A Comprehensive Review on Intrusion Detection System and Techniques

Dharmendra Kumar[1]
Department of Computer Science and Engineering
R.K.D.F. University, Bhopal, India
dharmendra_cse@rediffmail.com

Ravi Singh Pippal[2]
Department of Computer Science and Engineering
R.K.D.F. University, Bhopal, India
ravesingh@gmail.com

***Abstract*—Intrusion detection system (IDS) is an important component to maintain network security. As network applications grow rapidly, network security mechanisms require more attention to improve speed and accuracy. The evolving nature of new types of intrusion poses a serious threat to network security: although many network security tools have been developed, the rapid growth of intrusive activities is still a serious problem. Intrusion detection systems (IDS) are used to detect intrusive network activity. Machine learning and data mining techniques have been widely used in recent years to improve intrusion detection in networks. These techniques allow the automatic detection of network traffic anomalies. In this paper, provide an overview of the research progress using machine learning to the problem of intrusion detection. The goal of this paper summarized and compared research contributions of Intrusion detection system using machine learning, define existing research challenges and anticipated solution of machine learning. This paper discusses some commonly used machine learning techniques in Intrusion Detection System and also reviews some of the existing machine learning IDS proposed by researchers at different times.**

Keywords—component; formatting; style; styling; insert (key words)

I. INTRODUCTION

This template, modified in MS Word 2007 and saved as a “Word 97-2003 Document” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings.

Malwares are a big threat to modern computer world. These are mischievous programs crafted to prohibit the normal operations, gain unauthorized access to data and resources of the system that may lead to privacy violation and other abusive behavior. Poor separation between code and data is the foremost cause of malware. Intrusion detection is a technique that attempts to discover the unauthorized access to a computer by analyzing the malicious activities and negatively identifying all the non-attacks [1].

Nowadays, protection of network and system via malware and attacks is a very challenging task. The requirement of intrusion detection system depends upon the intrusion recognition, data collection and preprocessing. Intrusion recognition is most vital, among these. Observe data and detection model are compared and identify the patterns of intrusion behavior either its successful or unsuccessful. At the earlier age of intrusion detection system e focused only ho to implement the intrusion detection system, according to Denning [2] observed this major research identification in 1987. During 1980s to1990s, a hybrid expert system and statistical approaches were very much famous. Detection models were generated from the domain knowledge of security experts. Set of training data discovery methods for a set of training data is data classification, rule-based induction, and data-clustering. In intrusion detection problems, data are not trivial when the process of automatically constructing models [3].

There are several challenges during intrusion detection system such as distinguishes criteria for normal and abnormal behavior, dynamically update the process of learning, huge network traffic. Thus, at that criteria machine learning andartificialintelligenceunabletohandlethesolutionsocomputationalintelligence play measure role into this significance.

Machine learning techniques can be effective for detecting intrusions. Many Intrusion Detection Systems are modeled based on machine learning techniques. Learning algorithms are designed either on offline dataset or real data collected from university or organizational networks. Usually machine learning techniques is classified into two classes i.e. supervised Learning and unsupervised Learning [4]..

II. Intrusion detection

An intrusion is an unauthorized attempt to break into a computer system. Such a break may force the system to move into an insecure state. It is a deliberate attempt made by the intruder to gain access of, manipulate or misuse valuable resources. If successful, it may result in rendering the resources as unreliable or unusable.

In intrusion detection system, dynamically observe the system and network regarding a legitimate user of the system. In Fig. 1 denotes solid lines for data or control flow, dashed lines for intrusive activities. Intrusion detection is a type of security management system which is not only to identify an attack but also it includes the following [5]:

1. Monitoring the user and system activities
2. Analyzing system configurations and target vulnerabilities
3. Analysis of abnormal activity pattern
4. Accessing file integrity
5. Track of user policy violation
6. It provides the ability to recognize patterns typical of attacks by providing access to system and file integrity. If any target found vulnerable it would notify.

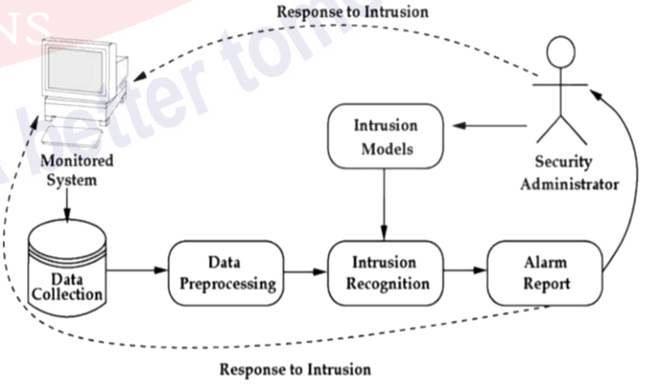

Fig. 1: **Fundamental block diagram of intrusion detection system**

Intrusion Detection System is a security system that dynamically monitors and observes the target system for any misuse and handles the abnormal activity either by itself or by raising an alarm [6]. It can be configured to respond to predefined

suspicious activities (e.g. when someone is trying to compromise the system's information through malicious activities) or it can even monitor the internet for latest attacks that could result into some future attack.

Intrusion detection system is categorized into several types which are discussed below:

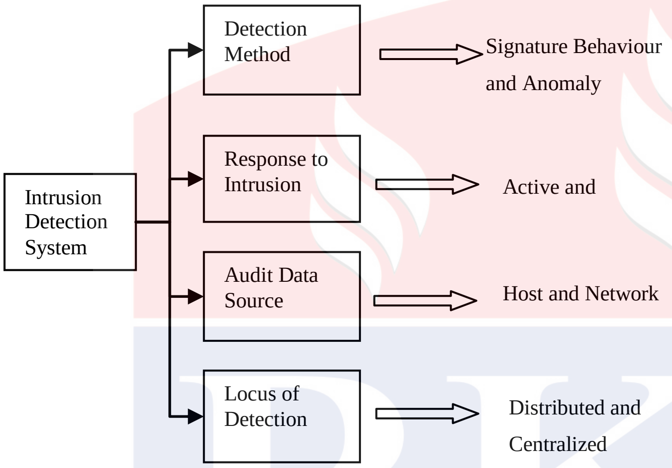

Fig. 2: Types of intrusion detection system

Signature based IDS: This is a primitive, simple and efficient technique of intrusion detection. This technique basically scans the malware program code and extracts its signature pattern [1]. Later it matches this signature pattern with the ones already fed in the database. During the extraction of signatures, all the system logs, executable files, records are taken into account. An alarm is raised immediately notifying the system user of the attack. The database is created by the antivirus developer, who analyses the new discovered malwares to find a specific pattern or a signature. Once the signature is extracted, it is then updated into the database. Since, detection rate and accuracy largely depends on the preexisting database, antiviruses need to be updated timely to provide better protection against the new upcoming latest malwares.

Anomaly based IDS: This technique overcomes the limitations of the signature based technique. The process of detection is divided into two phases namely, training phase and detection phase. In the training phase the system is trained about the normality whereas in the detection phase it compares the real data with the established profile to flag deviations and raise alerts [6]. It considers heuristics and artificial intelligence type techniques to differentiate between normal and abnormal activities. Its main advantage is that it can easily detect unknown viruses as they also produce a different, anomalous behavior. The only disadvantage with this technique is that there are a very large number of false positive attacks.

Behavior based IDS: It is used for detection of viruses in terminals like PCs and mobile phones. This technique observes behavior of various malwares and fetches information like its source and destination, attachment size, type etc. Based on this behavior; it can flag a particular source as an intruder. It is capable of detecting known and unknown viruses with self-reference replication behavior. The only disadvantage is generation of false positives and false negatives.

Active and passive IDS: IDS can also categorized by active IDS and passive IDS, where usually used and known Intrusion Detection and Prevention System (IDPS) is active Intrusion Detection Systems. It is designed such that it automatically blocks suspected attacks without any interference by an operator. In response to an attack IDPS has the advantage of providing real-time correction action. A passive IDS is a system that is configured to only monitor and analyze activities of network traffic and to generate alert for an operator to potential findings of attacks. Therefore passive IDS cannot perform any protection or correction functions on its own [7,8].

Host Based IDS: This type of IDS collects information from an individual computer system. It is the combination of signature based, rule based and heuristics based approach to detect intrusion. It monitors local host events, so can even detect attacks that a Network based IDS may miss. Events monitored include contents of operating systems, system and applications logs [9]. Its shortcoming is that it can easily be disabled by certain denial-of-service attacks. For example Tripwire.

Network Based IDS: This type of IDS collects information from the entire network. Sensors are placed that monitor all network traffic. Packets are collected and the captured data is then analyzed for predefined patterns and signatures to generate alerts [7]. There are 3 signatures that are very important [5]: a. String signatures b. Port signatures c. Header signatures This type of IDS does not actually impact the network performance, and are independent of the operating system. Its only when the networks are flooded, the packets are lost. For example SNORT [11].

Distributed IDS: A distributed IDS consists of multiple intrusion detection system over large network, all of which communicate with each other. A DIDS can be considered as "Meta-IDS" collecting data from widely dispersed IDS, correlating attacks and providing information for determining patterns and mitigation strategies across the internet as whole. It is designed for heterogeneous network.

III. Related Work

So far, it has been discussed in this paper about some of the existing approaches which are incorporating IDS. However, there is no universal efficient solution found yet. Each has some limitations. Some of the important contributions in the field of IDS are discussed below:

Author Name	**Approach Used**	**Average Detection Rate**
Sufyan T. Faraj Al-Janabi et al. [2]	The model used BPANN for classification of anomalous network traffic from normal traffic.	93 % (on testing)
Yinhui Li et al. [3]	K-means clustering is used to compact the dataset into 5 clusters. Ant Colony Optimization (ACO) algorithm was then used to select a small representative subset of the whole dataset. Further Gradually Feature Removal (GFR) is used to reduce the size of the feature set. At the final step, SVM classified the attack instances from benign data.	98.62 %
Feng et al. [4]	Introduced a new classification technique and utilized the advantages of SVM and Clustering based on Self-Organized Ant Colony Network.	94.86 %
Meng et al. [8]	Compared ANN, SVM and DT schemes for anomaly detection in an uniform environment and concluded that J48 algorithm of DT gives better performance than the other two schemes. The detection rate of low frequent attack types (U2R, R2L) was also high.	About 99 %
Manjula C. Belavagi et al. [10]	Classification and predictive models for intrusion detection are built by using machine learning classification algorithms namely Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest. Experimental results shows that Random Forest Classifier out performs the other methods in identifying whether the data traffic is normal or an attack.	About 99 %

Saad Mohamed et al. [11]	Presented a hybrid approach to anomaly detection using of K-means clustering and Sequential Minimal Optimization (SMO) classification.	97.36%
Horng et al. [14]	Proposed an IDS based on a combination of BIRCH hierarchical clustering and SVM technique	95.72%
Kuang et al. [15]	Proposed an IDS based on a combination of the SVM model with kernel principal component analysis (KPCA) and genetic algorithm (GA). KPCA was used to reduce the dimensions of feature vectors, whereas GA was employed to optimize the SVM parameters.	95.26%
Khadija Hanifi et al. [16]	In this work, to detect network attacks, used the k-means algorithm a new semi-supervised anomaly detection system.	80.119%.
Wathiq Laftah Al-Yaseen et al. [17]	Presented hybrid SVM and Extreme machine learning technique.	95.75%

The following are the most common shortcomings:

Lack of Efficiency: IDS are necessary to evaluate activities in real time. It is very difficult to manage a large number of events, as is the case with today's networks. As a result, HIDS generally slows down a system in which NIDS release network packets for which processing time is insufficient.

High Number of False Positives: Most IDS detects attacks on the enterprise by analyzing information from a single host, application, or network interface at many points in the network. False alarms are high and attack detection is not perfect. Lowering the thresholds to reduce false alarms increases the number of attacks not detected as false negatives. Improving the ability to accurately identify attacks is the main problem faced by IDS manufacturers today.

Burdensome Maintenance: There is a specific knowledge of the requirements and a considerable effort to configure and maintain IDS. Here are some examples, such as detection of abuses, which typically use expert system shells for their implementation that encode and map signatures using rule sets. The updating of these rule sets includes details specific to the expert system and its language for the expression of rule sets and can only allow an indirect specification of sequential relations between events. These considerations also apply to the addition of a statistical measure generally used to find unusual variations in nature.

Limited Flexibility: When an intrusion detection system is designed for a typical specific environment, it can be difficult to use in other environments, whether it is the same concerns or policies themselves. The mechanism used for detection can also be difficult to adapt to different usage patterns. The process of identifying personalization, particularly for the system with some yews and chains replacing them with time-enhanced detection techniques, is also problematic in many IDS implementations. Often, the IDS must be restarted completely for the changes and additions to take effect.

High Speed Communications: the increase in communication speed also changes the processing speed because the communication speed is directly proportional to the processing speed. It is therefore necessary to analyze the contents of the communication packages. This is where packet loss can occur. With NIDS, to observe different communication flows at the same time, the difficulty increases when communication is changed, i.e. Conventional communication is used instead of broadcast communication. So far, we have discussed some of the existing approaches which are incorporating IDS.
Here we summarize some common pros and cons of existing techniques of IDS:
Requires more training time and samples for detection accuracy.
Cannot be used for all types of attacks.
Computational overhead high.

IV. Machine learning Approach in IDS

The goal of machine learning algorithms is to learn automatically without human intervention. machine learning is the basic field of artificial intelligence, because learning is the heart of intelligence [2-6]. There are different types of machine learning techniques such as:

Supervised Learning: These algorithms make predictions for specific data samples. The entry is forming data and labels known as spam or non-spam. A model is prepared by a training process, they are made in forecasts and corrections if these predictions are false. This training process continues until the model achieves the required accuracy of the training data.

Artificial Neural Network (ANN): ANN are the computational models of neural structure of human brain. Neurons are the basic building blocks of human brain. An ANN is a layered network of artificial neurons. An ANN may consist of an input layer, one or more hidden layer(s) and an output layer. The artificial neurons of one layer are fully or partially connected to the artificial neurons of the next layer. Each of these connections is associated with a weight, and feedback connections to the previous layers are also possible [2].

Support Vector Machine: It is for classification and regression problems. SVM classifies data into different classes by identifying a hyperplane (line) that separates training data into classes. When the hyperplane that maximizes the distance between classes is identified, the probability of generalizing invisible data increases. SVM offers the best classification performance, i.e. The accuracy of the training set. It does not overflow the data. SVM does not make strong assumptions about the data. Show more efficiency for the correct classification of future data. SVM is classified into two categories, i.e. Linear and non-linear. In a linear approach, training data is represented by a line, i.e. hyperplane.

k-Nearest Neighbor (kNN): kNN is used for classification and regression problems. This is one of the simplest classification algorithms. Determines the parameter k, which is the number of the nearest neighbors. When a new data point is to be classified, its nearest neighbors are determined by the training data. The distance is calculated using one of the Euclidean distance measurements, the Minkowski distance, the Mahalanobis distance. The bigger the k, the better the classification.

Decision Tree Algorithms: The decision model is based on the actual values of the attributes in the data. The decision interval continues until a prediction decision is made for a given record. It has a default destination variable. Decision trees are trained in the data for classification and regression problems. Decision trees are popular in machine learning because they are often quick and precise. It works for categorical and continuous input and output variables. In this technique, the population or sample is divided into two or more homogeneous subpopulations or more based on the most significant fragment in the input variables. The decision is taken from the tree in the strategic division. This greatly affects the accuracy of the tree. This decision criterion differs for classification and regression trees. Decision trees use different algorithms to decide to divide a node into two or more subnodes. The trees break the nodes for all the available variables, then select the division that leads to the most homogeneous sub-nodes. The most common decision tree algorithms are: C4.5 and C5.0, CART (Classification and Regression Tree).

Unsupervised Learning: Unsupervised learning is a type of machine learning algorithm that is used to draw conclusions from data sets consisting of input data without marked responses.
The most common non-supervised learning method is cluster analysis, which is used to analyze exploration data to find hidden models or group data. Clusters are modeled using a similarity measure defined by metrics such as Euclidean or Probabilistic distance.
Common clustering algorithms include:
Hierarchical clustering: builds a multilevel hierarchy of clusters by creating a cluster tree.
K-Means clustering: partitions data into k distinct clusters based on distance to the centroid of a cluster.
Gaussian mixture models: models clusters as a mixture of multivariate normal density components.
Self-organizing maps: uses neural networks that learn the topology and distribution of the data.
Hidden Markov models: uses observed data to recover the sequence of states.

V. Proposed Model

The biggest challenge for today is to protect the users from Intrusion due to wide use of internet. Intrusion Detection Systems (IDS) are one of the security tools available to detect possible intrusions in a Network or in a Host. Research showed that application of machine learning techniques in intrusion detection could achieve high accuracy rate as well as low false alarm rate. Accurate predictive models can be built for large data sets using supervised machine learning techniques, that is not possible by traditional methods. IDS learns the patterns by the training data, so it can detect only the known attack, new attacks cannot be identified. This research work is based on designing an optimized feature based classifier and performing analysis on three different datasets. This section describes the proposed hybrid model for intrusion detection. The KDD-99 dataset is used as a benchmark to evaluate the performance of the proposed model.

The algorithm flow of the proposed method is described as follows:

Following steps will be used to build the proposed model for intrusion detection:

Step 1: Convert the symbolic attributes protocol, service, and flag to numerical.

Step 2: Normalize data to [0,1].

Step 3: Separate the instances of dataset into two categories: Normal, DOS, R2L, U2R and Probe.

Step 4: Feature Reduction and Extraction.

Step 5: Data Clustering

Step 6: The data set is divided as training data and testing data.

Step 7: Train classifier with these new training datasets.

Step 8: Test model with dataset.

Step 9: Finally computing and comparing Accuracy and FAR for different classifiers.

The proposed algorithm flow diagram of intrusion detection model is illustrated in figure 3. The proposed framework consists of three phases i.e. Preprocessing, Post Processing Phase and Intrusion Detection Phase. Below each stage is described individually in details.

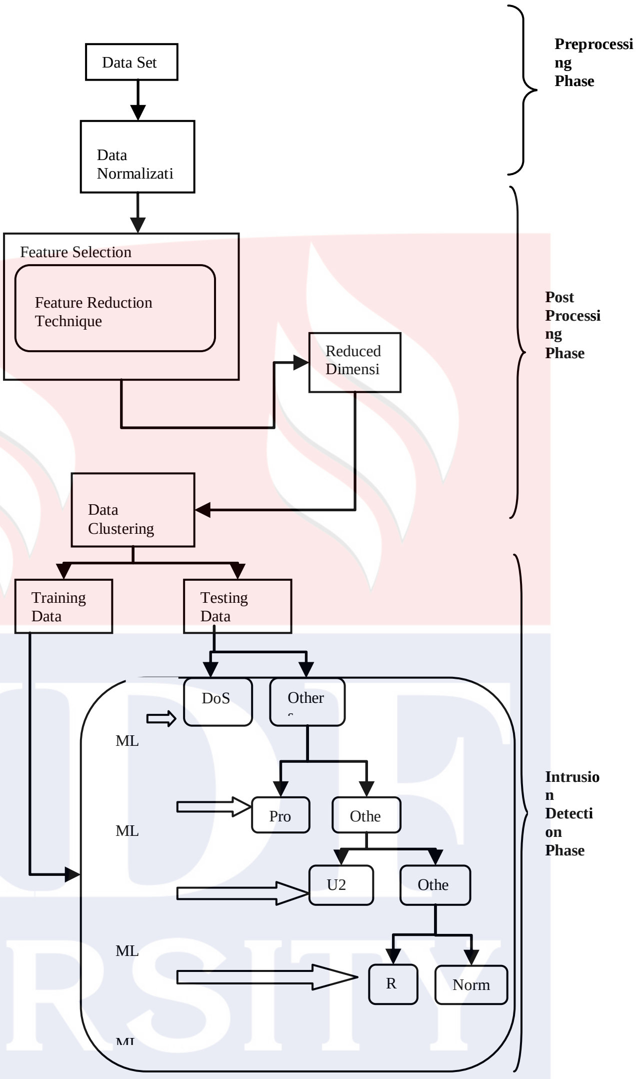

Figure 3: Proposed Flow Diagram of Intrusion Detection System

To evaluate the proposed algorithm, it is concentrated on three indications of performance: detection rate, accuracy and False Alarm Rate (FAR).

If one sample is an anomaly and the predicted label also stands anomaly, then it is called as true positive (TP).

If one sample is an anomaly, but the predicted label stands normal, then it is called as false negative (FN).

If one sample is a normal and the predicted label also stands normal, then it is true negative (TN).

If one sample is normal, but the predicted label stands anomaly, then it is termed as false positive (FP).

TP stands the number of true positive samples, FN stands the number of false negative samples, FP stands the number of false positive samples, and TN stands the number of true negatives.

To evaluate the proposed algorithm, it is concentrated on three indications of performance: detection rate, accuracy and False Alarm Rate (FAR).

If one sample is an anomaly and the predicted label also stands anomaly, then it is called as true positive (TP).

If one sample is an anomaly, but the predicted label stands normal, then it is called as false negative (FN).

If one sample is a normal and the predicted label also stands normal, then it is true negative (TN).

If one sample is normal, but the predicted label stands anomaly, then it is termed as false positive (FP).

TP stands the number of true positive samples, FN stands the number of false negative samples, FP stands the number of false positive samples, and TN stands the number of true negatives.

From equation (5.1) and (5.5), the accuracy, detection rate, False Positive rate (FPR), False Negative Rate (FNR) and False Alarm rate (FAR) is calculated.

Accuracy = (TP+TN)/(TP+TN+FP+FN)*100 (1)

Detection Rate =TP/(TP+FN)*100 (2)

False Negative Rate (FNR) = FN/(FN+TP) *100 (3)

False Positive Rate (FPR) =FP/(FP+TN) *100 (4)

False Alarm Rate (FAR)= (FPR+FNR)/2 (5)

VI. Conclusion

In modern society, the security of computer networks becomes an increasingly vital issue to be solved. Traditional intrusion detection techniques lack extensibility in face of changing network as well as adaptability in face of unknown attack type. Machine learning techniques are proved to be efficient for intrusion detection. High accuracy in intrusion detection can be achieved using machine learning techniques even though the detection accuracy depends on some other factors too. Some of them are selection of correct feature set, selection of appropriate training and testing data, etc. With the selection of the appropriate attributes for these factors, a higher performance could be achieved. This research work proposes a multi-level hybrid classification intrusion detection system. The proposed model may illustrate better performance than single level classification models. The clustering technique is used to pre -process training dataset and provides high accuracy and detection rate as compared to existing work

References

[1] Garcia-Teodoro, P., “Anomaly-based network intrusion detection: techniques”, systems and challenges. Comput. Security vol. 28.issue, pp. 18–28, 2009.

[2] Sufyan T Faraj Al-Janabi, Hadeel Amjed Saeed, “A neural network based anomaly intrusion detection system”, IEEE, 2011.

[3] Li, Y., “An efficient intrusion detection system based on support vector machines and gradually feature removal method”, Expert Syst. Appl. 39(1), 424–430, 2012.

[4] Feng, W., “Mining network data for intrusion detection through combining SVMs with ant colony networks”, Future Gener. Comput. Syst. 37, 127–140, 2014.

[5] P. L. Nur, A. N. Zincir-heywood, and M. I. Heywood, “Host-Based Intrusion Detection Using Self-Organizing Maps,” in Proceedings of the IEEE International Joint Conference on Neural Networks, pp. 1714–1719, 2002.

[6] K. Labib and R. Vemuri, “NSOM: A Real-Time Network-Based Intrusion Detection System Using Self-Organizing Maps,” 2000.

[7] Sharma, R.K., Kalita, H.K., Issac, B., “Different firewall techniques: a survey”, International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 2014.

[8] Meng, Y.-X., “The practice on using machine learning for network anomaly intrusion detection”, International Conference on Machine Learning and Cybernetics (ICMLC), vol. 2, IEEE, 2011.

[9] Sumaiya Thaseen Ikram, Aswani Kumar Cherukuri, “Intrusion detection model using fusion of chi-square feature selection and multi class SVM”, Journal of King Saud University –Computer and Information Sciences, 2016.

[10] Manjula C. Belavagi and Balachandra Muniyal, “Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection, Procedia Computer Science”, Elsevier, 2016.

[11] Saad Mohamed Ali Mohamed Gadal and Rania A. Mokhtar, “Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique”, International Conference on Communication, Control, Computing and Electronics Engineering, IEEE, 2017.

[12] Ibrahim, H. E., Badr, S. M., & Shaheen, M. A., “Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems”, International Journal of Computer Applications, vol. 56, issue 7, pp. 10–16, 2012.

[13] Wen Feng, Qinglei Zhang, Gongzhu Hu, Jimmy Xiang Huang, “Mining network data for intrusion detection through combining SVMs with ant colony networks”, Elsevier, Vol 37, pp 127-140, 2014.

[14] Shi-JinnHorng, Ming-Yang Su, Yuan-Hsin Chen, Tzong-Wann Kao, Rong-Jian Chen, Jui-Lin Lai, Citra Dwi Perkasa, “A novel intrusion detection system based on hierarchical clustering and support vector machines” Expert Systems with Applications, Elsevier, vol. 38, pp. 306–313, 2011.

[15] Kuang, F., Xu, W., & Zhang, S., “A novel hybrid KPCA and SVM with GA model for intrusion detection”, Applied Soft Computing Journal, vol. 18, pp. 178–184, 2014.

[16] Khadija Hanifi ve Hasan Bank “Network Intrusion Detection Using Machine Learning Anomaly Detection Algorithms” , IEEE, 2016.

[17] Wathiq Laftah Al-Yaseen , Zulaiha Ali Othman ,Mohd Zakree Ahmad Nazri, “Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System”, International Journal in Expert Systems With Applications, Elsevier, 2017.