Hybrid Model of Intrusion Detection using Neural Network and Neighbourhood Component Analysis

Pooja Agarwal[a], Dr. R. K.Srivastava[b]

[a] Research Scholar, Department of Computer Science, Dr Shakuntala Misra National Rehabilitation University, Lucknow, Uttar Pradesh, India.
[b]Head and Dean, Department of Computer Science, Faculty of Computer & Information Technology, Dr Shakuntala Misra National Rehabilitation University, Lucknow, Uttar Pradesh, India.

Abstract: The study of Intrusion detection system plays a vital role in research due to the increasing amount of security threats. In our proposed study, the training and testing process is executed on standard dataset NSL-KDD based on neighbourhood component analysis and neural network technique, where neural network used as a classification strategy and NCA as a feature selection technique to finalize dataset features. We have used neural network to classify the network traffic in attacks and normal conditions. The objective of this paper is to study the combined impact of feature selection and classification techniques. The study aims to improve the efficiency, recognition of malicious traffic exploratory assessment performed using the parameters, viz. false positive rate, detection rate & accuracy. The experimental results of proposed method show the improvement in the detection rate, accuracy as well as false positive rate.

Keywords: Network security, Intrusion detection system, Classification, NSL-KDD, Neighbourhood Component Analysis, False positive, Accuracy, Detection Rate, Neural Network, Feature Selection.

1. Introduction

The wide use of the internet and computer network have produced a serious issue of network security. Several intrusion detection models have been proposed to serve as a security feature. The first intrusion detection model was proposed by Denning in1987 [1], with an objective of how to build detection models efficaciously and accurately [2]. An intrusion detection system [3] is basically detecting, preventing and resisting unauthorized access to computer network. It is basically designed to analyse the anomalies of network traffic via different parameters and to achieve this target, it has to accomplish various phases [4]. The real issue with an intrusion detection system is the significant model overhead [5]. Accuracy turns into an essential concern [6], on the ground that intrusion detection system can recognize a wide assortment of interruptions progressively. The premier issue in planning an intrusion detection system is positioning and choosing the subset of appropriate features [7]. This paper proposes a methodology where feature selection using the Neighbourhood component analysis and neural network technique is combined for classification [8] of traffic on a computer network system.

In this paper, selection of the significant feature set is focused upon which affects the performance of the proposed model. In previous research, optimization and feature selection techniques have received due attention to find appropriate features. Feature selection is the process [9] which is used to reduce the vast input variable size while evolving an analytical model. Usually dataset includes insignificant, redundant, and strenuous features. Feature selection is used to select a subset of the prime features predicted on certain criteria. Feature selection methodology amazingly yields in choosing the noteworthy feature selection. To describe the dataset, it has a significant influence in picking the most relevant and noteworthy feature [10-11]. Classification involves finding rules that divide the data into distinct known groups. The input for the relegation is the training data set, whose class labels are already identified. Since the class field is identified, this type of relegation is known as supervised learning. A set of relegation rules are engendered by such a relegation process, which can be habituated to relegate future data and develop a better understanding of each class in the database. It distinguishes anomaly by breaking down the parameters of the system traffic information. Majority of times traffic is huge, and the contribution of specific feature builds the accuracy of interruption identification, while certain different features diminish the recognition precision. For an intrusion detection system, the selection of instructive feature is a contribution to the learning model in monitoring the events during the process of computer network.

In our paper, Segment II examines the framework of used technology in brief. Segment III depicts the technique executed, Segment IV portrays the model formation, Segment V examines the exploratory results gotten as well as comparing them with past existing result and Segment VI is the conclusion of the proposed concept.

2.Background

Intrusion detection is an issue on which several researchers worked but challenge to increase the reliability in detection rate and accuracy is still persist. In this section we are providing the brief technology background used in the proposed model.

2.1 Neighbourhood component analysis (NCA)

Neighbourhood component analysis (NCA) is a non-parametric and rooted technique for choosing highlights with the point of target of benefit from estimate precision of grouping calculations. It is a supervised learning method, used to classifying multivariate data into distinct classes as per the given distance metric over the data.

It is working to finding a linear transformation of input data by learning a distance metric. The average leave-one-out (LOO) [12] classification performance maximizes in the transformed space. The use of this technique is that the number of classes k can be determined as a function of matrix A, up to scale a scalar constant [13].

2.2 Neural Network

Neural Network is an enormously parallel appropriated data handling framework that has different performance factors which imitate organic neural systems of the human mind. The goal of an ANN is to generalize a relation in the form of Yj = F : (Xi). The way toward preparing a neural system includes tuning the estimations of the weights and biases of the system to enhance network execution [14].

Xi = x1; x2; ………………, xi (1)
Yj = y1; y2; ………………, yj (2)

where Xi is an i-dimensional input vector and Yj is a j-dimensional output vector. A neural system is indicated by its structure which displays the configuration of association between nodes, the method of ascertaining the weights and the activation function [15].

3.Proposed Methodology

The proposed model presents the neighbourhood component analysis and neural network based intrusion detection system model, which has been implemented to calculate the performance of the system, based on NSL-KDD dataset. Neural Network can implicitly sense complex nonlinear relationships between independent and dependent variables. Ability to learn and model non-linear relationships, which is useful to solve real time problems. The steps of the proposed method are as follows:

- Dataset taken from NSL-KDD repository.
- Pre-processing is applied to selected data.
- The feature selection method is applied for dimensionality reduction.
- Applied the Neural Network for training and testing on the NSL-KDD dataset selected features.
- Evaluated the performance of proposed model using performance metric e.g.: Detection rate, Accuracy and False alarm rate.

In our investigation, we apply the proposed strategy on NSL-KDD dataset [16]. We sorted the dataset based on target class and capabilities individually. In NSL-KDD, 20 percent training dataset, traits like protocol type, service, flag, and characterization of assault are in content arrangement. To convert them into numeric qualities, unique numeric code has been allotted to every possible estimation of the given characteristic. NSL-KDD is an advanced version of KDD 99 dataset [17]. Moreover, the quantity of records in the NSL-KDD prepares reasonable test sets and solved some KDD 99 dataset quandaries [18].

In our process, we analysed our proposed model on class 02 type attack, normal dataset. We analysed class 02 type attack, normal dataset and discovered encouraging results in particular outcomes. NCA performs feature selection by utilizing the labels and replications of standard dataset NSL-KDD training dataset [19]. In the process Neighbourhood component analysis starts with selecting features utilizing their weights and a relative threshold [20]. After the implementation we get the different class predicated data models as in Table 1.

Table 1. NSL-KDD Data Models of Selected Feature

S. No.	Data Model	Total Number of Selected Features	Class
1.	NCA25	6	2
2.	NCA28	9	2
3.	NCA214	15	2
4.	All	42	2

Interpretation of Table1.

In the above table, data models are presented which are selected from supervised NSL-KDD class 2 {normal, attack} data set.

4.Performance Criteria

In the paper we evaluated the experimental result on the basis of standard performance assessment criteria [21].

true positive: tp, false positive: fp, true negative: tn, false negative: fn.

Accuracy = tp + tn / (tp + tn + fp + fn)

Detection Rate = tp / (tp + fn)

False Positive Rate (FPR) = fp / (fp + tn)

where, false positive: the normal instances incorrectly classified as malicious.

true positive: represents the normal instances precisely classified as normal.

true negative: represents the anomalous instances precisely classified as malicious.

false negative: denotes the malicious instances incorrectly classified as normal.

Accuracy: it is the proportion of aggregate quantities of true positive in addition to add up to quantities of true negative separated by add up to number of false positive in addition to add up to number of false negative [22].

5.Implementation

In our test, we have utilized NSL-KDD 20percent preparing dataset for feature selection taken after by classification procedure to procure the more exact result to utilize and execution with examination of comes about have performed by utilizing MATLAB 2018b. During the method we watched the dataset models with diverse include set, removed by NCA based highlight choice method. NCA contains the dataset, fitting data, feature weights, and other parameters. Within the process it learns the feature weights using a inclining adjustment of NCA. Within the classification process to train our model we have used Neural Network Levenberg-Marquardt backpropagation algorithm [23]. Here Figure 1 is presenting the network structure of function fitting neural network.

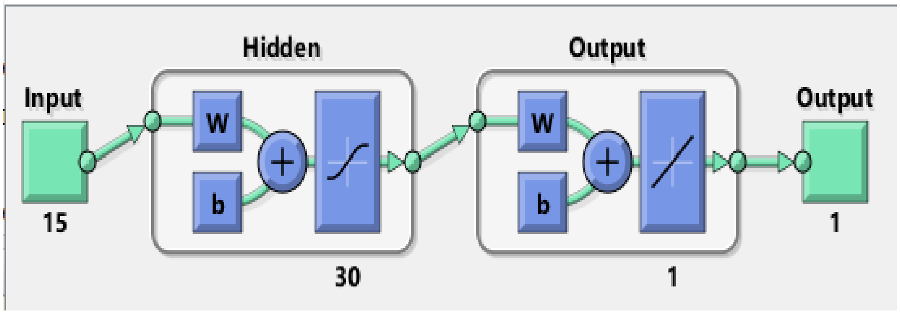

Figure 1. Function Fitting Neural Network

6.Result and Discussion

The experimental results are evaluated by benchmark measurements like false positive rate, detection rate & accuracy. In our study, we compared the effect of the proposed NCA and neural network model with similar model on the basis of false positive rate, detection rate & accuracy. We have categorized our result of proposed model of intrusion detection system in following four cases as follows [25]:

Case 1: In case 1, we had taken the NSL-KDD 20 percentage training data and divided it in 79:21 ratio used as training and testing respectively. On the same dataset we had applied NCA and by changed threshold, acquired the 6 features. Table 2

Table 2: Experiment Results - Applied on 6 Features dataset

Data Model	Class	Features	FP Rate	Accuracy	Detection Rate
NCA 25	2	6	0.0212	0.9807	0.9825
NCA 25	2	6	0.0248	0.9804	0.9850
NCA 25	2	6	0.0244	0.9800	0.9839

Interpretation of Table 2.

Table 2 showed the best three results of FPR, accuracy and detection rate which are taken from 6 features data model based experiment.

Case 2: In case 2, we have considered the NSL-KDD 20 percentage training data and divided it in 79, 21 ratio used as training and testing respectively. On the same dataset we have applied NCA and by changed threshold, acquired the 09 features. Table 3

Table 3: Experiment Results - Applied on 9 Features dataset

Data Model	Class	Features	FP Rate	Accuracy	Detection Rate
NCA 28	2	9	0.0143	0.9873	0.9887
NCA 28	2	9	0.0130	0.9873	0.9876
NCA 28	2	9	0.0183	0.9852	0.9883

Interpretation of Table 3.

Table 3 showed the best three results of FPR, accuracy and detection rate which are taken from 9 features data model based experiment.

Case 3: In case 3, we have considered the NSL-KDD 20 percentage training data and divided it in 79, 21 ratio used as training and testing respectively. On the same dataset we have applied NCA and by changed threshold, acquired the 15 features. Table 4

Table 4: Experiment Results - Applied on 15 Features dataset

Data Model	Class	Features	FP Rate	Accuracy	Detection Rate
NCA 214	2	15	0.0029	0.9967	0.9963
NCA 214	2	15	0.0037	0.9960	0.9956
NCA 214	2	15	0.0033	0.9956	0.9945

Interpretation of Table 4.

Table 4 showed the best three results of FPR, accuracy and detection rate which are taken from 15 features data model based experiment.

Case 4: In case 4, we have considered the NSL-KDD 20 percentage training data and divided it in 79, 21 ratio used as training and testing respectively. In the dataset we have taken all 42 features. Table5

Table 5: Experiment Results: Applied on 42 Features dataset

Data Model	Class	Features	FP Rate	Accuracy	Detection Rate
All	2	42	0.0012	0.9965	0.9945
All	2	42	0.0041	0.9961	0.9963
All	2	42	0.0033	0.9961	0.9956

Interpretation of Table 5.

Table 5 showed the best three results of FPR, accuracy and detection rate which are taken from 42 features data model based experiment.

After the experiment over all four case model, the values of false positive rate, detection rate and accuracy have obtained from the four different cases (Case 1, 2, 3 and, 4) as given in the Table 2, 3, 4 and, 5. Here the output in terms of false positive rate, detection rate and accuracy of four cases, establish the objective of proposed model. The different number of feature sets have used in proposed model to view possible deviation of result in terms of FP rate, detection rate and accuracy. One can see the performance of our proposed model that produced improved accuracy and intrusion detection rate as shown in Figure2, 3 and 4. The data used for training and testing was collected from https://www.unb.ca/cic/datasets/nsl.html. The result of various data models with best features are compared in Table 6.

Table 6: Comparison of different data models with best Accuracy and Detection Rate

Data Model	Class	Features	FP Rate	Accuracy	Detection Rate
NCA 25	2	6	0.0212	0.9807	0.9825
NCA 28	2	9	0.0143	0.9873	0.9887
NCA 214	2	15	0.0029	0.9967	0.9963
All	2	42	0.0012	0.9965	0.9945

Interpretation of Table 6.

Table 6 showed the best results of FPR, accuracy and detection rate which are taken from 6, 9, 15 and 42 features data model based experiment.

As shown in Table 6, It is clear by the comparison of various data models of class 2 proposed four cases of different data models using NCA and neural technique that NCA214 data model is giving better accuracy and detection rate among various iterations.The comparison chart of proposed model also confirmed it through Figure2 and 3.

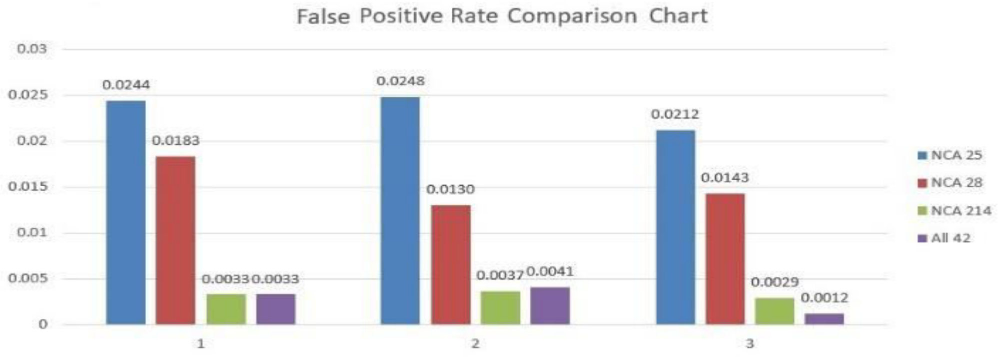

Figure 1. Proposed Model: False Positive Comparison Chart

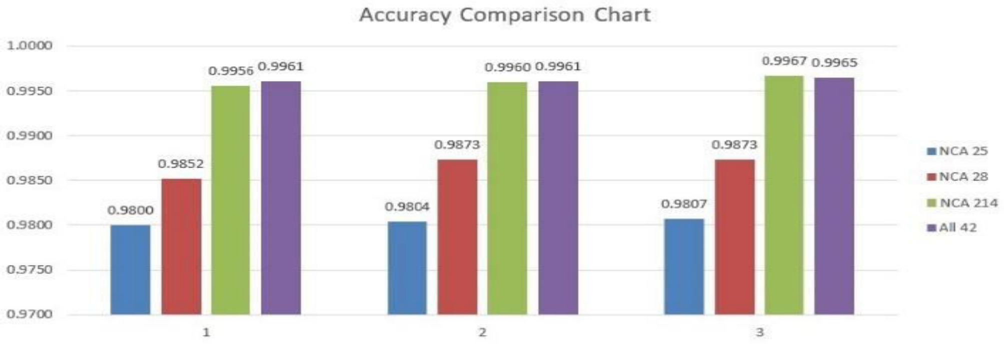

Figure 2. Proposed Model: Accuracy Comparison Chart

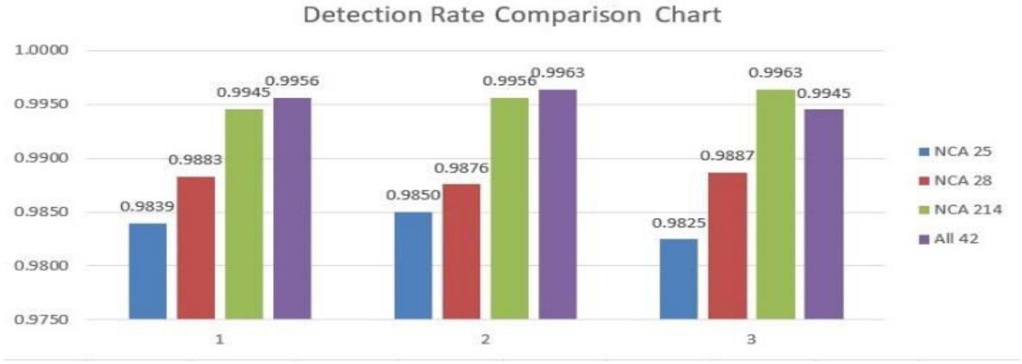

Figure 3. Proposed Model: Detection Rate Comparison Chart

It is also clear from Table 6 that when we are using 15 number of features, it is giving best result among considering 6, 9, 15 and 42 features, so proposed study shows best finding that data model NCA214 with 15 number of features is giving best results among others. Hence, it illustrates that the features should be in medium range that is 15 features which is neither low nor high in terms of selecting the features. We are also comparing our proposed model with similar research study of Subba Model [26], showing in following Table 7. In the Subba model, they used the feed forward along with back propagation algorithms besidevarieddifferentimprovement techniques to attenuatethe generalprocess overhead, whereas at an equivalent time maintain a high performance level.

Table 7: Comparison between Proposed Model and Subba Model Experiment Result

Data Model	Proposed Model			Subba Model		
	FPRate	Accuracy	Detection Rate	FPRate	Accuracy	Detection Rate
NCA 25	0.0212	0.9807	0.9825	0.0269	0.9540	0.9368
NCA 28	0.0143	0.9873	0.9887	0.0428	0.9626	0.9675
NCA 214	0.0029	0.9967	0.9963	0.0110	0.9856	0.9825
All	0.0012	0.9965	0.9945	0.0020	0.9888	0.9806

Interpretation of Table 7.

Table 7 showed the experimental results of Proposed Model and Subba Model and compared on the basis of FPR, accuracy and detection rate values.

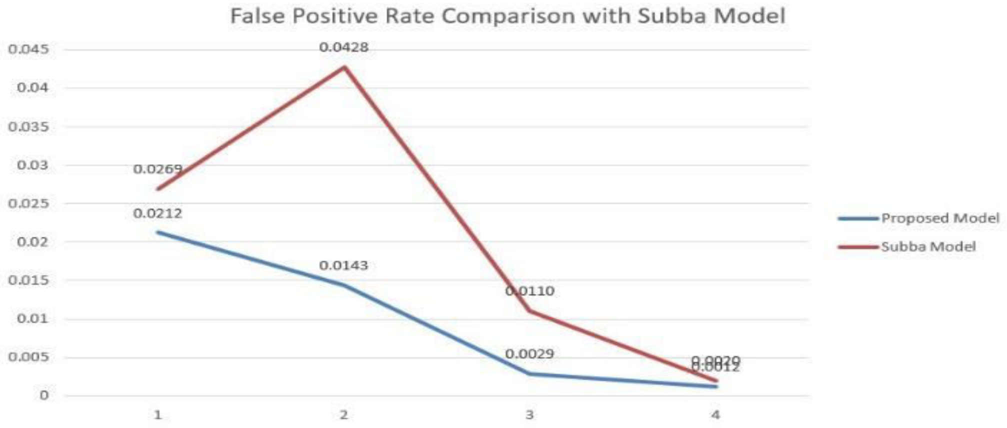

Figure 4. Proposed and Subba Model: False Positive Rate Comparison Chart

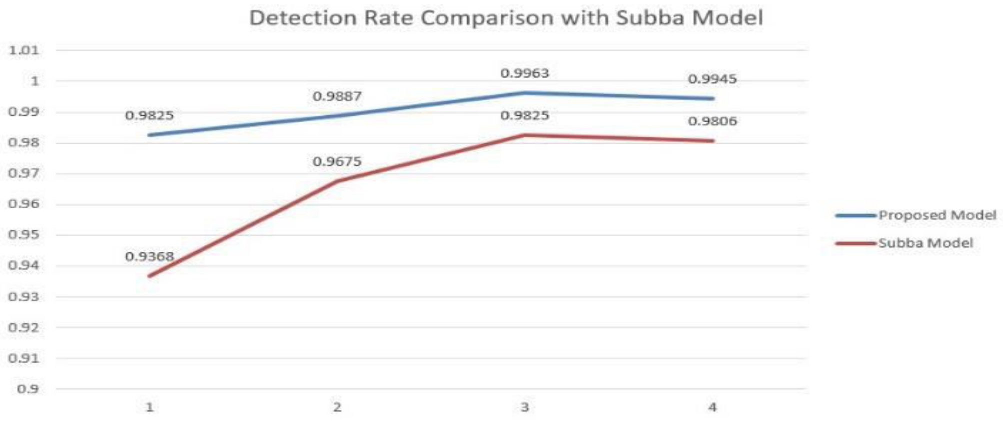

Figure 5. Proposed and Subba Model: Detection Rate Comparison Chart

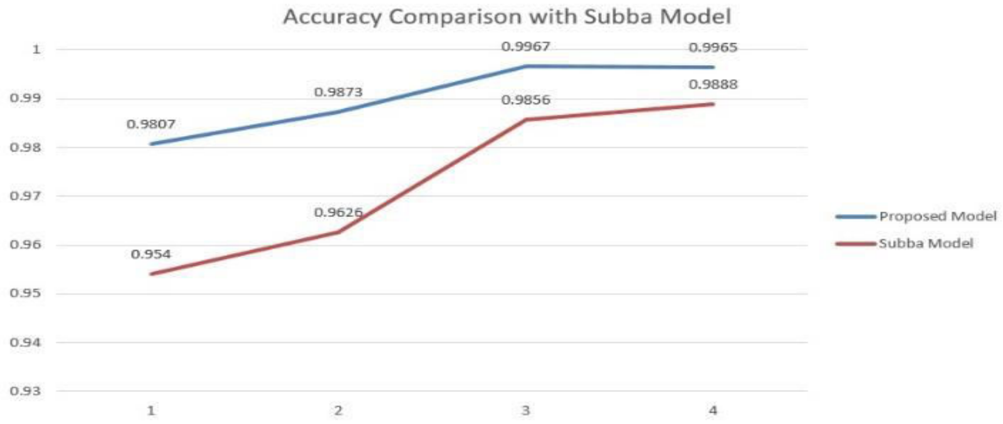

Figure 6. Proposed and Subba Model: Accuracy Comparison Chart

By comparing our proposed model with similar intrusion detection Subba Model shows the authority of our proposed model on basis of experimental results of detection rate & accuracy in all the considered data models e.g.: NCA25, NCA28, NCA214 and All. The main motivation of our proposed model have the improved experimental results of accuracy 0.9967 and detection rate 0.9963 for the NCA214 dataset, proves the effectiveness of model in comparison to Subba Model as shown in Figure5,6, and 7.

7.Conclusion

The motivation of our proposed model for intrusion detection classification model is the possibility of improvement in previous intrusion detection models. To fulfil the objective of our proposed model, NCA and artificial neural network technique has been used with the standard dataset for training and testing. We have calculated the FP rate, detection rate & accuracy by using class 02, different features and data models. Our study shows success to get the result in terms of better detection rate & accuracy. We have compared our proposed model with Subba Model and have got good success rate in comparison to earlier IDS model.

References

[1] Denning, D. E. (1987). An intrusion-detection model. *IEEE Transactions on software engineering*, (2), 222-232.

[2] Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, *36*(1), 16-24.

[3] Mukherjee, B., Heberlein, L. T., & Levitt, K. N. (1994). Network intrusion detection. *IEEE network*, *8*(3), 26-41.

[4] Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37.

[5] Pervez, M. S., & Farid, D. M. (2014, December). Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs. In *The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014)* (pp. 1-6). IEEE.

[6] Gao, N., Gao, L., Gao, Q., & Wang, H. (2014, November). An intrusion detection model based on deep belief networks. In *2014 Second International Conference on Advanced Cloud and Big Data* (pp. 247-252). IEEE.

[7] Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science*, *25*, 152-160.

[8] Malan, N. S., & Sharma, S. (2019). Feature selection using regularized neighbourhood component analysis to enhance the classification performance of motor imagery signals. *Computers in biology and medicine*, *107*, 118-126.

[9] Nie, F., Huang, H., Cai, X., & Ding, C. (2010). Efficient and robust feature selection via joint ℓ2, 1-norms minimization. *Advances in neural information processing systems*, *23*, 1813-1821.

[10] Hamid, Y., Sugumaran, M., & Balasaraswathi, V. R. (2016). Ids using machine learning-current state of art and future directions. *Current Journal of Applied Science and Technology*, 1-22.

[11] Ganapathy, S., Kulothungan, K., Muthurajkumar, S., Vijayalakshmi, M., Yogesh, P., & Kannan, A. (2013). Intelligent feature selection and classification techniques for intrusion detection in networks: a survey. *EURASIP Journal on Wireless Communications and Networking*, *2013*(1), 1-16.

[12] Goldberger, J., Hinton, G. E., Roweis, S., & Salakhutdinov, R. R. (2004). Neighbourhood components analysis. *Advances in neural information processing systems*, *17*, 513-520.

[13] Elisseeff, A., & Pontil, M. (2003). Leave-one-out error and stability of learning algorithms with applications. *NATO science series sub series iii computer and systems sciences*, *190*, 111-130.

[14] Hodo, E., Bellekens, X., Hamilton, A., Dubouilh, P. L., Iorkyase, E., Tachtatzis, C., & Atkinson, R. (2016, May). Threat analysis of IoT networks using artificial neural network intrusion detection system. In *2016 International Symposium on Networks, Computers and Communications (ISNCC)* (pp. 1-6). IEEE.

[15] Rezazadeh, N. (2017). Initialization of weights in deep belief neural network based on standard deviation of feature values in training data vectors. *Vol (6)*, (6), 708-715.

[16] Ingre, B., & Yadav, A. (2015, January). Performance analysis of NSL-KDD dataset using ANN. In *2015 international conference on signal processing and communication engineering systems* (pp. 92-96). IEEE.

[17] Aggarwal, P., & Sharma, S. K. (2015). Analysis of KDD dataset attributes-class wise for intrusion detection. *Procedia Computer Science*, *57*, 842-851.

[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] Tsai, C. F., & Lin, C. Y. (2010). A triangle area based nearest neighbors approach to intrusion detection. *Pattern recognition*, *43*(1), 222-229.

[20] Jirapummin, C., & Kanthamanon, P. (2002). Hybrid neural networks for intrusion detection system. In *Proceedings of the IEEK Conference* (pp. 928-931). The Institute of Electronics and Information Engineers.

[21] Thongkanchorn, K., Ngamsuriyaroj, S., & Visoottiviseth, V. (2013, October). Evaluation studies of three intrusion detection systems under various attacks and rule sets. In *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)* (pp. 1-4). IEEE.

[22] Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *Ieee Access*, *5*, 21954-21961.

[23] Sapna, S., Tamilarasi, A., & Kumar, M. P. (2012). Backpropagation learning algorithm based on Levenberg Marquardt Algorithm. *Comp Sci Inform Technol (CS and IT)*, *2*, 393-398.

[24] Ballabio, D., & Vasighi, M. (2012). A MATLAB toolbox for Self Organizing Maps and supervised neural network learning strategies. *Chemometrics and intelligent laboratory systems*, *118*, 24-32.

[25] AgarwalPooja, SrivastavaR.K (2021), “Intrusion Detection with Neighbourhood Component Analysis and Neural Network Classifiers”, *ICIC Express Letters.*

[26] Subba, B., Biswas, S., & Karmakar, S. (2016, March). A neural network based system for intrusion detection and attack classification. In *2016 Twenty Second National Conference on Communication (NCC)* (pp. 1-6). IEEE.