International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Least Square Privacy Preserving Technique for Intrusion Detection System

Krishna Mandala, Sobhan Babu Kappala, Uma Shankar Rao Erothi

Abstract:*Network intrusion is a foremost growing concern threat in the cyberspace, which can be damage the network architecture in a multiple ways by modifying the system configuration/parameters. Hackers/Intruders are familiar with signature based intrusion detection models and they are making successful attempts to crash the networks. Hence, it is necessary to preserve user privacy on intrusion data. PPDM techniques forms a necessary but existing techniques such as Encryption, Perturbation, Data Transformation, Normalization, L-Diversity, K-Anonymity methods forms excessive generalization and suppression problems. In this paper, LSPPM distortion technique using Least Square Method with ensemble classification model have been proposed for providing efficient privacy preservation on intrusion data. The proposed methodology is validated on benchmark NSL_KDD intrusion dataset. A comparative analysis of NSL_KDD class attributes is performed for better classification in terms of accuracy, FAR, F-Score and time taken to build LSPPM-NIDS. The experimental results of state-of-art PPDM methods are also analyzed before and after distortion, and privacy measures to ascertain the degree of privacy offered.*

Keywords:*Network Intrusion Detection System, PPDM Techniques, Least Square Method, NSL_KDD, Ensemble classifier, Machine Learning*

I. INTRODUCTION

The advent of digital age in to our lives has brought forward many changes in internet and network technologies. Data sharing among network/third-parties is a major challenge in several areas, including medicine,marketing, communication, national security,education, finance etc., steered towards the development of different privacy preserving Data Mining (PPDM) techniques [1].The illegal access over computer networks to obtain sensitive information is increasing in spite of the presence of several security methods. Enforcing network security is important to protect data or information in the computer network against attacks from intruders.Any access to a system which violates the availability, confidentiality and integrity of the data is termed as an attack [2]. Several PPDM approaches include Anonymization based (K-Anonymity & L-Diversity), Perturbation based (Adding noise & Randomization), Cryptography based (Pseudonymization& Secure Multi-Party Computation),
Normalization etc., have been studied to upheld privacy in the data.The existing approaches often very complex and time-consuming to execute and suffer from problems such as excessive generalization and suppression. In this paper, to protect sensitive data from individuals,the original values in the dataset are transformed using Least Square Method (LSM) named as Least Square Privacy Preserving Method (LSPPM). To allow the boundaries to be defined for each attribute, LSPPM has been applied for dataset with numeric (categorical/continuous) values. Instead of having different scalesamongattributes, LSPPMconverts the attribute values within the range-1.0 to 1.0. As many machine learning (ML) classifiers accepts only numeric data, the conversion process is necessary to improve the performance of classification model and it also needs minimum for pre-processing, training and testing time for ML, and also lower memory consumption compared to any other state-of-the-art methods shown in section VI. On the other side,it is very difficult to manually monitor the sheer volume of network connections for intrusion detection. NIDSis a major component for securingnetwork/information systems, it continuously monitors the logs of network traffic in real time to identify potential attacks (intrusions) in a network using appropriate ML algorithms.The NIDS architecture using anomaly-based approach is more effective in detecting known-attacks, but it leads to high false alarm rate (FAR), whereas signature-based approach detects only attacks which are pre-defined in its knowledge database. To overcome the problems of each approach makes research community as challenges to build intelligent NIDS framework. Many researchers proposed different ML techniques and tools for developing an effective NIDS usingNeural Networks,Support Vector Machines, Linear Regression, Decision Trees, K-NN, Naïve Bayes etc. The ML approaches have been engaged successfully earlier for detecting intrusions using DARPA'99, KDD_CUP'99, NSL_KDD etc., have been used as intrusion datasets [3]. These datasets includes four class of attributes shown in Table 4 and Table 5, millions of records with a variety of intrusions to provide a training and testing datasets for researchers shown in Fig.4 & Fig.5. The issues with using ML models on network-connection/intrusion-detection data includes heterogeneous i.e., symbolic, continuous and binary values. Hence, it is important to transform packet data into a format suitable for ML models, because most of the classifiers accepts only numeric type [4]. The "Curse of high dimensional nature of the intrusion datasets" may encounter many difficulties,

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Krishna Mandala*, Dept of Mathematics, RAGHU Institute of Technology, Visakhapatnam, India
Sobhan babu K, Dept of Mathematics, JNTUK, UCEN, Narsaraopet, India,
Uma Shankar Rao Erothi, Dept of CSE, RAGHU Institute of Technology, Visakhapatnam, India.

which could degrade the performance of the many traditional classifiers. To address this problem, the proposed architecture chosen ensemble model for NIDS shown in Fig. (1). To tackle aforementioned problems, this article proposed intelligent framework for both privacy preserving and intrusion detection named as LSPPM-NIDS. The remainder of the paper is organized as follows; Section IIreviews related work in the PPDM area. Section III includes methodology of the proposed LSPPM-NIDS architecture. Section IVexplains overview of NSL_KDD intrusion dataset used for experimentation. Section Vshows comparative results from state-of-art PPDM methods with our LSPPM. Finally, Section VIconcludes the features of LSPPM-NIDS and suggests for the extension.

II. RELATED WORK

Many research communities concluded that data perturbation methods are best for privacy preservation (PP)as follows: Muralidhar and Sarathy [5] compared linear and non-linear models for numerical converted data using data-shuffling and data-perturbation techniques based on statistical theory. Likewise, Privacy Preserving using normalization methods (Min-Max, Z-Score and Decimal Scaling) on four benchmark datasets from UCI ML Repository is compared w.r.t accuracy and Data Distortion measures by Jain and Bhandare [6]. Similar Data distortion measures can be used on a novel data transformation technique was proposed by Erothi and Rodda [1].The experiments were conducted on NSL_KDD intrusion dataset as well as six other UCI ML repository. The effectiveness of proposed method on intrusion domain is evaluated before and after transformation. To achieve privacy on sensitive categorical data in case of clustering, Rajalaxmi and Natarajan [7] proposed a hybrid data transformation (HDT) method for categorical data by combining Translation-and-Rotation (TR) and Scaling-and-Rotation (SR) method. The data utility of the proposed method is measured, in-terms of variance between actual and perturbed values. To solve the problem of privacy breach in regression analysis, Fang et.al [8] discussed various objective functions for perturbation methods, Differential Privacy Budget Allocation (DPBA) and mechanisms to reduce-noise in the process of building regression-models.Similarly, the distortion data within the range [0, 1] scale and Linguistic distortion values (Low, Medium, and high) using triangular membership function is presented by Erothi and Rodda [9]. The authors used the knowledge of linguistic values to define ECOC coding schemes. Experiments were conducted on highly imbalanced benchmark NSL_KDD dataset as well as six other UCI ML repository datasets. The attribute suppression-and-generalization technique for anonymization, Mohana et al [10] presented comparative result analysis of NaiveBayes classifier using k-anonymity (KA) and Particle swarm optimization (PSO) algorithm. The resultindicated that, KA outrages PSO in terms of accuracy, recall and precision. The disadvantages of KA, such as attribute disclosure, data utility, Identity disclosure, and the increase of computation complexity to achieve privacy is clearly outlined by Kumara et al. To preserve user privacy on benchmark NSL_KDD intrusion dataset, Rodda and Erothi [2] utilized anonymization data for ensemble classification model by combining decisions of C 5.0, SVM and NNs. The experiments were conducted on IBM-SPSS Premium modeler, the effectiveness of proposed ensemble model is tested using different evaluation measures. Bhandare [11] used tan hyperbolic (tanh) activation function for data distortion in privacy preserving data mining. This method normalizes an attribute value within the range [-1, 1] in a non-linear transformation. Experiments were conducted to evaluate the performance of tanh on four real life datasets. WEKA tool was used to test the accuracy before and after distortion and privacy parameters are measured using java code. A novel encoding method proposed by Vatsalan et al [12] for private information using Counting-Bloom-Filters (CBF) and scalable-protocol for privacy preserving record linkage. The experimental results conducted on standard datasets are analyzed in terms of privacy protection, scalability, and disclosure risk and linkage quality. Chew et al.[13] utilized truncation method to achieve privacy on sensitive information of the network traffic connection data. The proposed method is evaluated using WEKA tool on Guru KDD_CUP intrusion dataset. The performance of 10 ML classifiers by changing options of IP address using truncation method was performed. To give privacy breach guarantee under realistic network traffic flows, Riboni et al. [14] proposed a novel obfuscation technique for network flows collected from Italian Tier II Autonomous Systems. The experimental results from the proposed technique shows that the technique preserves the utility of network-flows. To improve the detection rate, balancing training samples, to increase the diversity of training samples for intrusion detection presented by Yang et al.[15]. This author proposed a novel intrusion detection system by utilizing improved Conditional-Variational-Auto-Encoder (ICVAE) with a deep neural network (DNN) namely ICVAE-DNN. The experimental results were compared with 9 state-of-the-art methods w.r.t Accuracy, DR and FAR. To identify important features in NSL_KDD intrusion dataset, Aggarwal and Sharma [16] performed a detailed analysis of NSL_KDD intrusion dataset w.r.t four class attributes shown in below Table 4.The experiments were conducted on different combinations of class attributes, to identify high detection rate and low FAR. Latha and Thangasamy [17] proposed a modified version of tan hyperbolic (tanh) function to avoid the complexity of the hampel estimator by using mean and standard deviation (SD).The experiments on multi-model biometric system w.r.t error measures such as False Acceptance Error (FAE), False Rejection (FRR), Equal Error Rate (EER), and ROC plots demonstrate the working of various normalization and fusion schemes. The appearance of masked data similar to original to the end user is presented by Sarada et al. [19].The authors discussed various practical solutions of data masking techniques such as Min-Max, S-Shaped Fuzzy function, Rail Fence Method and Map Range. The approaches for maintaining accuracy between original and the masked data for categorical and numeric is clearly outlined to overcome the limitations of traditional methods. Taxonomy based transformation for categorical and Fuzzy based PP for numerical attributes is presented by kumari et al [20].

The effect of ML models when class imbalance exits is clearly outlined in [21]. The author show results that the all traditional classifiers are good at detecting majority class but fails to detect minority class, i.e. U2R and R2L.

III. METHODOLOGY

Many real world problems in computer vision are n-dimensional. Modeling the real world problem in present scenario can be solved effectively using a well-designed mathematical approach, which is applicable for numeric type of data. Numerical methods for classification are best approach to find optimal solutions. Here, the input dataset D for the proposed method is viewed as a solution for linear equation $AX = B$. The original dataset D with n-dimensions are converted to numeric type (Assume authorized user).The non-numeric (categorical/nominal) attributes in D are replaced with integer values as shown in Table 1 and Table 2. The numeric dataset shown in Table 2 is considered as matrix A and its transformed matrix TA shown in Table3, both of which are of n*m size. Where n and m indicates number of rows and columns in a matrix. The values in original matrix A_{n*m} is transformed to obtain TA_{n*m}matrix by using least square method (LSM). This method is popularly used in linear algebra as a part of numerical analysis for many computer based applications. A model can be implemented based on the relationship between the independent variables $X_i = \{X_{i1}, X_{i2}, X_{i3}, \dots, X_{ip}\}$, and the dependent variable Y_i. Then, Y_i can be calculated from a linear combination of the input variables (X_i) as shown in eq. (1).

$$AX = B \quad (1)$$

The matrix representation for the above equation as shown below.

$$\begin{bmatrix} A_{11} & A_{12} \dots\dots & A_{1m} \\ A_{21} & A_{22} \dots\dots & A_{2m} \\ A_{n1} & A_{n2} \dots\dots & A_{nm} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_n \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \\ B_n \end{bmatrix}$$

WhereA_{ij}, denotes the values of i[th] row and j[th]column. Then, solution vector is calculate as eq. (2)

$$X = A^{-1} \times B \quad (2)$$

Since eq. (2) cannot be solved directly with their improper sizes. For example, if A matrix is of size 4*2, B matrix will be of size 4*1, but $(A^{-1})_{4*2} \times (B)_{4*1}$ is not possible. Hence, Eq.(2) is modified to Eq.(3) .

$$A^{-1} = (A^T . A)^{-1} . A^T \quad (3)$$

Let $= A^{-1}$. Substitute Eq. (3) in (2)

$$X = M \times B \quad (4)$$

Where X = $[c, X_1, X_2, \dots, X_n]$ is a solution vector of original matrix A for the prediction of trained or un-trained data as shown in Eq.(5).

$$y_n = c + X_1 * A_{i1} + X_2 * A_{i2} + \cdots . + X_i * A_{ij} \quad (5)$$

Here c represents intercept value and y_n indicates the response for the given independent variable X_i(row).

The prediction shown in Eq.(7) including Residual Error ($\pounds_{err}$) for the actual value (Y_i) and predicted value (y_i) as shown in eq.(6)

$$\pounds_{err} = \frac{1}{n}\sum_1^n (Y_i - y_i)^2 (6)$$

$$y_n = c + X_1 * A_{n1} + X_2 * A_{n2} + \cdots . + X_n * A_{nm} + \pounds_{err} \quad (7)$$

The transformed dataset TA is obtained by multiplying the coefficients of solution vector (X) by each column value in matrix A with sum of intercept and residual error as shown in Eq.(8). To obtain the original value from the distorted data as shown in Eq.(9).

$$TA_{11} = \frac{(X_1 * A_{11})}{(c + \pounds_{err})} (8)$$

$$A_{11} = \left(\frac{A_{11}}{X_1}\right) * (c + \pounds_{err}) (9)$$

$$TA = \begin{bmatrix} TA_{11} & TA_{12} \dots\dots & TA_{1m} \\ TA_{21} & TA_{22} \dots\dots & TA_{2m} \\ TA_{n1} & TA_{n2} \dots\dots & TA_{nm} \end{bmatrix}$$

For better understanding of data distortion, Table 1 (Original data), Table 2 (Discretized data) and Table 3 (LSPPM data) shows the input space conversions from pre-processing stage to ML model.

Table 1:Weather Dataset before LSPPM

Outlook	Temp	Humidity	Windy	Class
overcast	hot	High	FALSE	yes
rainy	cool	Normal	TRUE	no
rainy	mild	Normal	FALSE	yes
rainy	mild	High	TRUE	no
sunny	hot	High	FALSE	no
sunny	hot	High	TRUE	no

Table 2 :Weather Dataset after replacing nominal data

Outlook	Temp	Humidity	Windy	Class
1	5	7	9	2
2	4	8	10	1
2	6	8	9	2
2	6	7	10	1
3	5	7	9	1
3	5	7	10	1

Table 3:Weather Dataset after LSPPM

Outlook	Temp	Humidity	Windy	Class
-0.154	0.132	0.088	-0.220	2
-0.308	0.066	0.176	-0.440	1
-0.308	0.198	0.176	-0.220	2
-0.308	0.198	0.088	-0.440	1
-0.462	0.132	0.088	-0.220	1
-0.462	0.132	0.088	-0.440	1

The transformed dataset shown in table 3 is obtained from solution vector {2.249,-0.349, 0.150, 0.200, -0.499} and residual error is 0.0166.The distorted dataset thus obtained, is evaluated by popular classifiers. The overall process of proposedLSPPM-NIDS architecture is depicted in Fig. 1 and Fig. 2, and Algorithms 1 to 4. The detailed description of the data transformation/distortion process is presented in Algorithm 1, 2 and 3.The transformed dataset is then provided as input to the classification technique presented in Algorithm 4. The classification model thus obtained will be used to classify incoming traffic as normal or attack type. The performance of the LSPPM-NIDS is evaluated using various performance measures viz. accuracy, false alarm rate (FAR), and F-Score.

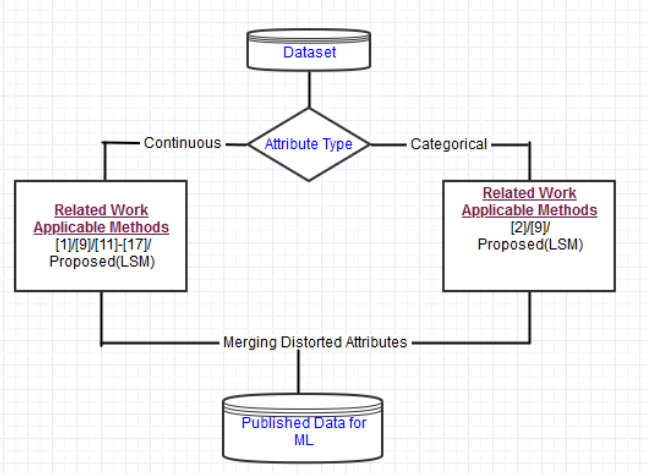

Figure 1:Attribute Transformation for Numeric and Categorical

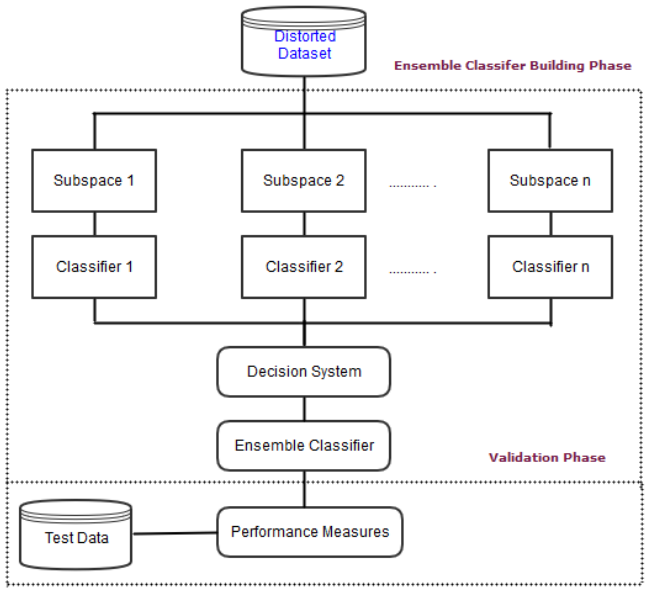

Figure 2: Architecture of LSPPM-NIDS for Classification

Algorithm 1: Proposed DistortionMethod (LSPPM)

Input : Original Dataset (D)
Output: Matrix for LSPPM (A)
Distorted Dataset (TA)
Evaluation Measures before and after

1 Read original dataset D
2 for each Attribute D_k in dataset D
where k indicates number of attributes
3 if D_k is Non-Numeric
4 update_Dk = {1,2,3…n} //Assign int value
5 end if
6 end for
7 return(A_{n*m}) for LSM
8 X_i = LSM(A_{n*m})
9 $\pounds_{err} = \frac{1}{n}\sum_{1}^{n}(Y_i - y_i)^2$
10 TA_{n*m} = Distort(A_{n*m})
11 Classification before and after
12 Evaluate measures

Algorithm 2: LSPPM(A)

Input : Original Dataset (D)
Output: Matrix for LSM (A)

1 Read input Matrix A
2 $AX = B$
3 Solve $X = (A.A^T)^{-1}.A^T \times B$
4 return(X)

Algorithm 3: Distort(A)

Input : Dataset (A_{n*m},X_i,$\pounds_{err}$)
Output: Transformed Dataset (TA_{n*m})

1 Read input Matrix A
2 for each i in n
3 for each j in (m-1)
4 $A_{ij} = [X_j * A_{ij} + c + \pounds_{err}]$
5 end for
6 Add row in TA_i
7 end for
8 return(TA_{n*m})

Algorithm 4 : Classification

Input - Training dataset S.,$S_i = \{X_i, Y_i\} \in S^n$, n is number of training samples
Output : Classifier model $\hat{Y}^b$ (X) = $[\hat{Y}^1, \hat{Y}^2, .. \hat{Y}^b]$
Final prediction , $y_i = [y_1, y_2, y_3, .., y_i]$

1 Construct classifier $C^b(X)$ for S
2 for each b: = 1 to n
3 $\hat{Y}^b(X_i) = C^b(X_i)$ //Training data
4 end for
Test data for classifier model
5 for each test instance X_i in S
6 $y_i = \hat{Y}^b(X_i)$ //invoke model to classify X_i
7 End for
8 $y_i = argmax_{y\in\{c_1,c_2,..,c_n\}} \sum_{b=\{1,2,..n\}} C^b(X)$
9 return (y_i) // final prediction
Evaluate Performance Measures

IV. PRIVACY MEASURES

To assess the workingof LSPPM, Data Utility (DU)and privacy measures (PMs) onNSL_KDD dataset is computed.To assess the performance of popular classifiersin weka librariesviz., Naïve Bayes (NB), Support Vector Machines (SMO), k-Nearest Neighbor(IBk) and Decision Tree (J48) on A_{n*m} and TA_{n*m} matrix is evaluated. The evaluation metrics of LSPPM-NIDS architecture remain un-changed with or with-out applying distortion techniques. Whenever data is converted toTA_{n*m}, DU and PMsmust be evaluated to appraise how the state-of-art methods (shown in section VI)affectsthe behavior of the original data.Generally, DU measure i.e., accuracy can be acceptablebetween A_{n*m} andTA_{n*m}. On the other hand, the value difference between A_{n*m} and TA_{n*m} matrix is computed using ValueDifference (VD).

Whereas the change in value positions are computed using RP, RK, CP and CK. The detailed description of these PMsare clearly outlined in [1].

V. DATASET DESCRIPTION

Since 1999; many researchers utilized KDD_CUP'99 dataset for misuse-based and anomaly-based detection approach. It has millions of records (normal and abnormal) collected from DARPA'98[14] dataset, which consists of 4GB tcp connection data. NSL_KDD [18] is another improvised version of KDD_CUP'99 intrusion dataset; it eliminates redundant connections from the training and test dataset. NSL_KDD consists of traffic connections belongs to normal and 22 attack types, each connection(data record) has 41 features (3-nominal,4-binary,34-numeric) categorized into Basic (BF), Content (CF), Traffic (TF) and Host (HF) features shown in Fig. 3 and Table 4. The number of connections utilized for different types of attacks (Dos, Probe, U2R and R2L) are presented in Fig. 4.The description of training (Fig. 5) and test instances clearly outlined in Fig. 6.

Table 4 : Summary of NSL_KDD Attribute class categories

Features	Description
BF (1-9)	duration, protocal_type, service, src_bytes, dst_bytes,flag,land,wrong_fragment,urgent
CF (10-22)	hot,num_failed_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creations,num_shells,num_access_files,num_outbound_cmds,is_hot_login, is_guest_login
TF (23-31)	count,serror_rate,rerror_rate,same_srv_rate,diff_srv_rate, srv_count, srv_serror_rate, srv_rerror_rate, srv_diff_host_rate
HF (32-40)	dst_host_count,dst_host_srv_count,dst_host_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_serror_rate,dst_host_srv_serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate
41	class

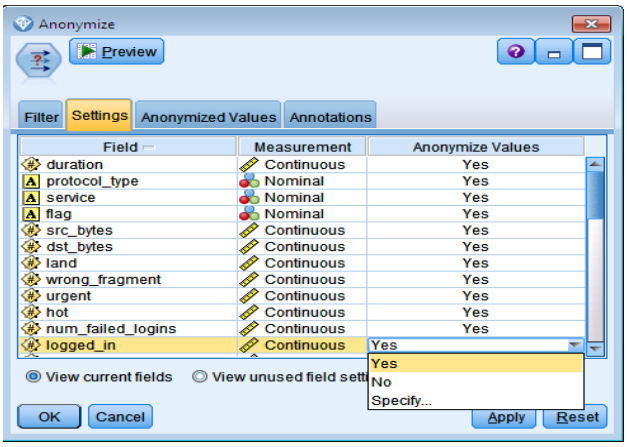

Figure 3 : Summary of NSL_KDD attribute types

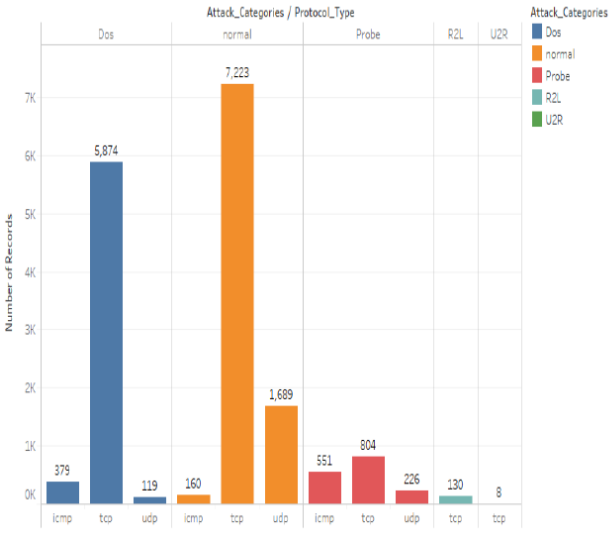

Figure 4 :Summary of Class Distribution in Overall NSL_KDD dataset

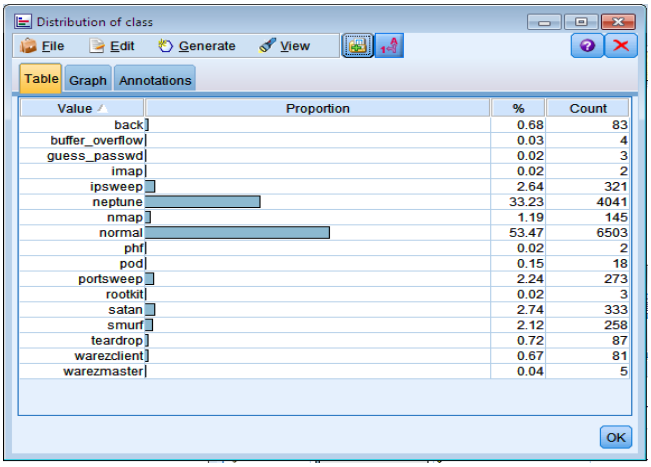

Figure 5: Summary of NSL_KDD Training Dataset

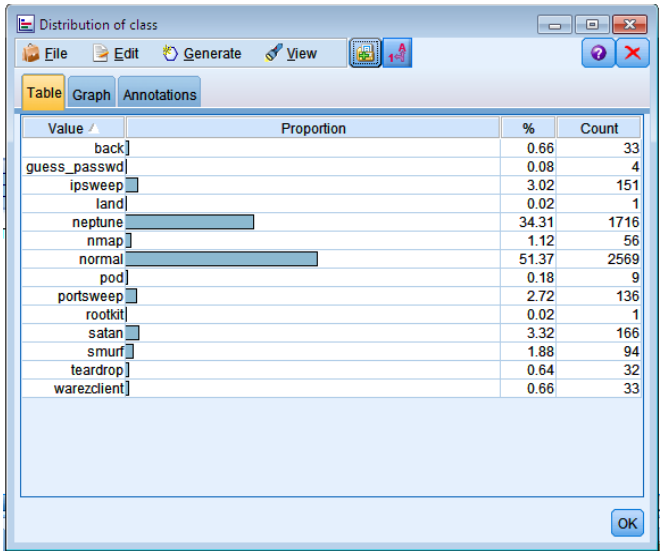

Figure 6: Summary of NSL_KDD Test Dataset

VI. EXPERIMENTAL RESULTS AND DISCUSSION

All the experiments were conducted on Intel® Core i3_5005U CPU @ 2.00GHz PC with 8GB RAM running on 64-bit Operating System and x-64 based processor. The LSPPM-NIDS is implemented using Java weka libraries, Python matplotlib for visualization, IBM-SPSS and Tableau for dataset analysis.The proposed LSPPM-NIDS approach is based on creating an ensemble classifier by training classifier on different subspaces of NSL_KDD attribute categories shown in Table 4.To evaluate the performance of LSPPM-NIDS are considered the measures provided in [3], from Eq. 2 to Eq. 8 in terms of confusion-matrix metrics TP, FN, FP and TN. The performance of the proposed method w.r.t NB, SMO, IBk and J48 on various state-of-the-art PPDM data distortion methods are compared. A comparative analysis is performed on various distortion methods such as Min Max (MM) [6], Map Range (MR) [19], Tanh [11], Anonymization (Anony) [2], Fuzzy [20] and find that the LSPPM distortion data exhibits best classification results shown in Fig. (7) to Fig. (9).

Fig.7 presents the variations of data utility measure (accuracy) obtained on NSL_KDD intrusion dataset. NB, SMO, IBk andJ48as base classifiers are evaluated before and after applying LSPPM with other existing methods. It can be observed that the performance of NB for MM, MR, Tanh and LSPPM is almost similar with slight variation and achieved higher accuracy compared to Anonymization and Fuzzy, but the minimum difference with original is achieved by Fuzzy method. Except for fuzzy, the other methods are almost near to original in the case of SMO and IBk.Whereas the maximum difference with original is shown in J48 classifier. Finally, the performance of SMO and IBk classifiers are best for most of the state-of-the-art PPDM methods compared to NB and J48. On examining the behaviors of considered PPDM input space with the considered classifiers, it can be concluded that SMO and IBk were best, due to presence of numeric type, whereas other classifiers failed to give similar performance w.r.t PPDM data utility measure.

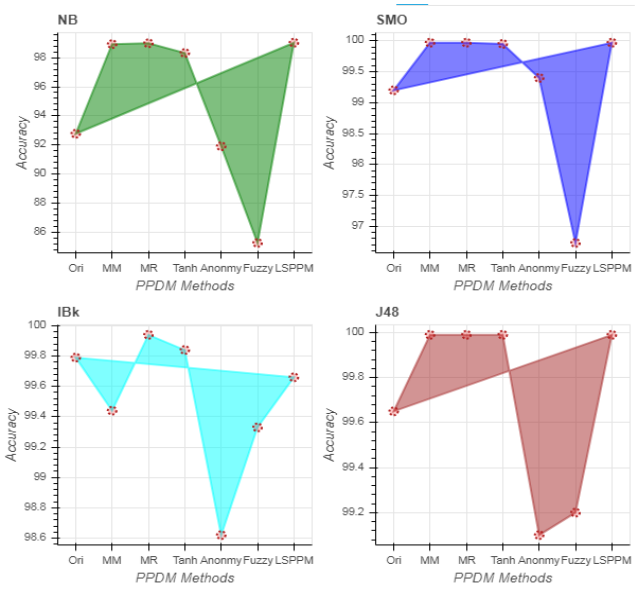

Figure 7: The change in Accuracy of state-of-art PPDM methods and Proposed LSPPM

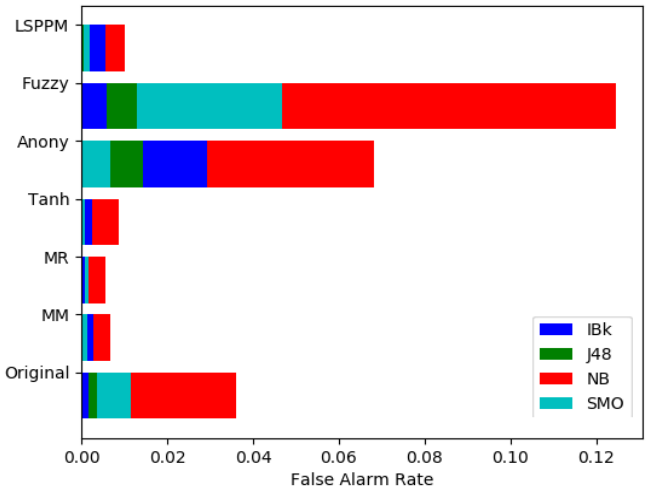

Figure 7: The change in FAR of state-of-art PPDM methods and Proposed LSPPM

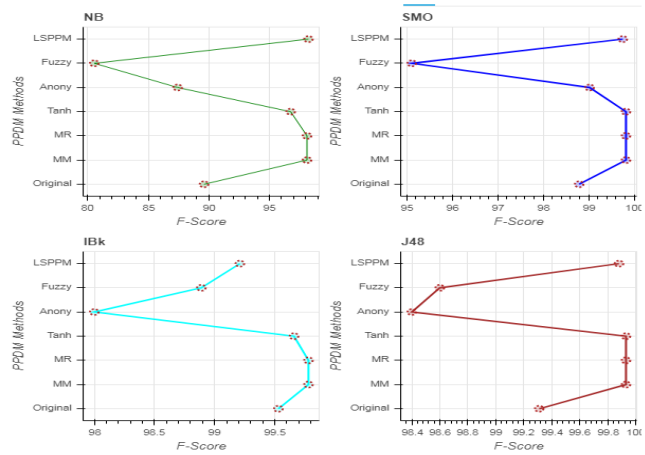

Figure 8:The change in F-Score of state-of-art PPDM methods and Proposed LSPPM

The change in F-Score indicated in Fig. 8, clearly observed that except for the IBk, LSPPM achieved highestF-Score when compared to other state-of-the-art methods. Generally, F-Score is best measure to study the performance of classifier when class imbalance is exist. The considered NSL_KDD dataset is highly imbalanced shown in Fig. (4). The F-Score with higher indicating that best in detecting minority classes. Similarly the performance of FAR is shown in Fig. (2). It is clearly observed that the lowest FAR is acheived by J48 for the most of the methods considered such as LSPPM, Tanh, MM and MR.The good-architecture for any NIDS should always need lowest FAR and higher detection rate. On the other hand, the IBk and SMO classifiers perform similar with slight changes in all the PPDM methods. The performance of NB classifier leads to high FAR in all the methods. Finally, concluded that after examining Fig. (7),Fig. (8) and Fig. (9), we can see that the LSPPM-NIDS with NB as base learner achieved poor results for all the PPDM methods and considered measures. The time taken from pre-processing to evaluation measures for LSPPM-NIDS is shown in Fig.(9). We have considered time as an important measure, because this model includes major steps such as data distortion methods (Fig. 1),

building ensemble classifier and testing ensemble classifier using majority voting scheme (Fig. 2), and evaluating measures.For any PPDM techniques for ML, the minor increase in time taken indicates that the distortion data are givingan additional overhead to the considered classifier. It is clearly observed that except for J48 for Anonymization, LSPPM distortion data consumed lower time for buiding ensemble classifier with individual base learners as NB,SMO, and IBk.

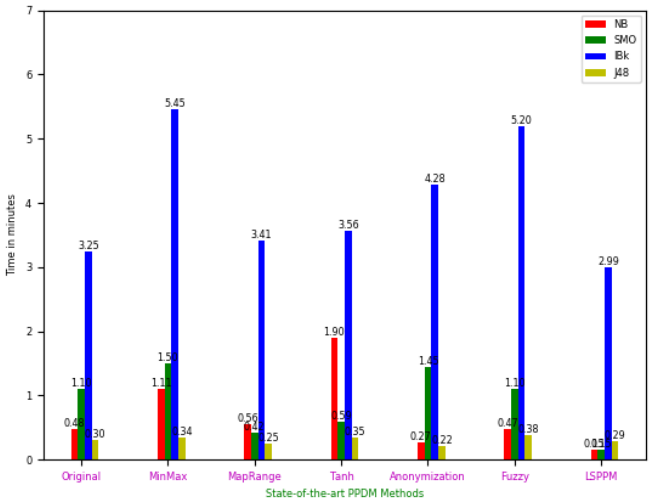

Figure 9:The change in time of state-of-art PPDM methods and Proposed LSPPM

Tabel 5 : Privacy measures of state-of-art PPDM methods and Proposed LSPPM

Methods	VD	RP	RK	CP	CK
Original	-	-	-	-	-
MM	0.4152	0.3313	0.7612	1.0	0.3321
MR	0.25	0.4434	0.6663	1.3233	0
Tanh	0.1856	0.1151	0.8782	0.3314	0.6566
Anony	-	-	-	-	-
Fuzzy	0.3535	0.4344	0.5456	0.0	1.0
LSPPM	0.4548	0.2212	0.7677	0.0	1.0

Table 5, shows the performance of PMs obtained on NSL_KDD intrusion dataset. The VD for the considered dataset with high value indicating, there is no correlation between A_{n*m} and TA_{n*m} values, and it offer more privacy to the intrusion data. It is clearly observed that LSPPM offers more privacy compared to other methods. Whereas, the RP values with minimum indicating, that the rank positions position of the NSL_KDD attributes in A_{n*m} and TA_{n*m} has not changed much due to considered PPDM methods. The Tanh method is obtained minimum value for the considered dataset. Similarly, the RK values close to 1 (Tanh, LSPPM and MM) indicating that the most of the values in all the NSL_KDD attributes maintained their ranks. On the other hand, the change of rank of attribute values close to 0 (Fuzzy and LSPPM)indicating best, it means that the rank of the average values of NSL_KDD attributes has not altered much during PPDM methods. The higher values of CK close to 1 (Fuzzy and LSPPM) indicate that most of the NSL_KDD attributes retain their ranks of the average value of attributes. The exceptions from different PPDM methods occurring in the values of VD, RP, RK and CP may be function mechanism. The function mechanism methods protect sensitive attributes from model inversion attack [ ], but they never considered the difference in sensitivity of each objective function. Finally, except for RP indicated in Table 5 shows that the proposed LSPPM provides minimal loss of information after considering values of VD,RK, CP and CK.

VII. CONCLUSION

This article analyzed a practical solutions of state-of-the-art PPDM methods for intrusion detection while preserving user privacy. The distorted network connections for ensemble training and testing the individual base classifiers viz. NB, SMO, IBk and J48 has obtained a similar results when compared to original w.r.t Accuracy, FAR and F-Score. The performance of LSPPM-NIDS for intrusion detection reduced the overall computational time. On toto, the LSPPM distortion outrages other considered methods w.r.t DU and PMs. The LSPPM with minimal loss of information from the observed experimental results motivated the research should be continued to reduce the overall computational time of LSPPM-NIDS using Feature Selection and Feature Extraction Techniques. The other popular intrusion datasets such as UNSW_NB15,TUIDS DDos, and SNMP_MIB should be considered for our next research to validate the consistency of LSPPM-NIDS architecture.

REFERENCES

1. Uma Shankar Rao Erothi, Sireesha Rodda, "Data Transformation Technique for Preserving Privacy in Data", International Journal of Computer Sciences and Engineering, Vol.6, Issue.5, pp.42-50, 2018.
2. Rodda, S., & Erothi, U. S. R. (2018). Network Intrusion Detection System to Preserve User Privacy. In Proceedings of International Conference on Computational Intelligence and Data Engineering (pp. 85-94). Springer, Singapore.
3. Rodda, S., & Erothi, U. S. (2018). A Roughset Based Ensemble Framework for Network Intrusion Detection System. International Journal of Rough Sets and Data Analysis (IJRSDA), 5(3), 71-88.
4. Salo, F., Nassif, A. B., & Essex, A. (2019). Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. Computer Networks, 148, 164-175.
5. K.Muralidhar and R.Sarathy, "Perturbation methods for protecting numerical data: Evolution and evaluation", Proceedings of the 5th Security Conference, 2006.
6. Jain, Y. K., & Bhandare, S. K. (2011). Min max normalization based data perturbation method for privacy protection. International Journal of Computer & Communication Technology, 2(8), 45-50.
7. Rajalaxmi, R. R., & Natarajan, A. M. (2008). An Effective Data Transformation Approach for Privacy Preserving Clustering 1.
8. Fang, X., Yu, F., Yang, G., & Qu, Y. (2019). Regression Analysis With Differential Privacy Preserving. IEEE Access, 7, 129353-129361.
9. Erothi, U. S. R., & Rodda, S. Fuzzy ECOC Framework for Network Intrusion Detection System.
10. Mohana, S., Sahaaya, S. A., & Mary, A. (2016). A comparitive framework for feature selction in privacy preserving data mining techniques using pso and k-anonymization. Iioab Journal, 7(9), 804-811.
11. Bhandare, S. K. (2013). Data Distortion Based Privacy Preserving Method for Data Mining System. International Journal of Emerging Trends & Technology in Computer Science, 2(3).
12. Vatsalan, D., Christen, P., & Rahm, E. (2017). Scalable multi-database privacy-preserving record linkage using counting bloom filters. arXiv preprint arXiv:1701.01232.
13. Chew, Y. J., Ooi, S. Y., Wong, K. S., & Pang, Y. H. (2019, February). Privacy Preserving of IP Address through Truncation Method in Network-based Intrusion Detection System. In Proceedings of the 2019 8th International Conference on Software and Computer Applications (pp. 569-573). ACM.

14. Riboni, D., Villani, A., Vitali, D., Bettini, C., & Mancini, L. V. (2012, March). Obfuscation of sensitive data in network flows. In 2012 Proceedings IEEE INFOCOM (pp. 2372-2380). IEEE.
15. Yang, Y., Zheng, K., Wu, C., & Yang, Y. (2019). Improving the Classification Effectiveness of Intrusion Detection by Using Improved Conditional Variational AutoEncoder and Deep Neural Network. Sensors, 19(11), 2528.
16. Aggarwal, P., & Sharma, S. K. (2015). Analysis of KDD dataset attributes-class wise for intrusion detection. Procedia Computer Science, 57, 842-851.
17. Latha, L., & Thangasamy, S. (2011). Efficient approach to normalization of multimodal biometric scores. International Journal of Computer Applications, 32(10), 57-64.
18. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (pp. 1-6). IEEE.
19. Sarada, G., Abitha, N., Manikandan, G., & Sairam, N. (2015, March). A few new approaches for data masking. In 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015] (pp. 1-4). IEEE.
20. Kumari, V. V., Rao, S. S., Raju, K. V. S. V. N., Ramana, K. V., & Avadhani, B. V. S. (2008). Fuzzy based approach for privacy preserving publication of data. International Journal of Computer Science and Network Security, 8(1), 115-121.
21. Rodda, S., & Erothi, U. S. R. (2016, March). Class imbalance problem in the network intrusion detection systems. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 2685-2688). IEEE.

AUTHORS PROFILE

Mr. M. Krishnapresently working as Associate Professor in Department of Mathematics at RAGHU Institute of Technology, Visakhapatnam, Andhra Pradesh, India.He received his M.Sc (Mathematics) degree in the year 2004. He received B.Sc (MPC) degree in the year 2002.His research interest Data Mining and Machine Learning Using Mathematics.

Dr. Sobhan Babu Kappala is a professor and HOD in Department of Mathematics, UCEN, JNTUK, Andhra Pradesh, India. He has published journals in international repute.

Mr. Uma Shankar Rao Erothiis currently an Assistant Professor in Department of CSE at RAGHU Institute of Technology, Visakhapatnam, Andhra Pradesh, India. He received his M.Tech (CSE) degree in the year 2013.He received B.Tech (CSE) degree in the year 2007. He published more than 6 papers in referred national and internal journals.His research interest Data mining, Image Processing and Data Structures.