International Journal of Engineering & Technology, *7 (3.34) (2018) 474-479*

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A novel Approach Using “Supervised and Unsupervised learning” to prevent the Adequacy of Intrusion Detection Systems

[*1]Pradeep Kumar Mallick, [2]Bibhu Prasad Mohanty, [3]Sudan Jha, [4]Kuhoo

*[*1]Department of Computer Science and Engineering, Vignana Bharathi Institute of Technology, Hyderabad.*
[2]Department of Computer Science & Applications, Utkal University Odisha
[3]School of Computer Engineering, Kalinga Institute of Industrial Technology Odisha
[4]Department of Mechanical Engineering, College of Engineering and Technology, Bhubaneswar, Odisha
** Corresponding author E-mail: [1]pradeepmallick84@gmail.com , [2]bpm.bibhu@gmail.com,[3]*
[4]jhasudan@hotmail.com,kuhoo.cet@gmail.com

Abstract

Countering digital dangers, particularly assault detection, is a testing region of research in the field of data affirmation. Intruders utilize polymorphic instruments to disguise the assault payload and dodge the detection methods. Many supervised and unsupervised learning comes closer from the field of machine learning and example acknowledgments have been utilized to expand the adequacy of intrusion detection systems (IDSs). Supervised learning approaches utilize just marked examples to prepare a classifier, however getting adequate named tests is lumbering, and requires the endeavors of area specialists. Notwithstanding, un-marked examples can without much of a stretch be acquired in some genuine issues. Contrasted with super-vised learning approaches, semi-supervised learning (SSL) addresses this issue by considering expansive number of unlabeled examples together with the marked examples to fabricate a superior classifier. In today’s age security is a big issue and every day when we are on the internet we are exposed to a huge number of threats where our personal information can be leaked. The information security and the Intrusion Detection System (IDS) play a critical role in the internet. IDS isan essential tool for detecting different kinds of attacks in a network and maintaining data integrity, confidentiality, and system availability against possible threats. In this paper, we are proposing a modified Elitist approach where the value of fitness is multiplied by the times a variable which is determined on the basis of the value of Kappa (K).

***Keywords**: Cyber Threats, Supervised & Unsupervised Learning, Machine Learning, Pattern Recognition, Intrusion Detection Systems (IDS).*

1. Introduction

In the vast network of interconnected devices, security is a major concern and certainly confidentiality of data is a big factor. History of various attacks on crucial systems leading to data breach shows users being dissatisfied and also the reputation of the company is affected.

Since intrusion can be external and internal, countermeasures regarding only to external intrusions is not adequate. Internal intrusion also cause damage to the system due to lack of care of the strength of passwords, careless granting of access to unauthorized users and/or faulty intrusion detection systems. Therefore, it is important to ensure a strong intrusion detection system to prevent misuse of data. Systems such as firewall can only do so much due to its limited capability. An IDS is a much powerful software having abilities to detect intrusions and act accordingly.

An intrusion detection system monitors and reports to the users about the suspicious activity and/or unauthorized access to the computer system. In addition to detection, the intrusion detection system can also track down the source of attack and take appropriate actions. Attacks on networks are numerous and their damage is insurmountable costing a lot of time and money. Before we move onto the types of IDS, a brief overview about the types of network attacks are given below: -

1. Remote to User (R2L): - Packet/s are sent over the internet from a remote computer to a local computer to which the attacker does not have access to. The motive of the attacker is to access confidential data on the local computer using the privileges it provides.
2. Denial of Service (DOS): - A DOS attack is one where an attacker tries to deny services to users by making the computer resources unavailable by making them serve non-legitimate requests.
3. Probing: - Here, a device or a machine is scanned in order to find vulnerabilities which can be later used to exploit the machine.
4. User to Root (U2R): - In this type of attack, an attacker starts as a normal user with login credentials and tries to gain administrative privileges by exploiting the system.
5. Smurfing: - A smurf attack is one where a large number of ICMP packets are broadcasted into a network using a spoofed source IP address.

The major classifications of the types of IDS are:

1. Active IDS
2. Passive IDS
3. Network IDS
4. Host IDS

5. Knowledge based IDS
6. Behavior based IDS

1. Active IDS: - An active IDS monitors attacks in real-time and takes action without any intervention from the user or the system administrator.
2. Passive IDS: - This type of IDS only informs the user of the threats and vulnerabilities the computer is facing and alerts the user or the administrator. The administrator can take appropriate actions in order to tackle the situation.
3. Network IDS: - A network IDS can be placed at a point in a network with a segment boundary, where the inbound and outbound traffic can be monitored and malicious activity can be monitored. A con of the network IDS is the fact that a central detecting device can be a bottleneck to the network thus increasing network traffic.
4. Host IDS: - A host IDS can be installed on a local computer system where monitoring is to be done locally and the entire network is not be monitored. Mainly workstations and servers implement a host IDS.
5. Knowledge based IDS: - These IDS use a database to match attack signatures. This is a very effective way of detecting attacks though the database must be constantly updated with new signature definitions.
6. Behavior based IDS: - A base pattern using statistical methods is used as a reference to actively identify intrusion attacks. Any deviation from this pattern causes an alarm to be triggered. Though complex in nature, it's dynamic nature of adapting to original attacks makes it very preferable.

In the early 2000's, IDS was the best approach to implement security in systems due to the incompetence of firewalls with regard to SQL injection attacks and deep packet inspection.

Gradually, intrusion prevention systems (IPS) were being adopted at a higher rate than before and subsequently a hybrid system was realized later. In the recent years, Next Generation Intrusion Detection Systems have been popular because of its ability to provide user control and application control. During 2011, IDS were being improved with immunity to phishing attacks. Also, checksum was being used as a matching protocol where the checksum of the packet would indicate the entry of a malware into the system.

Earlier, firewalls only used port, protocol or IP for intrusion detection. This was inadequate due to constant increase in immunity of attacks to the existing firewalls. With the advent of next generation firewalls and its adoption, organizations can now choose which features will be valuable to company because of the variety of the features provided by the next generation firewalls. As more and more original attacks are generated, the scope for improvement of the IDS is elevated inviting contributions for the betterment of the system.

2. Related Work

In the recent years, some work has been done on intrusion detection systems in an objective to improve existing systems. Graph based intrusion detection systems cluster the network into separate activity network providing monitoring control over the traffic. This proved to be a very effective approach for large networks. This work was done in the late 1990s.

Recent developments in intrusion detection systems take help of the modern data mining techniques to identify the hidden pertinent data quickly and efficiently. Approaches using Genetic Algorithm (GA) have also been carried out using a set of classification rules and a corresponding fitness function to detect intruders in a network. The genetic algorithm approach adapts dynamically to new situations but has a higher rate of false positives. These false positives can be pared down with the help of parameter tuning. Intrusion detection systems implemented in wireless sensors network employ a cluster based architecture to reduce energy consumption in combination with machine learning algorithms like Support Vector Machine to implement an anomaly based detection system. This kind of system has a lower false positive and a high accuracy.

In ICCC 2015, a new approach known as Outlier Detection approach was used to detect intrusion in a network. Performance was improved using a distributed computing environment. This detection method performs better than conventional machine learning algorithms and also has a good efficiency in terms of execution time required.

Another 2015 study involved the use of clustering center and nearest neighbors approach where a one-dimensional vector is formed by summing two distances; one being the distance between a node and the center, and the pother being the distance between nodes. The one-dimensional vector was used with the K-nearest neighbor algorithm which showed a performance increase than other machine learning algorithms like Support Vector Machine and also decrease in computation time. An ensemble approach to intrusion detection was thought to be powerful than traditional machine learning algorithm working alone. The KDD99 data set was used. A comparison of results was made between six-different SVM classifiers, six-different k-NN classifiers, an ensemble classifier based on particle swarm optimization, ensemble classifier based on local unimodal sampling improvement of the PSO algorithm and an ensemble classifier based on WMA.

An idea to distribute the workload of intrusion detection to other nodes in a network was published in 2016 to alleviate IDS pressure on a node. This approach aims to reduce latency and resource utilization by distributing IDS functions among nodes such that the resource utilization is below a certain threshold. This will prevent overloading a single node with all functions and keep the network stable.

The Random Forest Model was evaluated in relation to other machine learning models on the basis of its performance on the NSL-KDD data set. The model was found to be efficient with low false positives and high detection rate. This study focused on detecting four types of attacks, DOS, probe, U2R and R2L.On comparing with the j48 classifier, the proposed random forest model had an accuracy of 99.67% compared to 99.26% for j48. The objective was to tackle the non-linear problem of intrusion detection using a model with the Random Forest classifier. With the greater accessibility of more computing power, neural networks could be trained easily on general computers. A study involving computational neural network regression models were used to improve accuracy of host intrusion detection system.

A Generalized Regression neural network (GRNN) model and a Multilayer Perceptron Neural Network (MPNN) model were evaluated on the basis of accuracy, recall and precision. The results showed that GRNN performed had better metrics than MPNN with respect to detection accuracy and recall but MPNN had better metrics in case Precision. Nevertheless, the article concluded that both the neural network models could be used for host intrusion detection systems.

3. Literature Survey

In this section, we are explaining the most recent approaches that have been used by researchers and the also the research gaps that have been covered by researchers in the context of Intrusion Detection System. There are basically two types of Intrusion Detection System (IDS) - Host IDs and Network IDs. In IDS, there is a sensor that detects whether there has been an attack or not. A host based IDS generally uses the logging of the system to, system logs and other logs to detect whether there was an intrusion or not. Host based system mainly relies on the transactions made in the database for detection. A network based IDS unlike a host based IDS does not analyses each and every host in the network whereas it collects information from the network itself.

Enamul et al. [1] proposed a model where they first divided the dataset into different subgroups. The dataset used in the standard KDD 99 dataset. Each subgroup was consisted all the components that is all types of training data and all types of attacks. Considering the randomness of the dataset an optimized algorithm had been developed which selected the samples. Least Squares Support Vector Machine was applied on the extracted subgroups. In the paper, they have shown the amount of data required to give an accuracy of 95% and 99%. The size is being determined by samples which are pre-defined and they are using a formula to calculate the size of training and testing samples.

Tarfaet al. [2] proposed a new approach recursive feature addition and a bigram technique. The bigram technique is being proposed so that the encoded string features can be represented in a way for the ease and accuracy of feature selection. They have also proposed a new metric where the we get a combined value of the detection rate, false alarm rate and accuracy for determining the best approach with more confidence.

Mingming et al. [3] have used fuzzy logic and clustering unsupervised learning algorithm for an efficient algorithm on the implementation in a cloud storage environment. The recall of the proposed algorithm outperforms the ANN and gets a recall of around 80%. On comparison with other models the proposed algorithm gives better performance.

Abhishek Verma & Virender Ranga [4] have done a statistical analysis CIDDS-001 dataset and have applied KNN classifier and clustering algorithm. There main research lies on the use of the new dataset and they claim that this can also perform offline Intrusion Detection. They have achieved a maximum accuracy of 99.6% with 2NN and a minimum accuracy of 99.3%.

Nabila Farnaaz & M.A.Jabbar [5] have proposed an approach using the Random forest classifier on the NSL-KDD dataset. They have proposed this work for the classification of Network Intrusion Detection System. They have claimed that the model has a low false rate alarm and a high detection rate. They have found that the proposed approach gives an accuracy of 99%.

Sunil Kumar Gautam & Hari Om [6] have proposed a Generalized Regression Neural Network and Multi-Layer Perceptron Neural Network for the prediction of Host Intrusion Detection System. They have taken the og files from a personal computer where there has been multiple number of attacks in the past and they have used standard metrics for measuring the accuracy. The GRNN gave an accuracy of 98.46% and MPN gave an accuracy of 97.68%.

4. Comparative Analysis

Sl. No.	Year of Publication	Authors	Advantage	Limitations
1	2013[7]	Chun-Jen Chung, Pankaj Khatkar, Tianyi Xing, Jeongkeun Lee, Dijiang Huang	Proposed a multi-phase distributed vulnerability detection, measurement & countermeasure selection mechanism called NICE, which is built on attack graph based analytical models &reconfigurable virtual network based countermeasure.	NICE just examined the network IDS way to deal with counter zombie explorative attacks. Keeping in mind the end goal to enhance detection based accuracy, have based IDS arrangement are should have been fused and to cover the entire range of IDS in the cloud framework
2	2014[8]	Gideon Creech and Jiankun Hu	Introduced a new host-based anomaly intrusion detection methodology having discontiguous system call patterns, while trying to build detection rates while reducing false alert rates.	Researched the transference process further, alongside endeavors to diminish the training overhead and upgrade the inherent resilience of the new semantic feature to mimicry assaults.
3	2015[9]	KekeGai, MeikangQiu, Lixin Tao, Yongxin	Recognized and outlined the fundamental methods being executed in IDSs and mobile distributed computing with an investigation challenges for each technique.	Need solutions for (1) How would we be able to keep clients' protection spills while receiving the cloud-based IDS? (2) What is a protected information transmission strategy between end clients and cloud-based IDSs? (3) Whether we can build up a vitality mindful model for cell phones keeping in mind the end goal to guarantee they can completely use the advantages of heterogeneous 5G.
4	2016[10]	Robert Mitchell &Ing-Ray Chen	Proposed and dissect a behavior-rule specification-based strategy for intrusion detection of medical devices implanted in a medical cyber physical system (MCPS) in which the patient's security is absolutely crucial.	Need to break down the overheads of proposed detection procedures, for example, the distance-based based techniques in correlation with contemporary methodologies.
5	2017[11]	Rana Aamir Raza Ashfaq, Xi-Zhao Wang, Joshua Zhexue Huang, Haider Abbas, Yu-Lin He	Proposed a novel fuzziness based semi-supervised learning method by using unlabeled samples assisted with supervised learning methods to enhance the classifier's execution for the IDSs.	Need to apply this procedure to enhance the viability of IDSs for distinguishing numerous sorts of assaults.
6	2018[12]	L. Khalvati, M. Keshtgary and N. Rikhtegar	A hybrid approach is proposed towards achieving a high performance in IDS	Need to consider a hybrid approach that performs better in detecting the R2L, U2R, and Probe attacks and also need to find a new way to choose the number of clusters and also the initial cluster medoids.

5. Genetic Algorithm Approach

A Genetic algorithm is a programming technique that follows the methodology of natural selection of nature. The genetic algorithm approach uses a chromosome like data structure consisting of genes (features) along with some mutation and selection operators. The genes are encoded as bits to represent information. The process starts with an initial population generated randomly and are regarded as the candidate solutions. A fitness function is used to select chromosomes within the population to generate the next generation of chromosomes. Crossover and Mutation operators are then applied to the off-spring and consequently the fitness function is used to test the 'goodness' of the chromosomes from the new population which are fit to generate the next generation of chromosomes. To ensure survival, the fittest chromosomes are selected for reproduction of the next generation of species. The outcome of a genetic algorithm depends upon the representation of individuals, the fitness function and the parameters used in the genetic algorithm.

Genetic algorithm applied to intrusion detection provides very good results with regard to attacks like dos, smurf, U2R etc. A genetic algorithm based approach was described by Goyaland Kumar [13] in 2008 to classify all types of smurf attack with a detection rate of nearly 100% and a false positive rate of about 0.2%.

There are a few steps to be implemented while using Genetic Algorithm.

1. **Initial Population**: - An initial population is selected to start the algorithm. The population consists of chromosomes and each chromosome consists of many genes. The genes are encoded (Aggregated) into a chromosome. Each chromosome is also called as an individual in the population. Usually a binary string is used for representing an individual.
2. For example, the string '0111010100101' is an individual of arbitrary length and each bit within the individual is a gene. Many chromosomes congregate to form a population.
3. **Fitness function**: - The fitness function gives the measure of the chance of an individual being selected into the next generation. The more fit the individual is, the better it has a chance to reproduce for the next generation.
4. **Selection**: - Individuals are selected to perform crossover based on their fitness scores. The higher the fitness score, more the chance of crossover of genes. The fittest individuals (parents) are selected for the crossover process.
5. **Crossover**: - The crossover phase is a very significant phase of the algorithm. The individuals selected for mating are taken and a random point of partition is selected in the genome of both the parents. The bits are exchanged, until the crossover point, between the parents to generate the offspring. This offspring is added to the next generation.
6. **Mutation**: -Mutation is a process of introducing diversity into the population by incorporating randomness into the genome of the offspring. To implement mutation, some of the bits of the offspring are flipped with a low random probability. The process of mutation also prevents premature convergence.
7. **Termination**: - The algorithm terminates when the population of the current generation is similar to the population of the previous generation.

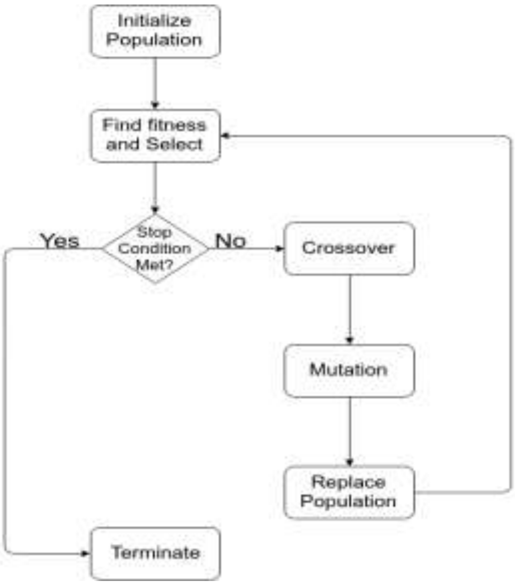


Fig: Genetic Algorithm approach

5.1. Elitist Approach

Elitist Genetic Algorithm(EGA) is a modification of the original genetic algorithm where from a particular generation best individuals are selected and they are taken over to the next generation without undergoing any mutation. This process continues in all generations till we are getting the best results and there cannot be any mutation possible. Elitism is one of the most practical variants of Genetic Algorithm. It has also been seen that the quality of the solution does not decrease from one generation to another. From the perspective of IDS each chromosome contains a part of the population and the fitness is calculated and it is checked that for which attributes we get the value of the fitness function is the highest. On the basis of the fitness value the population is divided into different ranks. The population consists of parents and child from the parents and child the individuals(attributes) having highest fitness is copied to the next generation and the rest are mutated and then they evolve to the next generation.

Steps: -

Step 1: Create initial the population
Step 2: Generate random population
Step 3: Generate Parents and initialize them
Step 4: Find Fitness value of all the individuals
Step 5: Divide the population into different ranks based on the fitness
Step 6: Copy the top ranks to the new generation along with some parents.
Step 7: Select the features
Step 8: Crossover the rest of the ranks followed by mutation
Step 9: Calculate the fitness values
Step 10: If the fitness value does not fit the condition go back to Step 5.

5.2. Proposed Work

In our proposed work, we have implemented a modified Elitist approach where the value of fitness is multiplied by the times a variable which is determined on the basis of the value of Kappa.

Pseudo Code: -

```
1.  K= Kappa ()
2.  K1 = K
3.  K= $K * (K/(1 - K)^K * 2$
4.  If K>1
```

a. K = K1

5. Return K

The main motivation for the modification to the algorithm is the fact that the parents while performing crossover may alter their genes to create new mutations within the parent's genes causing them to increase their fitness. However, the probability of gaining from the crossover is 0.5. Hence a randomization is implemented to actually modify the fitness value of the parent itself and consequently have another chance at crossover in the next generation.

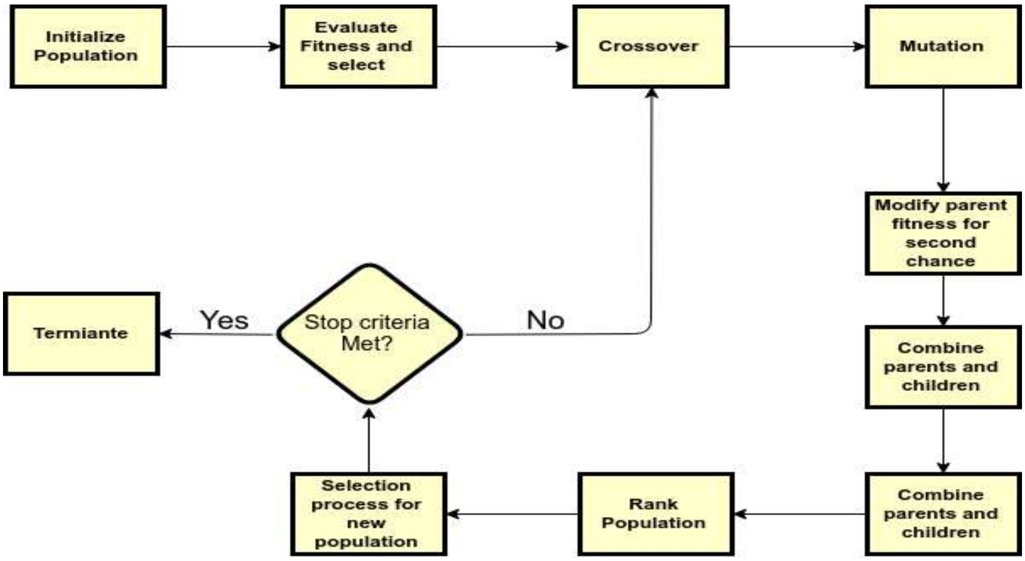

Initialize Population
Evaluate Fitness and select
Crossover
Mutation
Modify parent fitness for second chance
Combine parents and children
Combine parents and children
Rank Population
Selection process for new population
Stop criteria Met?
Yes
No
Termiante

Figure: he Schematic Diagram of proposed scheme using Elicits Approach

6. Results and Discussion

From the implementation results of the algorithms it can be inferred that the Proposed approach where the fitness function of the genetic algorithm is modified gives better results than the normal genetic algorithm. We have compared the accuracy of the genetic algorithm and the proposed approach. The classifier that was used in the study is ensemble classifier which is a combination of many classifiers, for example: Random forest, SVM etc. We have applied the algorithms on the NSL KDD 99 dataset which is a benchmark dataset. We have found that the Genetic algorithm gives an accuracy of 85.5% on an average and it depends on the number of iterations. Our proposed approach gave an accuracy of 87%.

Table 1: Confusion Matrix for Genetic algorithm

9454	257	3083	9750
9422	289	2979	9854

The last ten iterations (for simplicity) along with their accuracies are plotted below.

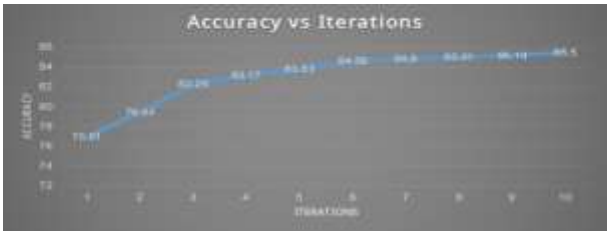
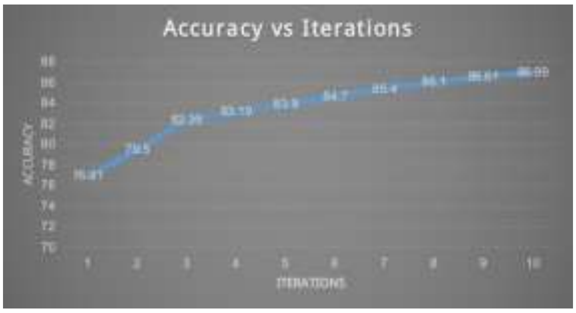


Table 2: Confusion Matrix for Proposed algorithm

9422	289	2979	9854
9757	244	2689	9854



7. Conclusion& Future Work

In this paper, we have proposed a modified Elitist approach where the value of fitness is multiplied by the times a variable which is determined on the basis of the value of Kappa (K).The main motivation for the modification to the algorithm is the fact that the parents while performing crossover may alter their genes to create new mutations within the parent's genes causing them to increase their fitness. However, the probability of gaining from the crossover is 0.5. Hence a randomization is implemented to actually modify the fitness value of the parent itself and consequently have another chance at crossover in the next generation. Our future research will be guided towards applying this technique to enhance the viability of IDSs for distinguishing different kinds of assaults by applying Fuzzy logic, Rough Set and Neutroscopic sets.

References

[1] Kabir, E., Hu, J., Wang, H., &Zhuo, G. (2017). A novel statistical technique for intrusion detection systems. *Future Generation Computer Systems*.

[2] Hamed, T., Dara, R., & Kremer, S. C. (2018). Network intrusion detection system based on recursive feature addition and bigram technique. *Computers & Security*, *73*, 137-155.

[3] Verma, A., &Ranga, V. (2018). Statistical analysis of CIDDS-001 dataset for Network Intrusion Detection Systems using

Distance-based Machine Learning. *Procedia Computer Science*, *125*, 709-716.

[4] Farnaaz, N., &Jabbar, M. A. (2016). Random Forest Modeling for Network Intrusion Detection System. *Procedia Computer Science*, *89*, 213-217.

[5] Gautam, S. K., & Om, H. (2016). Computational neural network regression model for Host based Intrusion Detection System. *Perspectives in Science*, *8*, 93-95.

[6] Chung, C. J., Khatkar, P., Xing, T., Lee, J., & Huang, D. (2013). NICE: Network intrusion detection and countermeasure selection in virtual network systems. *IEEE transactions on dependable and secure computing*, *10*(4), 198-211.

[7] Creech, G., & Hu, J. (2014). A semantic approach to host-based intrusion detection systems using contiguous and discontinuous system call patterns. *IEEE Transactions on Computers*, *63*(4), 807-819.

[8] Gai, K., Qiu, M., Tao, L., & Zhu, Y. (2016). Intrusion detection techniques for mobile cloud computing in heterogeneous 5G. *Security and Communication Networks*, *9*(16), 3049-3058.

[9] Mitchell, R., & Chen, R. (2015). Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems. *IEEE Transactions on Dependable and Secure Computing*, *12*(1), 16-30.

[10] Ashfaq, R. A. R., Wang, X. Z., Huang, J. Z., Abbas, H., & He, Y. L. (2017). Fuzziness based semi-supervised learning approach for intrusion detection system. *Information Sciences*, *378*, 484-497.

[11] Keshtgary, M., &Rikhtegar, N. (2018). Intrusion Detection based on a Novel Hybrid Learning Approach. *Journal of AI and Data Mining*, *6*(1), 157-162.

[12] Goyal, A., & Kumar, C. (2008). GA-NIDS: a genetic algorithm based network intrusion detection system. *Northwestern university*.