Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.2, 2019, pp. 105-110

DOI: https://doi.org/10.51983/ajcst-2019.8.2.2132

Intrusion Detection System Using Deep Learning

Irin Anna Solomon[1], Aman Jatain[2] and Shalini Bhaskar Bajaj[3]
[1]PG Student, [2]Assistant Professor, [3]Professor
[1,2&3]Department of Computer Science, Amity University Haryana, India
E-Mail: irinsolomon11@gmail.com, amanjatainsingh@gmail.com, sbbjaj@ggn.amity.edu
(Received 9 April 2019; Revised 20 April 2019; Accepted 3 May 2019; Available online 10 May 2019)

Abstract **- Intrusion detection system (IDS) plays a very critical part in identifying threats and monitoring malicious activities in networking system. The system administrators can use IDS to detect unauthorized access by intruders in different organizations. It has become an inevitable element to the security administration of every organization. IDSs can be generally categorized into two categories. The first group focuses on patterns/signatures of network packets/traffic and they identify network intrusions using rule-based matching. The second group uses machine learning (ML) based approaches such as supervised and/or semi-supervised learning and train IDS models on a collection of labeled and/or unlabeled network data. This method has obtained better detection compared to the previous method. This project paper's scope involves implementing an intrusion detection system using deep learning technology for efficient detection of intrusion and intrusive activities that can cause disruption in the networking system. We use a Feed-forward Neural Network, a deep learning based technique, on KDD99 CUP - a commonly used dataset for network intrusion. In this paper the performance of the proposed system is compared with the existing previous work.**
Keywords: **Intrusion, Intrusion Detection System, Dataset, Preprocessing, Deep Learning, Neural Network**

I. INTRODUCTION

Due to the increased dependence on Internet and intranet facilities, the system and its networks are under continuous risk. An intrusion into the system or device can compromise the security and confidentiality of the data through several means. Currently due to the fast increasing data transfer rate, proliferation of data and the unpredictable usage of internet have added to the security risks [8]. Number of solutions and techniques has been suggested in order to avoid such risks. Intrusion detection system has become an essential component for fighting against the security breach. It has become very popular commodity due to its capability to detect attacks whenever it is happening. Therefore it is high time that researchers develop more functional, reliable, and self-supervisory systems, which is capable of sorting out risks and carry out the activities free from the intervention of the human. By performing such attempts, failures of networking systems can be reduced. It is mainly operated to safe guard the company networks [11]. An IDS is a device or an application that monitors, detects and identifies network operations for unauthorized and malicious activities that end up in policy violation. They generally focus on detecting and identifying possible circumstances and then they log all the possible information and report attempts. In principle, an IDS has the capability of real-time detection of all possible intrusions and put into work the counter measures to stop the attack. Intrusion detection system can be classified into three categories [12]:

1. Host Based IDS: evaluate data obtained on a single or multiple host systems which includes contents of OS, system and applications.
2. Network Based IDS: evaluates information obtained from the communication between the networks, the obtained packets are analyzed. Sensors are used to capture the network traffic packets.
3. Vulnerability Assessment IDS: the vulnerability on the internal networks and the firewall are detected. Two primary models are available:
a. Misuse Detection Model
b. Anomaly Detection Model
Most IDS commercial tools refer to the misuse detection model, and signatures of intrusions must always be updated by vendors. IDS based on anomaly detection model have the ability to detect symptoms of attack without specifying model of attacks but they are very sensitive to false alarms.

A. Deep Learning: Deep learning originated from the rising technological growth in the field of Neural networks. Single hidden layered neural networks have various limitations and different methods are now proposed to overcome this limitation. Deep learning is a branch or a division of machine learning which falls under the heading of artificial intelligence [8]. As compared to the traditional shallow machine learning method, deep learning capabilities stand out in different key respects from allowing computers to determine the solution to solve a host of complex and difficult problems that otherwise cannot be tackled. In classical method of machine learning, the input features are designed on the manual basis and the system learns automatically by itself so that it is capable of mapping these features to an output. There are various levels features in deep learning and they are automatically discovered and are composed together in multiple levels to produce outputs. The features present in the previous levels are used to discover the abstract features of the successive levels and this is the case in each single level [9]. Deep learning uses multi-layered ANN to deliver optimum accuracy in various functions such as detection of object, speech recognition, language translation and others. Deep learning by itself learns to represent from various categories of data such as images, video or text, without introducing hand-coded rules /

human domain knowledge which helps it to stand out from traditional machine learning techniques. They have very flexible architectures which give the ability to learn directly from raw data and their predictive accuracy can increased. In our experiment, a simple deep neural network is constructed with an input layer, three hidden layers and an output layer as described in Fig.2. The hidden layers contain fifty, ten and one neurons respectively.

B. Dataset: The dataset used is KDD cup'99 dataset. The KDD Cup is the most commonly used data mining competition in the world. Millions of records are present in the dataset and they are labeled as normal or a particular kind of intrusion attack. These intrusions can be broadly classified into four main types and they are Denial of Service, Probe, User2Root, and Root2Local. Each with 24 sub-types in training set and an additional 14 sub-types in test set [13].

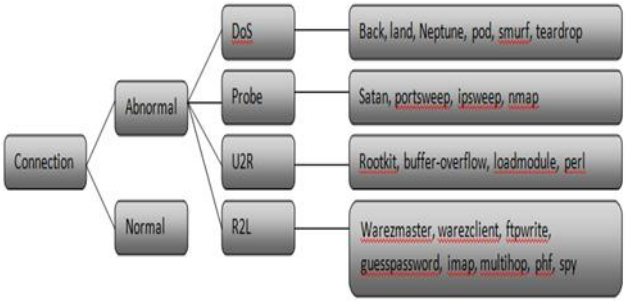


Fig. 1 Types of Attack

TABLE I DATA DISTRIBUTION IN KDD CUP 99 DATASET

S. No.	Categories	Attack Typres	Count
1	DOS	Back	2203
		Land	21
		Neptune	107201
		Pod	264
		Smurf	280790
		Teardrop	979
2	PROBE	Satan	1589
		Portsweep	1040
		Ipsweep	1247
		Nmap	231
3	U2R	Rootkit	10
		Buffer-Overflow	30
		Load_Module	9
		Perl	3
4	R2L	Warezmaster	20
		Warezclient	1020
		Ftpwrite	8
		Guess_Passwd	53
		Imap	12
		Multihop	7
		Phf	4
		Spy	2
5	NORMAL	Normal	97278

Root2Local attack is defined as a type of attack in which intruder tries different methods to gain access to network. In User2Root, attacker has got the access to machine which the victim but he still aims to get super-user privileges. In Probing attacker performs scanning so as to identify all the possible vulnerabilities in the victim system. The vulnerabilities that have been identified till act as the weakness of the machine and can be used to harm the system [9]. Denial of Service aims to make the resources inaccessible to authorized users. This can be attained by flooding systems or networks with huge amount of traffic, which disrupts the connection or services. This will lead to tardy and ineffective services. The dataset consists of 494,021 records articulated into 41 attributes. A wide variety of intrusions simulated in a military network environment is described in this dataset. The connection can be largely divided into normal and abnormal connections. The figure above explains the dataset in detail.

II. LITERATURE SURVEY

In real time machine objects can be recognized and the speech can be translated with the help of massive amount of computational power. AI is getting really smart. Many researches are taking place all around the world in the field of security with the help of machine learning algorithms. In this section some of the existing works are discussed.

Sodiya A.S *et al.,* [9] proposed a model that contains sensor, which collects the required information from the data source. A detector is used to for the analysis. Several sensors and detectors are used to build the model. It collects information in real time from all the available sources which are then analyzed by a detector. The detector recognizes known intrusion, it then learns new variety of intrusions, and takes actions based on events that occur, which if necessary produces an alarm. Multiple Self-Organizing Map is used for visualizing of the intrusion and Back-Propagation Neural Network is used classification of intrusions. When an intrusion is detected, the IDS inform the administrator on the intrusion and ask permission if it should allow the intruder to continue with the operations. If allowed and if it was a false alarm, then the IDS will lean from the event, otherwise, the system is shut down and the user comes out of the operation. The simulation results achieved 96 % detection rate and 3 % false alarm rate.

Inadyuti Dutt *et al.,*[1] proposed a model which would regularly analyze the vulnerabilities and weaknesses present in the system on the basis of the files that approach the system through network. The model employ statistical techniques and machine learning for structuring a frame work hybrid intrusion detection system. The proposed system identifies virus based on the definition which consist of characters or sequence of strings obtained from viruses in files. The main objective of this model is to design IDS which is capable of constantly monitoring the vulnerabilities of a system or network based on the files that approach in the system through network. The vulnerabilities present in

the system is observed keenly to identify if any intrusion is taking place. The results represent that the accuracy with respect to true positive rate was 92.65% with 7.35% false alarm rate.

Quamar Niyaz *et al.,* [7] proposed a model based on deep learning approach for building up a functional, efficient, flexible and a reliable NIDS. The proposed model contains sparse auto encoder and soft-max regression based NIDS. A benchmark dataset for network intrusion was used that is NSL-KDD dataset which can be used to evaluate anomaly detection accuracy. Normal/abnormal connections can be detected and classified when evaluated on the test data. The performance can be improved further by applying techniques such as Stacked Auto encoder, an extension of sparse auto encoder in deep belief nets, for unsupervised feature learning, and NB-Tree, Random Tree, or J48 for further classification. The efficiency obtained was appreciable and commendable.

Basant Subba *et al.,* [10] proposed an efficient model using the feed forward and the back-propagation algorithms to build an ANN based IDS model. The model contained three layers and they are input layer, a single hidden layer and an output layer. The number feature in the input feature vector is represented by using the number of nodes in the input and the hidden layers .Similarly, the desired output is represented by the number of nodes in the output layer. Additionally, the input layer and the hidden layers have a bias unit, whose value is set to constant +1. Different machine learning algorithms like Naïve Bayes, SVM etc have been compared along with ANN and ANN obtained detection accuracy of 95.03%.

Roy *et al.,* proposed a model for developing an IDS by exploiting deep learning models and the result show that a deep learning approach can enhance IDS performance. A DNN is selected which contains multi-layer feed forward neural network with 400 hidden layers. The output layer consists of Shallow models, rectifier and softmax. A benchmark dataset that is KDD 99 Cup dataset was used and the proposed model was compared with other classifiers and obtained a detection rate of 99.994%.

Another DNN based model was proposed by Polturi and Diedrich which mainly concentrated on enhancing the DNN implementation by using multi-core CPUs and GPUs. A deep learning based model using SAE was chosen to construct the DNN .Another benchmark dataset NSL-KDD dataset which is an improved version of KDD Cup dataset was used. The model consist of 20 neurons in the first hidden layer, 10 neurons in the second hidden layer and 5 neurons in the output layer which uses softmax activation function.

III. METHODOLOGY

A. Deep Learning Based Intrusion Detection System: The proposed model is a deep learning model containing input layer, multiple hidden layers and the output layer. The features are given into the input layer. There are four hidden layers in the proposed deep learning method. The first layer contains 10 neurons, the second contains 50 neurons, third layer contains 10 neurons and finally the last hidden layer contains only one neuron. The hidden layers have rectified linear unit (relu) and the output layer contains softmax activation function.

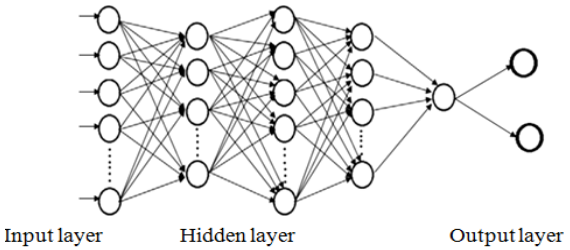


Fig. 2 Architecture of the deep learning model

The processes of the proposed deep learning based model can be split into number of different parts. The dataset used is KDD 99 Cup and the dataset requires data preprocessing in-order to transform the unlabeled and raw data into user understandable format. Further the unique features are extracted from the dataset and they will be used as the input data into the deep learning model. Different feature selection algorithms are available. The data is then split into two, one set for training and the other for validation or testing. The proposed model is trained with the data and later testing is performed in-order to validate whether the system is able to classify the data as normal and abnormal. The process is explained with the help of a diagram shown in Fig. 3.

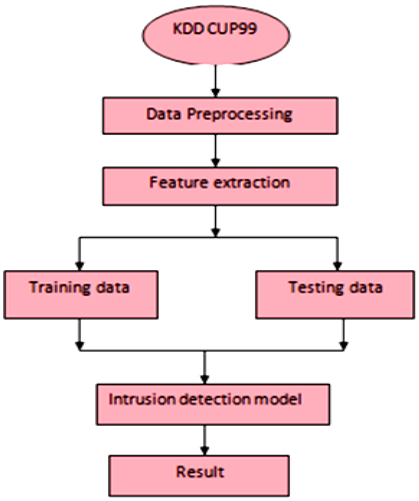


Fig. 3 Detection process

B. Data Pre-Processing: Data pre-processing is subset of data mining technique that deals with transforming raw data into a comprehensible format [10]. The data in the real world is majority of the time incomplete, unpredictable, lacking in some behaviors or trends, and it also contain many errors. Data preprocessing is a certified method for clarifying these above mentioned issues. Functions in data pre-processing [11].

1. Data cleaning
2. Data integration
3. Data transformation
4. Data reduction
5. Data discretization

C. Feature Selection: Feature selection also known as attributes selection or variable selection, is the process of selection of attributes in your data automatically that are appropriate to the process of prediction modeling problems. It is always a confusing fact that Feature selection and dimensionality are the same but to the contrary both are two different functions. Both approaches have the functionality to bring down the number of attributes or features in the dataset, but in the case of dimensionality reduction method, operation is done by creating new combinations of features whereas in the case of feature selection methods, the features or the attributes are included and excluded without changing the data [12]. Feature selection helps-

1. *Data Redundancy Reduction:* we can avoid unwanted calculations on the meaningless features by selecting only useful attributes.
2. *Enhancing Accuracy:* clearing out the deceptive or useless features can help in improving the system accuracy.
3. *Removing Over-Fitting:* By avoiding the correlated attributes, we can reduce the risk of over-fitting.

Feature selection methods can give relevant information on the importance of features for a complex problem. This information can be used to obtain a filtered version of dataset and improve the accuracy of the models.

a) Train/Test Split: The dataset that are used is usually divided into training and test data. The training dataset consist of known output and the developed model acquires knowledge on the data so that it will be comprehended to other data in the upcoming phase [13]. The subset test dataset is then used in to test the model's prediction function.

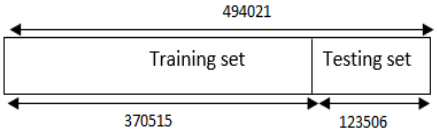


Fig. 4 Splitting of the KDD cup 99 dataset for training and validation

TABLE II KDD 99CUP DATASET

Attack Type	Training set %	Testing set %
DoS	79.24	73.90
Probe	0.83	1.34
U2R	0.01	0.07
R2L	0.23	5.21
Normal	19.69	19.48

D. Intrusion Detection Model: The model implemented is a deep learning based model containing more than one hidden layer. The training set data is used to train the system and the validation set data is used to test the system if it can classify the data as normal and abnormal [5]. The output of the implemented system will be either normal that indicate that the data is not malicious. But in case of abnormal output the data will fall under any one of the following attacks i.e. Denial of Service, probe, R2L and U2R. The result and discussion is performed in the next section. In training phase 370515 i.e. almost 75% of the sample data is used [6]. The deep learning model contains four hidden layers and these hidden layers use Rectified Linear Unit (relu). In a NN, the activation function plays an important role and is responsible for transforming the added input weights from the node into the activation of output for that input.

The rectified linear activation also called the relu function is a piecewise linear function that will provide the output for the given input directly if is positive and in the other case it will output zero. If the input is less than 0, the output will be 0 and raw output otherwise. If explained in other words, if the input value is greater than 0, the output is equal to the input value ie greater than 0. ReLUs' functionary is similar to a real neuron function in human body.

$$f(x) = max\ (x, 0) \qquad (1)$$

It acts as a default activation function for various types of NNs because such a model is easier to train and majority of the time achieves commendable performance more importantly in case of deep learning. Researchers that work in this field have proven that ReLUs provides much faster training result for huge networks. [15] The output layer uses softmax activation function. [14] This function calculates and evaluates the probability distribution of the different event over 'n' various events. In general, softmax function will work out the probabilities of each and every target class on top of all possible target classes. The output of this function is similar to a well-known categorical probability distribution; it provides the probability that any of the other classes are true. Mathematical representation of a softmax function is shown below, where z is the input vector assigned to the output layer . And again, j represents the output units, so j = 1, 2, ..., K.

$$\sigma(z)_j = \frac{e_j^z}{\sum_{k=1}^{K} e_k^z} \qquad (2)$$

The calculated probabilities will be helpful in later phase for evaluating the target class for the given inputs. The epoch value was set to 1000 and the optimization function used is Adam optimizer. It can be ued instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data. Adam is derived from adaptive moment estimation.

IV. RESULTS AND DISCUSSION

In the first phase of data pre-processing the redundancy, duplicate values etc is removed. The following table shows the resultant values. The total count reduces to 145586. Compared to the Table I, the values are reduced.

TABLE III DISTRIBUTION OF DATASET AFTER DATA PROCESSING

S. No.	Categories	Attack Types	Count
1	DOS	Back	2203
		Land	21
		Neptune	107201
		Pod	264
		Smurf	280790
		Teardrop	979
2	PROBE	Satan	1589
		Portsweep	1040
		Ipsweep	1247
		Nmap	231
3	U2R	Rootkit	10
		Buffer-Overflow	30
		Load_Module	9
		Perl	3
4	R2L	Warezmaster	20
		Warezclient	1020
		Ftpwrite	8
		Guess_Passwd	53
		Imap	12
		Multihop	7
		Phf	4
		Spy	2
5	NORMAL	Normal	97278

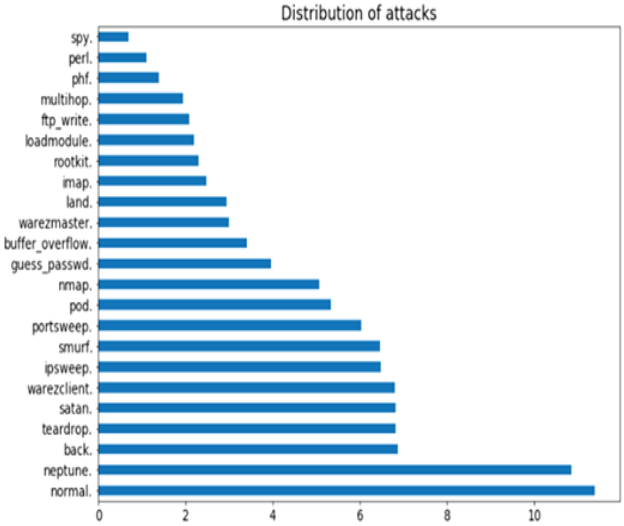


Fig. 5 Distribution of attack

Standardization, normalization and scaling is performed on the dataset. The graphical representation of each feature is described below in detail. Normalization often also simply called Min-Max scaling basically shrinks the range of the data such that the range is fixed between 0 and 1. This means that he largest value for each attribute is 1 and the smallest value is 0. The graphical representation of each feature is described below in detail. Standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and 1.

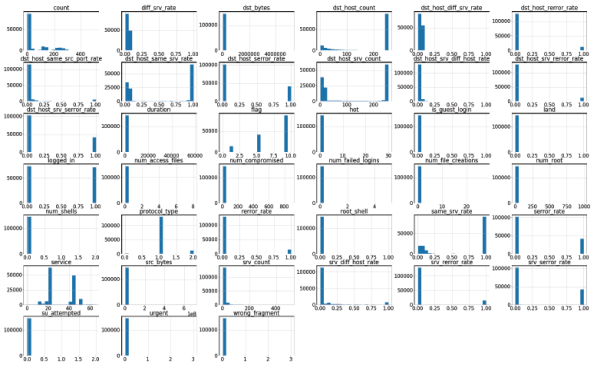


Fig. 6 Univariant Histogram

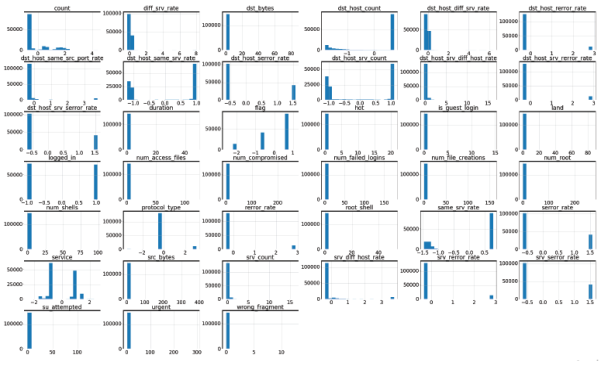


Fig. 7 KDD Standardizaton

The 41 features in KDD cup 99 dataset is standardized as well as normalized for better performance evaluation. All the redundant data has been removed and using feature selection the best features are slected and the rest are eliminated. Data scaling is also performed to standardize the range od independent variables or features of data.

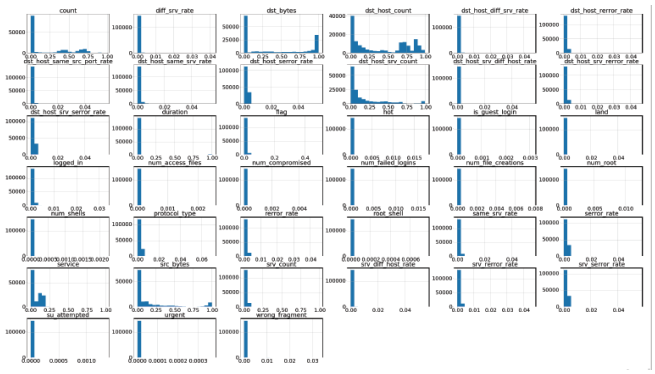
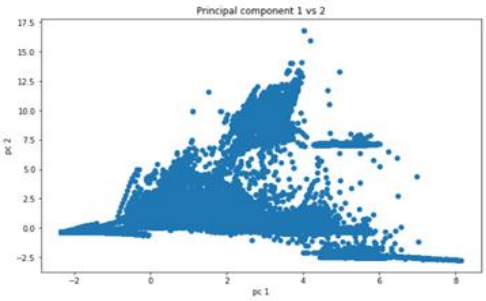


Fig. 8 KDD Normalization

Linear separability of various attack types is tested using the Convex-Hull method. The intersections between the hull boundaries of the classes normal and the two most frequent attack types neptune and smurf are visualized in a 2D plot against the first two principal components. This way it can be proved that the distinct attack classes are non-linearly separable. Similar to almost all of the existing deep learning research works that is going on; the developed system is a classification model implemented using TensorFlow. In order to perform the required evaluations, we have used the benchmark dataset i. e KDD Cup '99 datasets.



Out[61]:

	principal component 1	principal component 2	target
0	0.560535	1.428889	1.0
1	0.561570	1.404599	1.0
2	0.541429	1.380523	1.0
3	0.514420	1.354875	1.0
4	0.488805	1.329314	1.0

Fig. 9 Principle component 1vs 2

This dataset is considered as for evaluating within NIDS research. Further, using of these datasets helps in conducting a meaningful comparison with existing working models and research work. The detection rate obtained as part of the result showed that the proposed model obtained an accuracy of 99.87%.On the basis of the keen observation and evaluation; it is proved that our deep learning model has produced a promising set of results. As compared to the other machine learning algorithms like Naïve bayes, random forest, kNN using the same dataset the detection rate of the implemented method was exceptionally high. Another important matter is the time it takes to complete the entire function, the time required to train our model is drastically and commendably reduced. While classifying the KDD Cup '99 dataset, [3] Kim *et al.*, achieved an accuracy of 96.93%. Also, Gao *et al.*, [6] developed a deep learning based DBN model and it achieved an accuracy of 93.49%. As compared to both the methods, the proposed model obtained much better accuracy. These comparisons show that the result of the proposed is very promising and very much successful and when compared to other existing current deep learning based methods it is proved to have shown better efficiency.

V. CONCLUSION

In this paper a detailed description about Intrusion detection system and its different types have been discussed. A brief definition on deep learning is also included. The dataset that is employed is the KDD Cup dataset. Lot of issues persists in this dataset and these issues have been discussed in detail and the various methods that can be implemented in order to eliminate those problems have also been discussed. The result has been showcased in a graphical representation to get a clear understanding of the problem. A deep learning based model was proposed for the IDS and was implemented using tensorflow. The obtained results show an outstanding accuracy for the proposed system. The results have demonstrated that the proposed approach offers high levels of accuracy and precision along with reduced training time. As future scope, the detection can be done in real-time bases for a better performance based intrusion detection system. In real-time detection lot of advantages as well as disadvantages are present. If the intrusion detection system is able to detect the threat and reports it immediately, then the damage caused by the intrusion can be limited. In this particular work, intrusion detection is performed, as a future expansion of this work, prevention methods can also be included. Intrusion prevention system is able to block potential threats. They monitor log and report activities similar to IDS but they are also capable of stopping threats without the administrator getting involved.

REFERENCES

[1] Inadyuti Dutt, Samarjeet Borah, Indra Kanta Maitra, Kuharan Bhowmik, Ayindrilla Maity and Suvosmita Das *"*Real-Time Hybrid Intrusion Detection System Using Machine Learning Techniques", *Springer*, Vol. 462, pp. 885-894, 2018

[2] Jabez, and Dr. B. Muthukumar, *"*Intrusion Detection System (IDS): Anomaly Detection Using Outlier Detection Approach*"*, *Procedia Computer Science*, Vol. 48, pp. 338-346, 2015

[3] Dongseong Kim, Hanam Nguyen, Syngyup Ohn, and Jongsou Park, "Fusions of GA and SVM for anomaly detection in intrusion detection system", *Springer*, Vol. 3498, pp. 415-420, 2005

[4] Kwangjo, Kim, Muhamad Erza, Aminanto, Harry Chandra and Tanuwidjaja, "Network Intrusion Detection using Deep Learning ", *Springer Briefs on Cyber Security Systems and Networks*, pp. 978-981, 2018

[5] Wathiq Laftah Al-Yaseen, Zulaiha Ali Othman and Mohd Zakree Ahmad Nazri, "Real-Time Intrusion Detection System Using Multi-agent System", *IAENG International Journal of Computer Science*, pp.1-11, 2016

[6] N. Gao, L. Gao, Q. Gao and H. Wang, "An Intrusion Detection Model Based on Deep Belief Networks," *2014 2nd Intl.Conf. on Advanced Cloud and Big Data*, Huangshan, pp. 247-252, 2014.

[7] Quamar Niyaz, Weiqing Sun, Ahmad Y Javaid, and Mansoor Alam *"A Deep Learning Approach for Network Intrusion Detection System"*, BICT, 2016 pp. 03-05 ICST , December 2015.

[8] N. Shone, T.N. Ngoc, V.D. Phai and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection," in *IEEE Transactions on Emerging Topics in Computational Intelligence*, Vol. 2, No. 1, pp. 41-50, Feb. 2018.

[9] A.S Sodiya, O.A Ojesanmi, O.C Akinola. and O. Aborisade "Article: Neural Network based Intrusion Detection Systems*"*. *International Journal of Computer Applications*, Vol.106, pp. 19-24, Nov. 2018.

[10] B. Subba, S. Biswas and S. Karmakar, "Enhancing effectiveness of intrusion detection systems: A hybrid approach," *2016 IEEE International Conference on Advanced Networks and Tele communications Systems (ANTS)*, Bangalore, pp. 1-6, 2016

[11] Sumanta Kumar Deb, Ankan Bhowmik, Biswajit Maity, Abhijit Sarkar, and Amitava Chattopadhyay "Wi-Fi Optimization Using Parabolic Reflector and Blocking Materials in Intrusion Detection Systems,*" Emerging Technologies in Data Mining and Information Security*, Vol. 814, pp. 761–771, 2018.

[12] T.A. Tang, L. Mhamdi, D. McLernon, S.A.R. Zaidi and M. Ghogho, "Deep learning approach for Network Intrusion Detection in Software Defined Networking," *2016 Intl.Conf. on Wireless Networks and Mobile Communications*(WINCOM), pp. 258-263, 2016.

[13] Nabil El Kadhi, Karim Hadjar and Nahla El Zant, "A Mobile Agents and Artificial Neural Networks For Intrusion Detection", *Journal Of Software*, Vol. 7, No. 1, pp. 156-160, 2012

[14] Z. Wang "Deep Learning-Based Intrusion Detection with Adversaries", *IEEE Access*,Vol. 6, pp. 38367-38384, 2018

[15] Berlin H. Lekagning Djionang and Gilbert Tindo, "Network Intrusion Detection Systems based on Neural Network: A Comparative Study*"*, *International Journal of Computer Applications* Vol. 157, No. 5, pp. 42-47, 2017