Performance Evaluation Based Cross Entropy on Intrusion Deteection using Deep Neural Networks

K.Y.S.S. Manjusha[1], Dr.K.V. Krishnam Raju[2]

[1]M. Tech student of Department of computer science and engineering, S R K R Engineering College, Bhimavaram, AP, India.
[2]Associate Professor of Computer Science and Engineering, S R K R Engineering College, Bhimavaram, AP, India.

Abstract: Recent days all the activities or work are done in the internet. So the frequencies of cyber security attacks are increasing rapidly. Intrusion detection system (IDS) is utilized to examine the system and keep attention towards the unauthorized activities in the network. While monitoring the network large amount of data should be managed and analyzed. To handle that huge amount of data, Deep learning algorithms predict best performance and increase the true positive measures. This work uses the KDD Cup99 dataset which consists of 41 features. Deep Neural Networks (DNN) is used as classifier with five hidden layers and Rectified Linear Unit (ReLU), sigmoid are activation functions which are used to improve the algorithm performance. By using this model, 92.5% accuracy and 99.8% true positive rates are obtained.

KEYWORDS: Cyber Security, Attacks, Intrusion detection system, deep learning, KDD Cup99.

Check for updates

DOI of the Article: https://doi.org/10.46501/IJMTST0708018

Available online at: http://www.ijmtst.com/vol7issue08.html

As per UGC guidelines an electronic bar code is provided to seure your paper

To Cite this Article: K.Y.S.S. Manjusha and Dr.K.V. Krishnam Raju. Performance Evaluation Based Cross Entropy on Intrusion Deteection using Deep Neural Networks. *International Journal for Modern Trends in Science and Technology* 2021, *7*, 0708013, pp. 97-105. https://doi.org/10.46501/IJMTST0708018

Article Info. Received: 05 July 2021; Accepted: 28 July 2021; Published: 02 August 2021

INTRODUCTION

Intruders are internal and external. Internal intruders will be within the organization and External intruders will be outside the organization. The Network is the source for entering the intruders and intrusions, to overwhelm that two safeguard measures are used those are Intrusion Detection System (IDS) and Firewall. To configure the Firewall and IDS there are two ways of placing positions. They are placing before and after the internet connection is established. For more accurate attack detection placing the IDS before the internet connection is considerable as shown in fig1 and fig2 [1].

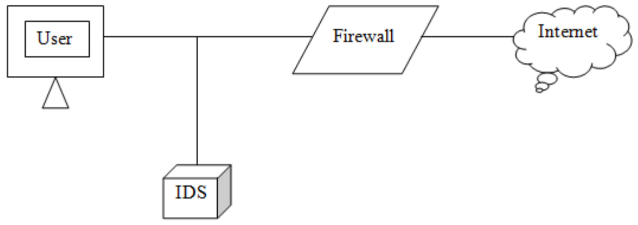


Fig 1: Role of Firewall based Network

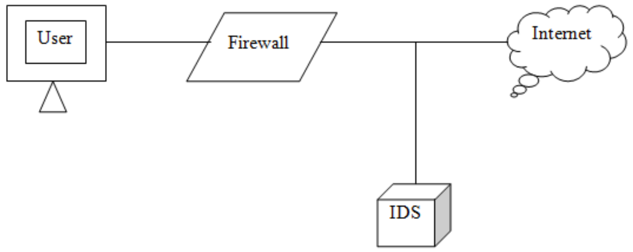


Fig 2: Role of an Intrusion detection system based Network

As shown in fig 3 the intrusion activities the IDS are divided into the Host-based Intrusion Detection System (HIDS) and Network-based Intrusion Detection System (NIDS). An IDS which has the property of network behavior is called NIDS. The network behavior consists of network devices such as tapes, switches, routers which are used to identify the attacks present in the network traffic. The IDS have some system activities which are in the form of log files that runs in the local host system to detect the intrusions is known as HIDS [2]. To combat the problem of HIDS and NIDS there are three solutions, one is the misuse detection which is very accurate for known patterns or signatures to find out the attacks in the network but cannot find out the new attacks [3]. The other method is anomaly detection which is used to find the unknown attacks with the help of known or existing patterns but while using this method false positive rates are high so to overcome this problem most of the organizations prefer a hybrid approach that binds misuse and anomaly detection methods [4].

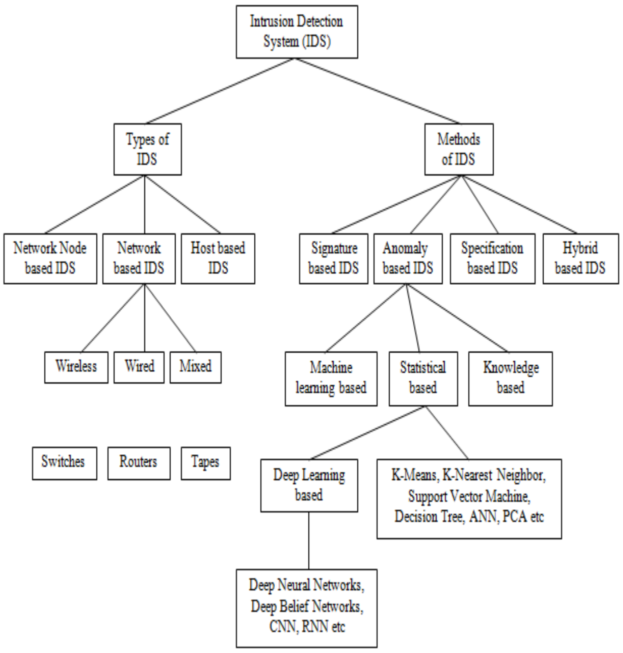


Fig 3: Taxonomy of Intrusion Detection System

In recent days the Deep Learning algorithms are using mostly and the publicly used benchmark datasets are mainly focused on Machine Learning algorithms. Deep Learning is the subpart of Machine Learning and it works on the functioning of brain neurons called Artificial Neural Networks, Deep Neural Networks, and Recurrent Neural Networks, etc [5]. Deep learning algorithms works on two or more hidden layers but in machine learning algorithms it works on a single hidden layer only. Compared to Machine learning strategies the Deep learning algorithms can deal with big data. By incurring distinct deep learning algorithms there are various uses to detect the attacks. A supervised algorithm predicts high accuracy with some labeled data whereas an unsupervised algorithm predicts low performance without labeling the data but it needs high training samples [6]. In hybrid approach the training samples are less and predict high performance but implementing the approach is very complex. Several Cyber security applications such as Botnet Protection, Attacks and Malware detection are using these Deep learning strategies and acquire the best detection rates as results [7-8]. And also using these Deep learning algorithms the weight values can change for the performance of back propagation learning and decreases the cost function [9-10].

LITERATURE REVIEW

Hindy et.al.[11] have described a taxonomy of intrusion detection with different situations of intrusions where it occurs in the network and the performance measures of intrusion detection and feature extraction techniques. And also they explained the different types of threats that occurred in which layer of the network and also represented the use of algorithm distribution in pie chart form.

Vinayakumar et.al., [12] have proposed a strategy called Deep neural network by using this DNN the test results of binary class classification and multi class classification are increased for every hidden layer, the hidden layers used are five. They also compared DNN results with other techniques like Random Forest, Decision Tree, etc with different datasets like KDD99, CICIDS2017, etc and in maximum situations the DNN result increases in every performance measures.

Resende et.al., [13] have reviewed different methods of Random Forest algorithm. Random Forest is the Ensemble method which decreases the training duration of the dataset and they also describe different attack types in detail. The comparison of different methods of random forest that describes a type of detection method, problem domain and parameters of the algorithm and they also use the feature selection method to select the efficient features from the dataset.

Khraisat et.al., [14] have reviewed intrusion detection techniques and its challenges to detect the attack which is present in the network. They briefly explain the types of network IDS, detection methods of IDS and benchmark datasets of NIDS. They also explained how the machine learning algorithms are used for detection intrusion, types of detection strategies and measure the performance of each technique namely detection rate, false-positive rate, and accuracies.

Taher et.al., [15] have described a novel supervised technique used to find out the intrusions network IDS. This ANN strategy uses 2 hidden layers with 17 characteristics of NSL KDD dataset and to gain the accuracy rate they proposed a feature extraction algorithm that is support vector machine and attain 94.02% detection accuracy, wrapper and filter methods are utilized for dataset training.

Sangkatsanee et.al., [1] have explained a real-time intrusion detection by machine learning approach. To classify the attack category, the authors used the Decision tree algorithm for real-time detection. There are three phases to detect the attack one phase is Preprocessing (Extracting features), the second phase is Classification (Classification result from Decision tree algorithm) and the third phase is Post-processing (Perform Voting for intrusion detection) and the dataset used is KDD99. After performing all these phases the real-time detection rate of the decision tree was 99.2% .

Aljawarneh et.al., [16] have used a hybrid efficient model to detect the anomaly-based intrusions present in the network. The hybrid model contains the following classifiers: J48, Random tree and Ada boost algorithms to gain the test accuracy of detecting the attacks with NSL-KDD dataset and attain 97% accuracy. As of using a single classifier, the accuracy is less compared to hybrid model. By using this hybrid model, they attain more accuracy in Normal, DOS and U2R attacks. Wrapper method is applied for feature extraction.

Abdulhameed et.al., [17] have used Deep neural networks and Random forest to detect the intrusions. They explained how these models are working with CIDDS-001 (Coburg intrusion detection dataset-001) by preprocessing imbalanced data and training DNN, RF for classification of attacks. By using these techniques, the results attained are 99.7% accuracy and 0.07 false positives and 10 features are used for the feature selection process.

Ahmim et.al., [4] have used a new hierarchical model to find both misuse and anomaly-based intrusions. They described that a single classifier will not give the best accuracy so they used this new model that combines three classifier models namely REP (Reduced Error Pruning) tree, JRip, and RF(Random Forest). The first classifier is used for preprocessing, the second method for categorize the connection as normal or benign, the last classifier is used to detect the attack class. CICIDS 2017 dataset which uses the best performance measures and attain 96.6% accuracy and 1.14% false-positive rate and 94.47% overall detection rate.

Liu et.al., [5] have explained the detailed information about the taxonomy of intrusion detection system for better understanding and define some differences between both the problems and solutions of intrusion detection system those are network based and host based IDS, misuse and anomaly detections for best idea about the survey on intrusion detection system using deep learning algorithms. They proposed deep learning algorithms so explained about the taxonomy of machine learning algorithms, advantages and disadvantages of shallow models in machine learning and described the comparison of different deep learning strategies which are applied on the different bench mark datasets. And they also illustrate regarding the research on machine learning based intrusion detection system.

Basnet et.al., [8] have used CSE-CIC-IDS 2018 dataset for cleaning up and preprocessing the records which consists of different features to detect the intrusions, this was the latest dataset using with latest attacks and also used internet of things for detection purpose. As it is the new dataset they briefly explained the features and different frameworks related to the deep learning frameworks. Each and every framework was explained briefly and compared the graphs associated and also given the information about the research challenges related to the two classification types those are binary-class classification and multi-class classification.

Shone et.al., [9] have surveyed on two datasets KDD Cup99 and NSL-KDD which are bench mark datasets, with these data the non-symmetric and stacked non-symmetric deep auto-encoders is used as training algorithms for classification. 5-class classification and 13-class classification are the different classification algorithms applied on KDD cup and NSL- KDD datasets are compared with Deep Belief Networks.

Jamadar et.al., [10] have surveyed on Auto-encoders, Recurrent Neural Networks, Cloud-based Real time Network based intrusion detection system, firefly algorithm based feature selection for Network based intrusion detection system used these type of deep learning algorithms for better accuracy and the algorithms are trained on different datasets and used feature extraction for best true positive values in the expected results.

Mishra et.al., [18] have proposed an ANN(Artificial neural network) algorithm, differentiate between both misuse and anomaly based detections and give a detailed information about the taxonomy of intrusions and explained the different features in the dataset and compare the flow of genetic algorithm k-means with the proposed algorithm and attain 97% accuracy with all 41 features.

Padke et.al., [19] have surveyed on different machine learning methodologies based on network intrusion detection system and applied on the different datasets related to the network connections and explained the features in depth and compare the results with different performance metrics and the algorithms are support vector machine, k-means clustering, intelligent intrusion detection system, artificial neural networks and back propagation neural network.

Kaur et.al., [20] have proposed a novel approach for detecting the attacks by exploring the ensemble classifiers in the area of smart grids. Smart grids means they taken one example as smart city, by using the wired area network (WAN) the central server will handle the whole city which includes houses, parks and libraries etc with the help of transmission substation and also a control server is also connected to WAN with some smart meters for reviewing the network connected. To take care of this smart city they explored the ensemble methodologies of machine learning algorithms and show the results in the form of accuracy and recall as confusion matrix. They applied these algorithms on different datasets because ensemble classifier combines different weak classifiers for better detecting of attacks.

Nadiammai et.al., [21] have explained a data mining approach on detection of attacks, evaluate different clustering approaches on the dataset of KDD Cup and also explained the features of dataset in detail. The clustering algorithms involved are fuzzy c means, hierarchical clustering and k-means. Out of these algorithms Fuzzy c means clustering attain highest accuracy applying on the dataset, also give the information about the features involved in that dataset.

Duan et.al., [22] have proposed Random forest algorithm which is a machine learning strategy used for detection of network security. The random forest algorithm here used as both for training and extraction of features, first the algorithm used to train the dataset that is KDD cup99 and they also proposed a feature extraction method for increasing the accuracy and to decrease the false positive rates occurred at the time of testing. For feature selection the performance measure used is information gain and by using this extraction method the accuracy values are increased.

PROPOSED ALGORITHM

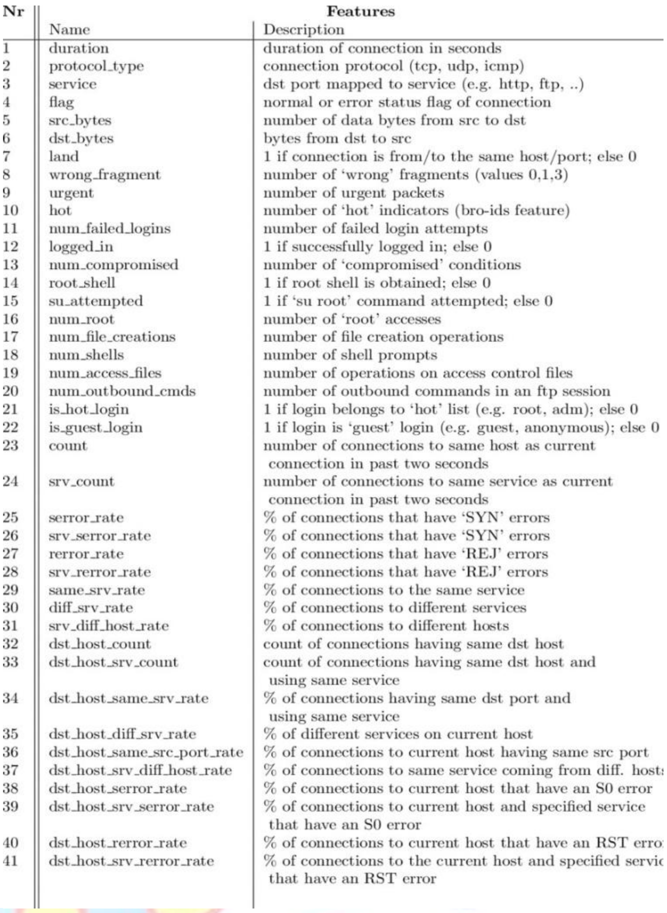
To implement the proposed algorithm the KDD Dataset is used as input. The dataset was built with tcp dump data that consists of 41 features with 4,94,021 records of training dataset and 3,11,029 records of testing dataset. The features are listed below in Fig 3.1. After giving the input as dataset, the required python packages should import for preprocessing the dataset, and to convert the feature values from string to numeric.

The proposed Deep learning algorithm uses Keras, Tensor Flow packages. There are some packages like sklearn.model_selection which is used for splitting the training dataset, testing dataset and also useful for converting the string values to numeric or floating values. The package keras.layers used for defining and embedding the network model as sequential and add the required dense, dropout features, initializing the input dimensions and defining the activation function as ReLU with sigmoid function.

ReLU(Rectified Linear Unit): ReLU is used for vanishing gradient boost problem when training the neural network with the back propagation method and it is faster than the non-linear unit. For optimizing the configurations two features are used those are loss and optimizer.

There are two callbacks are also available to fit the model. Both Check Pointer and CSV_Logger are used to monitor and compile the optimizer for model configuration.

Table 1: Features of Dataset



Nr	Features	
	Name	Description
1	duration	duration of connection in seconds
2	protocol_type	connection protocol (tcp, udp, icmp)
3	service	dst port mapped to service (e.g. http, ftp, ..)
4	flag	normal or error status flag of connection
5	src_bytes	number of data bytes from src to dst
6	dst_bytes	bytes from dst to src
7	land	1 if connection is from/to the same host/port; else 0
8	wrong_fragment	number of 'wrong' fragments (values 0,1,3)
9	urgent	number of urgent packets
10	hot	number of 'hot' indicators (bro-ids feature)
11	num_failed_logins	number of failed login attempts
12	logged_in	1 if successfully logged in; else 0
13	num_compromised	number of 'compromised' conditions
14	root_shell	1 if root shell is obtained; else 0
15	su_attempted	1 if 'su root' command attempted; else 0
16	num_root	number of 'root' accesses
17	num_file_creations	number of file creation operations
18	num_shells	number of shell prompts
19	num_access_files	number of operations on access control files
20	num_outbound_cmds	number of outbound commands in an ftp session
21	is_hot_login	1 if login belongs to 'hot' list (e.g. root, adm); else 0
22	is_guest_login	1 if login is 'guest' login (e.g. guest, anonymous); else 0
23	count	number of connections to same host as current connection in past two seconds
24	srv_count	number of connections to same service as current connection in past two seconds
25	serror_rate	% of connections that have 'SYN' errors
26	srv_serror_rate	% of connections that have 'SYN' errors
27	rerror_rate	% of connections that have 'REJ' errors
28	srv_rerror_rate	% of connections that have 'REJ' errors
29	same_srv_rate	% of connections to the same service
30	diff_srv_rate	% of connections to different services
31	srv_diff_host_rate	% of connections to different hosts
32	dst_host_count	count of connections having same dst host
33	dst_host_srv_count	count of connections having same dst host and using same service
34	dst_host_same_srv_rate	% of connections having same dst port and using same service
35	dst_host_diff_srv_rate	% of different services on current host
36	dst_host_same_src_port_rate	% of connections to current host having same src port
37	dst_host_srv_diff_host_rate	% of connections to same service coming from diff. host
38	dst_host_serror_rate	% of connections to current host that have an S0 error
39	dst_host_srv_serror_rate	% of connections to current host and specified service that have an S0 error
40	dst_host_rerror_rate	% of connections to current host that have an RST erro
41	dst_host_srv_rerror_rate	% of connections to the current host and specified servic that have an RST error

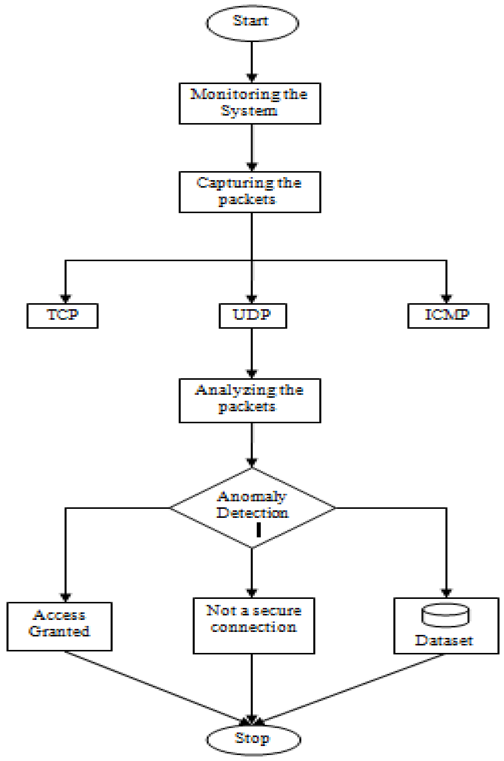


Figure 4: Flow Chart of the Proposed Algorithm

From the Fig 4, we can say that the flow of a proposed approach. First take the input as dataset and monitor that which is the form of packets because it works on network layer. After capturing the packets we should recognize the sender, receiver, number of bytes transferring from source to destination and protocol type. Protocol type which defines the connection protocol is either tcp, udp and icmp, the time taken for transferring the packet is also necessary when the connection is established.

They are some other specifications while transferring they are, how many urgent packets are available, wrong fragments, failed login attempts, duration of connection in seconds etc this information is called as packet analyzing. That analyzed packets will take as input for anomaly detection of intrusion detection system then with the help of dataset the machine learning algorithm is trained. By default we have the dataset i.e., existing patterns in the database. Through the existing patterns the algorithm will trained and it recognizes the intrusion and non-intrusion activities or attack and benign. After training the algorithm with deep neural networks the testing process starts. They are three ways for identifying the intrusions. If the new connection shows the matching then we can recognize it is an intrusion or not, if not matched permission is granted for allowing into the system, it didn't show either this two options then we can say that it is a non secure connection.

Deep Neural Networks (DNN):

Deep Neural Networks (DNN) is one of the deep learning algorithms with five layers those are input layer, output layer and three hidden layers. DNN uses the back propagation strategy for giving the best accuracy and less false positive rates. Before giving the dataset as input first preprocess the dataset while preprocessing the data the string values are changed to floating values, that preprocessed data is given as input to the input layer. The output from the input layer is given to first hidden layer and the process is continued till the last hidden layer that is fifth hidden layer values are given to output layer. The output layer values are calculated using sigmoid activation function ReLU from Fig 5.

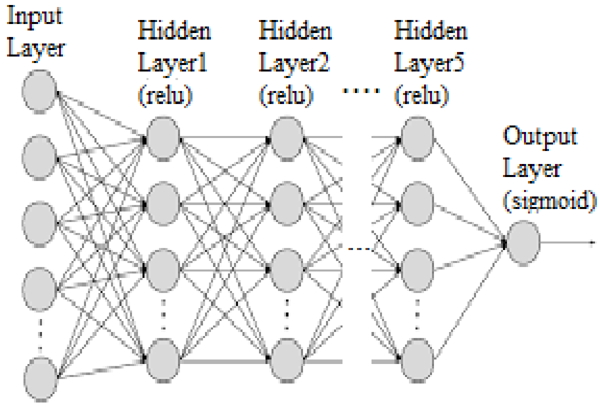


Figure 5: Architecture of Deep Neural Networks (DNN)

At last the output layer will give two values those are loss and accuracy. These loss and accuracy values will get through the epochs. Loss value is predicted based on the feature of cross entropy. An epoch is a hyper parameter which works number of times like cycles on the algorithm by taking input as dataset. This proposed algorithm uses 10 epochs in each layer. Each epoch produces the loss and accuracy values for all 10 epoch values in each layer. These training epoch values are given for testing and that test case predicts the confusion matrix as accuracy, precision, recall and f1-score values of the algorithm for each layer.

EXPERIMENTAL RESULTS

Performance Measures:

The Performance measures are more important in machine learning algorithms based on intrusion detection systems to predict statistical values.

TP (True Positive): Number of correctly detected intrusions.

FP (False Positive): Number of incorrectly detected intrusions.

TN (True Negative): Number of correctly detected non-intrusions.

FN (False Negative): Number of incorrectly detected non-intrusions.

Accuracy: Total number of correctly detected intrusions to the total number of all intrusions.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

TPR (True Positive Rate) or Precision: Total number of true positives to a true positive and false positive.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: Total number of True positive to the total number of positives (True Positive and False Negative).

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1 score: Harmonic mean of precision and recall.

$$\text{F1 score} = \frac{2TP}{2TP+FP+FN}$$

Table 2: Test Results for Deep Neural Networks

Method	Accuracy	Precision	Recall	F1 score
DNN 1 layer	0.923	0.996	0.907	0.950
DNN 2 layer	0.924	0.998	0.908	0.950
DNN 3 layer	0.913	0.998	0.894	0.943
DNN 4 layer	0.924	0.998	0.908	0.951
DNN 5 layer	0.925	0.998	0.909	0.951

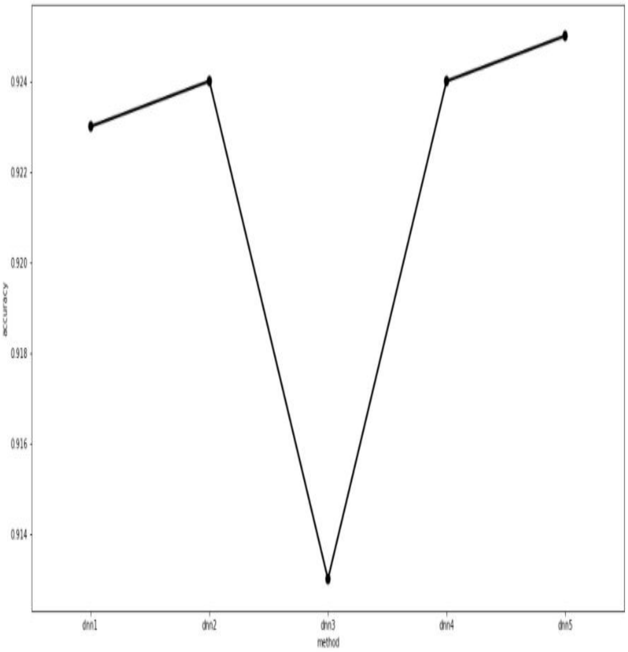


Fig 6: Graph for Accuracy of DNN algorithm

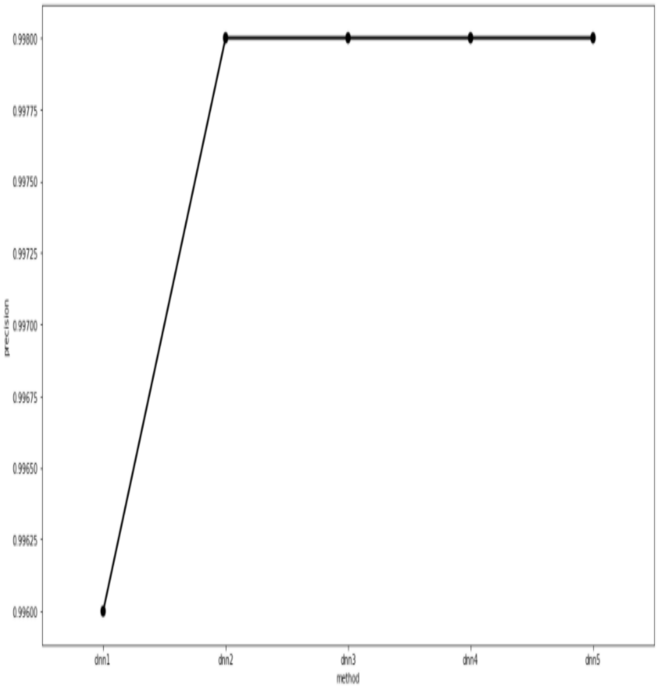


Fig 7: Graph for Precision of DNN algorithm

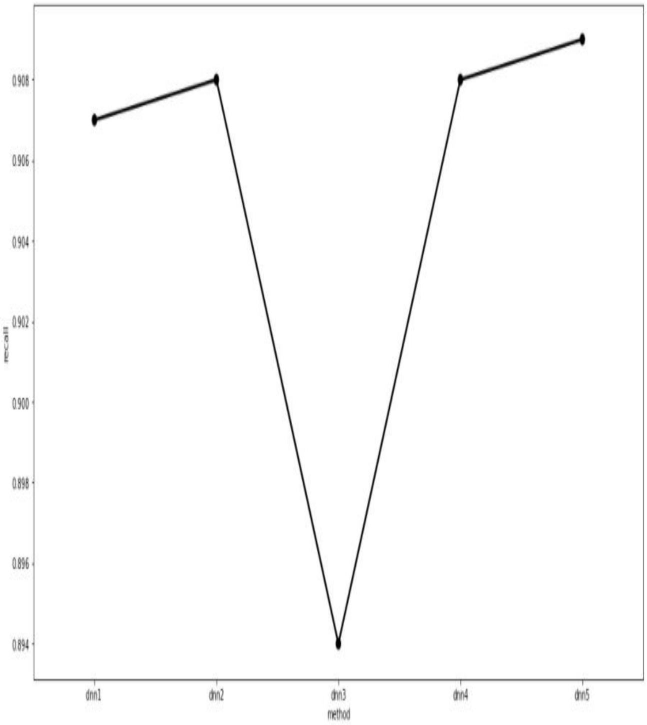


Fig 8: Graph for Recall of DNN algorithm

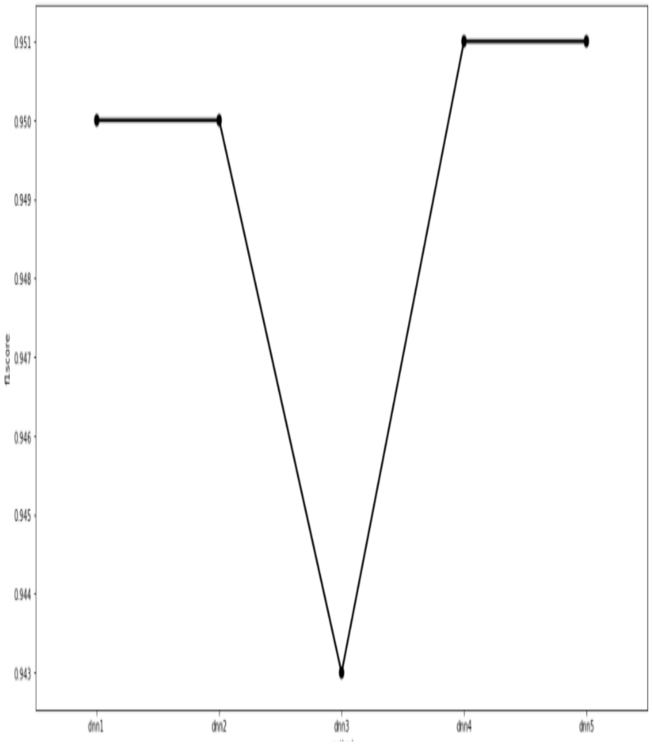


Fig 9: Graph for F1score of DNN algorithm

CONCLUSION

In this paper, we proposed a deep neural network for detection of network intrusion to achieve high accuracy and obtain high true positive rates. By using five hidden layers, in the third layer the accuracy, precision values decreases and the two parameters are increased gradually in fourth and fifth layer.

FUTURE WORK

This work includes binary classification only, by using this algorithm categorical classification can also achieve best accuracy rate and false positive rates should also be increased for best performance.

REFERENCES

1. P. Sangkatsanee, N. Wattanapongsakron and C. Charnisripinyo, “Practical real-time Intrusion Detection using Machine learning approaches”, Elsevier, pp. 2227-2235, 2011.
2. N. Chattopadhyay, R. Ghosh, S. Bhattacharya and A. Paal, “Data Intrusion Detection with basic Python coding and prevention of other intrusive manifestation by the use of intrusion application”, IEEE, pp. 1094-1099, 2018.
3. C. F. Tsai, Y. F. Hsu, C. Y. Lin and W. Y. Lin, “Intrusion detection by Machine learning: A review”, Elsevier, pp. 11994-12000, 2009.
4. A. Ahmim, L. Maglaras, M. A. Ferrag, M. Derdour and H. Janicke, “A Novel hierarchical intrusion detection system based on Decision tree and Rule-based models”, 15th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 228-233, IEEE 2019.
5. H. Liu and B. Lang, “Machine learning and Deep learning methods for Intrusion detection system : A survey”, mdpi, applied sciences, vol. 9, 4396, doi:10.3390/app9204396, 2019.
6. Y. Wu, D. Wei and J. Feng, “Network attacks detection methods based on Deep learning techniques : A survey”, Hindawi, Security and Communication Networks, Volume 2020, Article ID 8872923, 17 pages, https://doi.org/10.1155/2020/8872923, 2020.
7. S. Rawat, A. Srinivasan and R. Vinaykumar, “Intrusion detection systems using classical machine learning techniques versus integrated unsupervised feature learning”, https://arxiv.org/pdf/1910.01114., October 2019.
8. R. B. Basnet , R. Shash , C. Johnson , L. Walgren , and T. Doleck, “Towards detecting and classifying Network intrusion traffic using deep learning frameworks”, Journal of Internet Services and Information Security (JISIS), pp. 1-17, volume 9, number 4 , November 2019.
9. T. A. Tang, L. Mhamdi , D. McLernon , S. Ali Raza Zaidi and M. Ghogho, “Deep learning approach for Network intrusion detection in Software defined networking”, IEEE, 2016.
10. N. Shone, T. N. Ngoc, V. D. Phai, Q. Shi, “A Deep learning approach to Network intrusion detection”, IEEE Transactions on Emerging Topics in Computational Intelligence, November 2017.
11. H. Hindy, D. Brosset, E.Bayne, A.Seeam, C. Tactatzis, R. Atkinson and X. Bllekens, “A Taxonomy and Survey of Intrusion detection system design techniques, Network threats and datasets”, ACM, Vol.1, No. 1 Article, June 2018.
12. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat and S. Venkataraman, “Deep learning approach for intelligent intrusion detection system”, IEEE, pp. 2169-3536, Vol.7, April 2019.
13. P. Angelo Alves Resende and A.C. Drummond, “A Survey of random forest based methods for Intrusion detection systems”, ACM, Vol. 51, No. 3, Article 48, 36 pages, May 2018.
14. A. Khraisat, I. Gondal, P. Vamplew and J. Kamruzzaman, “Survey of intrusion detection systems: techniques, datasets and challenges”, Springer, 2019.
15. K. A. Taher, B. Md. Y. Jisan and Md. M. Rahmam, “Network intrusion detection using Supervised machine learning technique with Feature selection”, 2019 International Conference on Robotics, Electrical and signal processing techniques (ICREST), IEEE 2019.
16. M. Mazini, B. Shirazi and I. Mahdavi, “Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms”,

Elsevier, Computer and Information Sciences 31, pp. 541-553, March 2018.

17. R. Abdulhammed, M. Faezipour, A. Abuzneid and A. AbuMallouh, "Deep and Machine learning approaches for Anomaly-based intrusion detection of imbalanced Network traffic", IEEE Sensors council, Vol. 3, No. 1, January 2019.
18. P. Mishra, V. Varadharajan, U. Tupakula and Emmanuel S. Pilli , "A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection", pp. 686-728, IEEE Communications Surveys & Tutorials, Vol. 21, No. 1, First Quarter, IEEE Access 2019.
19. A. Padke, M. Kulkarni, P. Bhawalkar and R. Bhattad, "A Review of machine learning methodologies for Network intrusion detection", *Proceedings of the Third International Conference on Computing Methodologies and Communication,* pp. 272-275, *IEEE Xplore,* Part Number: CFP19K25-ART, ISBN: 978-1-5386-7808-4, 2019.
20. K. Kaur and A. Hahn, "Exploring the ensemble classifiers for detecting attacks in the smart grids", CyberSec 18, April 9–11, *Association for Computing Machinery (ACM)*, 2018.
21. G. Nadiammai and M. Hemalatha, "An Evaluation of clustering technique over Intrusion detection system", *International Conference on Advances in Computing, Communications and Informatics (ICACCI-2012),* pp. 1054-1060, August 3-5, *ACM*, 2012.
22. Y. Duan, Xin Li, X. Yang and L. Yang, "Network Security Situation factor Extraction based on Random forest of Information gain", pp. 194-197, *ICBDC 2019 Conference*, May 10–12, *2019 Association for Computing Machinery*, 2019.