Vol.10,03,JulSept2021

IFPRBS

INTERNATIONAL JOURNAL OF RESEARCH PEDAGOGY AND TECHNOLOGY IN EDUCATION AND MOVEMENT SCIENCES (IJEMS) ISSN: 2319-3050

Intelligent Intrusion Detection System Combining Misuse Detection and Anomaly Detection Using Random Forest Algorithm

[1]D V JEYANTHI, [2]DR. B.INDRANI

[1] Assistant Professor, Department Of Computer Science, Sourashtra College Madurai

[2]Assistant Professor, Department Of Computer Science, Dde Madurai Kamaraj University

ABSTRACT

Intelligent Intrusion Detection is a system that has the ability to provide security to data. It compares the existing attacks with the new attacks. This helps to secure data from the intruders and also monitors the intruder's system. The IDS generates alarm when a malicious attack is detected [4]. In this research, an Intelligent Intrusion Detection System is proposed which detects attacks with high accuracy and in fast response time. The proposed system has two phases, 1) Clustering Technique which process large amounts of data efficiently to identify outliers 2) A Classification Algorithm that classifies attacks and also avoids high false positives.

Key Words: *Intrusion Detection, Clustering, Classification, Outliers*

INTRODUCTION

Internet plays a major role in the software, hardware based works, medical field, political field, financial, business and many. In short, internet based works are ruling the world. It is the best idea to turn into digital but in other hand we have too known about our data security [4] also. When the world turns digital, attackers also updates them and becomes too smart in stealing. They can steal our data by phishing mail, fake ids, unknown hosts, software updates anything. So it's important to secure ourselves from intruders. Based on the usage and research, it is found that wireless sensor networks play a vital role. Because its necessity is more needed in the field of critical applications, cyber threats and data security [4] etc., it is becoming a challenging task day by day because of the birth of new attacks and insufficient approaches to that attacks. To know whether the attack has occurred or attempted by analyzing huge volume of data gathered from the network, host or file systems to find malicious activity.

WSN security is mainly categorized into four levels. (1) Escaping intrusion (2) Identifying intrusion (3) Providing response (4) Cryptography and Firewalls. The fourth level is considered to be more important. Many tools and software's are developed for providing security and controlling attacks. One of the tools is Intrusion Detection System [5, 6]. This proposed paper is categorized as follows. Section II details the survey report of related methods;

Section III explains the proposed methodology and flow of proposed work. Section IV represents the results obtained using the proposed methodology. And finally concluded the paper with discussion.

LITERATURE SURVEY

Internet based tasks are ruling the world. Our Everyday life depends on Internet based services. Meantime, Attackers are getting smart in stealing information from the systems. It is becoming a challenging task for the researchers to identity the new born attack and to find sufficient approaches to deal with those attacks. Therefore, Researchers are working to find intelligent intrusion detection system [5, 6] that detects new types of attacks and protect our systems with minimum computational cost.

To provide an efficient and secure intrusion detection system, some of the research is taken on the existing approaches.

Authors, Wen-Hui Lin1, Hsiao-Chung Lin, Ping Wang, Bao-Hua Wu, Jeng-Ying Tsai presented a paper, "Using Convolutional Neural Networks to Network Intrusion Detection for Cyber Threats" in the year 2018. The main aim of this paper is to provide security to cyber threats. The cyber threats are mainly happened using phishing emails, malicious software or with legal network protocols. The whole process is in hand of attackers using robust controllers. To improve the accuracy of controlling cyber threats, this crew developed a classifier learning model by training Convolutional Neural Networks [1, 2] (CNNs). The paper methodology CNN is explained by two examples. (1) Feature Extraction Phase – In this a set of reduced feature is selected using information speed to increase the speed and then converted the features into image matrix. (2) Model Learning Phase – In this phase, they classified by large amount of samples and 39 sub-categories of attacks. They used this approach to enhance the precision model.

"A Multi-Level Intrusion Detection System for Wireless Sensor Networks based on Immune Theory" paper is presented by the authors, Vishwa T. Alaparth and Salvatore D Morgera in the year 2018. This paper is based on immune system [3] (i.e biological models). This crew mainly talks about the development of WSN similar to immune cells in human body. For this they have considered size, deployment, power and density. They proposed HIS based IDS based on Negative Selection and Danger Theory. First involves the detection of self and non-self entities. Second is generated from dendritic cells as danger signals while intrusion is identified. In addition, they have used Clonal Selection. They designed this method to WSN and also to ADHOC Networks by doing some slight changes. Especially this model identifies the energy depleting attacks with high accuracy in limited time. This model is not attack specific and it classifies data by the signal received [10, 3].

In the year 2018, the researchers, "Mohammed Hasan Ali, Bahaa Abbas Dawood Al Mohammed, Alyani Ismail and Mohamad Fadli Zolkipli" presented a paper, "A New Intrusion Detection System based on Fast Learning Network and Particle Swarm Optimization". This paper talks about the security in the fields of computer network, computer vision, robotics, control and communication. They developed a learning model for Fast Learning Network (FLN) based on Particle Swarm Optimization (PSO). They used the dataset of KDD99. This paper provides accuracy based on leaning models. FLN is a parallel connection of SFLN and which is a Double Parallel Forward Neural Network. PSO-Based optimized FLN is trained based on selecting weights using particle swarm optimization [8, 9, and 10].

METHDOLOGY

The research work aims is to design an Intelligent Intrusion Detection System that would be accurate, low in false alarms, not easily cheated by small variations in patterns, adaptive and with minimal computational cost.

IDS include monitoring of unauthorized access and abuse of system by both penetrators. IDS generate alarm when a malicious attack is detected. IDS parameters are numerous and it presents uncertain and imprecise causal relationships which affects attack types. Many IDS focuses on signature detection and implemented using techniques like Data Mining, Fuzzy Logic, Neural Network, Bayesian Network etc [8, 9]. While focusing on signature detection, it recognizes only on known attacks and cannot detect novel attacks. The disadvantage of this is the intrusion signatures changes over the time and the system must be retrained.

Intrusion Detection [5, 6] helps very well in this purpose. Intrusion Detection can be defined as the process of identifying suspicious behavior that targets a network and its devices. Suspicious behavior is defined as a system which tries to attack or access other's system without authorization (i.e.) crackers and the privilege excess. Due to more unknown attacks, following Intrusion Detection System is a difficult one because of network connectivity and the need to update regularly.

So, the research work is on IDS with two phases. First phase is clustering and the second is the combination of misuse and anomaly detection using machine learning algorithm. The collected datasets for this work is from KDD Cup 1999 dataset [7] which is considered as the latest dataset. In below sections, we discuss in detail.

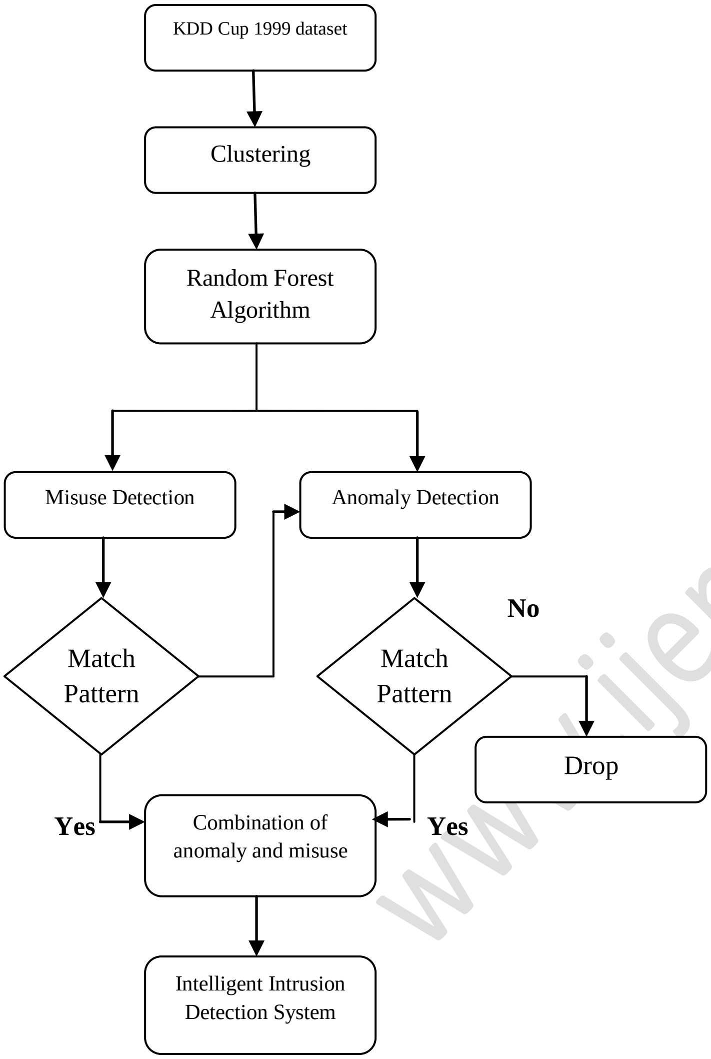
Dataset – Base of Proposed Work

The KDD Cup 1999 [7] dataset used for benchmarking intrusion detection issues is employed in our experiment. The dataset was a group of simulated raw protocol dump information over a period of nine weeks on a local area Network. The training data was processed to regarding 5 million connections records from seven weeks of network traffic [8,9] and time period of testing data yielded around two million connection records. The training information is formed from twenty two totally different attacks out of the thirty-nine present in the test data. The known attack types are those present within the trained dataset whereas the novel attacks are the additional attacks in the test datasets not available in the training data sets. The possible attacks are shortly described below.

- **Denial Service of Attack (DoS) -** DOS is an attack which essentially involves the resources are too busy to handle other requests or the attacker making use of specific resources to an extent that denied access for legitimate users.
- **User to Root Attack(U2R) -** It is a form of security exploitation, whereby the attacker would gain access to a normal user account, through conventional means, and thereafter proceed to attempt root access to the system through the exploitation of a vulnerability.
- **Remote to Local Attack (R2L) -** This is when an attacker attempts access to a system over a network. The attacker can only transmit data packets over the network, the attacker attempts to gain access to the machine, by exploiting some vulnerability.
- **Probing Attack (Prob) -** It is when an attacker attempts to acquire information from a network, for evading the systems, security protocols.

Since 1999, a large number of researchers assessed their IDS models using KDD Cup 99. This shows how KDD Cup 99 has been a working benchmark data set for over 15 years, and is still easily accessible and available today. The objective of the KDD 99 IDS competition is to make a standard data set for the surveying and evaluation of research in intrusion detection. Researchers found some difficulties or hurdles in training with KDD99 and analyzed that the KDD 99 data set for selecting a relevant feature. They proposed that some features or attributes were not related to any attack, and taken 10% of the whole data set to perform their analysis.

Flowchart of Proposed Methodology



Derived Features

One of the researcher defined higher¬level options that facilitate in characteristic traditional connections from attacks. There are several categories of derived features.

- The “same host” options examine solely the connections within the past 2 seconds that have an equivalent destination host because the current affiliation, and calculate statistics related to protocol behavior, service, etc.

- The similar ``same service" options examine solely the connections within the past 2 seconds that have the same service as the current connection.

"Same host" and "same service" options square measure along known as time¬based traffic options of the affiliation records.

Some inquisitory attacks scan the hosts (or ports) employing a a lot of larger amount than 2 seconds, for instance once per minute. Therefore, affiliation records were additionally sorted by destination host, and features were made employing a window of a hundred connections to an equivalent host rather than a time window. This yields a group of so¬called host¬based traffic options.

Unlike most of the DOS and inquisitory attacks, there appear to be no sequential patterns that are frequent in records of R2L and U2R attacks. This is as a result of the DOS and inquisitory attacks involve several connections to some host(s) in a very short period of time, but the R2L and U2R attacks are embedded in the data portions of packets, and usually involve solely one affiliation.

Useful algorithms for mining the unstructured information parts of packets mechanically are associate degree open research question. Researchers used domain information to feature options that search for suspicious behavior in the data portions, such as the number of failed login attempts. These features are called ``content" features.

First Phase – Clustering Algorithm

Clustering is a technique to identify “outliers” as anomalous. Clustering can be done by using k-means. Using the K-means algorithm it is easy to extract relevant data and free from false rates. Our proposed technique is same as k-Nearest Neighbor (k-NN) Algorithm. K-NN locates only the local outliers but we are using to find both local and global. It process large amounts of data efficiently. To identify unknown attacks, Anomaly detection needs extensive training data to identify new attacks. In Our proposed system, datasets preferred for this approach is KDD Cup 1999 dataset. The dataset is classified based on the classifier types. It helps to overcome the issue of false positives and false negatives.

Steps in Clustering

Step 1: initialize the set of clusters, CL, to ∅

Step 2: Creating clusters

Step 2.1: for s ∈ S

```
    Step 2.2: for $q \in Q$
    Step 2.3: if distance $(s, q) \leq T$, here s is assigned to q
    Step 2.4: if s is not assigned
    Step 2.5: create cluster cl' with s as the centroid and add cl' to CL
Step 3: Assigning data points to additional clusters
    Step 3.1: for $s \in S$
    Step 3.2: for $q \in Q$
    Step 3.3: if distance $(s,q) \leq T$, here s is not assigned to q
    Step 3.4: assign s to q
```

where, s is used to define data point, T is used to represent Width, Clusters as cl

S is preferred to represent dataset and CL to define set of clusters.

Here Input is S and Output is CL.

Before Clustering

The below figure 3.1 represents the preferred KDD Cup 1999 dataset. The relevant data's to anomaly, misuse and also irrelevant dataset. In the proposed methodology, the preferred input dataset is represented as S.

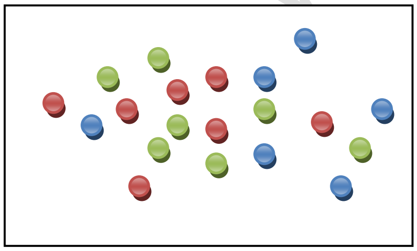
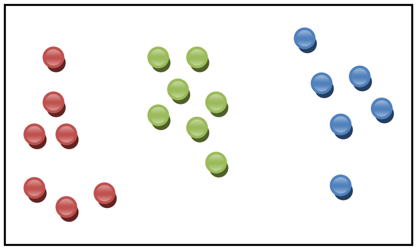


Figure 3.1 – KDD Cup 1999 Dataset

After Clustering

The random data is generated using k-NN of clustering algorithm. For example, if there are 20 data generated, it is categorized into multiple groups of same cluster as each of 10. The representation is presented in the figure 3.2.



– Clustered KDD Cup 1999 Dataset

In the proposed methodology, the obtained output dataset is represented as CL. The KDD Cup 1999 dataset is clustered to remove the irrelevant data's from the dataset and to avoid the time consumption at the last stage. It makes a sense to identify the data related to the proposed methodology.

Second Phase – Random Forest Algorithm

A random forest F is basically an ensemble of T independent decisions trees F = {F1, · · · , Ft , · · · , F_T }. As demonstrated by Breiman, replacing a single tree by an ensemble of decorrelated trees provides very good generalization. During the learning phase, randomness can be injected to achieve independence between trees constructed from the same training set.

To build decorrelated or independent trees based on a unique training set, several randomization approaches have been proposed. Breiman introduced the concept of bagging which comes from the combination of the term's "bootstrap" and "aggregating". Given a training set S = $\{X^{(n)}, Y^{(n)}\}_{n=1}^{N}$, a bootstrap is basically a subset St of the full training set, in which element has been randomly sampled using a uniform distribution, and this, with or without replacement. Each tree F_t of the ensemble is then trained using a different bootstrap S_t. Finally, predictions from all individual trees are aggregated together using averaging.

Randomization can be also injected in the node optimization. Indeed, as this phase relies on a greedy strategy, a set of splitting functions candidates is generated, and the best is then chosen according to a predefined objective function. Obviously, randomness can be injected in the generation of function candidates. Let us take the example of linear projections followed by a thresholding operation. First, the projection vector v can be randomly drawn using any kind of distributions. This encourages the trees to select different type of features and to weight them differently. Furthermore, the choice of threshold τ can be also randomized instead of optimizing it or taking the mean/median of the projected values.

The impact of injecting randomness in the tree training has several advantages: first, increasing the degree of randomness decreases the correlation between the different trees and provide thereby greater generalization, second it enables implicit feature selection if v is constrained to be sparse and third, it permits to gain independence from the training set, i.e. to gain robustness to noisy data.

Here RFA is used in the classification of 3classes into 6classes. By considering one as test data and 2 as training data, the evaluation is taken and vice-versa. By doing this we believe that, can evade intrusion from some new attacks. The performance and time consumption is also considered. This phase classifies the detection as misuse and anomaly.

In the misuse detection, it checks for the matching patterns, if matches it saves the data for next phase else sends to anomaly detection to check whether it belongs to that class or not. Same as misuse detection, anomaly detection checks for the pattern, if matches save else drops the data.

The below figure 3.3 represents the RFA structure.

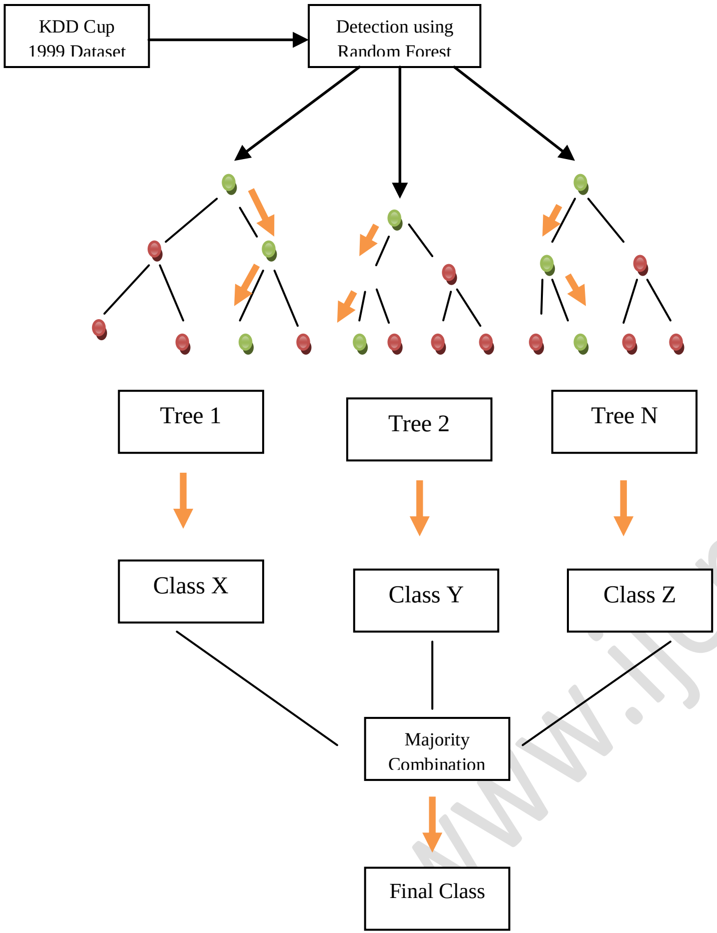


Figure 3.3 – Structure of Random Forest Algorithm

Instead of a choosing a classifier which doesn't provide good accuracy rate and making risk to solve the issue is a bad idea. By comparing the analysis result of the classifiers in the previous chapter, Random forest Classifier as the right choice to the development of this research work. Some reasons behind choosing RFC is as follows.

- **Statistical Reason**

The generalization performance of the classifier can be different and it is secure to prefer which gives the better accuracy rate. And also, it will decrease the risk of choosing the inadequate classifier and facing issues at the final stage.

- **Computational reason**

The best classifier can initiate a searching strategy by running the local search from the various initial points. This solves the local search of a inadequate classifier which faces issues in the local optimal search.

- **Representational Reason**

The classifier space that is considered for pattern recognition does not always contain the optimal classifier. For instance, a set of linear classifiers is chosen for a dataset that can best recognize by nonlinear classifier. This way an optimal classifier can never be obtained. On the contrary, an ensemble of linear classifiers can approximate a decision boundary with any predefined accuracy.

The matching patterns from these two detections are forwarded to the next phase. RFA achieves classification with perfection and in less period of time.

Third Phase - Combination of Misuse and Anomaly Detection

The two general approaches of IDS are misuse detection and anomaly detection. The misuse detection uses known patterns to detect attacks but it fails to detect unknown patterns which are not in the set. In other hand, anomaly detection identifies unknown patterns from where the deviation is but with high false positive rates. By considering these issues, we are combining misuse detection with anomaly detection. It classifies the attacks into normal, known and unknown attacks. It helps to find solution for more new attacks and also avoids high false positives.

To achieve this, we are classifying data into DR (Detection Rate), PFR (Positive False Rate – Rate of Normal Data as intrusion), NFR (Negative False Rate – Rate of Intrusion Data as Normal) and Accuracy (Rate of Correct Data). We can detect attacks with high accuracy in this way comparing to misuse IDS and anomaly IDS.

The data is processed based on the below calculation to achieve the best prediction than the existing methods and also in less time

(1) Detection Rate – The relationship or ratio of true positive and the **e** total nonself samples identified by detector set. TP and FN are represented as the true positive and false negative.

DR=TPTP+FN – (1)

(2) Positive False Rate – It is defined as the result of the detecting the positive as negative class or data.

(3) Negative False Rate - It is defined as the result of the detecting the negative as positive class or data.

(4) Rate of Correct Data or Accuracy – It is defined as the one metric in classification calculation. In other words, it can be said as the number of positives outcome achieved using our classifier. The calculation is done as follows.

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+TP+FN} - (2)$$

The result obtained using the proposed methodology is explained in the next section.

TABLE I: ATTRIBUTE VALUE TYPE

Type	**Features**
Nominal	Protocol_type(2), Service(3), Flag(4)
Binary	Land(7), logged_in(12), root_shell(14), su_attempted(15), is_host_login(21),, is_guest_login(22)
umeric	Duration(1), src_bytes(5), dst_bytes(6), wrong_fragment(8), urgent(9), hot(10), num_failed_logins(11), num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23) srv_count(24), serror_rate(25), srv_serror_rate(26), rerror_rate(27), srv_rerror_rate(28), same_srv_rate(29) diff_srv_rate(30), srv_diff_host_rate(31), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38), dst_host_srv_serror_rate(39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41)

WEKA:

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code.

Advantages of Weka include:

I. Free availability under the GNU General Public License

II. Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform

III. A comprehensive collection of data preprocessing and modeling techniques

IV. Ease of use due to its graphical user interfaces

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

RESULTS AND DISCUSSION

The designed clustering algorithm process large amounts of data efficiently, to identify unknown attacks. Extensive training dataset is provided to the proposed method for anomaly detection to identify the new attacks. Because traditional intrusion detection systems are generally limited to single host, or network architecture, which lacks the ability to detect large-scale networks and heterogeneous systems, we need to combine the sequential pattern mining with other intrusion detection techniques.

In this paper, the Machine learning Algorithm is implemented which helps in the detection of anomaly and misuse but also the relevant successive patterns on KDD Cup 1999 dataset to detect intelligent attack sequences on network traffic data. Time consumption of classification algorithm is reduced with the usage of RFA. The proposed system designed to detect new born attacks intelligently and alert the administrator with quick response time with minimal computational cost.

A. Experiment Setup Many standard data mining process such as data cleaning and pre-processing, clustering, classification, regression, visualization and feature selection are already implemented in WEKA.

The automated data mining tool WEKA is used to perform the classification experiments on the 20% NSL-KDD dataset. The data set consists of various classes of attacks namely DoS, R2L, U2R and Probe.

B. Pre-processing, Feature Selection and Classification The data set to be classified is initially pre-processed and normalized to a range 0 -1. This is done as a requirement because certain classifiers produce a better accuracy rate on normalized data set.

Correlation based Feature Selection method is used in this work to reduce the dimensionality of the features available in the data set from 41 to 6. Classification is done in this work by using KNN, SVM and Naïve Bayes algorithms

C. Result Analysis The experiments are carried out in WEKA and effectiveness of the classification algorithms in classifying the NSL-KDD data set is analyzed.

The accuracy rate in detecting normal and attack class of network connection is shown in the table II. This shows that when RFA is used for dimensionality reduction, RFA classifies the data set with a better accuracy rate.

TABLE II: ACCURACY IN DETECTION OF NORMAL AND ATTACK NETWORK FLOWS BY USING THE RFA, J48, AND SVM CLASSIFIERS

	Test Accuracy with 6 features		
Class Name	**Classification Algorithm RFA**	**Classification Algorithm J48**	**Classification Algorithm SVM**
Normal	**99.8**	**99.8**	**98.8**
DoS	**99.7**	**99.1**	**98.7**
Probe	**99.4**	**98.9**	**91.4**
U2R	**99.1**	**98.7**	**94.6**
R2L	**98.8**	**97.8**	**92.5**

I. Conclusion

The reason to propose this paper is to study about the detection system of intrusion and the efficient method to overcome the flaw in the existing methodologies and to secure from the intruders or malicious software's or hackers. The implementation of this method does not require any special software or hardware and computational time is also less. It can also be implemented without any special nodes or classification process. Due to this reason, the cost of implementation is low. The proposed work results with the accuracy rate are high and some special ways to stop from hacking or intrusion. In future, it can be implemented with some other methods and the comparative study can be presented.

References

[1] Wikipedia, Convolutional neural networks, available at https://en.wikipedia.org/wiki/ convolutional neural networks.

[2] "Using Convolutional Neural Networks to Network Intrusion Detection for Cyber Threats**"** Wen-Hui Lin1, Hsiao-Chung Lin, Ping Wang, Bao-Hua Wu, Jeng-Ying Tsai**,** 2018.

[3] S. A. Hofmeyr and A. S. Forrest, "Architecture for an Artificial Immune System," *Evolutionary Computation,* vol. 8, no. 4, pp. 443-473, December 2000.

[4] S. Forrest, A. S. Perelson, L. Allen and R. Cherukuri, "Self-Nonself Discrimination in a Computer," in *proceedings of the 1994 IEEE Symposium on Security and Privacy*.

[5] M. S. Hoque, M. A. Mukit, M. A. N. Bikas, and M. S. Hoque, ``An implementation of intrusion detection system using genetic algorithm," *Int. J. Netw. Secur. Appl.*, vol. 4, no. 2, pp. 109_120, 2012.

[6] J. Frank, ``Artificial intelligence and intrusion detection: Current and future directions," in *Proc. 17th Nat. Comput. Secur. Conf.*, Baltimore, MD, USA, Oct. 1994.

[7] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, Ocotber 2007.

[8] C. Chen, S. Mabu, K. Shimada and K. Hirasawa, "Network Intrusion Detection using Class Association Rule Mining Based on Genetic Network Programming", *IEEJ Transactions on Electrical and Electronic Engineering*(Conditionally Accepted)

[9] C. Chen, S. Mabu, C. Yue, K. Shimada and K. Hirasawa, "Network Intrusion Detection using Fuzzy Class Association Rule Mining Based on Genetic Network Programming", *In Proc. of the IEEE Internatinal Conference on Systems, Man and Cybernetics*, 2009 (Submitted).

[10] "A New Intrusion Detection System based on Fast Learning Network and Particle Swarm Optimization"**,** Mohammed Hasan Ali, Bahaa Abbas Dawood Al Mohammed, Alyani Ismail and Mohamad Fadli Zolkipli**,** 2018.