Design of Privacy Preservation Model for Data Stream Using Condensation Based Anonymization

*[1]Latha P, [2]Dr Thangaraj M

**[1]Reseach Scholar , Department of Computer Science, School of Information Technology, Madurai Kamaraj University, Madurai,India.

[2]Professor, Department of Computer Science, School of Information Technology, Madurai Kamaraj University, India.

*Corresponding Author Mail Id: lathaphd2023@gmail.com

*ORCID: 0009-0007-0210-2755

Abstract

In modern world, continuous data flows that may be analyzed in real-time and are created from multiple sources are referred to as data streams. As data streams become increasingly prevalent across various sectors, preserving the privacy of sensitive information while maintaining data utility is a significant challenge. Due to privacy concerns in most of the organizations, data are shared with third party or the public. In many cases, users are hesitant to provide personal information. This paper proposes a privacy-preservation model for data streams based on condensation-based anonymization. This model uses a condensation mechanism to reduce the granularity of data while ensuring that personally identifiable information and sensitive attributes are effectively anonymized. By leveraging a condensation approach, the model selectively compresses data, allowing for both privacy protection and the retention of essential data patterns for analysis. The condensation method minimizes lost information throughout the anonymization process in an effort to maintain the statistical features of the data. Using generalization and suppression approaches, k-Anonymity obscures identifying information in datasets, making it a useful tool for protecting individual privacy. The proposed model employs hybrid of K- Anonymity algorithm and condensation method to enhance the security and privacy of the data.

Keywords: Money Laundering, Privacy preservation, Data stream,K-Anonymity, Condensation Method.

1. Introduction

In recent era, many applications which need personal sensitive data of individuals. Thus people are more alarm about sharing their personal sensitive information due to upsurge of privacy intensions such as customer relationship management market based analysis, medical, financial. Etc., Privacy preserving is the procedure of hiding and defensive sensitive data of individuals. Data stream can be formulated as a continuous and transforming sequence of data that continuously arrive at a system to store or process. To protect such sensitive information, the anonymization process needs to be carried out and a concentration method to privacy preserving data mining methods are used in dynamic data update.

In today's digital world, financial landscapes are facing multiple issues such as financial crime threats, particularly money laundering looms huge[1, 2]. Money laundering is considered as transfer or conversion of property knowing that the property derived from offense activities. Generally, it can involve with three major steps called, placement, layering and integration[3]. Such type of illegal activity has crucial negative impacts on society and economy, as it put the financial sectors or individuals into critical at economic growth. Thereby, promotes corruption and crime ad undermined the efforts or legitimate private sectors [4][5].Significant obstacles stand in the way of privacy preservation in data streams, such as the concession between privacy and utility, the dynamic nature of data streams, the possibility of re-identification, insider threats, and implementation complexity[6]. These drawbacks highlight the requirement for more resilient and flexible privacy-preserving techniques created especially for the special qualities of ongoing data streams. Presently, manual reporting processes are often leading to delays, cumbersome and potential oversights[7]

This work proposes a new and adaptable approach for privacy preserving data mining which does not entail new problem specific algorithm since it maps the primary dataset into a current anonymised data set. These anonymised data closely match the features of the unique data counting correlations among the dissimilar dimensions and data are shortened into manifold groups of predefined size. Then they produce consistent pseudo-data by using statistics condensation tactic constructs forced clumps in data set and then result in simulated data from the statistics of these clumps. This technique assists in better privacy preservation as associated to other approaches as it uses pseudo data rather than adapted data.

In order to get k-anonymity, the condensation method groups' related records, suppresses or generalizes identifying characteristics, and minimizes information loss while maintaining identity protection. Anonymization

is the process of deleting personally identifiable information from datasets; it is necessary to preserve individual privacy in shared datasets and to comply with privacy rules. When combined, these ideas create a framework for protecting and managing personal information while extracting insightful information from ongoing data flows.

The major contributions of the present Privacy Preservation of Data Stream (PPDS) model areprovided below.

- To utilise a PPDS model for anomalies prediction with AMLdataset to improve the computation and accuracy in the proposed PPDS model.
- To deploy Optimized precision based Adaptive Beetle PSO algorithm for the process of feature extraction.
- To employ Modified Deep CNN-BiLSTM to accomplish money laundering classification to enhance security.
- To employ the proposed K-Anonymity algorithm and Condensation method for providing high level of privacy in optimized data for further classification process.
- To estimateefficiency of the proposed system with performance metrics like F1score, accuracy, precision and recall.

1.1 Paper Organization

The flow of the present model is given here: section 1 provides overview on the background of the proposed model. Section 2 reviews the conventional literatures related to anomalies detection and problem identification. Section 3 precisely describes about proposed methodology. Furthermore, section 4 provides table and graphical representation of data analysis. Section 5 discuss the results of the proposed model with traditional researches. Finally, section 6 concludes the present model along with future researches.

2. Related Work

This section provides about the analysis of various existing researches on anomalies detection along with other techniques for the prediction on anomaly classification system.

For an instance, the prevailing research has focused on detecting the illicit activities such as financing terrorism, Ponzi schemes and scams on the infrastructures of crypto currency at both transaction and account level. It has deployed Adaptive Stacked Extreme Gradient Boosting (ASXGB) model for the detection which has utilized Elliptic dataset. It has attained 0.961 of accuracy [13]. Contrarily, the conventional research has presented Amaretto which is active learning model for the detection of money laundering. It performed evaluation on synthetic dataset that stimulated the trading clients' profile. Due to the imbalanced data, accuracy is meaningless in this existing model [14]. Correspondingly, the prevailing method has examined the ML and DL techniques for AML in crypto currency. Random Forest (RF)[15], Deep Neural Network (DNN), Naïve Bayes (NB)[16] and K-Nearest Neighbours (KNN) have employed with bit coin elliptic dataset[17]. The experimental results have shown that the RF and NN classifier attained Better accuracy rates [18]. In parallel, Graph Neural Networks (GNN) has presented as solution for the detection of illegal transaction through the prevailing model. It has employed Node and Edge Neural Network (NENN) for classification process with transactions (edges) and bank accounts (vertices). Experimental results has attained better while transactions represented the nodes[19].

Similarly, the conventional research has developed an improved graph embedding technique [particularly for the detection of money laundering called GTN2vec. Through examining the records of Ethereum transactions, the algorithm concerns the money launderers' behavioural patterns comprehensively. The existing model has deployed a DNN technique for anomaly detection in Network Intrusion detection mechanism. It has utilised NSL-KDD dataset[25] and softmax layer has been used along with cross entropy loss mechanism for enforcing the conventional model in various classification comprising 5 labels like normal, DoS, R2L, Probe and U2L. It has attained better results [26]. Correspondingly, a conventional model has developed anomaly based intrusion prediction system for the IoT networks. Later, it has implemented utilising CNN in 1D, 2D and 3 D. It has validated on four datasets namely IoT Network Intrusion, BoT-IoT, IoT-23 intrusion detection and MQTT-IoT-IDS2020 dataset which have been attained satisfactory values of accuracy [8]. The existing research has presented effective anomaly detection mechanism which has been named as D-PACK. It comprises of unsupervised DL model like Auto encoder and Convolutional Neural Network (CNN) for auto profiling traffic patterns and extracting the abnormal traffic. It has used Mirai-CCU dataset and has attained satisfactory results [27]. Congruently, a 5-layer auto encoder based model for a network anomaly classification tasks. This model has evaluated on NSL-KDD dataset that outperformed. It has attained 90.61% of accuracy in detection performance [28]. In parallel, CSE-CIC-IDS2018 dataset has been employed to evaluate the traditional model. It has applied CNN, RNN, LSTM, DNN, CNN+RNN and CNN+LSTM models have been constructed to detect the anomaly attacks in network and has attained better results [29].

2.1 Problem Identification

This section describes that several conventional methods have been limited by predicting the anomalies which has several lacks.

- Though the conventional model has produced satisfactory results, it needs improvements on optimizing techniques [30].
- The traditional research has used imbalanced data from the used dataset, hence, accuracy metric is shown as meaningless [14][23].
- Accuracy is a significant metric to measure the efficiency of the model. However, several existing researches lacks in providing accuracy [11][13, 20][24][28].

3. Proposed Methodology

Money laundering accelerates the wide range of suspicious underlying crimes and offenses that ultimately threatens through financial line organisations or an individual. Therefore, the detecting of money laundering includes Manual reporting processes which are often leading to delays, cumbersome, potential oversights and vulnerable to various types of laundering-based attacks. To overcome the issue, several traditional models have focused on the money laundering detection however, lacks in accuracy, security and computation. To overcome this issue, the proposed model employs PPDS model based om privacy preservation techniques. The figure.1 shows the architecture of proposed DL model.

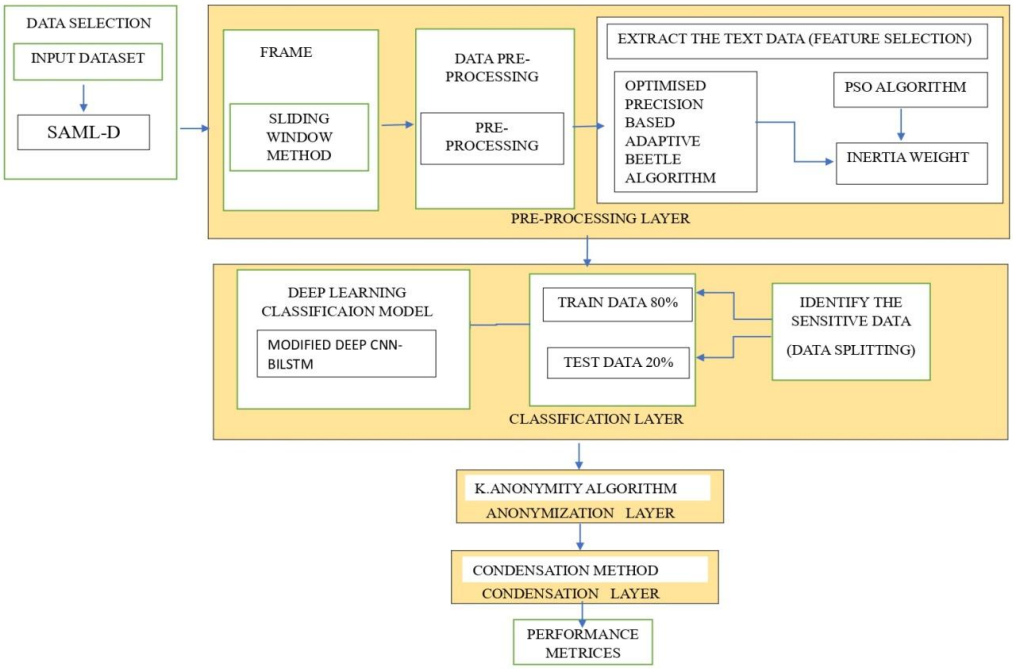

Figure.1 Architecture Diagram of Proposed PPDS (Privacy Preservation of Data Stream) Model

The figure.1 deliberates the proposed model comprises of selection of data, pre-processing of data, feature selection using optimised Adaptive Beetle PSO algorithm, splitting of data and eventually, classification process with Modified Deep CNN-BiLSTM with attention block, Anonymization and Condensation method for better data and performance metrics to find the results. The following sections deliberate the functions used in proposed model.

3.1Data Selection

A data stream serves as the artificial transaction monitoring dataset's source of data. This implies that the data is continuously created and updated, which is perfect for AML systems that must have the capacity to recognize suspicious activity immediately. The proposed PPDS model uses SAML-D dataset embraces transactions between two accounts and typologies. It incorporates of 28 typologies (where 17 suspicious and 11

normal) and 12 features. These features and typologies are selected form conventional datasets, researches and AML specialists' interviews. This dataset comprises of 9,504, 852 transactions, among them 0.1029% of transactions are suspicious. It also involved with 15 structures of graphical network to represent the flow of transaction among these typologies. Hence, the proposed model uses SAML-D dataset to provide advanced capabilities of security against money laundering.

3.2Pre-processing Layer

3.2.1 Data Pre-processing

Pre-processing is the important step to convert raw input data into required data. The proposed research utilises sliding window technique for handling continuous data streams andit is common in analysis of network traffic and detection systems. Sliding window method included in making overlapping data form a stream. Each window or a segment consists a fixed number of data points. The counter-based sliding window method preserves anonymity while allowing for real-time data stream monitoring timely insights without jeopardizing the protection of sensitive information by dynamically counting sensitive data points within a shifting time window. Besides, the pre-processing technique improves the detection performance of the proposed method and it is executed two significant data wrangling methods like checking missing values and data normalisation.

3.2.2 Feature Extraction-Optimized precision based Adaptive Beetle PSO algorithm

The proposed model used the feature extraction process for selecting the useful and most relevant data from the dataset to enhance the performance of the PPDS model. It involved in recognising and choosing the significant features which could contribute to the proposed detection model while reducing redundancy and noise. This method is aimed to improve the accuracy, interpretability and efficiency through focusing on relevant input data in AML system. For feature optimization process, the proposed method utilised Adaptive Beetle Algorithm which has a beetle group behaviour and incorporated PSO algorithm. Furthermore, the Beetle optimisation algorithm has been widely utilised for various optimisation problems which has better convergence speed and provide accurate results.

Pseudo code-Optimised Adaptive Beetle PSO algorithm

$\boldsymbol{Initialize the beetle Ai, and population velocity vel}$;
$\boldsymbol{Set step size, speed boundary, population size and maximum iteration and so on}$;
1. $Max_{Itr} \leftarrow Maximum number of iterations$
2. $n_{swa} \leftarrow Swarm size$
3. $for i = (1\ to n_{swa})\ \boldsymbol{do}$
    4. $A_i \rightarrow initialize the position of the ith particle$
5. $L^i_{bes} = T_i \triangleright \boldsymbol{used to obtain the positions of left and right antennas of beetles, respectively}$;
    6. $if f(L^i_{bes}) > f(G_{bes})\ \boldsymbol{then}$
        7. $G_{bes} = L^i_{bes}$
    8. $v_i \leftarrow initialize the velocity of the ith particle$
9. $\boldsymbol{while}(t < Max_{Itr})\ \boldsymbol{do}$
    10. $t = t + 1$
        11. $for i = (1\ to n_{swa})\ \boldsymbol{do}$
        12. $v_i \leftarrow Update the velocity of the ith particle using Equation\ 3.6$
        13. $A_{i(t)} = A_{i(t-1)} + vel_i \quad // Update the position of the ith particle$
        14. $if f(T) > f(L^i_{best})\ \boldsymbol{then}$
            15. $L^i_{best} = A_i$
            16. $if f(L^i_{bes}) > f(G_{bes})\ \boldsymbol{then}$
                17. $G_{bes} = L^i_{bes}$

$\boldsymbol{Output}$: G_{bes}

3.3 Classification Layer

3.3.1 Data Splitting

In Deep learning, this technique is used to eliminate the data over fitting issue. Essentially, PPDS uses the data splitting method to train the respective research where the train data is given to the proposed research for equipping the training stage parameters. After the training process, the test set data are measured to calculate the present model for handling the observations. In the proposed model, the original AML data is split into 2 sets in the ratio 80:20. The eighty percent of the new observations is utilised for the training and the remaining 20 percent of the data are applied for testing process to calculate the performance of the proposed research.

3.4Anonymization and Condensation

The hybrid model of anonymization and condensation is trained in the SAML-D dataset. This section deliberates the details of equations and classification mechanism of the proposed hybrid method and the Modified Deep CNN and BiLSTMmechanism.

3.4.1 Modified Deep CNN-BiLSTM with attention Block

Modified Deep CNN-BiLSTM with attention block is utilised to know the sequential features that are extracted to acquire the final prediction outcomes. The Modified Deep CNN-BiLSTM architecture is used to improve the data generalisation ability of the respective model.BiLSTMfunctions the sequential data in both forward and reverse directions. It captures the semantic dependences better in both directions. It ensures that it could better handle with huge data. The input gate allows the data, when the data is invalid the forget gate discards the input data from the input gate. If the data is valid, then it passes to the output gate. The number of layers of BiLSTMhave a crucial influence in detection accuracy.Theoptimal of activation function Rectified Linear Unit (ReLU) considerably impact the ability of network to learn complex patterns, generalize to unseen data, and differentiate among normal and anomalous data. For lengthy sequence data, Cyclic gate improved RNN models achieve better values.

$$\widetilde{h_t} = ReLU(W\mathrm{ta}_c + W\mathrm{tr}_\mathrm{t} + Wth_{t-1}) \quad (2)$$

$$h_t = \left(z \otimes \widetilde{h_t}\right) + \left((1-z)h_{t-1}\right) \quad (3)$$

Wherez, r update and reset gates while are$W(z, r)$ are corresponded to the weight matrices. At time $t, \widetilde{h_t}$ And h_t indicates the linear interpolation between current $\widetilde{h_t}$ and the activation states of previoush_t. Similarly, $\widetilde{h_t}$ the current memory gate having the activation function which is stated in equation (3). BiLSTM processes input data sequences in both directions along with 2 sub-layers. These layers compute forward and backward hidden sequences $\vec{h}and\overleftarrow{h}$ respectively. After that, they integrated to compute current hidden stateh_t. It specifically mentioned in the following equations (5), (6) and (7).

$$\vec{h} = LSTM(a_t, \vec{h}_t - 1) \quad (4)$$

$$\overleftarrow{h} = LSTM(a_t, \overleftarrow{h} - 1) \quad (5)$$

$$h_t = Wt^t\vec{h} + Wt^t\overleftarrow{h} \quad (6)$$

Where the function of LSTM denotes the nonlinear transformation of input data that are encoded in the corresponded hidden state of LSTM. Wt^tRepresents the coefficients of weights that corresponds to forward hidden state $\vec{h}$ and backward hidden state $\overleftarrow{h}$ in bidirectional LSTM respectively. A cyclic gate LSTM cell has single new gate which controls the output gate to the hidden state. $G_{t'}'and G_{t'}$ Represent the equations (7) and (8) applying in $g_{t'}$ as followed below.

$$G_{t'}' = g_{t'} \otimes W\mathrm{t}_z a_t + W\mathrm{t}_z h_{t-1} \quad (7)$$

$$G_{t'} = g_{t'} \otimes W\mathrm{t}_{ri} h_{t-1} + W\mathrm{t}_r a_t \quad (8)$$

3.4.2 Anonymization and Condensation

K-Anonymity is used to secure the individual privacy when still allowing for an effective detection of suspicious money transactions. The use of Anonymization methods, the k-anonymity is one of the popular techniques. Often, it requires data changing and hence necessarily affects the results of PPDS models trained in underlying data. It is used to achieve the data privacy preservation. It helps to mitigate the re-identification risk when permitting the data to be utilised for the analysis and models to recognise the suspicious activity. K-anonymity makes sure that no transaction could be identified uniquely when still preserving relationships and patterns in the data which could be used for money laundering detection. K-Anonymity is a privacy-preserving method that guarantees every record in a dataset may be identified from other records, at the very least. Creates generic data clusters that anonymised the dataset while weighing privacy concerns against information loss. It makes an effort to reduce the disparity between the anonymised dataset and the original dataset. By guaranteeing that the addition or deletion

of a single record in the dataset won't materially alter the model's output, it offers a mathematical assurance of privacy. In addition to upholding privacy regulations, they also call for striking a balance between model utility and privacy, frequently by carefully regulating the amount of noise or generalization added to the data. Similarly, Condensation method is an unsupervised learning technique which groups same transactions together to recognise the clusters of suspicious activity. It works through iteratively condensing data into lesser number of representative points, when preserving the relationships and structure in the data. Likely, it permits for anomalous patterns or transactions identifications which deviates from normal behaviour that can represent money laundering. K-Anonymity has comprised of three major stages such as initial population generation, selection and evaluation of the black hole and positions updating and stars fitness. Therefore, the following equation (1) is passed into the input of the proposed model.

$$a_{i(t+1)} = a_{i(t)} + rand * \left(a_{\{BH\}} - a_{i(t)}\right) i = 1,2,3,\cdots,N \quad (1)$$

Where $a_{i(t+1)}$ and $a_{i(t)}$ are known to be the locations of i^{th} star at iterations $t+1$ and$t$ respectively. Furthermore, $rand$ denotes the random number in the intervals (0, 1). $a_{\{BH\}}$Is the black hole location and N is the number of stars. In following, the condensation method constructed the constrained clumps in the data and creates simulated data from the clumps. It condenses the input data into various pre-defined size groups. Certain statistics is maintained for each group which has size$k$, an indistinguishability level of privacy securing method.

Algorithm for Anonymity Condensation Method

$AlgorithmCreateCondensedGroups(Indistinguish.Lvl.: k,$
$Database: D\ );$

$begin$
$whileDcontainsatleastkpointsdo$
$begin$
$RandomlysampleadatapointXfromD;$
$gr = a$
$Findtheclosest\ (k-1)\ recordstoXandaddtoG;$
$foreachattributejcomputestatisticsFs_{j(G)};$
$foreachpairofattributesi,jcomputeSc_{ij(G)};$
$Setn(gr) = k;$
$AddthecorrespondingstatisticsofgroupGtoH;$
$D = D - gr;$
$end;$
$Assigneachremainingpointin Dtotheclosestgroup$
$andupdatethecorrespondinggroupstatistics;$
end
$returnH;$
end

The co-variance matrix $C(G)$of $d * d$ has constructed for every group$G$. The ij^{th} entry of the matrix is among the attributes i and j of the records set in G.

4. Results and Discussion

This division represents and analyses the outcomes of proposed method that detects the launderings and typologies in network systems (data stream) toprovide privacy of sensitive data with proposed hybrid model called Anonymization and Condensation.

4.1Exploratory Data Analysis

In this section, proposed PPDS(hybrid) modelalong with existing methods are condensation technique and Anonymization technique are implemented through deep learning methods using python coding by applying AML dataset taken from https://www.kaggle.com/datasets/berkanoztas/synthetic-transaction-monitoring-dataset-aml.The dataset comprises a 9,504,852 transactions of which 0.1039% are suspicious. These transactions are representative are made in the domain of Anti Money Laundering.

In our work, the number of giga bytes is input to the proposed method. Based on that, we have to preserve the sensitive information. In our dataset, we consider the number of giga bytes in the range from 100, 200, 300,

400 and 500 for performing the experiments. In order to investigate the functionality of the proposed PPDS method, it is evaluated against with the two existing methods while taking into account with certain metrics. The performance of the proposed PPDS method and existing methods is determined using comparison tables and Zgraphs displaying its metrics values.

This section provides EDA of the AML dataset in the respective model. The EDA is applied to envision and inerpret the data in the generated dataset. Correlation matrix, statistical technique is utilised to calculate relationship among two variables in the dataset. Figure. 2 signifies the correlation matrix for the features in the proposed model.

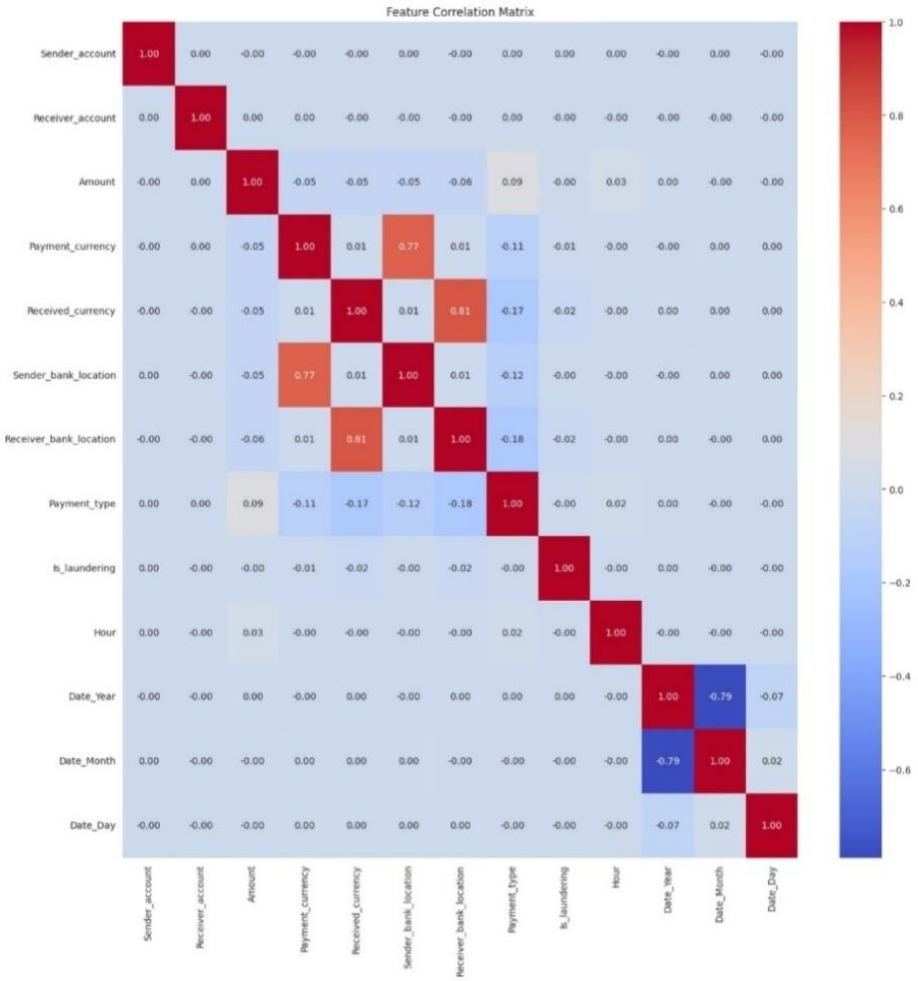

Figure. 2 Correlation Matrix for Features

Figure. 2 shows the feature correlation matrix which denotes the correlation coefficients among pairs of features (variables) in the dataset. Sender bank _location and Receiver Bank location shows high positive correlation (0.81). Therefore, the strong relationships between the features such as currencies and bank locations. Is laundering has no linear correlation with single feature, representing that the detection of money laundering could require modelling techniques. Congruently, the figure. 3 shows the class distribution.

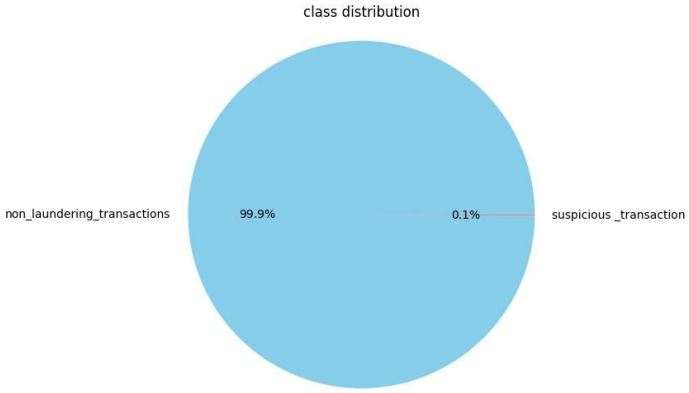

Figure. 3 Distribution of Class

In figure. 3, it understood that the distribution of class is of two types. Such as non_laundering_transactions and Suspicious transactions. Among these, suspicious transactions have 0.1% whereas, reaming part is covered with Non_laundering_transactions. Additionally, figure. 4 deliberates the typologies distribution.

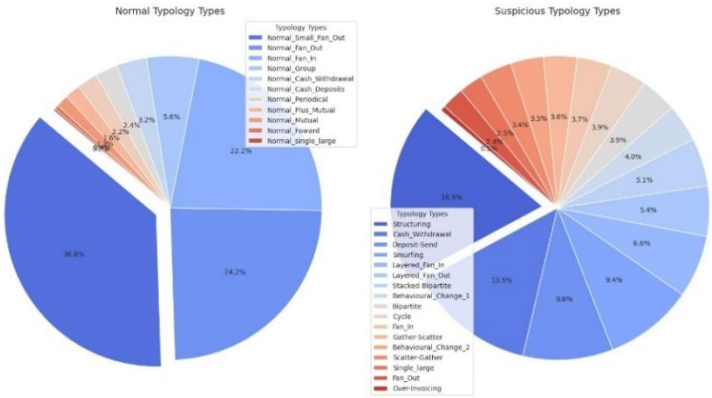

Figure. 4 Distribution of Typologies

Figure. 4 signifies the distributions of both suspicious and normal typology types. Likely, below figure. 5 depicts the Average Number of Transaction for a month.

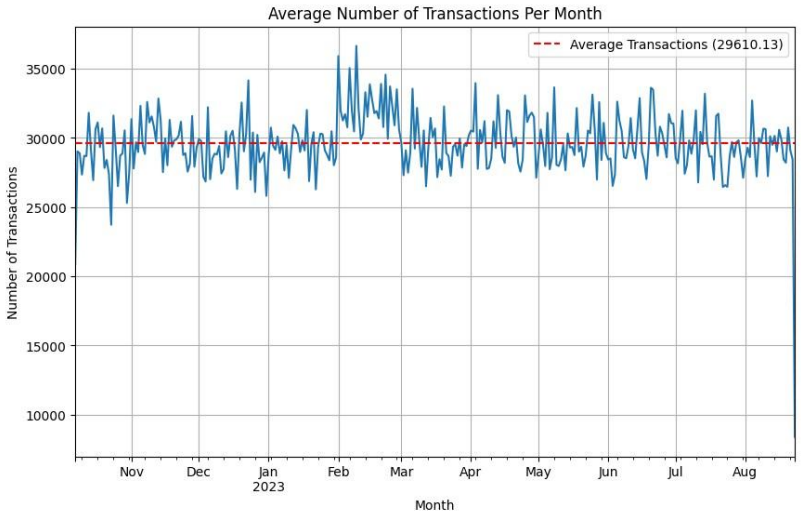

Figure. 5 Average Number of Transaction per Month

Figure. 5 exemplifies number of transactions per month in average. It is found to be 29610.13 transactions are made in November 2022 to August 2023. Among this, February 2023 is achieved with high number of transactions. Contrastingly, figure. 6 shows the laundering transactions.

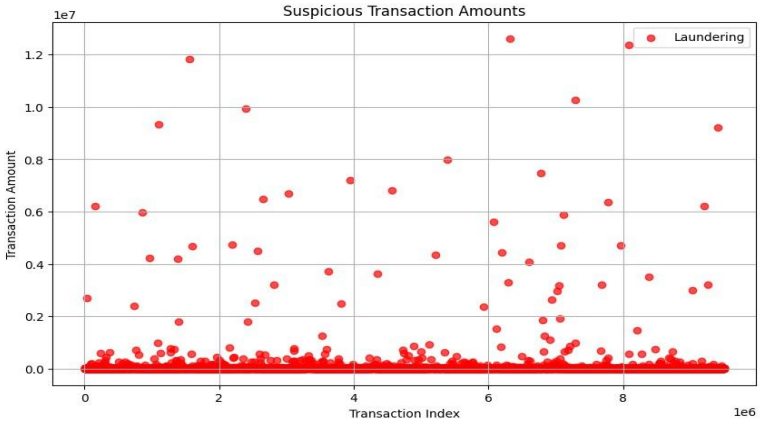

Figure. 6 Suspicious Transactions Amount

From figure.6, it understood that the laundering is happened in a smallernumber of transactions. While in high range of transaction amount contains less times of launderings made.

4.2 Performance Metrics

The metrics are predominantly utilized for observing the effectiveness of the respective PPDS model by applying various metrics such as recall, Precision, Accuracy and F1 score.

1. Recall Metric

Recall is utilized to examine the data percentage which is correctly detected in the respective model. The recall formula is defined in the following equation (9),

$$\text{Recall} = \frac{\text{True\_Pos}}{\text{False\_Neg+True\_Pos}} \tag{9}$$

2. Accuracy Metric

Accuracy is an important metric used to analyse the number of estimates which are approximately correct in the proposed model. The accuracy formula is illustrated in equation (10),

$$\text{Accu} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{10}$$

FP, TN, TP, FN Are False Positive, True Negative, True Positive and False Negative.

3. F1-Score Metric

F1-score is used to analyse the predictions in the proposed model that are made for positive class. The f1score formula is mentioned in equation (11),

$$\text{F1 Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall+Precision}} \tag{11}$$

4. Precision Metric

Precision metric is indicated as the method's covariance unit. It is resulted through suitably predictable cases (True_Pos) to the entire group of cases which have been precisely considered(True_Pos + False_Pos). It comprises repeatability and producibility of the capitals. Equation (12), depicts the formula for precision.

$$\text{Precision} = \frac{\text{True\_Pos}}{\text{False\_Pos+True\_Pos}} \tag{12}$$

4.3 Performance Analysis

Performance of proposed PPDS model is considered utilising several metrics like Precision, Recall, Accuracy and F1-score. Similarly, Confusion Matrix (CM) is utilised for recognising the performance. It summarises and visualizes the performance of the proposed algorithm which signifies how many predictions are 0 and 1 as per the class. Likely, ROC curve is a graphical representation of classification performance at various thresholds. It is widely utilised to measure the accuracy rates of the classification tests. The ideal ROC curve has an AUC of 1.0. Figure. 7 illustrates the ROC curve and CM for proposed PPDS method.

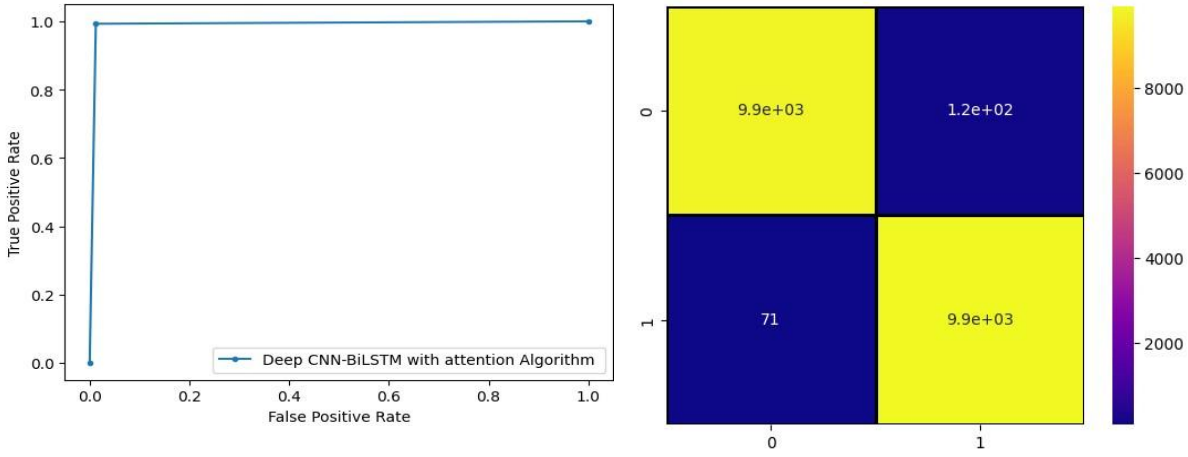

Figure. 7 ROC Curve and CM for Proposed CNN-BiLSTM

The upper left corner of the graph indicates the Specificity= 1 (false positive rate =0) and sensitivity =1 (True Positive rate=1). Therefore, above figure.7contains ROC curve for proposed model which is produced efficient results. The CM comprises 0 and 1 of laundering classification. It deliberates the attacks (launderings) in used dataset. Correspondingly, figure. 8 signifies the output of proposed model.

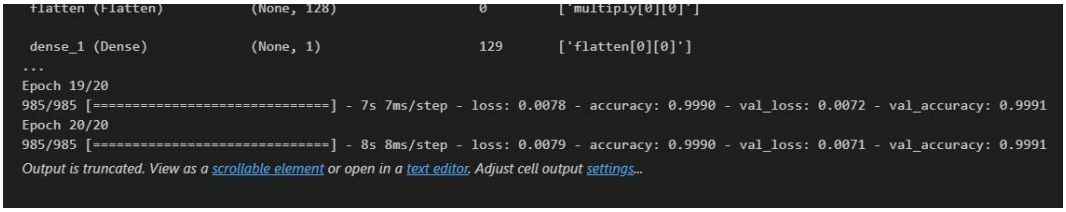

Figure. 8 Output of Proposed PPDS Model

Figure.8 shows the accuracy value and loss value at epochs count. The accuracy value is 0.9991 and loss value is 0.0072 at epoch 19 out of 20. Likely, the accuracy value is 0.9991 and loss value is 0.0071 at epoch 20 out of 20. Therefore, the proposed model is considered as the effective model in money laundering detection.

4.4 Comparative Analysis

The section exemplifies comparative analysis of the proposedPPDS model in accordance with various performance metrics.

Table 1 comparison of accuracy

Number of Giga Bytes	PPDS Proposed Model	Condensation technique	Anonymization technique
100	99	95	96
200	98.43	95.22	96.14
300	97.68	94.11	96.36
400	97.44	94.02	96.21
500	97.34	93.89	96.11

Table 1 depicts the performance outcomes of accuracy versus a number of Giga Bytes that are taken as input while utilizing the three different techniques including existing condensation technique**,**Anonymization Technique and proposed PPDS (hybrid)method. From experimental results, the proposed PPDS method

effectively performs to preserve the sensitive information with higher accuracy when compared to the other methods. The graphical illustration of accuracy is illustrated in **Figure 9** while taking into account the values shown in **Table 1.**

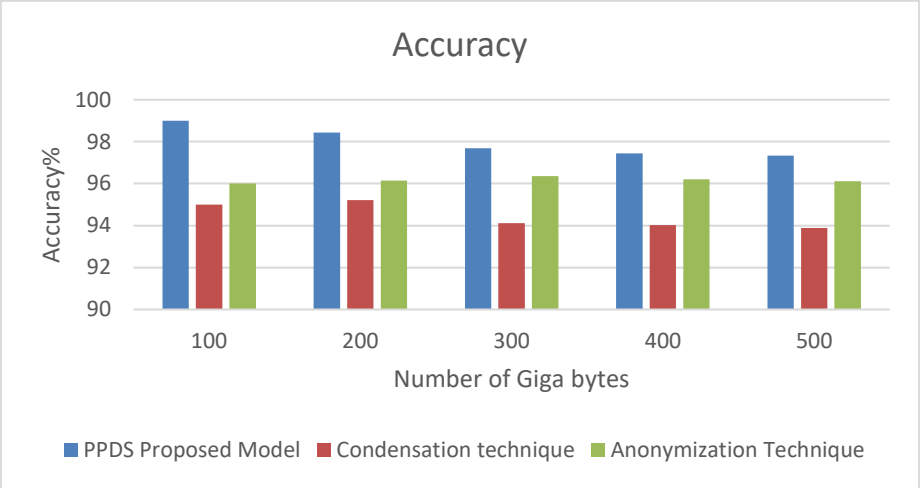

Figure 9 graphical representation of accuracy

Figure 9 illustrates the experimental results of accuracy versus the number of giga bytes, scope from 100 to 500. In graphical representation, the number of giga bytes is plotted on the horizontal axis, while the accuracy results are depicted on the vertical axis. To calculate the accuracy, 100 giga bytes were considered. By applying the PPDSmethod, an accuracy of 99% was observed. Similarly, using methodaccuracies of 96%, 97% were obtained, respectively. This process was repeated up to 500 giga bytes for each method, and the results were compared. The average of these compared results reveals that the accuracy of the PPDSmethod is significantly increased by 4% compared to condensation technique and 3% compared to Anonymization technique.

Table 2 comparison of precision

Number of Giga Bytes	**PPDS Proposed Model**	**Condensation technique**	**Anonymization technique**
100	0.99	0.95	0.96
200	0.986	0.952	0.962
300	0.975	0.948	0.957
400	0.972	0.942	0.952
500	0.971	0.944	0.95

Table 2 shows precision using the three methods. Contrary to two existing methods, PPDSachieved higher precision performance. By applying the PPDSmethod, a precision of 0.99 was observed. Similarly, using the methods of condensation and anonymizationtechniques value for precision is 0.97, 0.98 were obtained, respectively.

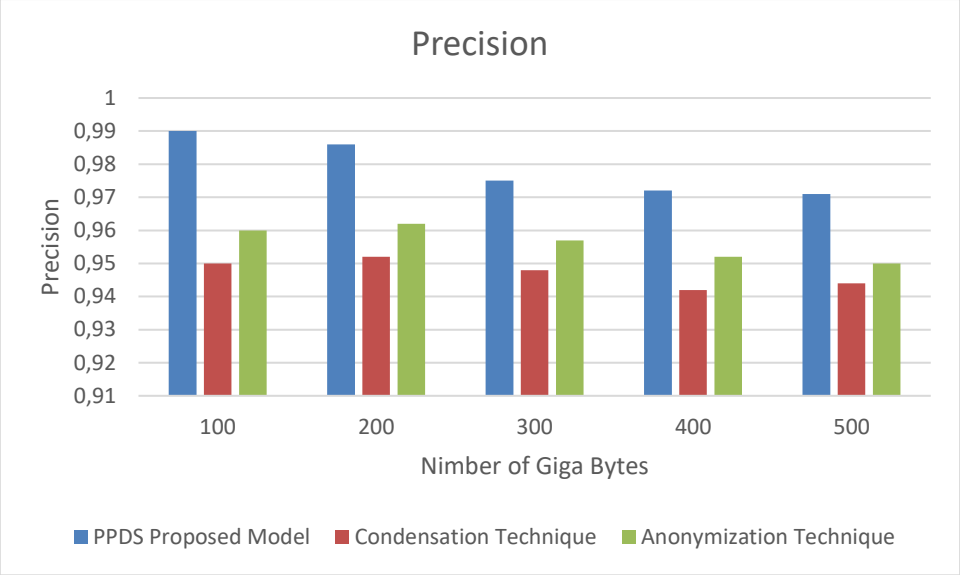

Figure 10 graphical representation of precision

Figure 10 depicts a performance analysis of precision in preserving sensitive information versus the size of data set in giga bytes taken in the range from 100giga bytes to 500giga bytes. Among the three methods, PPDSdemonstrates improved precision performance. On average, the comparison of five results reveals that the precision performance during using PPDS by 4% compared to condensation technique and 3% compared to anonymization technique.

Table 3 comparison of recall

Number of Giga Bytes	PPDS Proposed Model	Condensation technique	Anonymization technique
100	0.99	0.956	0.968
200	0.986	0.949	0.966
300	0.978	0.944	0.957
400	0.972	0.948	0.951
500	0.971	0.947	0.949

Table 3 illustrates the performance analysis of recall for three different methods with respect to the anti-money laundering (size is Giga Bytes) taken as input. The observed results indicate that the recall of the PPDS method is relatively higher than the existing methods.

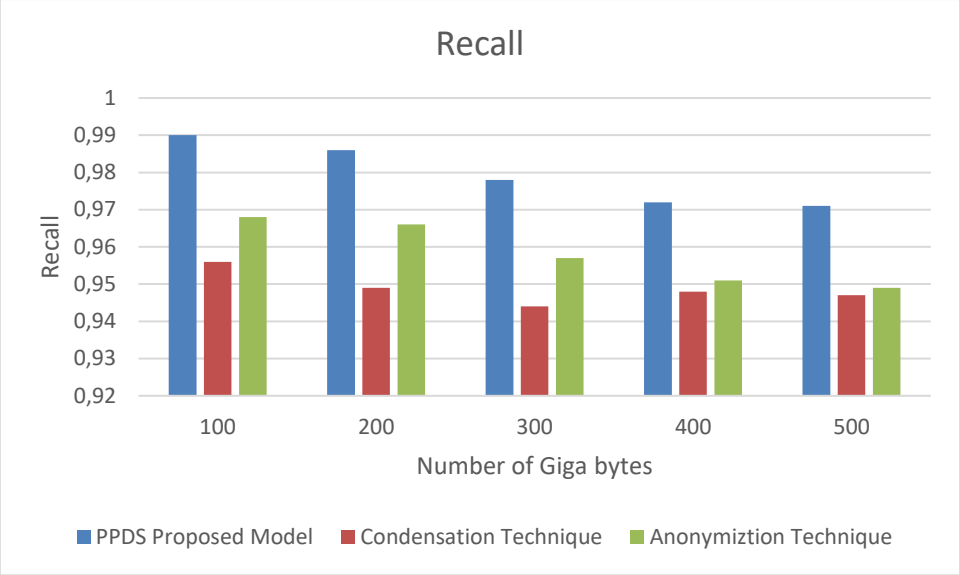

Figure 11 graphical representation of recall

Figure 11 displays the performance results of recall. The observed results indicate that the recall of the PPDS method is found to be higher than the two existing methods. Specifically, the PPDs method achieves a higher recall of 4%, 3% when compared to condensation technique and anonymization technique respectively. This is because PPDS achieves accurate relevant information retrieval results with higher true positives and minimal true negatives.

Table 4 comparison of F1-score

Number of Giga Bytes	PPDS Proposed Model	Condensation technique	Anonymization technique
100	0.99	0.96	0.95
200	0.98	0.961	0.955
300	0.982	0.965	0.949
400	0.97	0.962	0.943
500	0.971	0.96	0.94

As indicated in table 4, three distinct methodologies are implemented in the simulation process to determine the results in terms of F1-score. The average of 05 companion findings demonstrates that the proposed PPDS method greatly enhances the F1-score compared to the two existing methods. While taking into account the numbers found in Table 4, a graph is created between the quantity of anti-money laundering inputs and F1-score, as shown below in Figure 12.

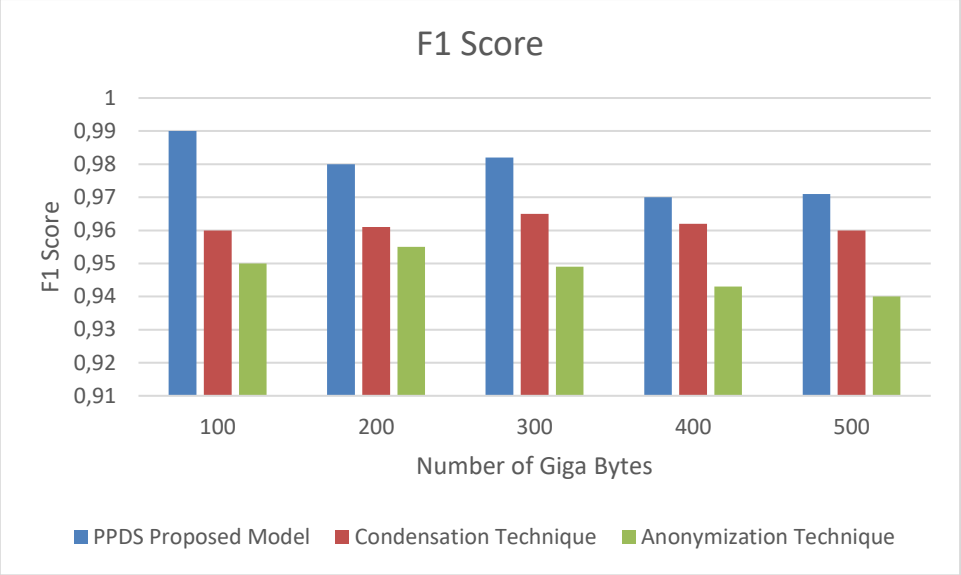

Figure 12 graphical representations of F1-score

The performance results of the F1-score for preserving sensitive information in data stream versus the size of data set in giga bytes are depicted in Figure 12. This improvement is attributed to the PPDS method enhancing both precision and recall in anti-money laundering. The average of five comparison results indicates that the F1-score of the PPDS method is considerably improved by 3%, 4% compared to existing methods condensation technique and anonymization technique respectively. This improvement highlights the effectiveness of the PPDS method in achieving a balanced trade-off between precision and recall for preserving sensitive information.

5. Conclusion

K-anonymity ensures that, through methods like generalization and suppression, each person's information cannot be distinguished from at least k−1 others, protecting data privacy. Data aggregation is one example of a condensation technique that summarizes individual records to mask identities while maintaining broad trends, improving privacy compliance, and protecting sensitive data. By means of recognising and categorising laundering anomalies, security teams can quickly identify and alleviate potential security risks, thereby reducing the likelihood of a successful attacks. Hence, an effective anomalies classification is significant to ensure the financial security. However, providing accuracy by human professionals is considered to be time-consuming and limited. To solve this challenge, the proposed research of hybrid of K- Anonymity algorithm and condensation method to enhance the security and privacy of the data. For classification, CNN-BiLSTM with attention block model focused to overcome the limitations and improves the detection accuracy for money laundering classification. The process of feature selection utilised Optimized Adaptive Beetle PSO algorithm. The proposed research utilised data stream to demonstrate the effectiveness. The proposed K-anonymity and condensation methods are used to achieve the data privacy preserving. In future, the present approach could be applied on various money laundering datasets for prediction purpose in order to enhance the entire financial security among unreliable accesses and malicious attacks. It could assist in further researches of effective money laundering detection models.

Declaration

Conflict Of Interest: The author reports that there is no conflict of interest

Funding: This research received no external funding.

Acknowledgement: None

Data Availability: Data sharing not applicable to this article as no datasets were generated.

Ethical statement for human participant: Not applicable for this research

Clinical trail number: Not applicable

REFERENCES

[1] H. Gandhi, K. Tandon, S. Gite, B. Pradhan, A. J. I. J. o. S. S. Alamri, and I. Systems, "Navigating the Complexity of Money Laundering: Anti–money Laundering Advancements with AI/ML Insights," vol. 17, no. 1, 2024.

[2] M. Tiwari, J. Ferrill, and D. M. J. J. o. A. L. Allan, "Trade-based money laundering: a systematic literature review," 2024.

[3] T. Pousette and A. Rosendal, "Explainable Anti-Money Laundering," 2024.

[4] F. J. E. E. A. i. A. D. A. Wójcik, "An Analysis of Novel Money Laundering Data Using Heterogeneous Graph Isomorphism Networks. FinCEN Files Case Study," vol. 28, no. 2, pp. 32-49, 2024.

[5] M. Nazzari, "THE BEHAVIORAL HETEROGENEITY OF MONEY LAUNDERING," 2024.

[6] C.-H. Tai and T.-J. Kan, "Identifying money laundering accounts," in *2019 International Conference on System Science and Engineering (ICSSE)*, 2019, pp. 379-382: IEEE.

[7] A. J. I. M. l. j. Kotagiri and C. Engineering, "AML Detection and Reporting with Intelligent Automation and Machine learning," vol. 7, no. 7, pp. 1-17, 2024.

[8] I. Ullah and Q. H. J. I. A. Mahmoud, "Design and development of a deep learning-based model for anomaly detection in IoT networks," vol. 9, pp. 103906-103926, 2021.

[9] R. J. A. J. o. A. I. Avacharmal and S. Development, "Leveraging Supervised Machine Learning Algorithms for Enhanced Anomaly Detection in Anti-Money Laundering (AML) Transaction Monitoring Systems: A Comparative Analysis of Performance and Explainability," vol. 1, no. 2, pp. 68-85, 2021.

[10] D. V. Kute, B. Pradhan, N. Shukla, A. J. I. J. o. S. S. Alamri, and I. Systems, "Explainable deep learning model for predicting money laundering transactions," vol. 17, no. 1, 2024.

[11] M. E. J. J. o. A. S. R. Lokanan, "Predicting money laundering using machine learning and artificial neural networks algorithms in banks," vol. 19, no. 1, pp. 20-44, 2024.

[12] J. F. M. Pazos, J. G. González, D. B. Lorenzo, and J. A. R. J. J. o. E. C. T. Morales, "Fraud Transaction Detection For Anti-Money Laundering Systems Based On Deep Learning," vol. 3, no. 1, pp. 29-34, 2024.

[13] D. Vassallo, V. Vella, and J. J. S. C. S. Ellul, "Application of gradient boosting algorithms for anti-money laundering in cryptocurrencies," vol. 2, no. 3, p. 143, 2021.

[14] D. Labanca, L. Primerano, M. Markland-Montgomery, M. Polino, M. Carminati, and S. J. I. A. Zanero, "Amaretto: An active learning framework for money laundering detection," vol. 10, pp. 41720-41739, 2022.

[15] R. I. T. Jensen and A. J. I. A. Iosifidis, "Fighting money laundering with statistics and machine learning," vol. 11, pp. 8889-8903, 2023.

[16] A. Kumar, S. Das, and V. Tyagi, "Anti Money Laundering detection using Naıve Bayes Classifier."

[17] M. Weber *et al.*, "Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics," 2019.

[18] J. Alotibi, B. Almutanni, T. Alsubait, H. Alhakami, A. J. I. J. o. A. C. S. Baz, and Applications, "Money Laundering Detection using Machine Learning and Deep Learning," vol. 13, no. 10, 2022.

[19] Í. D. G. Silva, L. H. A. Correia, and E. G. Maziero, "Graph Neural Networks Applied to Money Laundering Detection in Intelligent Information Systems," in *Proceedings of the XIX Brazilian Symposium on Information Systems*, 2023, pp. 252-259.

[20] J. Liu *et al.*, "Graph embedding-based money laundering detection for Ethereum," vol. 12, no. 14, p. 3180, 2023.

[21] M. Çağlayan and Ş. J. G. U. J. o. S. Bahtiyar, "Money Laundering Detection with Node2Vec," vol. 35, no. 3, pp. 854-873, 2022.

[22] G. Yang, X. Liu, B. J. C. Li, and Security, "Anti-money laundering supervision by intelligent algorithm," vol. 132, p. 103344, 2023.

[23] N. Bakhshinejad, "A Graph-Based Deep Learning Model for Anti-Money Laundering," 2023.

[24] T. Senior, "Enhancing Anti-Money Laundering Efforts with AI and ML A Comprehensive Approach to Financial Crime Prevention," 2021.

[25] R. A. R. Mahmood, A. Abdi, and M. J. B. S. J. Hussin, "Performance evaluation of intrusion detection system using selected features and machine learning classifiers," vol. 18, no. 2 (Suppl.), pp. 0884-0884, 2021.

[26] Z. J. B. S. J. Hussien, "Anomaly detection approach based on deep neural network and dropout," vol. 17, no. 2 (SI), pp. 0701-0701, 2020.

[27] R.-H. Hwang, M.-C. Peng, C.-W. Huang, P.-C. Lin, and V.-L. J. I. A. Nguyen, "An unsupervised deep learning model for early network traffic anomaly detection," vol. 8, pp. 30387-30399, 2020.

[28] W. Xu, J. Jang-Jaccard, A. Singh, Y. Wei, and F. J. I. A. Sabrina, "Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset," vol. 9, pp. 140136-140146, 2021.

[29] Y.-C. Wang, Y.-C. Houng, H.-X. Chen, and S.-M. J. S. Tseng, "Network anomaly intrusion detection based on deep learning approach," vol. 23, no. 4, p. 2171, 2023.
[30] L. Ashiku and C. J. P. C. S. Dagli, "Network intrusion detection system using deep learning," vol. 185, pp. 239-247, 2021.
[31] P. Jing, Y. Gao, and X. J. a. p. a. Zeng, "A Customer Level Fraudulent Activity Detection Benchmark for Enhancing Machine Learning Model Research and Evaluation," 2024.