated and compared.

2.4.1 Random Forest (RF)

RF consists of a collection of decision trees, each of which is independent of the others. A decision tree is a tree structure with a root node as a starting point, each non-leaf node represents a feature test, each branch represents the outcome of a feature test and finally the leaf node is a category with a decision outcome [19]. The outcome of RF algorithms is determined by the most votes of the decision tree in a forest as a final prediction. Each decision tree within a forest is formed through a random subset of samples and features from the training dataset [20].

2.4.2 Naïve Bayes (NB)

NB is a simple probabilistic classifier that uses Bayes' theorem and assumes independence between the features. It assumes that the effect of a predictor value on a specified class is independent of the other predictor values [21]. The outcome of NB algorithms is determined by the class with the highest posterior probability as the

prediction [22]. Based on Bayes' theorem, the conditional probability of each class given the features is calculated as in equation (1):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{1}$$

where $P(H/X)$ is the probability that the hypothesis H holds given the observed training data X, $P(X/H)$ is the probability of observing training data X, given that the hypothesis H holds. $P(H)$ is the prior probability of hypothesis H which is the initial probability before observing the training data X and $P(X)$ is the probability that training data X is observed.

2.4.3 K-Nearest Neighbour (KNN)

KNN is classified as a lazy learning algorithm since it predicts through real-time analysis of the data until new instances are introduced [23]. The new instances can be determined by the majority vote of its k neighbours, where k is a positive small integer [24]. Before applying the KNN model, data normalisation was required to change the numeric values into a fixed range. The formula of data normalisation is shown in equation (2):

$$v' = \frac{v - v_{min}}{v_{max} - v_{min}} \tag{2}$$

where v' is the normalised value, v_{min} is the minimum value from each data point, v_{max} is the maximum value from each data point.

The k closest neighbours are calculated using Euclidean distance [20]. The Euclidean distance is used as a distance metric to calculate the distance between two points. The formula of Euclidean distance is indicated as in equation (3):

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{d} (x_{2i} - x_{1i})^2} \tag{3}$$

Based on equation (3), x_{1i} is the data sample, x_{2i} is the testing data, i is the variable of the data and d is the number of dimensions in the data.

2.5 Evaluation Metrics

In this study, confusion matrix, accuracy, recall, precision, F1-score, Matthews Correlation Coefficient (MCC), Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Diagnostic Odds Ratio (DOR) are used to evaluate the performance of the supervised learning algorithms. The algorithm with the highest value in the evaluation metrics is considered the best classifier for detecting money laundering activities.

There are four terms in the confusion matrix, which are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP describes the number of samples with correct prediction as positive, while TN describes the number of samples with correct prediction as negative. FP is the number of samples with incorrect prediction as positive, while FN is the number of samples with incorrect prediction as negative [25].

3. Results and Discussion

3.1 Descriptive Analysis

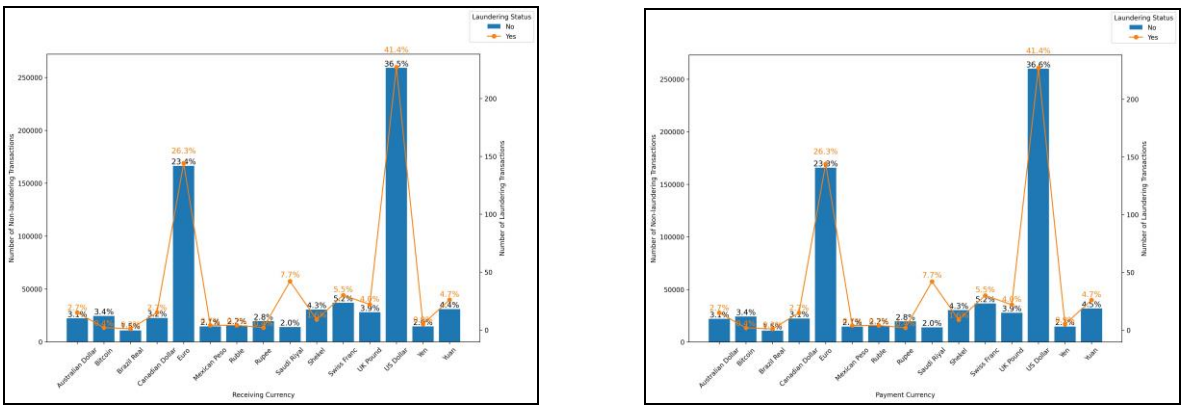
The descriptive statistics was applied to identify the laundering and non-laundering transactions based on receiving currency, payment currency and payment format using dual axis bar chart, clustered column chart and heatmap.

3.1.1 Laundering and Non-laundering Transactions with Receiving and Payment Currency

Fig. 1(a) and Fig. 1(b) show the number of laundering and non-laundering transactions with receiving currency and payment currency. Based on Fig. 1(a), the highest number of laundering transactions is from US Dollar which is 227 equals to 41.4%, while the second highest is from Euro which is 144 equals to 26.3%. The US Dollar has the highest number of non-laundering transactions which is 259102 equals to 36.5%, then followed by the Euro which is 166402 equals to 23.4%. This result indicates that the US Dollar is the most dominant currency that used for international transactions, then followed by the second largest currency, Euro.

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) 397-408

Based on Fig. 1(b), the highest number of laundering transactions is from US Dollar which is 227 equals to 41.4%, while the second highest is from Euro which is 144 equals to 26.3%. The US Dollar achieves the highest number of non-laundering transactions which is 260020 equals to 36.6%, then followed by the Euro which is 165778 equals to 23.3%. This result indicates that the transactions are mostly paid by US Dollar compared to the Euro.



(a) (b)

Fig. 1 Dual axis bar chart (a) Receiving currency; (b) Payment currency

3.1.2 Laundering and Non-laundering Transactions with Payment Format

According to Fig. 2, at the status of laundering, the number of ACH transactions is the highest which is 461 equals to 84.1%, while the number of bitcoin transactions is the lowest which is 2 equals to 0.4%. The second highest number of laundering transactions is cheque which contains 46 transactions equals to 8.4%. The credit card contains 23 laundering transactions equals to 4.2%, and the cash contains 16 laundering transactions equals to 2.9%. The result shows that the ACH involves more laundering transactions compared to the other payment format because it is convenient to transfer the money between the bank accounts.

At the status of non-laundering, the reinvestment achieves the highest number of transactions which is 241702 equals to 34.0%, then followed by the cheque which is 179959 equals to 25.3%. There are 134433 non-laundering transactions equals to 18.9% from the credit card, 62764 non-laundering transactions equals to 8.8% from the ACH, 48221 non-laundering transactions equals to 6.8% from the cash, and 24272 non-laundering transactions equals to 3.4% from the bitcoin. Lastly, the lowest number of non-laundering transactions is from wire transfer which is 18936 equals to 2.7%. The result shows that the reinvestment involves more non-laundering transactions because it is a standard practice of using interest and dividends to do the investment.

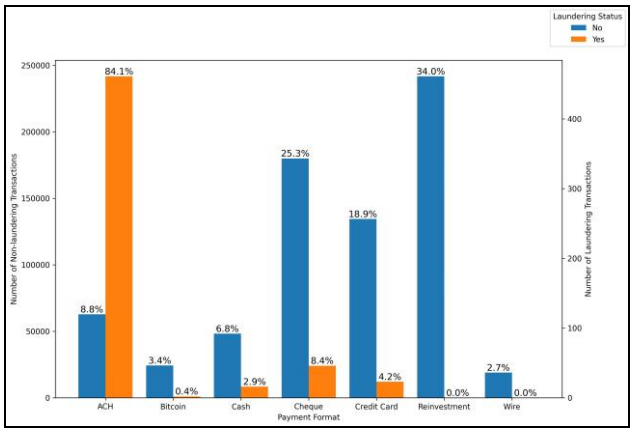
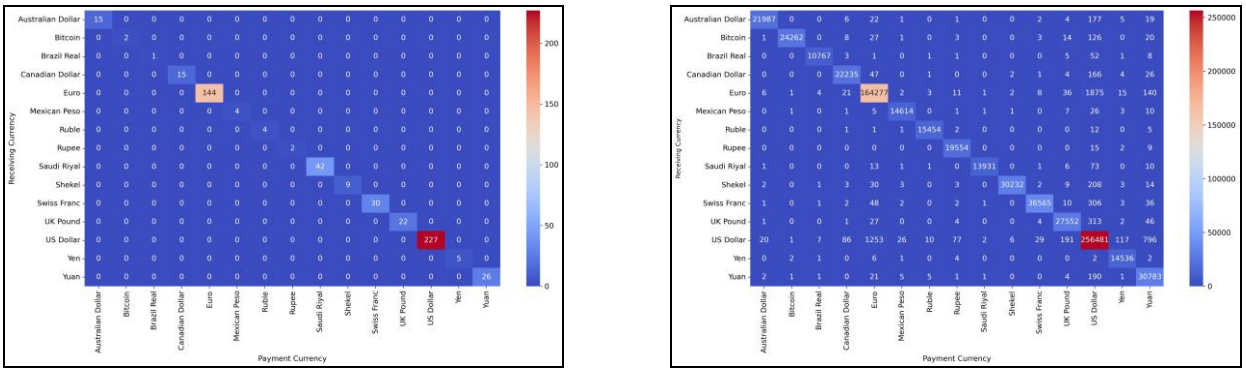


Fig. 2 Dual axis clustered column chart for the payment format

3.1.3 Laundering and Non-laundering Transactions with Currency Pairs

Fig. 3(a) and Fig. 3(b) show the number of laundering and non-laundering transactions with different currency pairs. The number of transactions is divided into colour gradient with blue and red to identify the transaction pattern. The colour of dark blue represents low transactions, light blue represents medium low transactions, light red represents medium high transactions and dark red represents high transactions.



(a) (b)

Fig. 3 Heatmap for the currency pairs (a) Laundering transactions; (b) Non-laundering transactions

3.2 Classification Performance of Random Forest (RF) Model

Based on the Fig. 4, features such as payment currency, payment format, amount received and receiving currency are used to split the data. The payment currency (bitcoin) is positioned at the top of the tree as the root node because splitting on this feature obtains the lowest value of Gini impurity which is 0.001, indicating a highly pure split and strong class separation at the initial stage of the decision tree. If the payment currency is bitcoin, then continue split to the right branch which is the laundering status is classified as no. If the payment currency is not bitcoin, then split to the left branch. Next, the left branch is receiving currency (Shekel). If the receiving currency is Shekel, then split to the right branch (amount received is 43152.875). If the receiving currency is not Shekel, then split to the left branch (payment currency is Saudi Riyal).

At the third level of split, the payment currency of Saudi Riyal splits into two nodes, payment format (cheque) and payment format (cash). If the payment currency is not Saudi Riyal, then split to the payment format (cheque). If the payment is Saudi Riyal, then split to the payment format (cash). On the other sides, if the amount received is not less than 43152.875, then split to the payment format (cheque). If the amount received is less than 43152.875, then split to the payment format (wire).

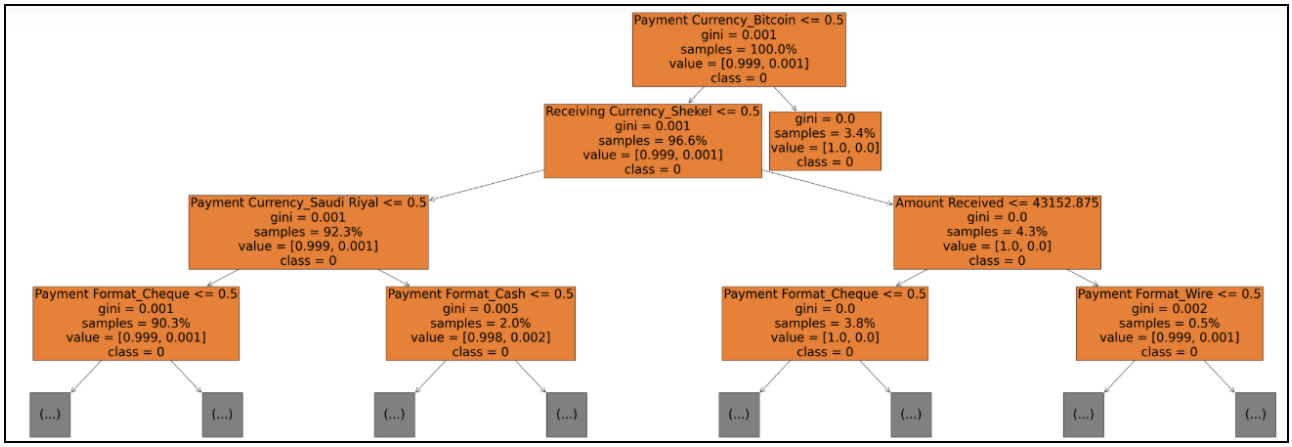


Fig. 4 First decision tree from the RF model

According to Table 2, three observations were correctly predicted as laundering transactions and 141961 observations as non-laundering transactions. However, there were 90 observations where non-laundering transactions were incorrectly predicted as laundering and 113 observations where laundering transactions were incorrectly predicted as non-laundering.

Table 2 Confusion matrix for the RF model for each type of transaction

Actual class	Predicted class	
	Laundering	Non-laundering
Laundering	3	113
Non-laundering	90	141961

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) 397-408

3.3 Classification Performance of Naïve Bayes (NB) Model

Based on the Table 3, non-money laundering transaction had the highest a-priori probability of 0.9992, while money laundering transaction had the lowest a-priori probability of 0.00076. The result indicates that the probability of a transaction being identified as money laundering is very low which is close to 0%, and the probability of a transaction being identified as non-money laundering is extremely high which is almost 100%.

Table 3 A-priori probability of Naïve Bayes model

Money laundering	Yes	No
A-priori probability	0.00076	0.9992

According to Table 4, 142051 observations were correctly predicted as non-laundering transactions. However, there were 116 observations where laundering transactions were incorrectly predicted as non-laundering.

Table 4 Confusion matrix for the NB model for each type of transaction

Actual class	Predicted class	
	Laundering	Non-laundering
Laundering	0	116
Non-laundering	0	142051

3.4 Classification Performance of K-Nearest Neighbour (KNN) Model

The feature scaling used in this dataset is normalisation which is min-max scaling. Firstly, the dataset is normalised to change the values in the numerical column to a fixed range, typically between 0 and 1 because the values in different scales will affect the distance calculation. In order to identify the distance between the observations and target data, the amount received and amount paid data have been normalised using min-max normalisation to ensure the values fall within the range of 0 to 1. The categorical variables such as receiving currency, payment currency and payment format have been transformed into dummy variable that takes the binary value of 0 or 1. The value of 1 represents the category is present, and the value of 0 represents the category is absent.

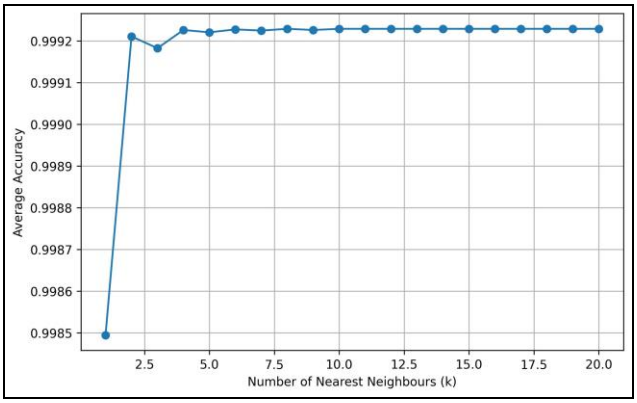


Fig. 5 Average accuracy using stratified 5-fold cross validation

Based on Fig. 5, the curve begins with the lowest average accuracy which is below 0.9985. The average accuracy increases sharply at k = 2, then falls down at k = 3 and increases back to the value which is greater than 0.9992. After that, the curve becomes flat until k = 20. The maximum average accuracy is at k = 8 which is 0.999229, so the value of k was specified as 8, indicating that the predictions are made based on the 8 nearest training observations with the test sample.

Table 5 Confusion matrix for the KNN model for each type of transaction

Actual class	Predicted class	
	Laundering	Non-laundering
Laundering	0	116

Non-laundering	1	142050

According to Table 6, 142050 observations were correctly predicted as non-laundering transactions. Among these, there is no observation was correctly predicted as laundering transaction. However, there were 1 observation where non-laundering transaction was incorrectly predicted as laundering and 116 observations where laundering transactions were incorrectly predicted as non-laundering.

3.5 Performance Comparison

According to Table 7, the RF model achieves the highest recall value with 0.0259 for laundering transactions. The result indicates that the RF model correctly identified 2.59% of actual laundering transactions. For non-laundering transactions, the NB model achieves a perfect recall value of 1.00. This implies that the model is good at correctly identifying non-laundering transactions.

The RF model yields a higher precision value of 0.0323 for laundering transactions compared to the recall value of 0.0259. This shows the model only predicts the transactions as laundering but misses many actual laundering transactions. The RF model had the highest precision value with 0.0323, indicating that 3.23% of the laundering transactions correctly predicted as laundering were correctly classified. For non-laundering transactions, the precision value is 0.9992 in the RF, NB and KNN models. This describes that these models are excellent in correctly predicting non-laundering transactions.

In terms of F1-score, the RF model achieves the highest F1-score with 0.0287 for laundering transactions compared to other models. This implies that the RF model's ability to correctly identify transactions as laundering is poor. For non-laundering transactions, the NB model achieves the highest F1-score of 0.9996. This indicates that the NB model performs well in identifying non-laundering transactions correctly.

Table 6 Summary of recall, precision and F1-score

Model	Class	Recall	Precision	F1-score
RF	Laundering	0.0259	0.0323	0.0287
	Non-laundering	0.9994	0.9992	0.9993
NB	Laundering	0.0000	0.0000	0.0000
	Non-laundering	1.0000	0.9992	0.9996
KNN	Laundering	0.0000	0.0000	0.0000
	Non-laundering	0.9992	0.9992	0.9992

According to Table 8, the RF model achieves the highest macro-average recall value with 0.5126 compared to the NB and KNN models. This shows that the RF model correctly identified 51.26% of the transactions in each class on average. The performance is moderate for each class because a high recall value in the laundering class, while a low recall value in the non-laundering class.

Apart from that, the RF model has the highest macro-average precision value with 0.5157. This describes that 51.57% of the transactions correctly predicted on average for each class. In terms of the macro-average F1-score, the RF model achieves the highest value of 0.5140. This implies that the RF model has moderate performance in detecting the transactions correctly on average for each class.

Furthermore, the RF model achieves an accuracy value of 0.9986, while the NB and KNN models achieve an accuracy value of 0.9992. Since there are zero values exist in the confusion matrix for the NB and KNN models, so their high accuracy values are misleading because there are only non-laundering transactions correctly predicted. Hence, the accuracy value of 0.9986 in the RF model is reasonable because there is no zero value exist in the confusion matrix, indicating that the RF model can predict the laundering and non-laundering transactions. This result indicates that all the RF model correctly classified more than 99.86% of the transactions in the dataset.

Table 7 Summary performances of the RF, NB and KNN models

Model	Macro-average recall	Macro-average precision	Macro-average F1-score	Accuracy
RF	0.5126	0.5157	0.5140	0.9986
NB	0.5000	0.4996	0.4998	0.9992
KNN	0.4996	0.4996	0.4996	0.9992

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) 397-408

According to Table 9, the RF model achieves the highest MCC value which is 0.0287 and DOR value which is 41.88 compared to the NB and KNN models. This indicates that the RF model has a weak performance in predicting the laundering status of transactions and able to distinguish between laundering and non-laundering. In terms of AUC value, since the NB and KNN models have zero MCC and DOR values, its high AUC values are misleading because the models will predict only non-laundering transactions but it cannot predict any laundering transactions since their TP values in the confusion matrix are zero. The RF model achieves the moderate AUC value of 0.55, indicating that it is weak at differentiating between laundering and non-laundering classes.

Overall, the RF model achieves the highest values in MCC and DOR, indicating that it has limited predictive performance and weak discrimination, although its AUC value is low. Therefore, the RF model is not the most effective model for classifying the laundering status of transactions compared to the NB and KNN models. In this study, the models' high accuracy values are misleading to reflect their ability to detect money laundering transactions due to the class imbalance problem, while MCC and DOR are used to verify whether the models' high accuracy reflect the true detection or always classify the majority non-laundering class.

Table 8 Summary values of MCC, DOR and AUC

Model	MCC	DOR	AUC
RF	0.0287	41.88	0.55
NB	0.0000	0.0000	0.89
KNN	-7.5789×10^{-5}	0.0000	0.56

3.6 Discussion

Based on the results obtained, the RF model achieves low values of recall, precision and F1-score for the laundering class, while achieves high values of recall, precision and F1-score for the non-laundering class. The RF model also achieves a high accuracy. A study discovered that the RF obtained an accuracy of 99% on the majority class and only 2% on the minority class. This indicated that RF was more biased to the majority class, and less effective in detecting the minority class. The RF also obtained a very low recall value of 0.021, a precision value of 0.1304 and a F1-score of 0.0361, but obtained a high accuracy of 0.96. The results indicated that RF was good at identifying the majority class but weak at identifying the minority class [26]. Hence, this study explains the class imbalance problem that causes the performance of RF to be affected by the majority class.

Besides, Ali obtained the macro-average recall, precision and F1-score of 0.70 for the low and high performers classes. The results indicated that the RF had less ability to classify the high performers correctly, which is common in an imbalanced dataset [27]. Another study noticed that the RF achieved the macro-average recall of 0.53, macro-average precision of 0.75 and macro-average F1-score of 0.51. This study concluded that the RF was more robust than KNN in managing class imbalance since the KNN has a poor performance in macro-average recall, precision and F1-score [28]. So, these studies justify the moderate performance of macro-average recall, precision and F1-score in the RF, while the RF outperformed other models in the class imbalance situation.

Moreover, Chicco & Jurman found that the MCC values close to zero when applied to the positively and negatively imbalanced datasets, which indicated the poor performance of the machine learning method with a low quality of prediction. The MCC only produces high value when the positive and negative classes are classified correctly based on the values in the confusion matrix. The study also revealed that the use of MCC is better than F1-score and accuracy in imbalanced datasets because its value is obtained based on the ability of the classifier to predict correctly on both the majority of the positive and negative classes [29]. Therefore, this study justifies that the MCC gives a reliable result on whether the model can classify correctly the majority and minority classes.

A study conducted the classification of machine learning techniques in unbalanced and balanced datasets. The results indicated that the RF achieved a low DOR value of 25.397 in the unbalanced dataset, while the DOR value became large when the balancing techniques such as Random Oversampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), SMOTETomek and Adaptive Synthetic Sampling Algorithm (ADASYN) are used [30]. So, this study explains the comparison results of the DOR value between unbalanced and balanced datasets, which reflected the effect of the class imbalance.

A study found that there is a high AUC value in the million datasets with different numbers of positive classes, because the AUC value can be misleading in very imbalanced datasets [31]. The Area Under Precision Recall Curve (AUPRC) is more effective than AUC because the classifiers obtain high AUC scores but obtain low

AUPRC scores since the random under-sampling is used to make the size of the majority class closer to the minority class. The high AUC scores become misleading because the large size of the majority class dominates the false positive rate, and fails to show the negative impact of random under-sampling [32].

4. Conclusion

In this study, there is a large gap between laundering and non-laundering transactions in the receiving currency and payment currency, with many non-laundering transactions and very few laundering transactions, which causes imbalanced data. Among the three models, the RF model was not the most effective model to classify the laundering status in the transactions because it did not achieve the highest values for accuracy and AUC despite having the highest values of MCC and DOR which are remain low values.

In the future work, researchers should consider a dataset with more balanced classes by including a greater number of laundering cases in the transactions or applying resampling techniques to balance the majority and minority classes. Secondly, obtaining the additional information such as the customer's nationality, occupation, income level, and transaction frequency to capture the possibility of a money laundering transaction. Thirdly, expanding the dataset to cover multiple years, such as two to five years, to identify the transaction pattern. Next, applying more advanced machine learning models such as Support Vector Machine, XGBoost, and Artificial Neural Networks would give more accurate predictive analysis, increasing the effectiveness of the money laundering system. Lastly, replacing the AUC metric with other evaluation metrics such as Area Under Precision Recall Curve (AUPRC), Geometric Mean (GM), and balanced accuracy which would produce reliable results when applied to an imbalanced class dataset.

Acknowledgement

The authors would thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Christine Yap En Xi, Rohayu Mohd Salleh;* ***data collection:*** *Christine Yap En Xi;* ***analysis and interpretation of results:*** *Christine Yap En Xi, Rohayu Mohd Salleh;* ***draft manuscript preparation:*** *Christine Yap En Xi, Rohayu Mohd Salleh. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Chen, J. (2024, July 24). *What is money laundering*? Investopedia. Retrieved April 20, 2025, from https://www.investopedia.com/terms/m/moneylaundering.asp

[2] Bank Negara Malaysia. (2022). *Money laundering & terrorism financing*. Anti-Money Laundering & Counter Financing of Terrorism. Retrieved April 20, 2025, from https://amlcft.bnm.gov.my/what-is-money-laundering

[3] United Nations Office on Drugs and Crime. (n.d.). *Overview of money laundering*. United Nations: Office on Drugs and Crime. Retrieved April 20, 2025, from https://www.unodc.org/unodc/en/money-laundering/overview.html

[4] Burnett, S. (2025, June 16). *Money laundering statistics 2025: key global trends and insights*. CoinLaw. Retrieved November 18, 2025, from https://coinlaw.io/money-laundering-statistics/

[5] Ong, E. (2023, March 18). The scourge of modern day scams. *Borneo Post Online*. https://www.theborneopost.com/2023/03/18/the-scourge-of-modern-day-scams/

[6] Financial Crime Academy. (2025, October 20). *The 1MDB money laundering scandal and corrupt politicians*. Financial Crime Academy. Retrieved November 19, 2025, from https://financialcrimeacademy.org/the-1mdb-money-laundering-scandal-and-corrupt-politicians/

[7] Khan, A., Jillani, M. A. H. S., Ullah, M., & Khan, M. (2025). Regulatory strategies for combatting money laundering in the era of digital trade, *Journal of Money Laundering Control, 28(2)*, 408-423, https://doi.org/10.1108/JMLC-07-2024-0113

[8] Raiter, O. (2021). Applying supervised machine learning algorithms for fraud detection in anti-money laundering, *Journal of Modern Issues in Business Research, 1(1)*, 14-26, http://dx.doi.org/10.17613/2g0z-0814

[9] Alarab, I., Prakoonwit, S., & Nacer, M. I. (2020, June 19-21). *Comparative analysis using supervised learning methods for anti-money laundering in bitcoin* [Paper presentation]. 2020 5th International Conference on Machine Learning Technologies, (ICMLT), Beijing, China. https://doi.org/10.1145/3409073.3409078
[10] Pathak, S., & Shrestha, B. (2020). Financial crime detection: A machine learning approach to anomalous financial transaction detection using SWIFT synthetic dataset, 1-6, https://www.researchgate.net/publication/377776311
[11] Altman, E. (2025, July 8). *IBM Transactions for Anti-Money Laundering (AML)*. Kaggle. Retrieved November 19, 2025, from https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml
[12] Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining, *Journal of Engineering and Applied Sciences, 12(16)*, 4102-4107, https://doi.org/10.3923/jeasci.2017.4102.4107
[13] García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4
[14] Zlobin, M., & Bazylevych, V. (2025). Development of a preprocessing methodology for imbalanced datasets in machine learning training. *Technology Audit and Production Reserves, 3(83)*, 55-61, https://doi.org/10.15587/2706-5448.2025.330639
[15] Birba, D. E. (2020). *A comparative study of data splitting algorithms for machine learning model selection*. Degree Project in Computer Science and Engineering, 1-23, https://www.diva-portal.org/smash/get/diva2:1506870/FULLTEXT01.pdf
[16] Khan, A. A. (2022). Balanced split: A new train-test data splitting strategy for imbalanced datasets. *arXiv preprint*, https://doi.org/10.48550/arxiv.2212.11116
[17] Sivakumar, M., Parthasarathy, S., & Padmapriya, T. (2024). Trade-off between training and testing ratio in machine learning for medical image processing. *PeerJ Computer Science, 10*, 1-17, https://doi.org/10.7717/peerj-cs.2245
[18] Kaur, P., Stoltzfus, J., & Yellapu, V. (2018). Descriptive statistics, *International Journal of Academic Medicine, 4(1)*, 60-63, https://doi.org/10.4103/IJAM.IJAM_7_18
[19] Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm, *Procedia Computer Science, 162*, 503-513, https://doi.org/10.1016/j.procs.2019.12.017
[20] Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview, *Babylonian Journal of Machine Learning*, 69-79, https://doi.org/10.58496/bjml/2024/007
[21] Vembandasamyp, K., Sasipriyap, R. R., & Deepap, E. (2015). Heart diseases detection using naïve Bayes algorithm, *International Journal of Innovative Science, Engineering & Technology, 2(9)*, 441-444, https://ijiset.com/vol2/v2s9/IJISET_V2_I9_54.pdf
[22] Alahmar, M. (2023). Naïve Bayes algorithms, *Antifiction Intelligent Project*, 1-14, https://doi.org/10.13140/RG.2.2.15378.73921
[23] Halder, R. K., Uddin, M. N., Uddin, Md. A., Aryal, S., & Khraisat, A. (2024). Enhancing k-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications, *Journal of Big Data, 11(113)*, 1-55, https://doi.org/10.1186/s40537-024-00973-y
[24] Khamis, H. S., Cheruiyot, K. W., & Kimani, S. (2017). Application of k-nearest neighbour classification in medical data mining in the context of Kenya, *JKUAT Annual Scientific Conference*, 990-1000, http://journals.jkuat.ac.ke/index.php/jscp/index
[25] Kumar, V. V. (2024, February 5). *Evaluating machine learning models: metrics and techniques*. AI Accelerator Institute. Retrieved April 19, 2025, from https://www.aiacceleratorinstitute.com/evaluating-machine-learning-models-metrics-and-techniques/
[26] Raza, A., Javeed, M. U., Javed, M., Aslam, S. M., Saleem, R., Irshad, W., Maqbool, H., & Ejaz, G. (2025). A hybrid machine learning framework for personalized news recommendation, *International Journal of Advanced Computing & Emerging Technologies, 1(3)*, 49-62, https://doi.org/10.64796/qwjyd222
[27] Ali, S. M. (2025). Impact of background music on student task performance: A statistical and machine learning analysis, *Journal of Emerging Technologies and Innovative Research, 12(7)*, 483-525. https://www.jetir.org/papers/JETIR2507258.pdf
[28] Li, P., Lyu, S., & Niu, P. (2025). *AI-driven lung cancer screening: A comparative analysis of machine learning models*. Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management, 310-314. https://doi.org/10.5220/0013329700004558
[29] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics, 21(6)*, 1-13. https://doi.org/10.1186/s12864-019-6413-7
[30] Ormeño, P., Márquez, G., & Taramasco, C. (2024). Evaluation of machine learning techniques for classifying and balancing data on an unbalanced mini-mental state examination test data collection applied in Chile, *IEEE Access, 12*, 49376-49386. https://doi.org/10.1109/ACCESS.2024.3383837

[31] Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2020). Investigating class rarity in big data, *Journal of Big Data, 7(23)*, 1-17. https://doi.org/10.1186/s40537-020-00301-0
[32] Hancock, J. T., Khoshgoftaar, T. M., & Johnson, J. M. (2023). Evaluating classifier performance with highly imbalanced big data, *Journal of Big Data, 10(42)*, 1-31. https://doi.org/10.1186/s40537-023-00724-5