# Network Intrusion Detection System using Improved K-Mediod Clustering Approach

Rajat Bajwa[1], Rasneet Kaur[2]
[1]*Research Scholar,* [2]*Assistant professor*
[12]*Department of Computer Science Engineering,*
[12]*Shaheedudham Singh college of engineering and technology*

***Abstract-*** With the Rapid advancement in technology and in data processing applications areas and looking after permeate all features of business and, therefore, that leads to an increase in the development of strategic ways to mount malicious attacks on various systems. The various techniques has been developed for the detection of the attacks and malicious attacks .The security of administrated computer networks has been improved with the development techniques using Intrusion Detection System (IDS).Network traffic anomaly may indicate a possible intrusion in the network and therefore anomaly detection is important to detect and prevent the security attacks. Network intrusion plays a vital role in the security attacks. The application areas in the network intrusion detection are monitoring exposure, risk analysis, prevents damage to attacks, security management systems and the alterations in the files and directories are easily detected. In the network intrusion detection, the basic approach on signature based may become a basic issue in the anomaly detection. The main issues in the signature based method are database signature in network intrusion and network anomaly detection. In the existing research, the different machine learning techniques based on the entropy approach. The computation of the data set for the evaluation of the performance of the various algorithms. The research work, describe the precision rate of the data set using Receiver Operating Curve (ROC) metric, Radial Basis Function (RBF) In proposed research, the different algorithms used are k-mean, k-medoid and improved k medoid. Precision recall, accuracy rate increased and the error rate decreed using improved k medoid algorithm. In research method improve the accuracy rate, precision and recall rate as compared to the existing method. In research method accuracy rate 98%, Precision rate value is achieved 99% and Recall rate value is 98% and existing ensemble method value is 96%. Experiment simulation tool used MATLAB with GUI project application designed and 2% improvement as compared to the existing method.

***Keywords-*** Intrusion Detection System, ROC, GUI, MATLAB and Radial Basis Function.

## I. INTRODUCTION

Data mining is the method of extraction of the hidden probable data from the database which is valuable to people. The extraction can be extracted from randomized information, fuzzy dataset. The growth of the technology has created maximum amount of the dataset in various regions. This method is also known as knowledge detection method, mining knowledge from information, knowledge extraction or data pattern analysis. With the advent in technology, there has been vast growth of internet and network in the field of administration, social sites and industries. The presence of the complex network needs security on the system to achieve better communication. Network intrusion helps in protection against the attacks [1].This network is a software device used to monitor the system against malicious action or program violation. If any malicious activity is detected then it is reported to supervisor or may be gathered at centre through authenticated data and event supervision scheme. In computer network, an act of bypassing the security method is known as intrusion. Intrusion detection system is an automated method that detects the probable intrusion. This method is based on the security purposes which are, monitoring, detecting and responding to illegal activity by the organisation in and outside intrusion [2].

Intrusion detection system aimed at detection of the attacks, misused system, and sent and alert message to the system. On other hand, if a specific operation is assumed to contain safety event, an alarm message will be determined by that event. Some intrusion detection scheme is capable of distributing an alert message, so that supervisor of Intrusion detection may get an alarm message [3][4].

In proposed methods used these clustering methods and classification method. KNN K-nearest neighbour is simpler learning algorithm placed in equipment. The whole approach is computed distance among different points and removal the k point with closest data points and then counting k data points that is related to classifier. The largest k data point with closest neighbour is based on classifier [5].

Support vector machine is kind of the classification approach that is used as dual classifier technique established by Vapnik and Bell laboratories. The main benefit of the support vector machine can be acquired through optimum result under definite samples to get global result in absence of local optimum. Support vector machine is the dual classification of analysing the information and determine the pattern. The main objective of the trained vector is done in various classes. The linear classification is determined as the optimum separation of the hyperplane. The vectors can be separated through linear and nonlinear process [6].

Naïve Bayes is simple approach based on Bayes theorem which is also called as principle of possible to search the most required classifier. This algorithm presents the method to link the previous possible and provisional possible in unique formula to compute the possible classification. It is the principle based on statistical features that aimed to decrease the possible incorrect choice [7]. This algorithm is also called as simpler Bayes and self-regulating Bayes. Naïve Bayes is simple to construct it. Naïve Bayes is the common method that is utilised in classifier issue and depends on Bayes theorem. This method is utilised for the classification of the text and also presents the medical diagnosis and management scheme.

FCM clustering is a method which gives individual segmented data to represent twice or more clusters. This technique allocates specific data point related to every cluster on standard distance among cluster centre and data point. Additional data is near to the cluster-centre more is its membership towards the particular cluster centre [8].

In research method using Clustering is dependent on different kinds of the data objects variance and utilise displacement functional value, rules and classify the models. The classifier is making a variance along with the dispersion of the pattern character values. In case the assistance of the values is having the similar samples and displacement for the classification of the statistical features with different groups. The Eigen vector of all sampling is done through pattern gathering and distribution of the space. The displacement functional value is measuring of the data patterns. In accordance to contiguity of the production of the measurement can classify in accordance to pattern analysis [9]. The optimisation of the resulted values aims to divide k model to satisfy specific model. Initially, selection of some methods as initialised cluster based data values. Other is gathering of the remained data samples to focalised facts related with the method of decreased displacement and after than initialised organised structure for the modification clustered focal data points. K mean approach is dependent on the division of the specific type of the cluster approach [10]. Hence, in this approach searching of the gradients technique towards the directional value of power reduced that deal with the initialised clusters and all data points is related to the minimum value.In this technique, before the calculation of the displacement of the data values to centroid of the clusters, k clustering is random selection of the n data values so that initialised data segmentation is built related to nearness of the data object to clustered centroid from starting of the segmented information. In addition, selection values are engaged regularly unless desired suitable segmentation data value is acquired. In this technique, after the selection, the data objects from every clustered sample are selected that is dependent on the enhancement of the clustered superiority [20].

Intrusion detection system permits the administrators of the networks able to recognise the variation of the programmes in the security structure. In IDS the different type of the security rates had been increased . The main limitations of the IDS system are[11]:-

(i) Excess Data:- The excessive data is the main issue present in the intrusion detection system. In IDS, the data source evaluates the data with different volume for effective study[12].

(ii) Fake Positives:- The main issue is the prediction of the false intrusion attacks. The normal attack predicts as malicious attack, if the rate is high. The decreasing value in the false positive rate is the complex tasks.[13]

(iii) Fake Negatives:- The IDS will not produce any alert message when the fake negative rated is high.[14]

(iv) Real Positive:- The IDS will raise an appropriate alarm if the real attacks occurs.

(v) Real Negative:- IDS will not raise an alarm if there is no attack[15,16]

Section I shows about an introduction intrusion detection system, clustering and classification methods. Section II defines the prior work and various techniques and issues. III describes section the research proposal in intrusion detection systems. Section V and IV defines the result explanation with comparative analysis and conclusion, future scope.

## II. PRIOR WORK

**Ran, Z et al.,2012[17]** proposed a combined intrusion detection method that depends on multiple agents. In this method, four types of the representatives were described and that were organised in hierarchical arrangement. The main representative in each host of the subgroup was accountable for performing the detection and reply job. The complex combined task was separation of the representative which was accountable for the conduct at low level identification of the job. The cooperative representative was assigning the job at low level connection. On the basis of the hierarchical organisation, the descripted format was determined. Moreover, this research described the cooperative domain of the capability for the supervision of the combined detection.**Sadeghi, Z.,andBahrami, A. S. et al., 2013[18**] proposed research on detection of anomaly intrusion that depends on selection which may be reliable for improving the speediness of intrusion detection scheme along with that decreases the amount of detectors. The two method used are clustered and updated clusters. The two methods help in reduction of the detectors. Experimental analysis was done to improve the speed of detection up 51%.**Liao, H. J., Lin, C. H. R., Lin, Y. C. and Tung, K. Y et al.,2013[19]** structured and executed a host-based interruption recognition framework, which joins two location advances, one was log document investigation innovation and the other was BP neural system innovation. Log document examination was a methodology of abuse recognition, and BP neural system was a methodology of abnormality location. By mix of these two sorts of discovery advancements, the HIDS that they had actualized successfully improved the productivity and precision of interruption recognition**.Arashloo, S. R., Kittler, J and Christmas, W et al.,2017[21]** proposed research in detection of anomaly intrusion. Initially, evaluate protocol to revise the consequence of the unexpected attacks to train and test the information. The novel proposed protocol was used to reveal the actual situations in spoofing challenges in which attacker originate from spoofing. Secondly, novel formulated method for detection of spoofing issue that depends on anomaly detection in which training information determined from positive class. The testing information originates from desirable and undesirable class. The single course preparation requires the obtainability of the undesirable trained samples that represent probable spoofing. In final process, calculation and comparison analysis was done using 20 dissimilar single value and dual class scheme of video series of desired dataset established to examine the single and dual class interpretations. Experimental analysis was done in which predictable dual method superior than anomaly method.**Azid, A., Juahir, H., Toriman, M. E. and Osman, M. R et al., 2014 [22]** demonstrated intrusion detection technique in clouding method. This method explained techniques of data mining using SVM(support vector machines) and RBM(restricted Boltzmann machines). The planned method used hybrid intrusion detection technique depends on linking of two techniques utilising hyperbolic Ricci flows to the Poincare disk. The dataset KDD-99 was used for testing dataset explaining four groups of intrusion and comparison analysis was done to improve the accuracy , false negative rate ad false negative rate.

## III. RESEARCH METHODOLOGY

Step 1:- Upload the dataset from the given KDD 1999 and KYOTO 2006+ dataset in intrusion Detection Systems.Step 2:- Pre-processing phase to remove the unwanted data in the given datasets.Step 3:- Clustering Process: -The clustering process is the main task of separating the given relevant data points into a number of cluster or groups such that all data values in the similar to other data values in the similar cluster than those in other groups.Step 4:- K mediod: K-medoids is generally location of data objects of the group of clusters with less mean displacement related to other data objects. The approach based on the segmentation of the database in to groups. Medoids is the group of the definite database with finite set of information that may not be same to decreased data values. The K-medoids Clustering is the process of the partitioning of the clusters through clustering algorithm. The k- medoids algorithm is more prone to noise and data points are required in the k- medoids. The medoids is the object of the cluster where all the objects are different between the mean and the medoids. The medoids algorithm calculates the average value of the data sets of the objects in the data items.

Step 5:- Improved K-mediod Clustering: The selection of objects is clusters called as medoidsand k mean clustering is called as Improved k medoids [19][20]. In this clustering mechanism there is partitioning algorithm where the number of clusters is recognised in advance and then partitioned in to the number of the clusters. Step 6. Performance metrics:- Evaluate and compare the performance metrics such as FAR, FRR, accuracy, Precision and Recall.

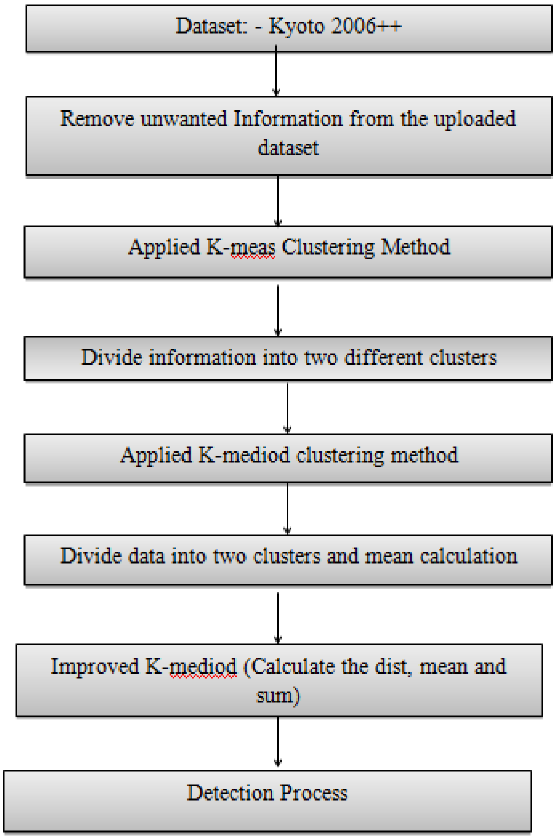

Dataset: - Kyoto 2006++
Remove unwanted Information from the uploaded dataset
Applied K-meas Clustering Method
Divide information into two different clusters
Applied K-mediod clustering method
Divide data into two clusters and mean calculation
Improved K-mediod (Calculate the dist, mean and sum)
Detection Process

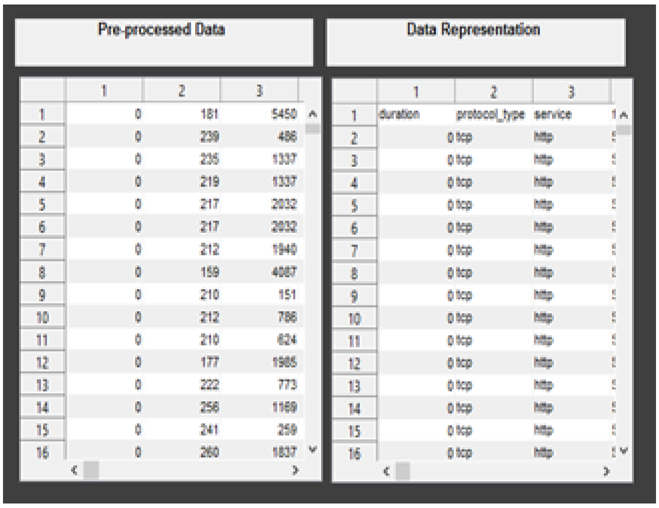
Fig.1: Proposed Flow Chart

## IV. EXPERIMENT ANALYSIS

In this section, a main interface using Network Intrusion Detection in Data Mining. In this interface shows upload the dataset from the .csv and .xls file. Apply data pre-processing phase to remove the unwanted information in the uploaded dataset. The clustering process has been implementing to divide the information into sub-groups or clusters in the NID system. K-mediod Clustering method has implemented to divide the information into two sub-groups with mean calculation. Detection of the network issues using improved K-mediod Clustering method. Evaluate the performance metrics with accuracy rate and ROC etc.

Upload the dataset from the dataset and used Kyoto 2006++ and KDD cup dataset.

Dataset: The new trained data was about four gigabyte of the compress double TCP dump data through 7week network system traffic. This was controlled in 5 million linking measures. The two week of the testing data consists 2 million scheduled at the similar time. The connection is sequence of the TCP information data frames initialised and stops at certain time period where data flows in direction of IP address of source to destination by regular protocol. Every connection is labelled as equal normal and along special attack. Every connection score consist of 100 bytes.



Pre-processed Data

|  | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 0 | 181 | 5450 |
| 2 | 0 | 239 | 486 |
| 3 | 0 | 235 | 1337 |
| 4 | 0 | 219 | 1337 |
| 5 | 0 | 217 | 2032 |
| 6 | 0 | 217 | 2032 |
| 7 | 0 | 212 | 1940 |
| 8 | 0 | 159 | 4087 |
| 9 | 0 | 210 | 151 |
| 10 | 0 | 212 | 786 |
| 11 | 0 | 210 | 624 |
| 12 | 0 | 177 | 1985 |
| 13 | 0 | 222 | 773 |
| 14 | 0 | 256 | 1169 |
| 15 | 0 | 241 | 259 |
| 16 | 0 | 260 | 1837 |

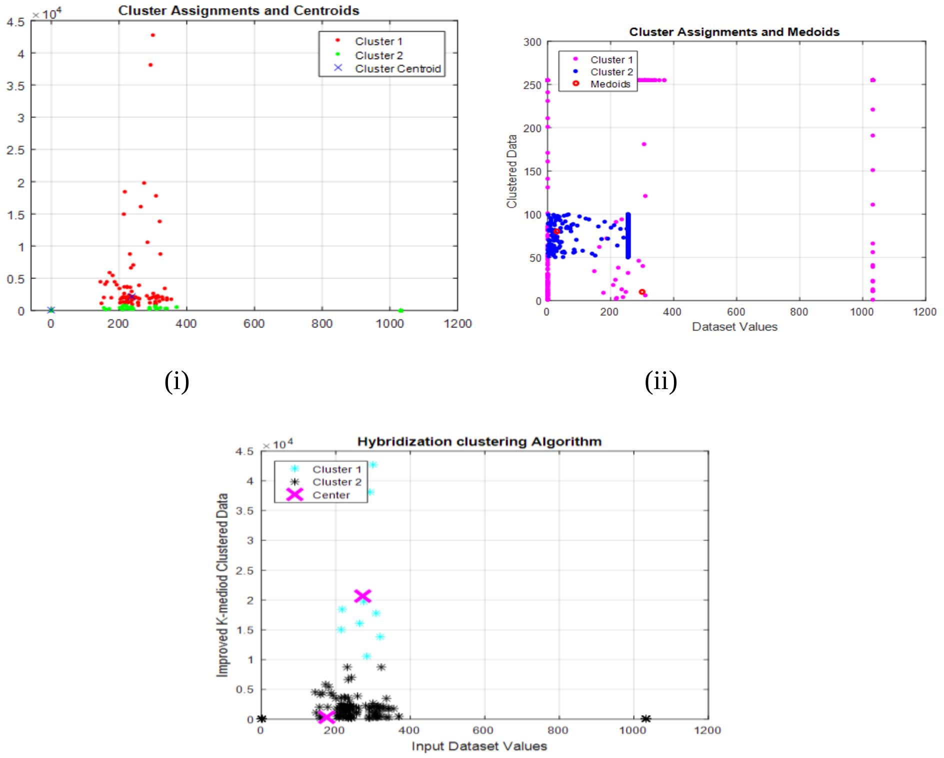
Data Representation

|  | 1 | 2 | 3 |
|---|---|---|---|
| 1 | duration | protocol_type | service |
| 2 | 0 | tcp | http |
| 3 | 0 | tcp | http |
| 4 | 0 | tcp | http |
| 5 | 0 | tcp | http |
| 6 | 0 | tcp | http |
| 7 | 0 | tcp | http |
| 8 | 0 | tcp | http |
| 9 | 0 | tcp | http |
| 10 | 0 | tcp | http |
| 11 | 0 | tcp | http |
| 12 | 0 | tcp | http |
| 13 | 0 | tcp | http |
| 14 | 0 | tcp | http |
| 15 | 0 | tcp | http |
| 16 | 0 | tcp | http |

Fig.2: Upload dataset

Fig 2 demonstrated the pre-processed data in internal dataset. It is the data technique that contains the transformation of the raw dataset in to data based on actual format. The real database is insufficient, reducing in unique exclusions. It is resolving method and consist the raw data for image processing. The figure explains the graphical representation that determines whole matrix information and plots the data values in graphical format.

Fig 3(i) below shows K means clustering is dependent on the division of the cluster approach and invented by J B Macqueen. This approach is unverified that is utilised in data mining process detection. It aims at decreasing of the cluster presentation index value, squaring fault value and fault method of this approach. The optimisation of the resulted values aims to divide k model to satisfy specific model. Initially, selection of some methods as initialised cluster based data values. Other is gathering of the remained data samples to focalised facts related with the method of decreased displacement and after than initialised organised structure for the modification clustered focal data points. Fig 3(ii) defines in this technique, before the calculation of the displacement of the data values to centroid of the clusters; k clustering is random selection of the n data values so that initialised data segmentation is built related to nearness of the data object to clustered centroid from starting of the segmented information. In addition, selection values are engaged regularly unless desired suitable segmentation data value is acquired. In this technique, after the selection, the data objects from every clustered sample are selected that is dependent on the enhancement of the clustered superiority.Fig 3(iii) shows this algorithm is also called as partitioning approach in which the amount of the clusters is determined previous and that segmented into amount of the clusters. In improved K mediod clustering, the method is appropriate for larger dataset.



(i)

(ii)

(iii)

Fig.3: (i) K-means (ii) K-mediod and (iii) Improved clustering Methods

Detection System of standard database and intrusion information data values in uploaded database.

Actual information value is 503; Intrusion information value is 202;

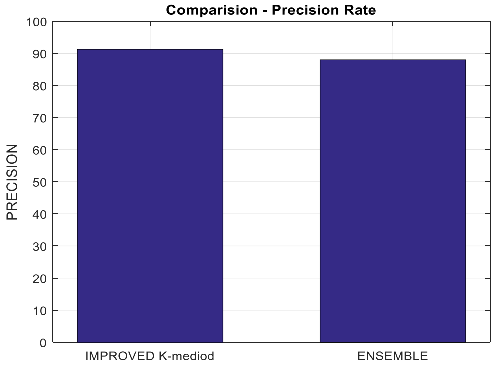


Fig.4: Comparison – Precision rate

Figure 4 shows the comparison between proposed and existing methods with precision rate. In precision rate calculates the high level performance of the network system and Precision value is 99.1%.

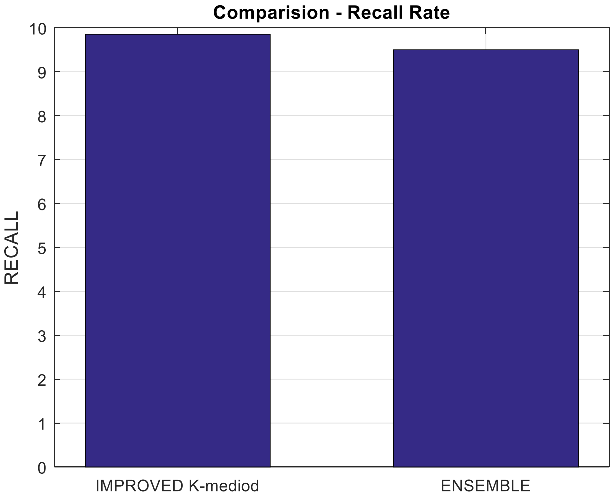


Fig.5: Comparison –Recall rate

Here fig 5 shows the comparison proposed and existing methods with recall parameter. In Recall parameters to detect the true negative rate as compared to existing one. In recall value is 98% achieved as compared to existing method.

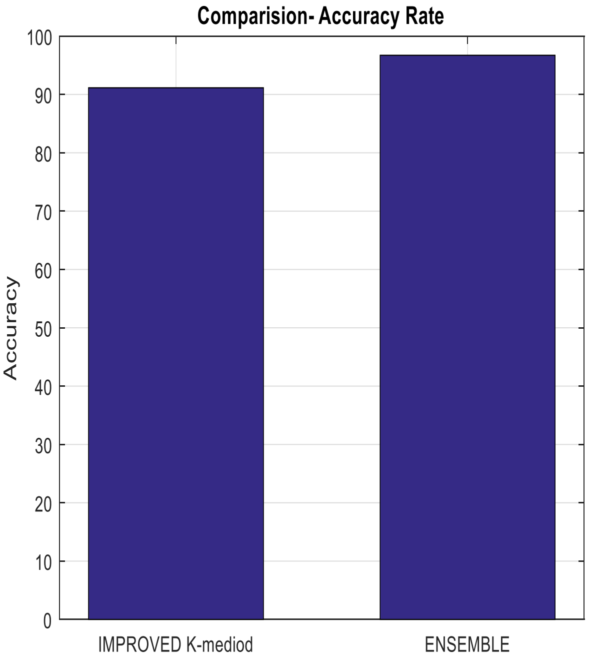


Fig.6: Comparison – Accuracy Rate

Above fig 6 defines the comparison between proposed and existing work parameters. In comparison shows the accuracy rate value achieved 98%.Proposed method improves the accuracy rate as compared to existing method.

Table 1. Comparison Analysis

| Parameters | Improved k-mediod clustering | Ensemble |
|---|---|---|
| Acc | 98 | 96 |
| Pre | 99 | 88 |
| Rec | 98 | 95 ~9.5 |

Table 2 shows the comparison between proposed and exitsing methods with accuracy rate, precision and recall. In proposed method achieved the accuracy rate is 98%, precision 99 % and recall is 98%.

In existing method performance metrics with accuracy rate is 96 %, Precision rate value is 88 % and recall value is 95%.

## V. CONCLUSION AND FUTURE SCOPE

Many areas like health system, business , retail stores link the data mining applications with information, design appreciation, and other significant equipment to accomplish data analytics. Generally an intrusion detection scheme is required to be developed to protect against the cyber threats. The main application area in intrusion detection system is dependent on the data mining method and it solve the issue of the large quantity of data. Generally, threat has the capability to determine the system traffic that has become the possible intrusion. Moreover, unconfirmed network have the possibility of the unidentified threats. Moreover, an intrusion detection system is required to be implemented to improve the security by detecting the threats. In proposed approach, an intrusion is detected by intrusion detection system. The dataset s uploaded to extract the unidentified in database though pre-processing step. After that, clustering method is the main work of the separation of the related data into clusters, so that whole values are same to other data values in same cluster than other groups. In addition, segmentation of the database in to groups takes place in clustering process. The method is dependent on the distribution of the dataset into clusters. K-mediod algorithm computes the average value of the dataset of data items. The collection of the data objects is clusters then portioning the clusters is called as improved k mediod clustering. Experimental analysis is done to evaluate the performance like as FAR, FRR , accuracy, precision and recall rate.In research method has concluded accuracy achieved value is 98%, Precision rate value is 99% and recall value is 98% as compared to the existing ensemble method.

Future scope is emphasis on the enhancement of the processing time of detection of intrusion using machine learning approaches. In addition, Convolutional Neural Network can be applied on dataset for evaluating the performance and compared with the proposed detection approach. Moreover, the performance of the anomaly detection technique can be improved using machine learning with SVM method.

## VI. REFERENCES

[1]. Vokorokos, L., Kleinová, A. and Latka, O. (2006). Network security on the intrusion detection system level. In *2006 International Conference on Intelligent Engineering Systems*(pp. 270-275). IEEE.

[2]. Titorenko, A. A. andFrolov, A. A. (2018). Analysis of modern intrusion detection system. In *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)* (pp. 142-143). IEEE.

[3]. Raghunath, B. R.andMahadeo, S. N. (2008). Network intrusion detection system (NIDS). In *2008 First International Conference on Emerging Trends in Engineering and Technology* (pp. 1272-1277). IEEE.

[4]. Asif, M. K., Khan, T. A., Taj, T. A., Naeem, U and Yakoob, S. (2013). Network intrusion detection and its strategic importance. In *2013 IEEE Business Engineering and Industrial Applications Colloquium (BEIAC)* (pp. 140-144). IEEE.

[5]. Nayak, B. S. (2013). Research on application of intrusion detection system in data mining.

[6]. Brutch, P and Ko, C. (2003). Challenges in intrusion detection for wireless ad-hoc networks.In *2003 Symposium on Applications and the Internet Workshops, 2003.Proceedings.* (pp. 368-373). IEEE.

[7]. Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G. and Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers& security*, *28*(1-2), 18-28.

[8]. Latha, S andPrakash, S. J. (2017). A survey on network attacks and Intrusion detection systems.In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1-7).IEEE.

[9]. Qin, S. J., Wen, Q. Y and Zhu, F. C. (2007). An external attack on the Brádler–Dušek protocol. *Journal of Physics B: Atomic, Molecular and Optical Physics*, *40*(24), 4661.

[10]. Scott, A. D. (2014). *U.S. Patent No. 8,712,596*. Washington, DC: U.S. Patent and Trademark Office.

[11]. Rowland, C. H. (2002). *U.S. Patent No. 6,405,318*. Washington, DC: U.S. Patent and Trademark Office.

[12]. Jaiganesh, V., Mangayarkarasi, S. and Sumathi, P. (2013). Intrusion detection systems: A survey and analysis of classification techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(4), 1629-1635.

[13]. Saxena, A. K., Sinha, S and Shukla, P. (2017). General study of intrusion detection system and survey of agent based intrusion detection system. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 471-421).IEEE.

[14]. Horng, S. J., Su, M. Y., Chen, Y. H., Kao, T. W., Chen, R. J., Lai, J. L. and Perkasa, C. D. (2011). A novel intrusion detection system based on hierarchical clustering and support vector machines. *Expert systems with Applications*, *38*(1), 306-313.

[15]. Wei, K., Huang, J. and Fu, S. (2007). A survey of e-commerce recommender systems.In *2007 international conference on service systems and service management* (pp. 1-5).IEEE.

[16]. Suykens, J. A and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), 293-300.

[17]. Ran, Z. (2012). A model of collaborative intrusion detection system based on multi-agents.In *2012 International Conference on Computer Science and Service System* (pp. 789-792).IEEE.

[18]. Sadeghi, Z. and Bahrami, A. S. (2013). Improving the speed of the network intrusion detection.In *The 5th conference on information and knowledge technology* (pp. 88-91).IEEE.

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

[20]. Arashloo, S. R., Kittler, J. and Christmas, W. (2017). An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE Access*, *5*, 13868-13882.

[21]. Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C and Osman, M. R. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, & Soil Pollution*, *225*(8), 2063.

[22]. Jin, S., Jiang, Y and Peng, J. (2018). Intrusion Detection System Enhanced by Hierarchical Bidirectional Fuzzy Rule Interpolation.In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 6-10).IEEE.