An Efficient Rapid Intrusion Detection Method For Detecting Intrusions In Networks

Ch. Kiran Kumar, M.Govindarajan

Abstract: The utilization of networks and particularly the Internet has turned into a major task of day by day life. As indicated by fast advancement and across the broad utilization of network frameworks, mixed interested methodologies have developed broadly in the ongoing years. Numerous security strategies have been utilized so as to deal with the security network dangers. In this manner, the utilization of intrusion identification frameworks as an extra resistance network is practically imperative. An Intrusion Detection System (IDS) progressively screens the activities occurring in a framework, and chooses whether these activities are symptomatic of an interruption or establish an authentic utilization of the framework. In this proposed work, an efficient Rapid Intrusion Detection Method (RIDM) is proposed to discover best subset of dataset to prepare our expectation model. The proposed method is compared with the existing IDS model and the results proved that the False Positive Rate (FPR) is very low for the proposed work and the Detection Rate (DR) is high in the proposed work. The proposed model performance and accuracy is high than the traditional methods..

Index Terms: Intrusion Detection, Networks, Data security, Attacks, classification, feature extraction..

1. INTRODUCTION

The tremendous development in communication technology has empowered the vision of consistent connectivity. This has provided for inter-connectivity between wired and wireless networks, infrastructure based and ad- hoc networks, thus facilitate heterogeneous devices to communicate with each other. With the number of devices and applications expanding exponentially, networking infrastructure has to deal with excessive amount of data traffic. This information will consist of helpful as well as malicious information[1]. And with increased advantages the burden of increasing number of types of attacks and threats is also to be faced in the network[2]. This threat is manifold and must be mitigated with deployment of new security techniques. These techniques must ensure security of data as well as administrative and legitimate privileges of users[3]. A large portion of the attacks are basically intrusions into the system in the form of malware, bots, viruses, worms and Trojans. With the huge growth in technology research is going on to create various security methods to make security of network non -vulnerable and to ensures the privacy and data of user from attackers but attackers comes with various different complex ideas to crack those mechanism[4]. So we need to develop something different and complex and non-vulnerable security mechanism and security models so we can ensure individual privacy[5]. For building a prediction model the data plays an important role. The data which we use for our prediction model is KDD Cup 99 dataset that is the publicly available data set used for the detection of intrusions. The dataset is huge, it has various features but it not necessary that all the features are important[6]. Some features in the dataset may be irrelevant for our model because it degrades the performance of security model and increases the false rate[7]. So before use of data for training purpose, we have to use data preprocessing techniques to remove the noise from data.

- *[1]Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar – 608002, Tamil Nadu, India. E-mail: kirankumar.chanumolu@gmail.com*
- *[2]Assoc.Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar – 608002, Tamil Nadu, India.*

E-mail: govind_aucse@yahoo.com

An IDS is used for identifying abnormal activities by examining various constraints, which are listed as below:

- Traffic in network
- Utilization rate of CPU
- I/O utilization
- Location of User
- Log file, etc.

The IDS system gives reliable response to prevent the intrusions in the system[8]. It acts like a defensive method against many types of attacks in distributed networks[9]. IDS is defined as a system or an application, which observes a network system in which any malicious thing arise or not and provides report to central station. Figure 1 depicts IDS model.

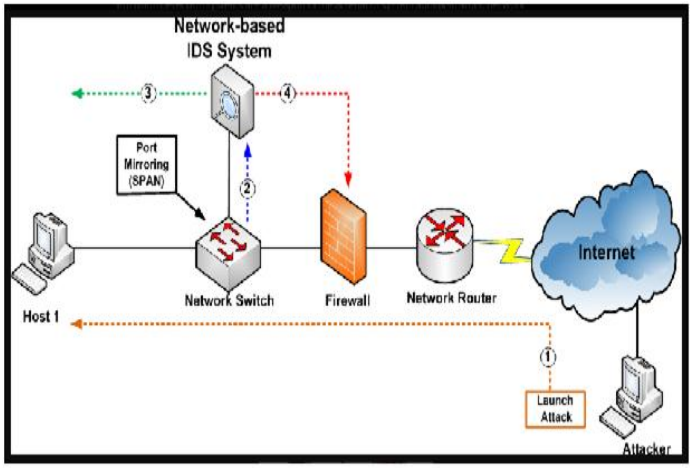

***Fig-1** Intrusion Detection System*

The network traffic is monitored by network IDS. It analyzes the malicious activity of the network and makes the decision on the purpose of traffic[10]. The network IDS were easy to act on a network and monitors the traffic line from multiple systems at a time.

2. LITERATURE SURVEY

Different types of intrusions are listed below.

- Attempted break-ins: It's detected by unusual behavior or if there is violation of security rules.
- Masquerade attacks: In this type of attack, attacker acts as authentic user [11].
- Control system get penetrated: It is detected by examining the specific activity patterns.

The improved IDS developed by O. Y. Al-Jarrah et al [1] with the help of Triangle Area based KNN and Information Gain (IG). The combination of Greedy k-means clustering algorithm and SVM classifier are used to select the features for accurate detection of attacks. In feature selection, the best attributes were selected using a statistical property, namely, Information Gain (IG). The global solution is obtained by processing a global k-means algorithm from greedy k-means method. K. Wang et al [2] have presented a feature selection method using rough set approach based on reduction to increase the accuracy and performance of IDS. A dataset was selected and classified into two datasets are training and testing data set. Using the simulation tools the table was converted into appropriate file formats. The NSL_KDD dataset was provided as input to generate the reduction. The dataset contained 42 attributes, whose split factor was 0.5 and a set of reduction was obtained as output. K. Wang et al [3] recommended a hybrid feature selection algorithm to enhance the detection rate. The crossover and mutation was applied to evaluate the fitness values. The highest fitness value, namely, Queen Bee Evolution (QBE) was introduced in LGP to increase the detection rate. The strong mutation of QBE reduced the probability of premature convergence O. Joldzic et al [5] analyzed the integrated feature selection method by combining the classification methods to provide the knowledge of finding the best solution. In the testing phase, the features were classified based on the types of attacks using the knowledge provided in the previous phase. The SVM classifier obtained the average fitness value to prevent over fitting problem. D. Papamartzivanos et al [6] utilized feature selection techniques such as C5.0 and Artificial Neural Network (ANN) to segregate the data either as a normal or attack data. The ANN was the popular classification technique used to mine a large amount of data, whereas the C5.0 was used to improve the classification accuracy. The neural network was composed of a set of neuron that connected the preceding and the succeeding layers. All the neurons received many inputs and produce a single output signal. Evaluate the KDD dataset in terms of accuracy, sensitivity, and specificity are used by the Clementine software. J. Kim et al [7] have designed to predict the intrusions of the networks using a set of rules in the rule base using fuzzy logic-based system.. The fuzzy rules were generated from the definite rules automatically. For identifying the important attributes for definite rule generation, single-length frequent items were mined from both the attack and normal data. The deviation method was applied on the attributes obtained as a result of mining to frame a set of definite and indefinite rules. M. Du et al [8] have proposed a Fuzzy Genetic Algorithm (FGA) for detecting various attacks by improving the IDS using genetic algorithm. Selection, crossover and mutation for the population generated in a random manner were the process of genetic algorithm. The if-then rules were encoded as strings and 16 bit vectors were converted into one of the five linguistic values such as M: Medium, MS: Medium Small, S: Small, , L: Large and ML: Medium Large. The fuzzy set was projected as graph by calculating the membership function for each attribute. P. Mishra et al [10] have proposed to show maximized detection rate, minimized false alarm rate and time complexity were used by an IDS using fuzzy logic and clustering techniques. A multi valued logic allowed intermediate values for describing the fuzzy logic. The fuzzy inference system used the Mamdani system to map the set of inputs to outputs using if-then rules. The proposed system included the components such as packet sniffer, fuzzy inference system controller, fuzzy inference systems, decision evaluator engine and a database. The data was captured by the packet sniffer from the network traffic and preprocessed to reduce the dimensionality of the database which improve the proposed IDS evaluation.

2.1 Description of Existing Algorithms

2.1.1 Genetic Algorithm

Genetic calculation is a group of computational models dependent on advancement and common choice. A Genetic calculation is a programming system, which mirrors biological development as a critical thinking approach[17]. The Genetic calculations use strategies roused by biological ideas like inheritance, mutation, selection, and crossovers. Genetic calculations are actualized as chromosome-like information structures. A Genetic calculation has numerous parameters, administrators and procedures which choose its appearance to an ideal arrangement. The structure of the Genetic Algorithm is depicted in Figure 2.

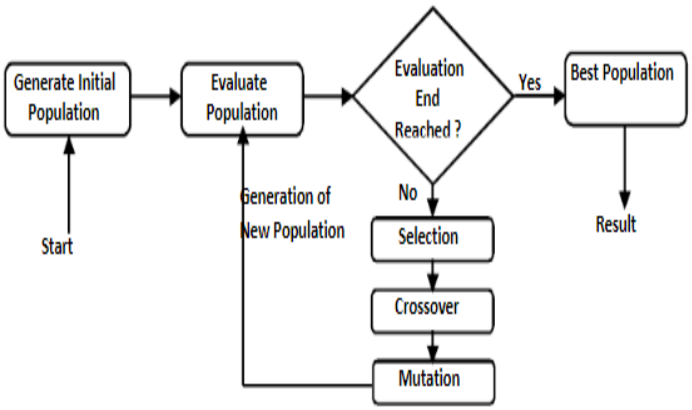

Fig 2 : *Genetic Algorithm Structure.*

2.1.2 Particle Swarm Optimization in Intrusion Detection

Particle Swarm Optimization (PSO), initially presented by Eberhart and Kennedy in 1995, is a class of random optimization calculation roused by swarm knowledge. It has been demonstrated productive at comprehending worldwide enhancement and designing issues. The benefits of PSO over numerous calculations are its usage effortlessness and capacity to join to a sensibly arrangement rapidly. PSO has been effectively utilized in numerous applications, likewise in the standard extraction for intrusion discovery[18]. In any case, a major issue of IDSs is an excessive number of manual processing. In this procedure, every individual reference to the data from different individuals from the gathering and its own experience to pick the following best searching destinations. With Multiple Iterations, every individual can pick a best site. However, the PSO method consumes time for providing better results. The process involved in PSO is depicted in Figure 3.

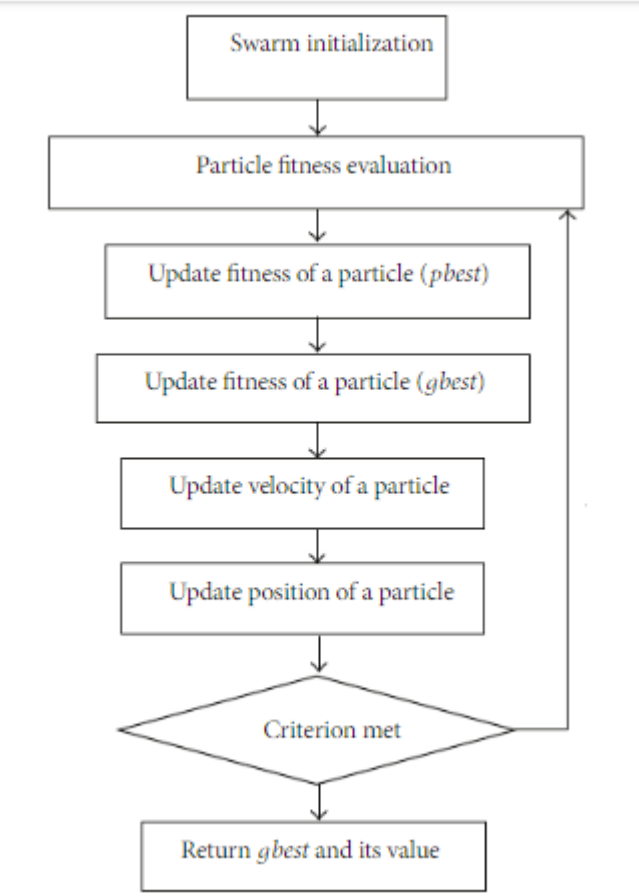

***Fig 3**: The Process of PSO*

2.1.3 Hidden Markov Model

Hidden Markov Model (HMM) is a generative model, for demonstrating input information. The model is proposed to profile TCP based communication channel for intrusions. HMM is utilized to profile source isolated traffic and the model accordingly constructed is utilized to arrange test traffic. Intrusion identification frameworks target distinguishing attacks against PC frameworks and systems or as a rule, against data frameworks. A high positive rate is the point at which the IDS says there is a security string , yet the traffic isn't vindictive or was never proposed to be malignant. The HMM architecture using forward algorithm and backward algorithm is depicted in figure 4.

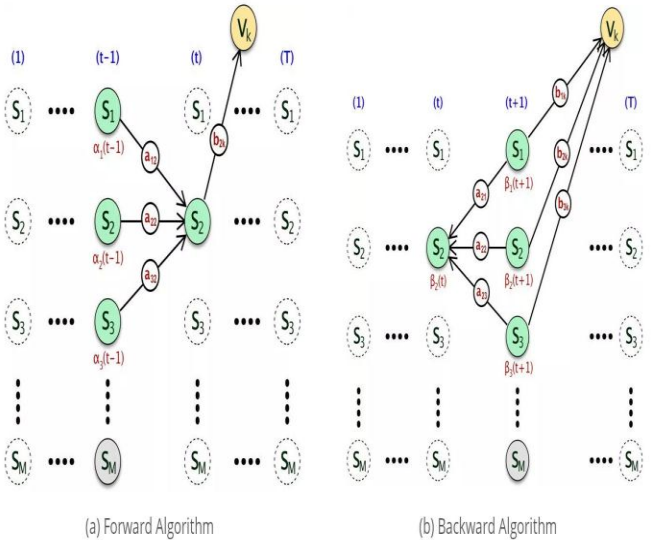

***Fig 4:** HMM Architecture*

3. Proposed Method

In the proposed work, RIDM continuously monitors the incoming and outgoing packets which pass through network for suspicious activity and enables the alarm to alert network administrator for abnormal pattern or suspicious activity of user to take indirect action to protects the network and maintain security. An RIDM can be categorized in various ways based on their working location and the way it trained with dataset. A rapid intrusion detection system takes advantage of existing and pre-defined attacks signature[12] of known attacks for training, the prediction model to protect the network. A RIDM system has less false alarm rate but it unable to detects those threats which are not pre-defined or whose signature are not available while training of model. In the proposed work a RIDM method is introduced that build a wrapper based feature selection algorithm and use those selected features to identify intrusions in the network. Initially the dataset is used for building a prediction model, this dataset contain improper values or noise. So to proceed forward data preprocessing techniques are applied to clean the data and to remove noise the data. Once we get noise free data, feature selection techniques are applied because this dataset contains both relevant and irrelevant features, which reduce the performance of IDS prediction model. The proposed model framework is depicted in Figure-5.

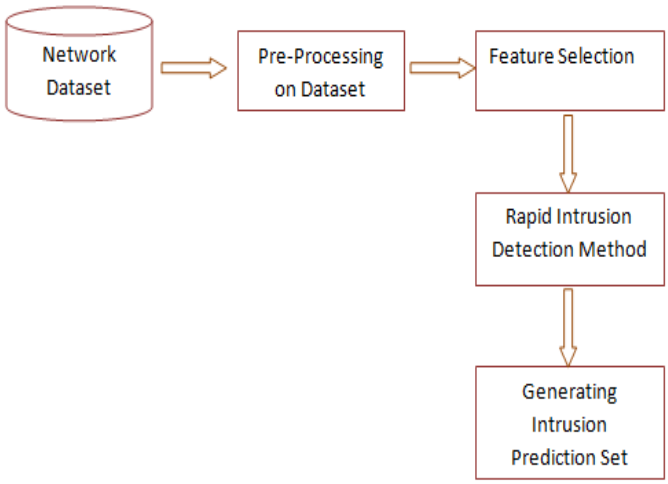

Fig-5 *General architecture of proposed model*

Feature selection is the significant and mostly used data preprocessing method to find the appropriate selection of nodes[13]. Either the intrusion prediction accuracy is improved or the size of the structure is reduced by processing the feature selection techniques which detects the feature subset [14]. Thus, this system improves the accuracy and size by using selected features without minimizing the accuracy of the classifiers.
The proposed method composed of three phases:

1. Preprocessing of Data
2. Feature Selection
3. RIDM

Generally, the limitation of KDD CUP99 dataset is the availability of repeated records which produces improper feature selection results. So, it is required to apply an appropriate preprocessing technique on the dataset [15][16]. During the communication of data in the network, a type of attacker is predicted in the system by computing the rules with the data. If any attack is identified between the nodes, then the connection is blocked to reduce the unwanted packet transmission [17]. If any node is acted as trusted one, them it is process and the log history is analyzed. If any misbehavior is achieved, then the connection base information is included into dataset [18]. After the dataset is processed, feature selection technique is applied. In the proposed work the process of feature selection is performed in the following stages. Consider the top N features utilizing attribute based selection techniques independently utilizing highlight ranker and then select the regular features that are recovered by all the characteristic feature methods and then choose the features with higher ranks and remove the features with ranks less than threshold value. Let the quantity of features be X. 3. Rehash above process for various estimations of N to get various arrangements of X features and then select the arrangement of X features which results great discovery rate and low flase positive rate as the applicant set of features. The proposed RIDM algorithm for detecting of intrusions is represented as

Algorithm RIDM

Input: Data set- KDD Cup 99
Output: Prediction set of Intrusions.

Step I: Initially load the dataset records.
Step 2: Perform Pre-Processing of data for removal of noise and unwanted data.
Step3: Perform clustering of data

For performing clustering of data a Intelligent Expectation Maximization Algorithm (IEMA) is introduced.

IEMA Algorithm

Input: Data sets
1. Load the records of the dataset DS.
2. Calculate the mean value of every record R(c1) -- -- R(cn), r <=th <= N, where r is initial record, th is threshold and N is the max count.
3. Mean= R'(th)(N − r) * DS(r1..rn)
4. For each cluster of parameter type
   For each data record do
   Calculate Intrusion posteriori distributions
   $$IPd = \frac{\exp\{Rth,N,r\}}{v\exp\{cnth\}} * Mean$$
   until R'(i)= = Empty Set
   end for
   end for
5. To Parameter P_x generate cluster set CS_i
   do
   $CS_x(P) = P_x$;
   ` done
6. Update cluster CS_x.
   $M_x(CS) = CS_x$;

Output: Cluster set of Network Data Records
The above calculation takes the first dataset as info and after that the each record closeness set is related to the given parameters and on the off chance that equivalent class is recognized, at that point they are assembled as a group. The mean estimation of each record is distinguished dependent on the limit esteem that r <=th <= N, where r is introductory record, th is edge and N is the maximum check. The intrusion posteriori dispersions are determined for each parameter which distinguishes the parameter to shape a group.
Step 4: Perform Feature selection on datasets.
The parameter weight (pw) of all features pw (f1) to pw (fn) is calculated as:

$$weight(pw(f1)\leftrightarrow pw(fn)) = \sum^{N}_{x=1}\{pw(\{fi\exists CS(x)\in DS\})$$

Step 5: Continuously monitor the data packets of the network.
Step 6: Record the network activities for rapid identification of intrusions.
Output: Display Intrusion Prediction set.

The proposed method identifies the rapidly identifies the intrusions in the network by continuously monitoring network activities and to avoid intrusions in the network.

4. METRICS

Confusion matrix is a matrix that represents result of classification. It represents true and false classification results.

Different Metrics Illustrated in the proposed work are

False Positive Rate (FPR) : It is calculated as

$$FPR = \frac{\text{Number of Normal Instances Detected as Attacks}}{\text{Total Number of Normal Instances}} = \frac{FP}{FP + TN}$$

The Detection Rate is the ratio between the number of correct classification of intrusion and the total number of intrusion available in the network. The mathematical formulation for calculation of detection rate (DR) is derived in the following equation:

$$DR = \frac{\text{Exact Attacks classified from available Attacks}}{\text{Total Attacks in Dataset}} * 100$$

The other definition of detection rate is the ratio of the accurate prediction of the intrusion attack events to the total attacks [19]. It is calculated using the following equation.

$$Detection\ rate = \left[\frac{TA - FN}{TA}\right] * 100$$

Where, TA is the Total Attack, FN is the False Negative. The Total Attack (TA) is the count of the number of attack events in the intrusion detection systems[20].

5. RESULTS

The analysis is performed by using proposed RIDM algorithm and the proposed method is implemented in ANACONDA SPYDER using Python. Using proposed clustering algorithm, the performance and accuracy of the method is improved from 78% to 88%. This indicates that Rapid Intrusion Detection Method is better than traditional methods.

In the proposed work datasets are used from http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. The Dataset contains 41 features representing parameters like source id, destination id, port number, log file id, attack type, is_hot_login, srv_serror_rate , dst_host_same_srv_rate etc. From the considered features only 16 features are extracted which plays a key role in detecting outliers.

The feature selection stage of the proposed method is higher than the existing techniques. The Figure 6 illustrates the True Positive Rate.

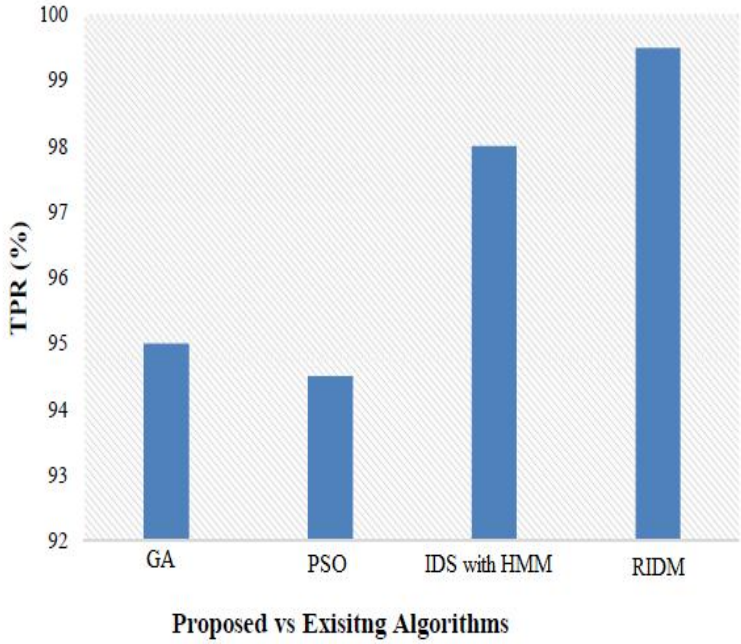

***Fig** 6: Comparison of True Positive Rate Estimation by various algorithms*

The classification performance is improved in RIDS by obtaining True Positive Rate (TPR), Higher Detection Percentage (DP) values and lower Error Percentage (EP) values.

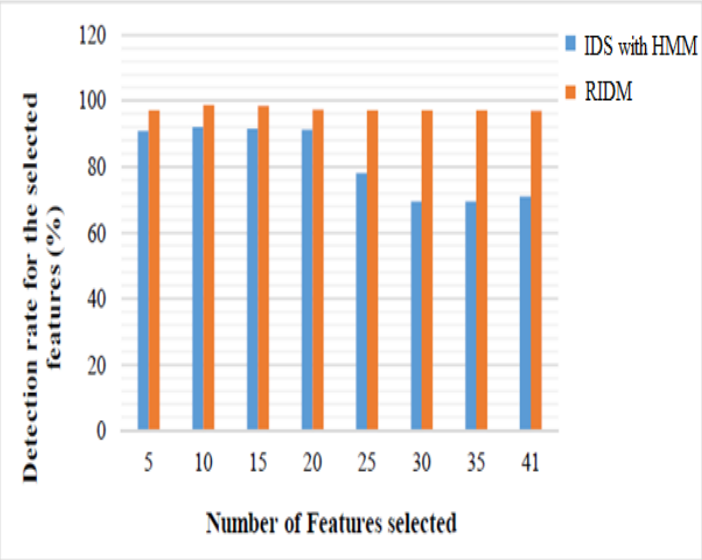

***Fig 7**: Detection rate (%) for the Selected Features*

Figure 7 shows the detection percentage for the selected features using RIDM algorithm. The DP is high, when the number of features is low. The DP is 90 % for 5 features and for 41 features, the DP achieved is 72%. Figure 8 shows the percentage of error for the selected features using the proposed algorithm.

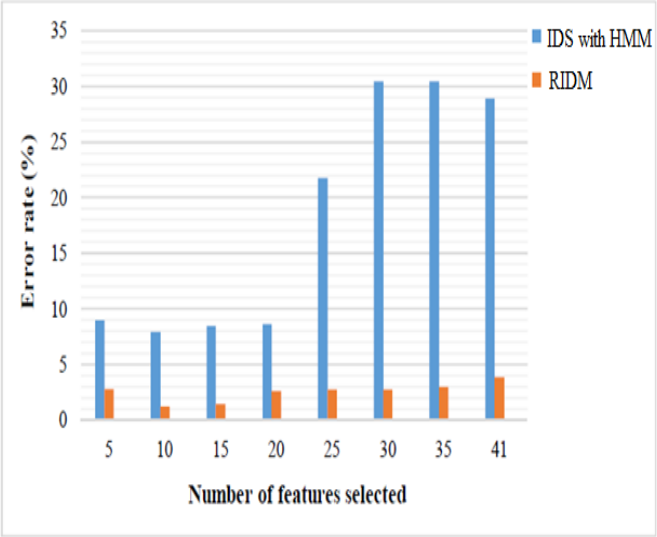

***Fig 8**: Error rate for the Selected features*

The proposed RIDM method has less False Positive Rate which indicates that the accuracy in detecting intrusions is more than the traditional methods. The False Positive Rate of the proposed and traditional methods are represented in Table-1.

METHOD	DETECTION RATE	FALSE POSITIVE RATE(FPR)
Rapid Intrusion Detection Method (RIDM)	87%	0.3
Random Forest Based Detection	65%	1.9
Density Based Intrusion Detection	68%	1.63
Genetic Algorithm for Intrusion Detection	71%	1.2
Hidden Markov Model for Intrusion Detection	73%	1.4

***Table-1**: False Positive Rate Levels*

The accuracy rate of the proposed method is depicted in the figure. The proposed method accuracy rate is high than the existing method.

6. CONCLUSION

In this proposed work a rapid intrusion detection method is introduced for identification of intrusions in the network. Traditional methods are not as efficient as our proposed model. In this proposed work, pre processing is initially done on the dataset for eliminating noise and then feature selection method is done for extracting only relevant features. The proposed model removes the redundant and unnecessary features and improves the performance and accuracy of rapid intrusion detection model and also reduces the false alarm rate by continuously monitoring the network activities. The time and resources required to build intrusion detection model is also less when compared to the traditional model. The analysis is performed on the KDD CUP99 datasets. In the future work, we plan to optimize the current model to get better performance and accuracy by improving the selected features. In future machine learning algorithms are used for detecting intrusions in the network for better results.

REFERENCES

[1]. O. Y. Al-Jarrah, O. Alhussein, P. D. Yoo, S. Muhaidat, K. Taha, and K. Kim, “Data randomization and cluster-based partitioning for botnet intrusion detection,” IEEE transactions on cybernetics, vol. 46, no. 8, pp. 1796–1806, 2016.

[2]. K. Wang, M. Du, Y. Sun, A. Vinel, and Y. Zhang, “Attack detection and distributed forensics in machine-to-machine networks,” IEEE Network, vol. 30, no. 6, pp. 49–55, 2016.

[3]. K. Wang, M. Du, D. Yang, C. Zhu, J. Shen, and Y. Zhang, “Gametheory- based active defense for intrusion detection in cyber-physical embedded systems,” ACM Transactions on Embedded Computing Systems (TECS), vol. 16, no. 1, p. 18, 2016.

[4]. K. Wang, M. Du, S. Maharjan, and Y. Sun, “Strategic honeypot game model for distributed denial of service attacks in the smart grid,” IEEE Transactions on Smart Grid, vol. 8, no. 5, pp. 2474–2482, 2017.

[5]. O. Joldzic, Z. Djuric, and P. Vuletic, “A transparent and scalable anomaly-based dos detection method,” Computer Networks, vol. 104, pp. 27–42, 2016.

[6]. D. Papamartzivanos, F. G. M´armol, and G. Kambourakis, “Dendron: Genetic trees driven rule induction for network intrusion detection systems,” Future Generation Computer Systems, vol. 79, pp. 558–574, 2018.

[7]. J. Kim, J. Kim, H. L. T. Thu, and H. Kim, “Long short term memory recurrent neural network classifier for intrusion detection,” in 2016 International Conference on Platform Technology and Service (PlatCon). IEEE, 2016, pp. 1–5.

[8]. M. Du, K. Wang, Y. Chen, X. Wang, and Y. Sun, “Big data privacy preserving in multi-access edge computing for heterogeneous internet of things,” IEEE Communications Magazine, vol. 56, no. 8, pp. 62–67, 2018.

[9]. M. Du, K. Wang, Z. Xia, and Y. Zhang, “Differential privacy preserving of training model in wireless big data with edge computing,” IEEE Transactions on Big Data, 2018.

[10]. P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, “A detailed investigation and analysis of using machine learning techniques for intrusion detection,” IEEE Communications Surveys & Tutorials, 2018.

[11]. S. Aljawarneh, M. Aldwairi, and M. B. Yassein, “Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model,” Journal of Computational Science, vol. 25, pp. 152–160, 2018.

[12]. J. Mi, K. Wang, P. Li, S. Guo, and Y. Sun, “Software-defined green 5g system for big data,” IEEE Communications Magazine, vol. 56, no. 11, pp. 116–123, 2018.

[13]. S. Maza and M. Touahria, “Feature selection algorithms in intrusion detection system: A survey.” KSII Transactions on Internet & Information Systems, vol. 12, no. 10, 2018.

[14]. V. Hajisalem and S. Babaie, “A hybrid intrusion detection system based on abc-afs algorithm for misuse and anomaly detection,” Computer Networks, vol. 136, pp. 37–50, 2018.

[15]. H. Hota and A. K. Shrivas, “Decision tree techniques applied on nsl-kdd data and its comparison with various feature selection techniques,” in Advanced Computing, Networking and Informatics-Volume 1. Springer, 2014, pp. 205–211.

[16]. A. J. Malik, W. Shahzad, and F. A. Khan, “Network intrusion detection using hybrid binary pso and random forests algorithm,” Security and Communication Networks, vol. 8, no. 16, pp. 2646–2660, 2015.
[17]. Lakshman Narayana Vejendla and A Peda Gopi, (2019),” Avoiding Interoperability and Delay in Healthcare Monitoring System Using Block Chain Technology”, Revue d'Intelligence Artificielle , Vol. 33, No. 1, 2019,pp.45-48.
[18]. A Peda Gopi and Lakshman Narayana Vejendla, (2019),” Certified Node Frequency in Social Network Using Parallel Diffusion Methods”, Ingénierie des Systèmes d' Information, Vol. 24, No. 1, 2019,pp.113-117. [FREE SCOPUS INDEXED JOURNAL]. DOI: 10.18280/isi.240117
[19]. M. Abdullah, A. Balamash, A. Alshannaq, and S. Almabdy, “Enhanced intrusion detection system using feature selection method and ensemble learning algorithms,” International Journal of Computer Science and Information Security (IJCSIS), vol. 16, no. 2, 2018.
[20]. V. Bol´on-Canedo, N. S´anchez-Maro˜no, and A. Alonso-Betanzos, “Feature selection for high-dimensional data,” Progress in Artificial Intelligence, vol. 5, no. 2, pp. 65–75, 2016.