HYBRID INTRUSION DETECTION SYSTEM MODEL USING IMPROVED J48 AND IMPROVED K-MEANS CLUSTERING ALGORITHM

[1]Sonali Soni ,[2]Pooja Sharma
[1]Student in Dept of CSE at IKG Punjab Technical University main campus Kapurthala, Punjab,India
[2]Assistant Professor in Dept of CSE at IKG Punjab Technical University main campus Kapurthala, Punjab,India

Abstract –The region of intrusion detection is the focal idea in general system and computer security engineering. It is a critical innovation in business segment and additionally in explore zone. By observing the computer and system assets, Intrusion Detection System (IDS) identifies any of the abuse or unapproved get to which is fundamentally an attack to these assets. At that point it alarms and advises head for event of an intrusion. A few techniques can be utilized to recognize an intrusion. In this paper a new intrusion detection system is proposed using improved K-Means clustering and E-J48 algorithm. Experimental results demonstrate that the proposed technique outperforms the existing technique.

Keywords – Intrusion detection system, network security, k-means clustering, E-J48 algorithm.

I.INTRODUCTION

Intrusion Detection Systems will be frameworks that screen PC framework occasions to find the noxious or suspicious exercises in the framework and issue caution when such thing happens. In the present undertaking condition, hurtful exercises are expanding step by step. In this way IDSs have turned into an imperative and basic piece of the associations [1]. Over the previous years, arrange security issues have been raised because of the fast development of systems. Association's system framework is normally shown to the expanding number of dangers from the outside programmers and additionally inside programmers. The outcomes can be changes of business information, block in administrations and infringement of the security arrangements, i.e., Confidentiality, Integrity and Availability (CIA) [2].

In a wide range of system, security is an essential issue particularly in huge associations as they have critical and secret information which if get hacked will cut down organization's profile. For the most part, we secure our frameworks by building firewalls or utilize some confirmation instruments, for example, passwords or some encryption strategies which make a defensive covering around them. All the above strategies give a level of security however they can't give assurance against noxious codes, inside assaults or unsecured modems. We require greater security components, for example, IDS since firewalls can't distinguish attacks inside the system since they are for the most part conveyed at the limit of the system, and in this way just control movement entering or leaving the system. In any case, a tremendous level of interruptions might be from inside the system and IDS can screen and break down different occasions in the system and if the framework has been abused it gives quick answer to the overseer [3].

First IDS at any point created was have based, since they are utilized to break down the framework log i.e., for dissecting the working framework log by assessing the marks with the little arrangement of patter or model. This prompts the improvement of different IDS, yet the security blemishes were in expanded numbers; the framework execution were corrupting because of these security imperfections. This situation prompts undeniable improvement of IDS at the soonest. Later IDS increased all the convention mindfulness, used to dissect the parcel, bundle structures and so forth., which predicts the known bundle movement characterized to be noxious. Presently multi day's ongoing IDS can anticipate the different attacks or sorts of interruption through the system in light of the fact that the ongoing advancement in IDS prompts have correspondence, in fabricated security highlights, convention mindfulness, parcel analyzing and so on [3].

II. RELATED WORK

Patil et al. incorporated examination of such Hybrid frameworks which are actualized by utilizing the benchmark dataset. Intrusion Detection System (IDS) are said to be more powerful when it has both high interruption recognition (genuine positive) rate and low false alert (false positive). In any case, current IDS when actualized utilizing information mining approach like grouping, arrangement alone can't give 100 % recognition rate consequently need adequacy [4]. So as to beat these challenges of the current frameworks, numerous scientists executed intrusion discovery frameworks by incorporating grouping and arrangement approach like k-implies and Fuzzy rationale, K-implies and hereditary calculation, a portion of the specialist additionally attempted utilization of Decision tree and Neural Network to recognize obscure assaults.

Algaith et al.introduce an approach for dissecting guard top to bottom, and represent the utilization of the approach with an exact investigation in which authors have surveyed the discovery capacities of interruption identification frameworks when sent in different, two-rendition, parallel safeguard inside and out arrangements [5]. Barrier top to bottom is a term regularly utilized as a part of security writing to signify structures in which numerous security insurance frameworks are sent to safeguard the profitable resources of an association (e.g. the information and the administrations). The designs have been surveyed in settings that support recognition of attacks, and also settings that support genuine activity.

Aljawarneh et al.built up an upgraded J48 calculation, which utilizes the J48 calculation for reinforcing the discovery exactness and the execution of the unconventional IDS method. This upgraded J48 calculation supposedly allows in a effective discovery of potential assaults which could imperil the system classification [6]. For this motive, the scientists utilized severa datasets by incorporating distinct methodologies like the J48, Naive Bayes, Random Tree and the NB-Tree. A NSL KDD interruption dataset turned into related while finishing all trials. This dataset changed into isolated into 2 datasets, i.E., making ready and checking out, which relied on the statistics handling. From that factor, an element choice approach in view of the WEKA application became utilized for assessing the viability of the

giant quantity of highlights. The results received endorse that this calculation tested a advanced, precise and more efficient execution without utilising the previously cited highlights whilst contrasted with the thing choice approach.

Singh et al. characteristic the quality first-rate desire method which enhances the precision of the calculations. The best worry of Network is safety. Introduction finds the traps and contraptions of the Attackers. Information Mining techniques clearly absorb the example of the tuples and Intelligent desire are made [7]. Managed gaining knowledge of strategies finds the attack in view of beyond statistics and obscure attacks are recognized via utilising Unsupervised gaining knowledge of. Dos, Probe and Normal facts are efficaciously diagnosed by most excessive Data Mining calculations, even as True Positive Rate of R2L and U2R are low. The Hybrid strategies are utilized to enhance True Positive Rate of R2L attacks. NSL_KDD Training and Testing dataset are utilized.

Kaur et al. proposed a hybrid K-way and Support Vector Machine calculation for malady expectation. Hybrid K-means and Support Vector Machine calculation for illness forecast. The proposed 1/2 and half of K-implies calculation is useful in selecting beginning centroids, range of bunches and furthermore to decorate the productivity of K-implies calculation. The hybrid K-implies calculation is utilized for dimensionality diminishment of the dataset that is given as a contribution to Support Vector Machine classifier. The pastime is achieved in MATLAB and from the outcomes it's been broke down that the exactness of the order is improved and the getting ready time to get the closing yield is reduced [8].

Chitrakar et al. apply hybrid learning technique with the aid of joining ok-Medoids primarily based grouping machine took after by Naïve Bayes characterization strategy. The part of Intrusion Detection System (IDS) has been unavoidable inside the area of Information and Network Security – highly to construct a respectable gadget safeguard foundation. Oddity based totally interruption place approach is one of the building squares of such an status quo. Due to the way that ok-Medoids bunching strategies communicate to this present truth scenario of information dissemination, the proposed upgraded method will gather the whole statistics into bearing on businesses extra exactly than ok-Means with the cease purpose that it brings about a advanced characterization [9]. An evaluation is carried out so that you can assess execution, precision, location charge and fake fantastic charge of the order plot. Results and investigations reveal that the proposed technique has improved the identity rate with least false high quality fees.

Kumar et al. analyze a kind model for misuse and anomaly attack detection using decision tree set of rules. Intrusion Detection System (IDS) is the most effective device that might cope with the intrusions of the computer environments with the useful resource of triggering indicators to make the analysts take moves to save you this intrusion. IDS's are based absolutely on the notion that an outsider's conduct is probably relatively wonderful from that of a legitimate man or woman [10]. A sort of intrusion detection structures (IDS) had been employed for defensive computer systems and networks from malicious attacks with the aid of using conventional statistical strategies to new information mining strategies in ultimate a long term. However, trendy commercially available intrusion detection systems are signature primarily based that are not able to detecting unknown attacks.

Ullah et al. introduce a clean out-based characteristic desire model for anomaly-primarily based intrusion detection structures. Feature choice is an crucial element in modeling anomaly-based totally intrusion detection systems. An beside the point characteristic can result in overfitting and feature an impact at the modeling electricity of sophistication algorithms [11]. The purpose of function selection is to take away irrelevant and redundant attributes from the dataset to enhance the predictive electricity of a category set of regulations. The proposed model evaluates the talents based totally on records gain via considering consistency, dependency, facts, and distance of every characteristic. The experimental consequences display that our proposed model has a key effect in decreasing computational and time complexity. The accuracy of the proposed model turn out to be measured as ninety nine.70 % and ninety nine.Ninety% for the ISCX and NSL-KDD datasets respectively.

Coppolino et al. advise a hybrid, mild-weight, allocated Intrusion Detection System (IDS) for wi-fi sensor networks. This IDS uses each misuse-based totally completely and anomaly-based totally totally detection techniques. It is composed of a Central Agent, which performs surprisingly correct intrusion detection thru using records mining strategies, and some of Local Agents walking lighter anomaly-based absolutely detection strategies on the motes. Decision timber were followed as class set of policies inside the detection device of the Central Agent and their behaviour has been analysed in selected attacks scenarios [12]. The accuracy of the proposed IDS has been measured and validated via an intensive experimental marketing campaign. This paper gives the outcomes of these experimental assessments.

Gadal et al. proposes a hybrid machine learning approach for community intrusion detection based totally on aggregate of Kmeans clustering and Sequential Minimal Optimization (SMO) kind. It introduces hybrid method that able to lessen the charge of fake nice alarm, fake terrible alarm fee, to beautify the detection rate and locate 0-day attackers. The NSL-KDD dataset has been used inside the proposed technique [13]. The elegance has been finished with the aid of the use of Sequential Minimal Optimization. After schooling and trying out the proposed hybrid machine gaining knowledge of technique, the effects have validated that the proposed approach (K-propose + SMO) has achieved a powerful detection rate of (99.Forty eight%) and decrease the fake alarm price to (1.2%) and completed accuracy of (ninety seven.3695%).

Mathew et al. depicts a targeted writing review of machine learning and facts digging techniques for virtual investigation in assist of interruption location. An interruption discovery framework is programming that displays a solitary or a device of PCs for noxious sports which can be long gone for taking or controlling facts or debasing gadget conventions. Most approach utilized as a part of the existing interruption reputation framework isn't always geared up to manage the dynamic and complicated nature of virtual attacks on PC systems [14]. Despite the reality that effective versatile techniques like special structures of gadget getting to know can bring about better identity quotes, bring down false warning fees and practical calculation and correspondence price. With the usage of information mining can bring about everyday instance mining, characterization, bunching and smaller than anticipated statistics move. In light of the amount of references or the significance of a growing approach, papers talking to each method have been recognized, perused, and condensed. Since records are so critical in system gaining knowledge of and records mining processes, a few extremely good virtual informational collections applied as a part of gadget studying and statistics digging are depicted for virtual safety is exhibited, and some suggestions on whilst to utilize a given method are given.

Rutravigneshwaran et al. investigate the skillability of device getting to know techniques in interruption identity framework, collectively with arrangement tree and bolster vector system, with the anticipate of given that reference for building up interruption area framework in destiny. IDS is a product result screens the mortification or conduct in addition to investigate any unsuitable pastime gift itself. Incredible increment and custom of internet brings concerns up in connection to a way to guard and impart the automated all collectively in a sheltered method [15]. These days, programmers utilize specific types of attacks for getting the profitable statistics. Contrasted and in addition interrelated works in records mining primarily based interruption identifiers precision, reputation price, false alert charge.

III. PROPOSED METHODOLOGY

The proposed system consists of following steps:

- A novel hybrid intrusion detection system based on data mining techniques is proposed. NSL-KDD dataset is used for evaluation of results and training and testing of proposed hybrid intrusion detection system.
- In this system first Misuse based intrusion detection model is designed using data mining algorithm enhanced k-means clustering which is used to make clusters with training dataset.
- First collect raw data from NSL-KDD dataset.
- Preprocess the raw data and filter the data to remove missing values and remove noise if present.
- Design Misuse detection model with enhanced k- means clustering algorithm. In misuse detection model uses training dataset of NSL-KDD train.csv file which is used to train our misuse detection phase.
- The size of cluster is fixed and the output of the first phase forms initial clusters.
- The cluster sizes vary and the output of this phase are the finalized clusters. Initial clusters are inputs for this phase.
- Then Anomaly detection phase is designed based on output of misuse detection phase. The output of misuse detection model is used as training data to anomaly detection model.

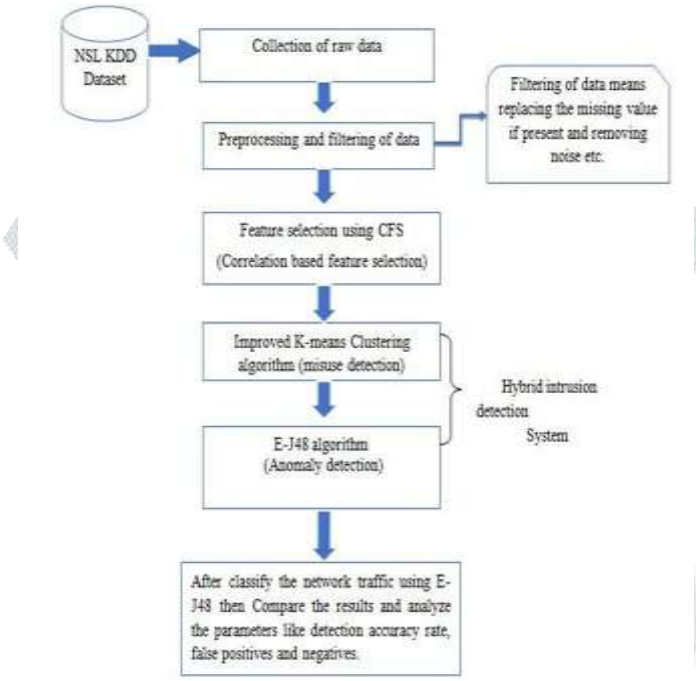

Figure 1: Flowchart of Proposed Technique

IV.EXPERIMENTAL RESULTS

This section presents experimental results of the proposed technique.
Proposed technique, improved intrusion detection system using improved K–Means clustering and E-J48 algorithm. Performance of proposed technique is evaluated on basis of various parameters:

- Accuracy
- Root Absolute Error
- Root Mean Squared Error
- Kappa Statistics
- Correctly Classified Instances
- Incorrectly Classified Instances

Accuracy

Accuracy is the basic measure for arrangement execution. Exactness can be estimated as accurately ordered occurrences to the aggregate number of occasions, while mistake rate utilizes inaccurately arranged cases rather than effectively grouped examples.

$$\text{Accuracy}=\frac{\text{True Positive}+\text{True Negative}}{\text{True Positive}+\text{False Positive}+\text{True Negative}+\text{False Negative}}$$

Table 1: Comparison of proposed technique with existing techniques on basis of accuracy

J48	Improved J48	Proposed Technique
97.81	98.67	99.43

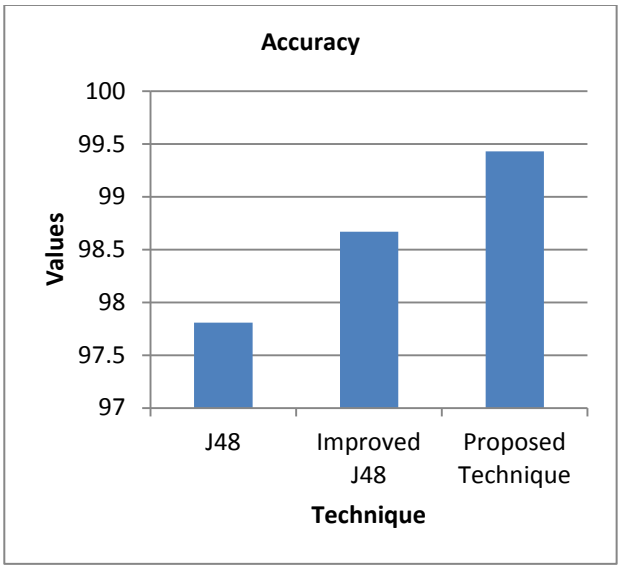

Figure 2: Accuracy comparison of proposed technique with existing techniques

Root Absolute Error and Root Mean Squared Error

Root absolute error

The Mean Absolute Error (RMAE) is the average of the absolute value of the error and very similar to the RMSE but is less sensitive to large errors.

$$RMAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Root mean squared error

The Root Mean Squared Error (RMSE) is the square root of the average squared distance of a data point from the fitted line.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Table 2: Comparison of proposed technique with existing techniques on basis of root absolute error and root mean squared error

Parameter/Technique	J48	Improved J48	Proposed Technique
Root Absolute Error	0.0302	0.0205	0.0083
Root Mean Squared Error	0.1356	0.1101	0.0706

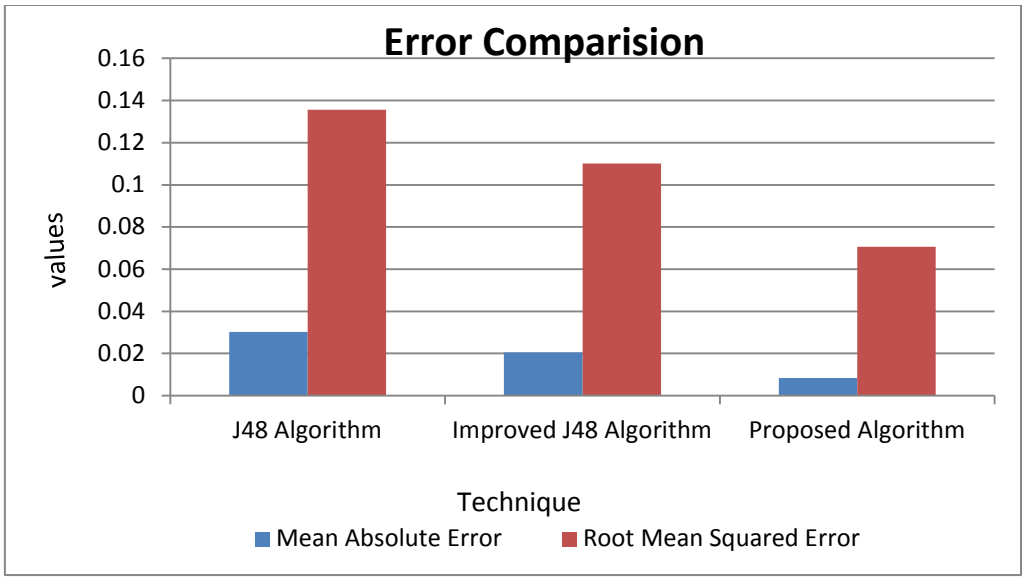

Figure 3: Error comparison of proposed technique with existing techniques

Kappa Statistic

Measures the relationship between classified instances and true classes. It usually lies between [0, 1]. The value of 1 means perfect relationship while 0 means random accuracy.

Kappa Statistic= total accuracy-random accuracy/1-random accuracy

Table 3: Comparison of proposed technique with existing techniques on basis of Kappa Statistics

J48	Improved J48	Proposed Technique
0.956	0.9732	0.99

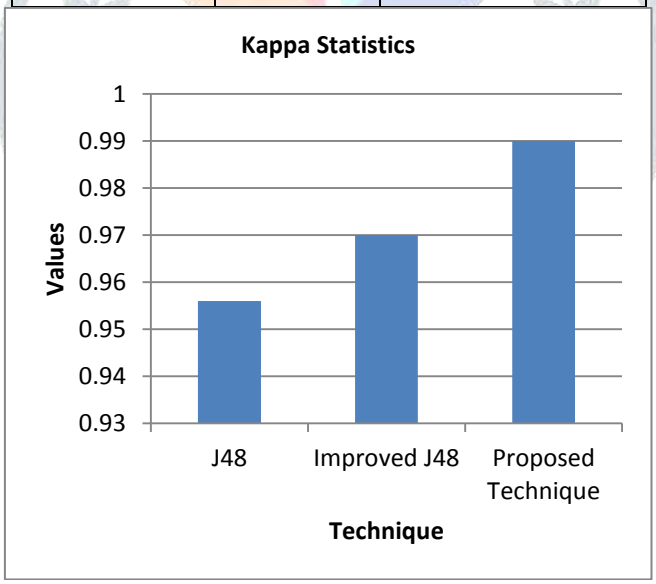

Figure 4: Kappa Statistics comparison of proposed technique with existing techniques

Correctly Classified Instances and Incorrectly Classified Instances

Correctly classified instances

Correctly classified instances are the sum of True positive (TP) and True Negative (TN) classes and the number of correctly classified instances divided by the total number of instances gives the accuracy.

Correctly classified instances = TP+TN

Incorrectly classified instances

Incorrectly classified instances are the sum of False Positive (FP) and False Negative (FN).

Incorrectly classified instances = FP+FN

Table 4: Comparison of proposed technique with existing techniques on basis of correctly classified instances and incorrectly classified instances

Parameter/Technique	J48	Improved J48	Proposed Technique
Correctly Classified Instances	97.81	98.67	99.43
Incorrectly Classified Instances	2.19	1.33	0.57

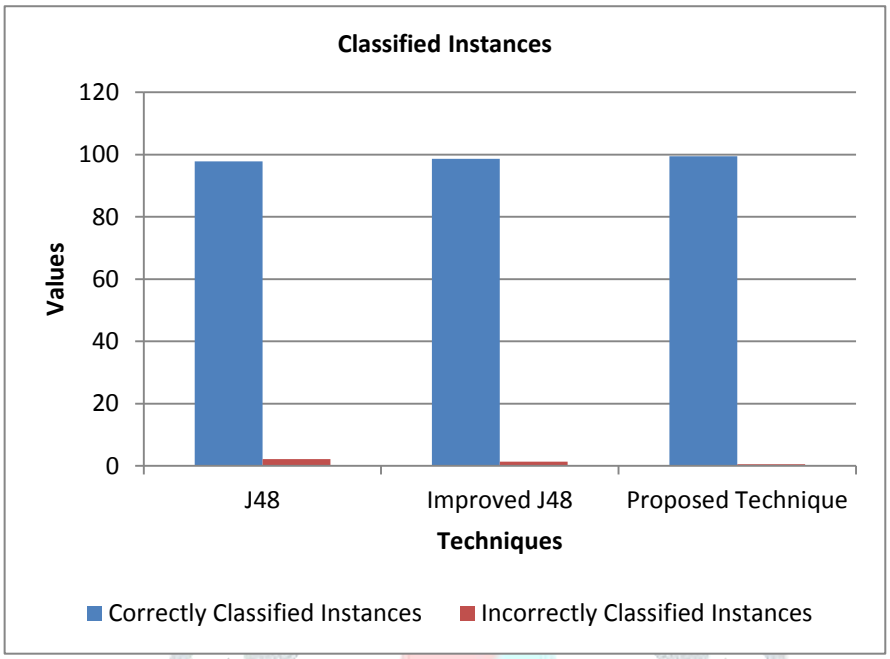

Figure 5: Classified Instances comparison of proposed technique with existing techniques

V. CONCLUSION AND FUTURE WORK

In this paper a new intrusion detection system is proposed using improved K-Means clustering and E-J48 algorithm. A novel hybrid intrusion detection system based on data mining techniques is proposed. NSL-KDD dataset is used for evaluation of results and training and testing of proposed hybrid intrusion detection system. Performance of proposed technique is evaluated on basis of various parameters like accuracy, root absolute error, root mean squared error, kappa statistics, correctly classified instances, and incorrectly classified instances. Experimental results

REFERENCES

[1] Kajal Rai, M. Shyamala Devi, "Intrusion Detection Systems: A Review", Journal of Network and Information Security, Vol. 1, Iss. 2 , December 2013, pp. 15-21.
[2] Sheenam, Sanjeev Dhiman, "Comprehensive Review: Intrusion Detection System and Techniques", IOSR Journal of Computer Engineering, Vol.18, Iss. 4,2016, pp. 20-25.
[3] Ajay kaurav, S.Sibi Chakkaravarthy, Pravin R.Patil, M.Vimal Karthik, "Intrusion Detection system: A Review of the state of the art", IOSR Journal of Computer Engineering, Vol. 16, Iss. 1,2014, pp. 108-112.
[4] Purushottam R. Patil, Yogesh Sharma, Manali Kshirasagar, "Performance Analysis of Intrusion Detection Systems Implemented using Hybrid Machine Learning Techniques", International Journal of Computer Applications, Vol. 133, No. 8, January 2016, pp. 35-38.
[5] Areej Algaith, Ivano Alessandro Elia, Ilir Gashi, Marco Vieira, "Diversity with Intrusion Detection Systems: An Empirical Study", pp. 1-5.
[6] Shadi Aljawarneh, Muneer Bani Yassein, Mohammed Aljundi, "An enhanced J48 classification algorithm for the anomaly intrusion detection systems", Springer, 2017.
[7] Varsha Singh, Shubha Puthran, Avanish Tiwari, "Intrusion Detection Using Data Mining with Correlation", IEEE, International Conference for Convergence in Technology, pp. 620-625, 2017.
[8] Sandeep Kaur, Dr. Sheetal Kalra, "Disease Prediction using Hybrid K-means and Support Vector Machine", IEEE, 2016.
[9] Roshan Chitrakar, Huang Chuanhe, "Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k- Medoids Clustering and Naïve Bayes Classification", IEEE, 2012.
[10] Manish Kumar, Dr. M. Hanumanthappa, Dr. T. V. Suresh Kumar, "Intrusion Detection System Using Decision Tree Algorithm", IEEE, pp. 629-634, 2012.
[11]Imtiaz Ullah, Qusay H. Mahmoud, "A Filter-based Feature Selection Model for Anomaly-based Intrusion Detection Systems", IEEE, International Conference on Big Data, pp. 2151-2159, 2017.
[12]Luigi Coppolino, Salvatore D'Antonio, Alessia Garofalo, Luigi Romano, "Applying Data Mining Techniques to Intrusion Detection in Wireless Sensor Networks", IEEE, International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, pp. 247-254, 2013.

[13] Saad Mohamed Ali Mohamed Gadal, Rania A. Mokhtar, "Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique", IEEE, International Conference on Communication, Control, Computing and Electronics Engineering, 2017.
[14]Jithin Mathew, S. Ajikumar, "Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Vol. 2, Issue. 2, pp.92-97, March-April.2017.
[15]P. Rutravigneshwaran, "A Study of Intrusion Detection System using Efficient Data Mining Techniques", Int. J. Sci. Res. in Network Security and Communication, Vol. 5, Issue. 6, pp. 5-8, Dec 2017.