Artificial Neural Network based System for Intrusion Detection using Clustering on Different Feature Selection

Wasima Matin Tammi
Dept. of Computer Science & Engineering
Ahsanullah University of Science & Technology

Noor Ahmed Biswas
Dept. of Computer Science & Engineering
Ahsanullah University of Science & Technology

Ziad Nasim
Dept. of Computer Science & Engineering
Ahsanullah University of Science & Technology

Khadizatul Zannat Shorna
Dept. of Computer Science & Engineering
Ahsanullah University of Science & Technology

Faisal Muhammad Shah
Dept. of Computer Science & Engineering
Ahsanullah University of Science & Technology

ABSTRACT
Intrusion Detection System (IDS) is an example of Misuse Detection System that works for detecting malicious attacks. This can be defined as software for security management. Many researchers have proposed the Intrusion Detection System with different techniques to achieve the best accuracy. In this paper it is projected that intrusion detection system with the amalgamation of k-means clustering and artificial neural network to improve the system. To obtain a better result benchmark dataset was split into training and testing part and then cluster the dataset into five different divisions. After getting the cluster data it has been trained by the different Artificial Neural Networks functions as- Feed Forward Neural Network (FFNN), Elman Neural Network (ENN), Generalized Regression Neural Network (GRNN), Probabilistic Neural Network (PNN) and Radial Basis Neural Network (RBNN). After implementing these functions we have proposed a comparative analysis between them and choose the best accuracy rate among them. Here, it has been proved that, using the clustering technique a better accuracy rate can be found that improve the system with the best neural network functions which is the probabilistic neural network. It is also important to select efficient feature sets for better accuracy.

Keywords
Intrusion Detection System, K-means Clustering, Artificial Neural Network, FFNN, ENN, GRNN, PNN, RBNN.

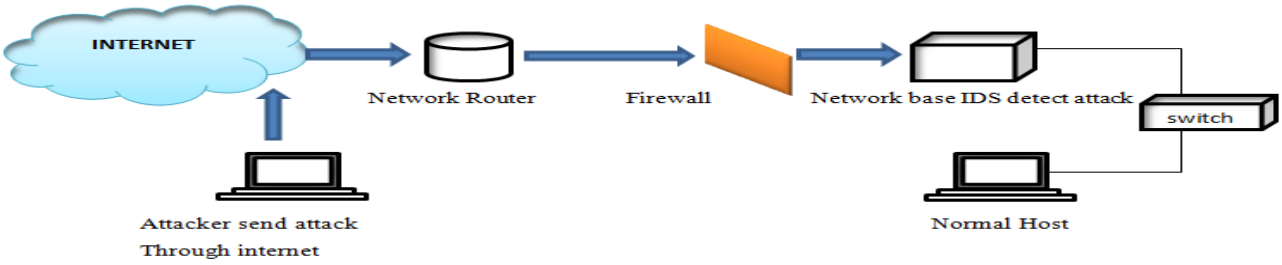


Figure 1: How Intrusion Detection Works

1. INTRODUCTION
Intrusion Detection System is one of the common phenomena in this early age of internet security. IDS works for known and unknown attacks or it can be said that it can also be worked for novel attacks also [1]. But this paper worked for known attacks only. Intrusion Detection System monitors the network traffic and warns for the suspicious activities that enter into the system [2]. There are two kinds of IDS and they are- 1. Network Based IDS and 2.Host Based IDS. Network based IDS detects network attacks as payload is analyzed and Host based IDS detects local attacks before they hit the network. IDS can be more understandable by its precision rate, the harmonic mean of sensitivity and the misclassification comparison rate [3]. In data mining, clustering is an unsupervised technique that main advantage is splitting the dataset into different groups with its similarities [4, 5, 6]. Among many clustering process, the k-means clustering is selected which is the common and simplest calculation technique [7].

After doing the clustering part the dataset will be alienated into different cluster which will be trained by the different artificial neural network functions. But before training the ANN with the clustered data, feature selection process has been applied to reduce the irrelevant features to get a proper result.

Intrusion detection system with artificial neural network is one of the widely used data mining techniques that proves

how it is worthy while doing the complex terms in an easy way [8]. And we have merged the IDS, clustering and ANN to get a better result for our proposed system. Though clustering and ANN have both some drawbacks, but the combination of these two can improve the performance of IDS. In k-means clustering the low capability to pass the local optimum and strong sensitivity are the main shortcomings [9]. On the other hand, lower detection precision for low number of attacks and weaker detection stability are the main disadvantages of artificial neural networks [1]. But when we cluster a dataset into several cluster and train individual neural networks with each cluster the precision for low frequent records improves up to maximum. This research work was tested on NSL-KDD dataset where there are five types of data. They are- Normal, Probe, Dos, U2R, and R2L. Low number of attacks means the remote to local (R2L) and user to root (U2R) data's. These data are called the low-slung frequent attacks.

Solving these problems, the idea of feature selection is used by selecting the random features and make different types of dataset to get a comparison analysis and normalization of dataset [10]. So the whole process can be summed up by the following steps which will be described thoroughly in the rest of the other parts of the paper.

Figure 2: Summary of the Proposed Model

2. RELATED WORKS

IDS is the most used and developed system that can detect attack. Intellectual IDS can perform as a dynamic defensive system which is capable of adapting dynamically changing traffic pattern. Many researchers have worked on Intrusion Detection System to give the preeminent output through their system. It can be more categorized where researchers have mentioned the IDS system with the merging of the artificial neural network and clustering techniques [1, 11, 12]. They showed the performance with other well-known methods such as decision tree, naïve byes, in terms of detection precision and detection stability. They mainly used the clustering techniques to generate different training subsets. And based on these subsets, different ANN modules are trained to formulate different base models. Finally they used an aggregation model to aggregate the result. They reduced the complexity of each of the sub training set and correspondingly they increase the performance of the detection. Moreover, many of them proposed a survey report on where they have cited the detection of accuracy from IDS using only the artificial neural network classifier [13]. Using different ANN functions they disclosed the accuracy system for the intrusion detection system that will detect on which technique of the ANN function will give the highest accuracy rate. Researchers have worked for the anomaly detection system using both the ANN and Decision Tree techniques for improving the system. Their experimental skills demonstrate that, while neural networks are vastly successful in detecting known attacks, decision trees are more attention-grabbing to detect new attacks [1]. In [14], they analyzed artificial intelligence approaches for application in IDS. An elaborated study is given in [15], where the importance of neural network in IDS is discussed.

Researchers have introduced decision tree based light weight intrusion detection using a wrapper approach [10]. They removed the rudundant instances that causes the learning algorithm to be unbiased by generating the genetic algorithm. They have focused on the feature selection and finally they have generated a neurotree to achieve the better detection accuracy. They also focused on the precision rate, true positive rate , false negative rate for different classes of attacks. Some of them have hosted K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset [6]. In their work they bring up only the dataset working with the unification of k-means clustering. As clustering algorithm proves to be very useful having large unlabeled dataset, the main objective of this paper is to provide a complete analysis of the NSL-KDD dataset and the attacks presented. They used K-means algorithm for this purpose and also represented the distribution of instances in clusters providing better illustration of the instances and making it clearer to understand [30]. In [16] they proposed hybrid framework based on neural network MLP and K-means Clustering. In [17] they introduced Enhanced K Means Clustering using Improved Hopfield Artificial Neural Network and Genetic Algorithm.

However, some have provided the combined approach for anomaly detection using clustering and neural networks techniques [18]. Using the modified SOM or self-organizing map is used to create the network with the help of distances threshold, connection strength and neighborhood functions and k-means clustering algorithms groups the nodes in the network with the help of similarity measures. It disclosed when the learning rate increases the number of output nodes decreases. And many have provided, Intrusion Detection using Fuzzy Clustering and Artificial Neural Network where it has also proved the better accuracy for using these two techniques [4] where they proved the better accuracy rate for each of the attack in the system. Some are worked for detecting novel attacks like [19, 20].

So all in all it can be articulated that the main purpose is to prove the best accuracy using the different techniques that are being used in intrusion detection system. The more the techniques the more the efficient results have been established. From the above state of art it can be explained that each and every paper has been provided with different models where few of them have implemented too many techniques. But in this proposal two different techniques i.e. clustering and neural network are implemented with feature selections to improve the output so much prominent than others.

3. DESIGN OF PROPOSED MODEL

The proposal is divided into three stages. They are-

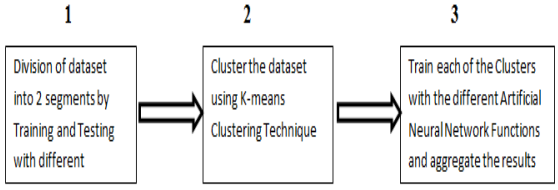


Figure 3: Outline of the Proposed Model

3.1 Proposed Architecture

This proposed system has 3 stages. The system is described in detail. Then we will discuss the modules of the system later. Primarily the dataset is divided into train and test set with ratio of 70:30 respectively and then cluster the train set with k-means clustering. Then train different ANN for different clusters, and cumulate the ANN on the last stage.

- Phase 1: Divide the dataset into two sets- (a) Training dataset and (b) Testing dataset. Cluster the training dataset with K-mean, as there are five different classes so the number of cluster will be five.
- Different Artificial Neural Networks will be trained by the clustered data sets.
- Phase 3: Finally aggregate the Artificial Neural Network and improve the target result from the different neural network functions.

The Architecture is demonstrated in figure 4.

3.2 Division of Dataset

While working with the dataset we have used the NSL-KDD99 Dataset and improvise the dataset as an advantage of calculation. In the dataset the nominal data are changed with some numeric value to be more effective while calculating. There are four attack types and they are- Probe, Dos, U2R (User to Root) and R2L (Remote to Local) and a Normal class. These names are initiated with the numeric value from 1-5. This dataset consists of 41 features and the features in columns 2, 3, and 4 in the KDD99 dataset are the protocol type, the service type, and the flag, respectively. The value of the protocol type may be tcp, udp, or icmp, the service type could be one of the 66 different network services such as http and smtp and the flag has 11 possible values such as SF or S2. For instance, in the case of protocol type feature, 0 is assigned to tcp, 1 to udp, and 2 to the icmp symbol [21]. After changing the dataset we have created 4 types of dataset using some algorithms that can be entitled by the feature selection methods. Using these we have point out 9 features, 16 features, 36 features and 41 features [10, 22].

3.3 Clustering the Dataset

As it is mentioned that clustering is the best process for a huge dataset and there are many types of clustering techniques. Among them we have designated K-means clustering which is the most effortlessly clustering approach for calculation [9].

3.3.1 K-means Clustering

K-means clustering is one of the simplest unsupervised data mining procedures. It is simplest to be implemented. It is fast, robust and understandable and so it is used in data mining procedure so efficiently [7]. K-means clustering give the best result when data set is distinct or well separated from each other [23, 24] and it can be summarized by the following algorithm.

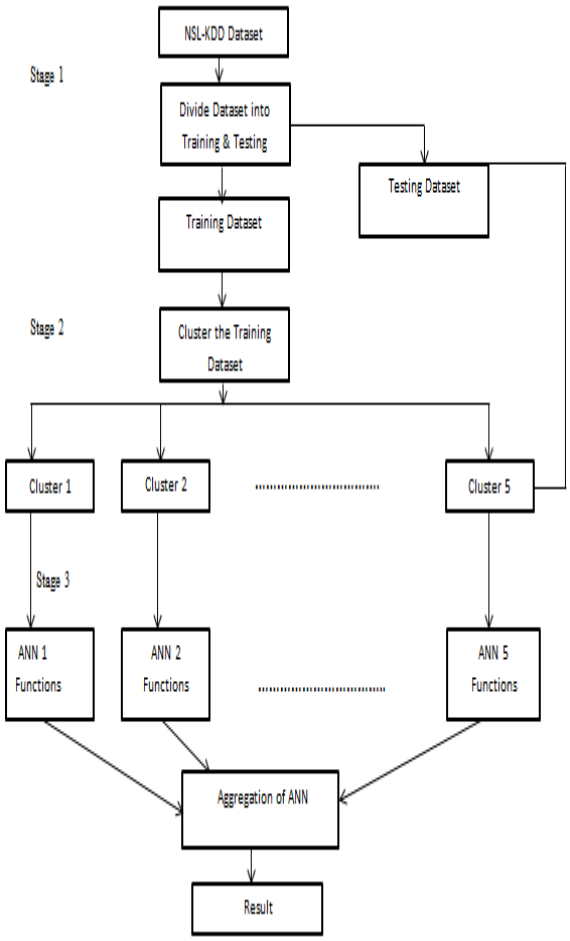


Figure 4: Proposed Architecture

Algorithmic steps for k-means clustering [25]

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is the minimum of all the cluster centers

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'c_i' represents the number of data points in i number of clusters.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3.

3.4 Artificial Neural Network

Artificial neural network invented by the idea of the human brain that deals with visual data and learned to distinguish objects. A neural network is a set of alarmed units following a particular topology [26]. It creates connections between many types of processing elements where each parallel to a single neuron in a biological brain. Neural networks have been extensively used in anomaly detection system as well as misuse detection. Its main advantage is it's a nonparametric model and it is easy to understand compared to statistical methods. There are many artificial neural network functions that are used in our proposed model [27]. They are Feed forward neural network, Elman neural network, Generalized Regression neural network, Probabilistic neural Network and Radial Basis neural network.

3.4.1 Characterization of different Neural Network Functions

Neural networks are very wide-ranging and can capture a variability of patterns very accurately [27, 28, 29]. So each and every function has its own appearances that can be summed up as following.

Table 1: Comparison of different Neural Network Functions

Function Name	Definition	Equation	Advantages	Disadvantages
Feed Forward Neural Network	It can also be defined as the multilayer perceptron's. An assortment of neurons connected together in a network can be represented by a directed graph	Arithmetically, the functionality of a hidden neuron is labeled by $\sigma(\sum_{j=1}^{n} w_j x_j + b_j)$ Where weights (w_j, b_j) are indicated with the arrows feeding into the neuron	➢ It's a most common and widely used function ➢ Networks are easy to maintain. ➢ It can be applied into many large problems and to generate non-linear dependencies	➢ Its error varies, depending upon the architecture.
Elman Neural Network	Elman neural network is one kind of feed forward network in addition of layer recurrent connections	$E=\sum_{i=0}^{T} E_p$ Where p indexes over all the patterns for the training set in the time interval [0, T].	➢ Elman Neural Network is most used to Patio-temporal pattern recognition, its result is better than another one. ➢ It can reduce data redundancy.	➢ Their inputs are all instantaneous constant values
Generalized Regression Neural Network	It is often used for function approximation. It is also a one-pass learning algorithm with highly parallel structure.	$E[y\|X] = \frac{\int_{-\infty}^{\infty} yf(X,y)dy}{\int_{-\infty}^{\infty} f(X,y)dy}$ Where, X is a particular measured value of random variable x and density f(x, y) is unknown.	➢ It can be used for any regression problem in which assumption of linearity is not justified ➢ There is no requirement of iterative algorithm	➢ Requires substantial computation to evaluate new points.
Probabilistic Neural Network	It is a feed forward neural network that is derived from the Bayesian network.	$P_i(x) = L_i(x) / \sum_{j=1}^{M} L_j\ (x)$ Where, $P_i(x)$is the conditional probability function and i is the number of iteration.	➢ Faster than multilayer perceptron networks. ➢ More accurate. ➢ Networks are relatively insensitive.	➢ Slower than multilayer perceptron networks at classifying new cases. ➢ Requires more memory space.

Radial Basis Neural Network	It uses radial basis functions as activation functions. The output network of this function is linear combination of radial basis functions.	$\varphi(x) = \sum_{i=1}^{N} \alpha_i \rho(\|\|x - c_i\|\|)$ Where, N is the number of neurons, C_i is the center of vector, for neuron i and α_i is the weight of neuron.	➢ Easily designed ➢ Strong tolerance to input noise ➢ Online learning ability	➢ It is much slower than any other functions while training.

4. RESULT AND DISCUSSION

This proposed system explained a comparison with different dataset where we have followed the following algorithm and it has been implemented in Weka and rest of the other part has been implemented in Matlab. The rest of the result part has been mentioned in the following segments.

4.1 Dataset & Selecting Features

In experiment, NSL-KDD99 dataset was selected evaluation. There are total 41 features which this research worked for. And to show comparison of different datasets for differently selected features some feature selection algorithm is used [10]. These algorithms are the following:

Table 2: Making Different Dataset with Different Algorithms [10]

Number of Features	Selected Algorithms
09 Features	**GreedyStepWise+CfsSubsetEval**
16 Features	**Genetic Algorithm (GA)**
36 Features	**Ranker+ChiSquaredAttributeEval**
41 Features	**NIL**

4.2 Experimental Work

As previously discussed on different dataset earlier and this research has worked for the comparison of different neural network functions, the following tables can be articulated and figure are given explanations. In table 4 and figure 3 the accuracy rate for 9 features is explained. From the given table 4 and figure 5 it can be mentioned that the accuracy for the probabilistic neural network is high than others and the accuracy for radial basis neural network is low. So it can be said that for reduced feature set almost all the function performs better. Correspondingly, for 16 features which are being selected by GA. PNN also gives highest accuracy.it is demonstrated in figure 4. For larger features sets like 36 and 41 features the performance of PNN is far better than other functions. It is here to be mentioned that for larger feature sets GRNN performance is lagged behind at a huge margin than the i.e. 73.54% for 36 features and 69.86% for 41, whereas the accuracy for PNN is 96.49 and 96.16 respectively. The graphical representations are shown in figure 5, 6. So one thing that is common for all the dataset, is the accuracy for probabilistic neural network is preeminent than others.

Table 4, the percentage for all the dataset accuracy has been given to state that the probabilistic neural network is more dominant than the others. And for a general discussion, it can be said, that the accuracy for generalized regression neural network is the lowest one which is 69.86% for 41 features. Moreover, the accuracy for probabilistic neural network is highest which 96.16% for 41 features is.All the graphical representation shows the same thing where the X axis for the accuracy and the Y axis for the network classifier.

Table 4: Performance comparison of different ANN functions for different features set

Features Number	Features Name	Accuracy for FFNN	Accuracy for ENN	Accuracy for GRNN	Accuracy for PNN	Accuracy for RBNN
09	flag,src_bytes, wrong_fragment,hot, num_access_files, diff_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate	95.81	90.91	95.88	96.57	79.31
16	protocol_type, service, flag, src_bytes,dst_bytes, wrong_fragment,hot, logged_in,srv_count, serror_rate, same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_srv_serror_rate, dst_host_rerror_rate	93.59	89.52	84.65	97.89	87.83

36	src_bytes, dst_bytes, service, wrong_fragment,flag, dst_host_srv_diff_host_rate, dst_host_diff_srv_rate, count,diff_srv_rate, srv_serror_rate, hot,dst_host_serror_rate, srv_count,same_srv_rate serror_rate, dst_host_srv_count, dst_host_same_srv_rate dst_host_same_src_port_rate, dst_host_srv_serror_rate, rerror_rate protocol_type, dst_host_rerror_rate, dst_host_count num_compromised, dst_host_srv_rerror_rate, logged_in, srv_rerror_rate, land, srv_diff_host_rate, duration, num_failed_logins, root_shell, is_guest_login, num_file_creations, num_access_files, num_root	94.29	89.94	73.54	95.61	90.05
41	All Features in NSL-KDD dataset	93.21	88.67	69.86	97.06	92.05

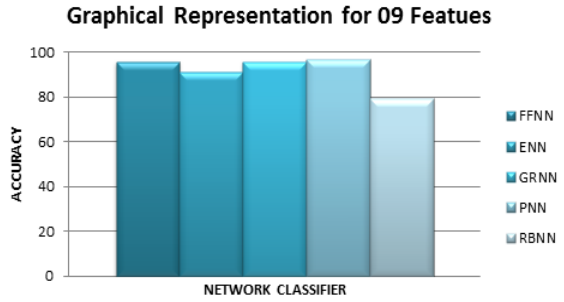


Figure 5: Accuracy for different ANN functions using 9 features

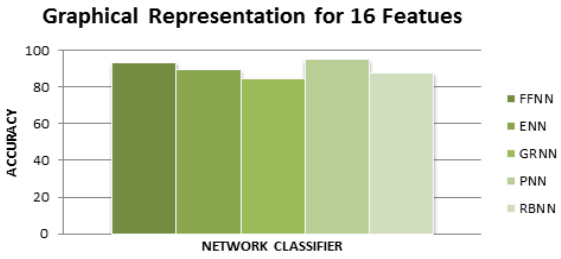


Figure 6: Accuracy for different ANN functions using 16 features

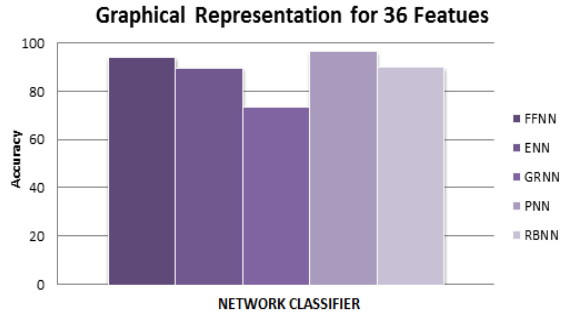


Figure 7: Accuracy for different ANN functions using 36 features

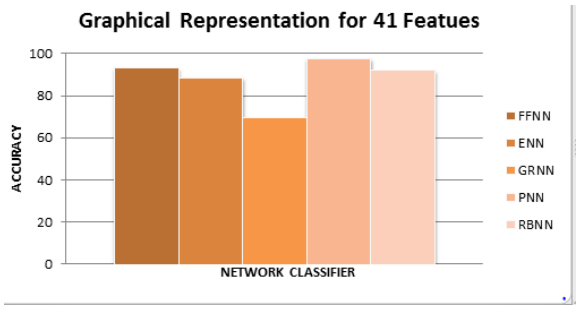
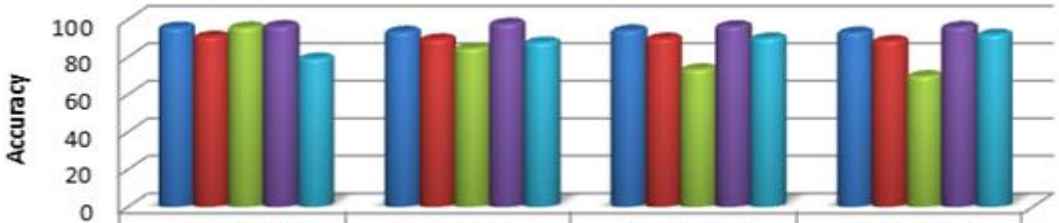


Figure 8: Accuracy for different ANN functions using 41 features



	09 Features	16 Features	36 Features	41 Features
FFNN	95.81	93.59	94.29	93.21
ENN	90.91	89.52	89.94	88.67
GRNN	95.88	84.65	73.54	69.86
PNN	96.57	97.89	96.49	96.16
RBNN	79.31	87.83	90.05	92.05

Figure 9: Summarization of the Research Results

5. FUTURE WORK

In this paper it has been proposed that using the clustering techniques with neural network results is higher accuracy compare to other model and in detecting the attacks it is important to get the leading result. Here, it is seen that, imposing features selection over the clustering process has solved the redundancy complexions so far. Clustering is the mainstream for detecting attack of low frequent data. And as the k-means clustering is the most easiest and effective clustering so the output has verified this characteristics very flourishing. Though we have selected those features selection method from some previous research, it can be more improvised by the other methods of feature selections. In future it could be more effective and accurate if we concentrate on the following future works:

- This research is functioned for the known attacks and in future it will be highly constructive for this paper to work for the novel attacks by upgrading the tentative model.
- Different types of clustering methods can be applied to improve our work. Here, the k-means clustering is mentioned and in future it could be a virtuous assessment if we calculate the accuracy for different cluster. And prepared it as a comparison analysis as we have done for different neural network functions.

6. CONCLUSION

In this study, it has been proved that, the Probabilistic Neural Networks provide better accuracy over other Neural Network functions i.e. Feed Forward Neural Network, Elman Neural Network, Generalized Regression Neural Network and Radial Basis Neural Network. And the reduction of feature matrix makes the performance enhanced. So improving the accuracy efficient feature selection techniques can be applied to improve the accuracy. We have used the clustering technique for identifying the low frequent data for more specification. And thus prove a better accuracy rate with the proposed techniques.

7. REFERENCES

[1]. Bouzida Y., Cuppens F., 2006, Neural networks vs. decision trees for intrusion detection, In IEEE / IST Workshop on Monitoring, Attack Detection and Mitigation.

[2]. Shrivas A.K., Dewangan A. K., 2014 An Ensemble Model for Classification of Attacks with Feature Selection based on KDD99 and NSL-KDD Data Set, International Journal of Computer Applications (0975 – 8887) Vol 99 – No.15.

[3]. Elhamahmy M. E., Hesham N. E. and Imane A. S., 2010 A New Approach for Evaluating Intrusion Detection System, CiiT International Journal of Artificial Intelligent Systems and Machine Learning, Vol 2, No 11

[4]. Surana S. 2013 Intrusion Detection using Fuzzy Clustering and Artificial Neural Network, Advances in Neural Networks, Fuzzy Systems and Artificial Intelligence, ISBN- 978-960-474-379-7.

[5]. Osoba O., Kosko B., 2013 Noise-enhanced clustering and competitive learning algorithms, Neural Networks 37 (2013) 132–140.

[6]. Kumar V., Chauhan H., Panwar D., 2013 K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4.

[7]. Wu, Junjie, Advances in K-means Clustering, A Data Mining Thinking, Springer Press, ISBN: 9783642298073

[8]. Wang G., Hao J., Ma J., Huang L. 2010 A new approach to intrusion detection using Artificial Neural Networks and Fuzzy clustering, Expert system with applications, vol 37, pp. 6225-6232.

[9]. Xu R., Wunsch D. C., Clustering, Wiley, IEEE Press, ISBN-10: 0470276800

[10]. Sindhu S. S. S., Geetha S., Kannan A. 2012 Decision tree based light weight intrusion detection using a wrapper approach, Expert Systems with Applications 39 129–141.

[11]. Gaikwad D.P., Jagtap S., Thakare K., Budhawant V. 2012 "Anomaly Based Intrusion Detection System Using Artificial Neural Network and Fuzzy Clustering", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 1 Issue 9.

[12]. Du K. L. 2010 Clustering: A neural network approach, Expert system with applications, vol 37, pp. 6225-6232.

[13]. Devaraju S., Ramakrishnan S. 2013 Detection of Accuracy for Intrusion Detection System using Neural Network classifier, International Journal of Emerging Technology and Advanced Engineering, Volume 3 Special Issue 1.

[14]. Novikov D., Roman V., Yampolskiy, and Reznik, L. 2006 Artificial Intelligence Approaches For Intrusion Detection, IEEE computer society.

[15]. Beghdad R. 2008 Critical Study on neural network in detecting intrusions. Computers and Security, 27(5-6)186-175.

[16]. Lisehroodi M. M., Muda Z., and Yassin W. 2013 A hybrid framework based on neural network MLP and K-means Clustering for Intrusion Detection System, 4th International Conference on Computing and Informatics, ICOCI.

[17]. Sakthi M., Thanamani A. S. 2013 An Enhanced K Means Clustering using Improved Hopfield Artificial Neural Network and Genetic Algorithm, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-3.

[18]. Aneetha A.S., Bose S. 2012 The Combined Approach for Anomaly Detection using Neural network and Clustering Techniques, Computer Science & Engineering: An International Journal (CSEIJ), Vol.2, No.4

[19]. Gaddam, Shekhar R., Kiran P., Vir V., & Balagani. 2007 A novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods, IEEE Transactions on Knowledge and Data Engineering, 19, 3

[20]. Al-Subaie M. 2006 The power of sequential learning in anomaly intrusion detection, degree master thesis, Queen University, Canada.

[21]. Bahrololum M., Salahi E. and Khaleghi M. 2009 Anomaly Intrusion Detection System using hybrid of Unsupervised and Supervised Neural Network, International Journal of Computer Networks & Communications (IJCNC), Vol.1,Issue No.2

[22]. Siddiqui M. K., Naahid S. 2013 Analysis of KDD CUP 99 Dataset using Clustering based Data Mining, International Journal of Database Theory and Application Vol.6, No.5.

[23]. Kanungo T., Mount D. M. 2002 An Efficient k-means Clustering Algorithm: Analysis and Implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence Vol: 24 , Issue: 7

[24]. Everitt B. S., Landau S., Leese M., Stahl D., Cluster Analysis, 5th Edition, Wiley. ISBN : 978-0-470-97844-3

[25]. Venables W. N., and Ripley B. D. 2002 Modern Applied Statistics with S, Springer-Verlag.

[26]. Rogas R., Neural Network- A Systematic Introduction, 3rd Edition- Springer Press, eISBN: 978-3-642-61068-4

[27]. Kriesel D., A Brief Introduction to Neural Network, - Zeta 2 Edition,

[28]. Kukielka P. and Kotulski Z. 2008 Analysis of different architectures of neural networks for application in intrusion detection systems, In proceeding of the international multiconference on computer science and information technology, pp. 807-811.

[29]. Moradi M. and Zulkernine M. 2004 A Neural Network based system for intrusion detection and classification of attacks, Queen University, Canada.