MACHINE LEARNING APPROACH FOR DETECTING THE NETWORK INTRUSION WITH FEATURE SELECTION

[1] Dr. Yenna Geetha Reddy, Associate Professor, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad.
[2] Y.Mahitha, B.Tech, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad.
[3] T.Ankitha, B.Tech, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad.
[4] P. Soumya Sri, B.Tech, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad.
[5] Y. GowThami, B.Tech, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad.

Abstract: A novel supervised machine learning system is developed to classify network traffic whether it is malicious or benign. To find the best model considering detection success rate, combination of supervised learning algorithm and feature selection method have been used. Through this study, it is found that Artificial Neural Network (ANN) based machine learning with wrapper feature selection outperform support vector machine (SVM) technique while classifying network traffic. To evaluate the performance, NSL-KDD dataset is used to classify network traffic using SVM and ANN supervised machine learning techniques. Comparative study shows that the proposed model is efficient than other existing models with respect to intrusion detection success rate.

1. INTRODUCTION

With the wide spreading usages of internet and increases in access to online contents, cybercrime is also happening at an increasing rate [1-2]. Intrusion detection is the first step to prevent security attack. Hence the security solutions such as Firewall, Intrusion Detection System (IDS), Unified Threat Modeling (UTM) and Intrusion Prevention System (IPS) are getting much attention in studies. IDS detects attacks from a variety of systems and network sources by collecting information and then analyzes the information for possible security breaches [3]. The network based IDS analyzes the data packets that travel over a network and this analysis are carried out in two ways. Till today anomaly based detection is far behind than the detection that works based on signature and hence anomaly based detection still remains a major area for research [4-5]. The challenges with anomaly based intrusion detection are that it needs to deal with novel attack for which there is no prior knowledge to identify the anomaly. Hence the system somehow needs to have the intelligence to segregate which traffic is harmless and which one is malicious or anomalous and for that machine learning techniques are being explored by the researchers over the last few years [6]. IDS however is not an answer to all security related problems. For example, IDS cannot compensate weak identification and authentication mechanisms or if there is a weakness in the network protocols. Studying the field of intrusion detection first started in 1980 and the first such model was published in 1987 [7]. For the last few decades, though huge commercial investments and substantial research were done, intrusion detection technology is still immature and hence not effective [7]. While network IDS that works based on signature have seen commercial success and widespread adoption by the technology based organization throughout the globe, anomaly based network IDS have not gained success in the same scale. Due to that reason in the field of IDS, currently anomaly based detection is a major focus area of research and development [8]. And before going to any wide scale deployment of anomaly based intrusion detection system, key issues remain to be solved [8]. But the literature today is limited when it comes to compare on how intrusion detection performs when using supervised machine learning techniques [9].networks against malicious activities anomaly-based network IDS is a valuable technology. Despite the variety of anomaly-based network intrusion detection techniques described in the literature in recent years [8], anomaly detection functionalities enabled security tools are just beginning to appear, and some important problems remain to be solved. Several anomaly based techniques have been proposed including Linear Regression, Support Vector Machines (SVM), Genetic Algorithm, Gaussian mixture model, knearest neighbor algorithm, Naive Bayes classifier, Decision Tree [3,5]. Among them the most widely used learning algorithm is SVM as it has already established itself on different types of problem [10]. One major issue on anomaly based detection is though all these proposed techniques can detect novel attacks but they all suffer a high false alarm rate in general. The cause behind is the complexity of generating profiles of practical normal behavior by learning from the training data sets [11]. Today Artificial Neural Network (ANN) are often trained by the back propagation algorithm, which had been around since 1970 as the reverse mode of automatic differentiation [12]. The major challenges in evaluating performance of network IDS is the unavailability of a comprehensive

network based data set [13]. Most of the proposed anomaly based techniques found in the literature were evaluated using KDD CUP 99 dataset [14]. In this paper we used SVM and ANN – two machine learning techniques, on NSLKDD [15] which is a popular benchmark dataset for network intrusion.

2. LITERATURE SURVEY

1. Supervised Feature Selection Techniques in Network Intrusion Detection: a Critical Review

Machine Learning (ML) techniques are becoming an invaluable support for network intrusion detection, especially in revealing anomalous flows, which often hide cyber-threats. Typically, ML algorithms are exploited to classify/recognize data traffic on the basis of statistical features such as inter-arrival times, packets length distribution, mean number of flows, etc. Dealing with the vast diversity and number of features that typically characterize data traffic is a hard problem. This results in the following issues: i) the presence of so many features leads to lengthy training processes (particularly when features are highly correlated), while prediction accuracy does not proportionally improve; ii) some of the features may introduce bias during the classification process, particularly those that have scarce relation with the data traffic to be classified. To this end, by reducing the feature space and retaining only the most significant features, Feature Selection (FS) becomes a crucial pre-processing step in network management and, specifically, for the purposes of network intrusion detection. In this review paper, we complement other surveys in multiple ways: i) evaluating more recent datasets (updated w.r.t. obsolete KDD 99) by means of a designed-from-scratch Python-based procedure; ii) providing a synopsis of most credited FS approaches in the field of intrusion detection, including Multi-Objective Evolutionary techniques; iii) assessing various experimental analyses such as feature correlation, time complexity, and performance. Our comparisons offer useful guidelines to network/security managers who are considering the incorporation of ML concepts into network intrusion detection, where trade-offs between performance and resource consumption are crucial.

2. Network Data Feature Selection in Detecting Network Intrusion using Supervised Machine Learning Techniques

Network attacks have become necessary in today's time due to increased network traffic. To determine whether network traffic is normal or anomalous a supervised machine learning system is developed. A network intrusion detection system (IDS) is a must-have piece of a security system. This proposed study aims to discover new patterns automatically from substantial quantities of network data, reducing time manually compiling intrusion and normal behavior patterns. The best model in terms of detection success rate was discovered using a supervised learning algorithm and feature selection method. AdaBoost outperforms Neural Network, kNN, and Naive Bayes in supervised machine learning with feature selection in this study, with a detection accuracy of 100.00%, 99.30%, 91.60%, and 99.70%, respectively. The Network Intrusion Detection dataset is used to classify network intrusions to evaluate the study and it has also been used in past studies. On the other hand, the proposed model proved to be more effective than other studies in terms of intrusion detection. The proposed approach can be used in various fields, including finance, health, and transportation. Furthermore, additional parameter tuning could be added, and different feature selection techniques could be used to improve the performance of the classifiers

3. A Hybrid Machine Learning and Data Mining Based Approach to Network Intrusion Detection

This paper outlines an approach to build an Intrusion detection system for a network interface device. This research work has developed a hybrid intrusion detection system which involves various machine learning techniques along with inference detection for a comparative analysis. It is explained in 2 phases: Training (Model Training and Inference Network Building) and Detection phase (Working phase). This aims to solve all the current real-life problem that exists in machine learning algorithms as machine learning techniques are stiff they have their respective classification region outside which they cease to work properly. This paper aims to provide the best working machine learning technique out of the many used. The machine learning techniques used in comparative analysis are Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) along with NSLKDD dataset for testing and training of our Network Intrusion Detection Model. The accuracy recorded for Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN) and Support Vector Machines(SVM) respectively when tested independently are 98.088%,

82.971%, 95.75%, 81.971% and when tested with inference detection model are 98.554%, 66.687%, 97.605%, 93.914%. Therefore, it can be concluded that our inference detection model helps in improving certain factors which are not detected using conventional machine learning techniques.

4. Enhancing Classification of Network Intrusion Attacks using Feature Reduction

In recent years Intrusion detection systems (IDS) considered an important approach to secure the network. The importance of IDS is due to the increasing of unauthorized access and policy violations. Machine learning approaches have been used in recent years in the field of network intrusion detection. These approaches can classify anomalous and normal patterns. Most of the databases used in the intrusion detection systems contain duplicates and irrelevant records. To improve detection systems and learning rate feature selection or feature reduction has been used in most approaches. In this paper NSL-KDD and UNSW-NB15 datasets have been used to evaluate the performance. Correlation and information gain have been used as feature selection method. Comparative study shows that the detection accuracy by UNSW-NB15 dataset is better than NSL-KDD dataset. WEKA tool has been used as simulation tool.

5. Heterogeneous Ensemble Feature Selection for Network Intrusion Detection System

Intrusion detection systems get more attention to secure the computers and network systems. Researchers propose different network intrusion detection systems using machine learning techniques. However, the massive amount of data that contain irrelevant and redundant features is still challenging the intrusion detection systems. The redundancy and irrelevance of features may slow the processing time and decrease prediction performance. This paper proposes a Heterogeneous Ensemble Feature Selection (HEFS) method to select the relevant features while achieving better attack detection performance. The proposed method fuses the output feature subsets of five filter feature selection methods, using a union combination method, to obtain an ensemble features subset. HEFS method uses merit-based evaluation to avoid the internal redundancy of the obtained ensemble features subset and acquire the final optimal features. We evaluate the HEFS method with random forest, J48, random tree, and REP tree. In a multi-class NSL-KDD dataset, the experimental results show that the proposed method achieves better prediction performance than the specific feature selection methods and other frameworks.

3. EXISTING SYSTEM

An intrusion detection technique that considers various issues like hugeness of network traffic dataset, feature selection, low accuracy and high rate of false alarms [2]. Online Sequential Extreme Learning Machine (OS-ELM) is used to process network traffic dataset to detect intrusions [5]. It is fast and accurate single hidden layer feed forward neural network (SHLFN) which can process network instances one by one or in chunks. It has proved its applicability in classification by performing in single iteration.

DISADVANTAGES OF EXISTING SYSTEM

1. The feature selection method is not good, irrelevant and redundant features are present.

2. The classifier does not work well for limited training data set.

4. PROPOSED SYSTEM

The promise and the contribution machine learning did till today are fascinating. There are many real life applications we are using today offered by machine learning. It seems that machine learning will rule the world in coming days. Hence we came out into a hypothesis that the challenge of identifying new attacks or zero day attacks facing by the technology enabled organizations today can be overcome using machine learning techniques. Here we developed a supervised machine learning model that can classify unseen network traffic based on what is learnt from the seen traffic. We used both SVM and ANN learning algorithm to find the best classifier with higher accuracy and success rate.

ADVANTAGES

- The new proposal was innovative as Hidden Naïve bayes which shows more advantage then traditional naïve bayes
- Which analyzes the large volume of network data and considers the complex properties of attack behaviors to improve the performance of detection speed and detection accuracy

SYSTEM ARCHITECTURE

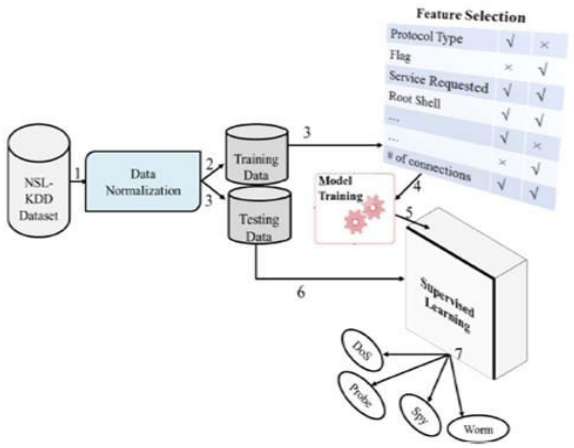

Fig 1: System Architecture

5. DATASET:

We can collect the dataset from the kaggle.com site and placed into our project folder used for detecting the network intrusions.

duration,protocol_type,service,flag,src_bytes,dst_bytes,land,wrong_fragment,urgent,hot,num_failed_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creations,num_shells,num_access_files,num_outbound_cmds,is_host_login,is_guest_login,count,srv_count,serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_host_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate,label
0,tcp,ftp_data,SF,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,150,25,0.17,0.03,0.17,0.00,0.00,0.00,0.05,0.00,normal
0,udp,other,SF,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,13,1,0.00,0.00,0.00,0.00,0.08,0.15,0.00,255,1,0.00,0.60,0.88,0.00,0.00,0.00,0.00,0.00,normal
0,tcp,private,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,123,6,1.00,1.00,0.00,0.00,0.05,0.07,0.00,255,26,0.10,0.05,0.00,0.00,1.00,1.00,0.00,0.00,anomaly
0,tcp,http,SF,232,8153,0,0,0,0,1,0,0,0,0,0,0,0,0,0,5,5,0.20,0.20,0.00,0.00,1.00,0.00,0.00,30,255,1.00,0.00,0.03,0.04,0.03,0.01,0.00,0.01,normal
0,tcp,http,SF,199,420,0,0,0,0,1,0,0,0,0,0,0,0,0,0,30,32,0.00,0.00,0.00,0.00,1.00,0.00,0.09,255,255,1.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,normal
0,tcp,private,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,121,19,0.00,0.00,1.00,1.00,0.16,0.06,0.00,255,19,0.07,0.07,0.00,0.00,0.00,0.00,1.00,1.00,anomaly
0,tcp,private,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,166,9,1.00,1.00,0.00,0.00,0.05,0.06,0.00,255,9,0.04,0.05,0.00,0.00

Table 5.1: Network Intrusion data

6. UML DIAGRAMS

1. CLASS DIAGRAM

The cornerstone of event-driven data exploration is the class outline. Both broad practical verification of the application's precision and fine-grained demonstration of the model translation into software code rely on its availability. Class graphs are another data visualisation option.

The core components, application involvement, and class changes are all represented by comparable classes in the class diagram. Classes with three-participant boxes are referred to be "incorporated into the framework," and each class has three different locations:

• The techniques or actions that the class may use or reject are depicted at the bottom.

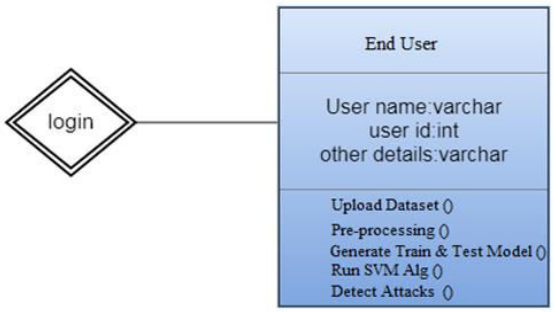

Fig 6.1 shows the class diagram of the project

2. USECASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

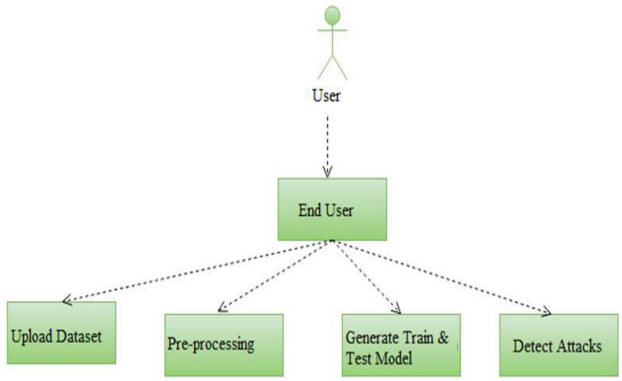

Fig 6.2 Shows the Use case Diagram

3. SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

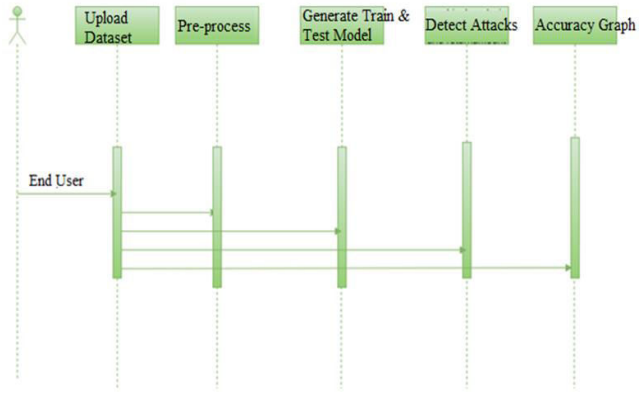

Fig 6.3 Shows the Sequence Diagram

7. RESULTS

7.1 Output Screens

Fig 7.1 Upload the Dataset

In above screen click on 'Upload Train Dataset' button and upload dataset

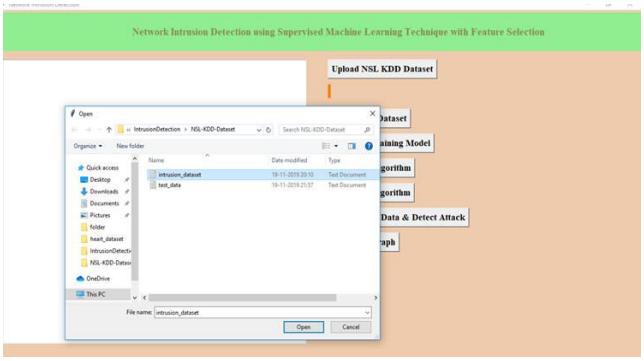

Fig 7.2 Uploading the Dataset File

In above screen I am uploading 'intrusion_dataset.txt' file, after uploading dataset will get below screen

Fig 7.3 Preprocess the dataset

Now click on 'Pre-process Dataset' button to clean dataset to remove string values from dataset and to convert attack names to numeric values

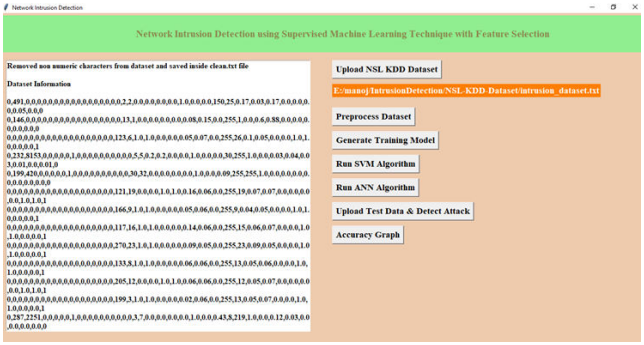

Fig 7.4 Preprocessing Results

After pre-processing all string values removed and convert string attack names to numeric values such as normal signature contains id 0 and anomaly attack contains signature id 1.

Now click on 'Generate Training Model' to split train and test data to generate model for prediction using SVM and ANN

Fig 7.5 Run the SVM Algorithm

In above screen we can see dataset contains total 1244 records and 995 used for training and 249 used for testing. Now click on 'Run SVM Algorithm' to generate SVM model and calculate its model accuracy

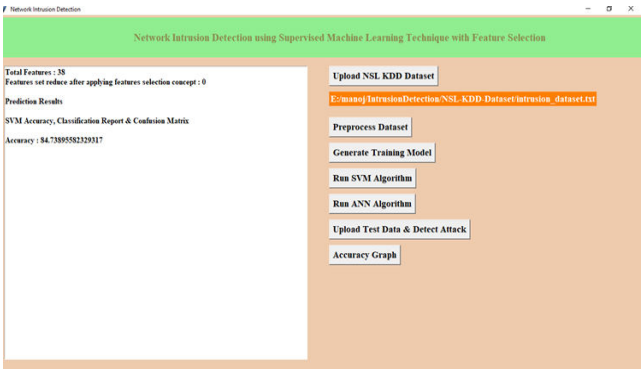

Fig 7.6 Run the ANN Algorithm

In above screen we can see with SVM we got 84.73% accuracy, now click on 'Run ANN Algorithm' to calculate ANN

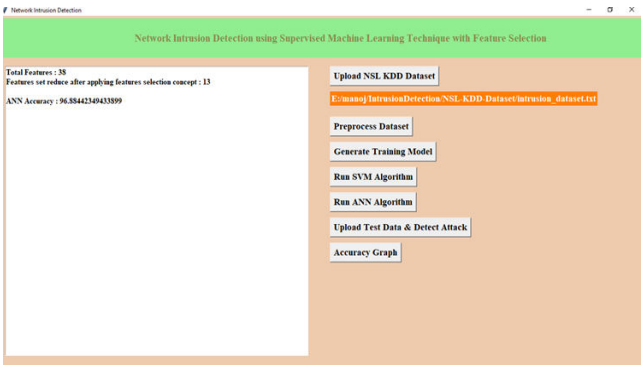

Fig 7.7 Accuracy ANN Algorithm

In above screen we got 96.88% accuracy, now we will click on 'Upload Test Data & Detect Attack' button to upload test data and to predict whether test data is normal or contains attack. All test data has no class either 0 or 1 and application will predict and give us result. See below some records from test data

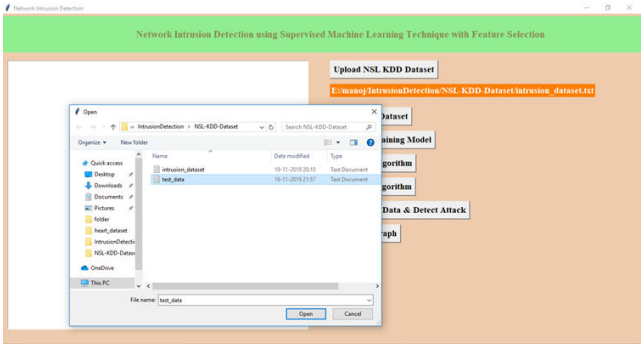

Fig 7.9 Import the test data

In above screen I am uploading 'test_data' file which contains test record, after prediction will get below results.

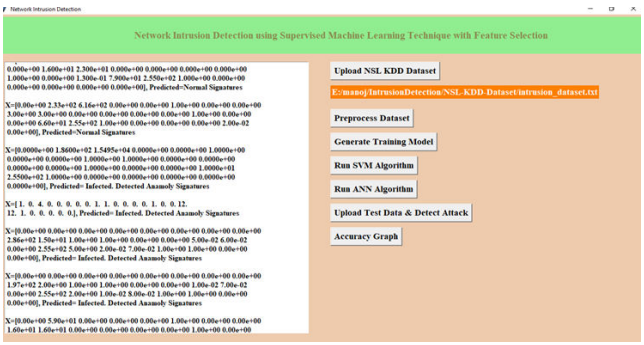

Fig 7.9 Intrusion anomaly Detection Results

In above screen for each test data we got predicted results as 'Normal Signatures' or 'infected' record for each test record. Now click on 'Accuracy Graph' button to see SVM and ANN accuracy comparison in graph format

8. CONCLUSION

In this paper, we have presented different machine learning models using different machine learning algorithms and different feature selection methods to find a best model. The analysis of the result shows that the model built using ANN and wrapper feature selection outperformed all other models in classifying network traffic correctly with detection rate of 94.02%. We believe that these findings will contribute to research further in the domain of building a detection system that can detect known attacks as well as novel attacks. The intrusion detection system exist today can only detect known attacks. Detecting new attacks or zero day attack still remains a research topic due to the high false positive rate of the existing systems.

9. REFERENCES

[1] H. Song, M. J. Lynch, and J. K. Cochran, "A macro-social exploratory analysis of the rate of interstate cyber-victimization," American Journal of Criminal Justice, vol. 41, no. 3, pp. 583–601.

[2] P. Alaei and F. Noorbehbahani, "Incremental anomaly-based intrusion detection system using limited labeled data," in Web Research (ICWR), 2017 3th International Conference on, 2017, pp. 178–184.

[3] M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, "Modeling and implementation approach to evaluate the intrusion detection system," in International Conference on Networked Systems, 2015, pp. 513–517.

[4] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, “Toward credible evaluation of anomaly-based intrusion-detection methods,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 5, pp. 516–524.

[5] A. S. Ashoor and S. Gore, “Importance of intrusion detection system (IDS),” International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp. 1–4.

[6] M. Zamani and M. Movahedi, “Machine learning techniques for intrusion detection,” arXiv preprint arXiv:1312.2177.

[7] N. Chakraborty, “Intrusion detection system and intrusion prevention system: A comparative study,” International Journal of Computing and Business Research (IJCBR) ISSN (Online), pp. 2229–6166.

[8] H. Song, M. J. Lynch and J. K. Cochran, "A macro-social exploratory analysis of the rate of interstate cyber-victimization", American Journal of Criminal Justice, vol. 41, no. 3, pp. 583-601.

[9] P. Alaei and F. Noorbehbahani, "Incremental anomaly-based intrusion detection system using limited labeled data", Web Research (ICWR) 2017 3th International Conference on, pp. 178-184.

[10] M. Saber, S. Chadli, M. Emharraf and I. El Farissi, "Modeling and implementation approach to evaluate the intrusion detection system", International Conference on Networked Systems, pp. 513-517.

[11] M. Tavallaee, N. Stakhanova and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods", IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews), vol. 40, no. 5, pp. 516-524,