IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Machine Learning in Intrusion Detection – A Survey

P. Anitha[1*], D. Rajesh[2], K. Venkata Ratnam[3]

[1,3]Department of CSE, Gayatri Vidya Parishad College for Degree and P.G Courses (A), Visakhapatnam, India
[2]Department of CSE, RISE Krishna Sai Prakasam Group of Institutions, Ongole

Corresponding author: anitha501p@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i3.112119 | Available online at: www.ijcseonline.org

Accepted: 07/Mar/2019, Published: 31/Mar/2019

Abstract*:* With the huge expansion of internet based services and important information on networks, network protection and security is a very significant task. Intrusion Detection system (IDS) is the standard component in network security framework and is essential to protect computer systems and network from different attacks. IDSs is designed to detect both known and unknown attacks in computer systems and networks. This paper presents different Machine Learning techniques of IDS for protecting computers and networks. This study analyzes different machine learning methods in IDS. It reviews related studies focusing on single, hybrid and ensemble classifiers with relevant datasets.

Keywords: Machine Learning, intrusion detection, Single Classifiers, Hybrid Classifiers, Ensemble Classifiers.

I. INTRODUCTION

With the advancement of internet, people enjoy the benefits of information technology, but the risk of network intrusion sharply increases. The rapid advancement in technology not only gives ease of access to the common people but also gives sophisticated techniques to the cybercriminals. This leads to the huge number of cyber-attacks on both individuals and organizations. The individuals and organizers, both need to protect credential data from the intruders. Generally, we secure our systems by building firewalls or employ some authentication mechanisms such as passwords or encryption techniques which create a protective covering around them. The above techniques provide a level of security but they cannot provide protection against inside attacks, malicious code or unsecured modems. We need more security mechanisms such as IDS because firewalls cannot detect attacks inside the network since they are mostly deployed at the boundary of the network, and thus only control traffic entering or leaving the network. An IDS is very helpful and act as a safeguard for data integrity, confidentiality and system availability for different kinds of attacks. An IDS is one of the framework security foundations that attempts to identify harmful activities.

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or forth coming threats of violation of computer security policies, acceptable use policies, or standard security practices. Such incidents have many causes, like malware, attackers gaining unauthorized access to systems from the Internet, and authorized users of systems who attempt to gain additional privileges or misuse their privileges for which they are not authorized [1]. The motive of the Intrusion Detection System (IDS) is to identify inner as well as outer attacks. Although many incidents are naturally malicious, many others are not. For instance, a person might mistype the address of a computer and accidentally attempt to connect to another system without authorization.

Therefore the aim of this paper is to review related studies published in the past by examining the techniques that have been used. Section II describes about IDS, Section III presents detection methodologies, section IV presents performance metrics, and section V presents performance metrics, section VI contains related work.

II. INTRUSION DETECTION SYSTEM

An intrusion detection system (IDS) is a software that automates the intrusion detection process. The term "Intrusion Detection" covers a wide range of technologies that are involved in the detection and reporting of network security events. These techniques can help to reduce the following type of threats [2].

- Unauthorized Access
- Data Destruction
- Buffer Overflow attempt
- System or Network Eavesdropping
- Denial of Service (DoS)

IDS can be a software or hardware or a combination of both that detects intrusions into a system or a network [3]. Active

IDS tries to block the attacks and counter measures or at the least alert administrators. Passive IDS just record the log intrusion details or create traces for audit. Based on the source of literature study information, IDS can be classified into two categories Host-based and Network-based intrusion detection systems.

Host-based IDS (HIDS): HIDS resides on the single host and scans activities. It is deployed on the systems which are more vulnerable to attacks such as web server. To perform intrusion detection, HIDS gathers information from its system calls, operating system audit trails, application logs, etc. HIDS stores information into a secure Database and compares to detect any malicious activity.

Network-based IDS (NIDS): NIDS is liable for detecting abnormal and unauthorized activity in a network. It is designed to monitor the packets on a network segment and detect suspicious activities in the network to prevent an illegal access of network resources.

1. Intrusion Detection Techniques

A) Misuse or Signature-based Intrusion Detection System: This detection system identifies abnormal behavior depending upon on the signatures of the known attacks and framework vulnerabilities. These attacks or vulnerabilities are already stored in the database. In this technique, abnormal activity compares with already known attack signature which is stored in the database. Misuse based IDS cannot detect a new attack [1] [4] [5].

B) Anomaly-Based Intrusion Detection System: Anomaly-based IDS detects the behavior of users and system activities. Initially, it creates profiles of users, hosts, network connections and applications. If the new activity deviates from the normal behavior, that activity is treated as an intrusion [5] [6]. It is able to detect novel attacks with the aid of anomaly based IDS.

III. DECTECTION METHODOLOGIES

Machine Learning Approaches

Machine learning is a special branch of artificial intelligence. It is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. Machine learning techniques are divided into three broad categories such as supervised, unsupervised, and reinforcement learning.

1) Supervised Learning

Supervised learning is also known as Classification. In supervised learning data, the learner is provided with two sets of data: a training set and a test set. The idea is for the learner to “learn” from a set of labeled examples in the training set, and it can identify unlabeled examples in the test set with the highest possible accuracy. That is, the goal of the learner is to develop a rule, a program, or a procedure that classifies new examples (in the test set). By analyzing examples, it has been given that test set already has a class label.

2) Unsupervised Learning

Unsupervised learning technique is known as Clustering. Unsupervised learning is having only input data and not having corresponding output variables. The goal of unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

3) Reinforcement Learning

Reinforcement learning means, a computer interacting with an environment to achieve a certain goal. It is at given a situation, it chooses possible actions in order to maximize reward.

To develop an Intrusion Detection System based on machine learning algorithms; three types of classifiers can be used. These classifiers are single classifiers, Hybrid classifiers, and ensemble classifiers.

Single Classifiers: One machine learning algorithm or technique for developing an intrusion detection system can be used as a stand alone classifier or a single classifier. Some of the single classifiers are as follows:

i) Decision Tree: Decision tree is one of the most popular method in data-mining and machine learning, for constructing prediction models from training datasets. It creating a classifier for predicting the value of the target class for an unseen test instance, based on several known instances. The optimal value is built by recursively portioning the training data space based on the values of one or more features which are represented in a tree structure. Each leaf in this tree represents a class, each non-leaf node represents a test as an attribute, and each branch represents an outcome of the test [7]. Decision trees are easy to interpret, computationally inexpensive, and capable of coping with noisy data [8]. It can be expanded in two types. These are: 1. classification tree, with a range of symbolic class labels 2. Regression tree with a range of numerically valued labels [9].

ii) Bayesian Decision Approach: It is the simplest and commonly used probabilistic classifier based on applying Bayes theorem with the assumption that all the features are conditionally independent for a given class label. Although this assumption may not always be precise, it results in significantly simplified scalable classification models that amazingly work well. The naïve Bayes classifier is applied to intrusion detection system either as a stand-alone classifier or in combination with other learning techniques.

iii) Genetic Algorithm: Genetic algorithm is work based on principles of evolution and natural selection. This algorithm converts the problem in a specific domain into a model by using a chromosome-like data structure and evolves the chromosomes using selection, recombination, and mutation operators. In genetic algorithm, the process begins with a randomly selected population of chromosomes. According to the attributes of the problem, different positions of each chromosome are encoded as bits, characters or numbers. These positions are sometimes referred to as genes and are changed randomly within a range during evolution. The set of chromosomes during a stage of evolution are called a population. An evaluation function is used to calculate the "goodness" of each chromosome. During the evaluation, two basic operators, crossover and mutation are used to simulate the natural reproduction and mutation of species. The selection of chromosomes for survival and combination is biased towards the fittest chromosomes [10].

iv) K-Means Clustering: K- Means clustering is a simple algorithm which aims to partition n data points to K clusters. The algorithm follows a simple iterative procedure. It starts with some initial guess of K centroids, each representing the center point (or mean) of one cluster. Next, each data point is assigned to the cluster of the nearest centroid. Then, after assigning all data points to their clusters, the position of the K centroids are recalculated as the means of all data points assigned to their clusters. Last two steps are repeated until there is no significant change in the location of the centroids [11].

v) Markov Models: A Markov model is a stochastic process that assumes the Markovian property i.e. the conditional probability distribution of future states of the process (conditional on both past and present states) depends only upon the present state, but not on the sequence of events that preceded it. Using the Markov chain model, anomaly-based intrusion detection involves two phases. These are: training the model and detecting anomalies. In the training phase, transition probabilities are estimated from the normal behavior of the monitored system [12]. An anomaly is detected by comparing the transition probability obtained for the observed sequence of activities with a fixed threshold obtained from the training phase. The Hidden Markov Model (HMM) is a statistical Markov model in which only the symbols that are consumed or produced by the transition are observable while the states of the Markov process are hidden

vi) K-Nearest Neighbor (KNN): K-Nearest neighbor is an old and simple method to classify samples. This classification algorithm is a data-mining algorithm which is theoretically mature with low complexity. The basic idea is that, in a sample space, if most of its K nearest neighbor samples belong to a category, then the sample belongs to the same category. The nearest neighbor refers to the single or multi dimensional feature vector that is used to describe the on the closest sample, and the closest criteria can be the Euclidean distance of the feature vector [13].

vii) Artificial Neural Networks: Artificial Neural Networks are methods proposed to solve problems, based on the behavior of human brain activities. It is a massively parallel computing system consisting of a large number of processing units called as "neurons" with many inter connections. It can be viewed as a weighted directed graph in which neurons are nodes, and connections between them are the directed weighted edges. There are many different types of known ANN, but perceptron network is the simplest form of classifying linearly separable patterns. The most widely used feed-forward architecture in many applications is the Multi-Layer Perceptron (MLP). In this type, neurons are organized into layers that have unidirectional connections between them from inputs towards the outputs [14]. MLP learns to approximate a particular function, forming the relationship between the inputs and outputs by adjusting the weights of the connections. The learning process is achieved iteratively so that the network can efficiently perform the specified task.

viii) Support – Vector Machines: The standard SVM is the most popular and successful classification algorithm in the data mining area that can perform pattern recognition and regression estimation tasks. SVM executes binary classification by finding a hyper-plane in a high dimension space, where there is a minimum error rate and maximizes the margin of separation between the data points of the two classes. The separating hyper plane is determined by a small subset of the training data (called support vectors) rather than the whole training samples which make it robust to outliers. Hence, SVM can be used to learn nonlinear decision functions by first, mapping the data to some higher dimensional space and then constructing a separating hyper plane in the resulting space[15]. Since mapping the data to a higher dimensional space can be time and memory consuming, a special type of function (kernel function) is used, which allows the construction of an optimal separating hyper plane without explicitly performing calculations in the higher dimensional space.

ix) Fuzzy Logic: It is also known as a fuzzy set theory, used for reasoning. Its value ranges from 0 to 1. It is a very effective and potential technique. It deals with human decision making and reasoning. It uses if then or else rules [9] .e.g, to rain is a natural event and its range can be from slight to violent. Fuzzy logic has been employed to handle the concept of partial truth, where the truth value may range between completely true and completely false. It is used in many engineering applications, but mainly in anomaly IDS. It is more effective in port scans and probes involving high resource consumption.

x) Radial–Basis Function Network: A Radial – Basis Function Network(RBFN) is a special type of ANNs that uses radial basis functions in the hidden layer as activation functions. An RBFN typically has three layers such as an input layer, a single layer with non-linear radial basis functions, and an output layer with a linear activation function. RBFNs have been used to build models for intrusion detection.

xi) Self- Organizing Maps: A Self – Organizing Map (SOM) is trained using unsupervised competitive learning technique. SOM algorithm can map a high dimension data into two dimension array. A Kohenen network is the most commonly used type of SOMs, having a feed-forward structure with an input layer and computational (output) layer. In this structure, each neuron is fully connected to all the nodes in the input layer. In the training mode, SOMs builds the map using input examples with the competitive learning process. Mapping automatically classifies a new input vector. The training phase starts by initializing all connection weights to small random values. Then each input pattern is fed to the network: its Euclidean distance to all weight vectors is computed. The neuron with the closest weight to the input pattern wins the competition and its weight as well as the weights of the neurons close to it in the computational layers are adjusted towards the input pattern. This procedure will cause the output units to self-organize into an ordered map such that units with similar weights will be placed nearby, after training. Hence, similar input patterns are mapped into the same or similar output units.

Hybrid Classifiers: Hybrid classifier is to combine two or more single classifier models into a single approach in order to allow the learning algorithm to find a more compact representation for the hypothesis, and therefore enhance its performance on unseen data. Use of some cluster-based algorithms for preprocessing samples in training data for eliminating non-representative training samples and then, the results of the clustering are used as training samples for pattern recognition, in order to design a classifier. Thus, either supervised or unsupervised learning approaches can be the first level of a hybrid classifier.

Ensemble Classifiers: The classifiers are performing slightly better than a random classifier is known as weak learners. When multiple weak learners are combined for the purpose of improving the performance of a classifier, it is significantly known as Ensemble classifier. Commonly used ensemble methods are bagging, boosting and stacking. Though it is known that the disadvantages of the component classifiers get accumulated in the ensemble classifier, it has been producing a very efficient performance in some combinations. So, researchers are becoming more interested in ensemble classifiers day by day.

NSL KDD Dataset: In order to create the accurate model, dataset for training and testing must have various attack types and distribution of those attacks must reflect the real world scenario. The first dataset was created in 1998 by Lincoln Laboratory in Massachusetts Institute of Technology called DARPA'99 which one year later was improved by introducing a KDD cup 99 data set [16].

The inherent drawbacks in the KDD cup 99 dataset revealed by various statistical analysis has affected the detection accuracy of many IDS modeled by researchers. NSL-KDD data set is a refined version of its predecessor KDD cup 99 data set. It contains essential records of the complete KDD data set. There are a collection of downloadable files at disposal for researchers [17].

IV. METRIC PERFORMANCE

In security system, IDSs is one of the essential elements which allows the network administrators to identify the policy variations. In IDS, different types of alarm rates were increased based on that anomaly and misuse based attacks. The aim of researchers is to calculate accuracy, detection rate, and false alarms and tabulate to evaluate the performance of the model. The Table shows the data classification. True positive (TP) means actual attack data are classified as attacks and False Positive (FP) means actual data are classified as an attack. Likewise, False Negative (FN) means actual attack data are classified as normal and True Negative (TN) means normal data are classified as normal.

TABLE: TYPES OF DATA CLASSIFICATION

		Predicted	
		Attack	Normal
Actual	Attack	TP	FN
	Normal	FP	TN

Accuracy: The rate at which data are correctly classified, that is the cases for which actual data are classified as attacks and normal data as normal.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

Detection Rate: The rate at which actual attack data are correctly classified as attacks.

$$\text{Detection Rate} = \frac{\text{TP}}{\text{TP+FN}}$$

False Alarm: The cases for which normal data are incorrectly classified as an attack.

$$\text{False Alarms} = \frac{\text{FP}}{\text{FP+TN}}$$

V. RELATED WORK

Intrusion detection classification using machine learning has been extensively researched for conventional networks. Numerous works have been surveyed covering various learning approaches such as supervised learning, unsupervised learning or clustering, reinforcement learning, hybrid, and Ensemble classifiers.

Hossein M. Shirazi [18] proposed anomaly detection engines based on K-NN and K-Means Clustering algorithms. SF-5NN and SUS-5NN models used for feature selection. These two selected engines based on best selected features models on KDD99 dataset, provided good classification results.

Ahmad et al. [19] have argued the importance of feature selection as a vital problem in intrusion detection and verified it using Genetic Algorithms (GA) for selecting an optimum set of feature leads to significant advancement in detection rates. The authors have taken into consideration the Principal Component Analysis, which transmutes the input sample into a fresh feature space, and Support Vector Machines (SVM) for classification. This newly crafted feature space is searched using GA to select a subset of the features. The experiments performed by the authors have shown substantial performance enhancements.

W L Al-Yaseen et al. [20] proposed hybridized algorithms to address the network intrusion detection. This paper contains a hybrid of modified k-means with C4.5 intrusion detection system in a multi agent system (MAS-IDS).The MAS-IDS consists of three agents namely coordinator, analysis, and communication agent. The coordinator agent constructs the clusters by using modified K-means. It subsequently applies C4.5 technique on each cluster to build the decision tree that will be used in testing phase. The analysis agent receives subset of data, centroids of clusters and trees from the communication agents then analyze the data, returns the results to the coordinator agents by the communication agent. KDD Cup 1999 dataset is used for evaluation. This method is developed in JADE platform. The results were compared with the other methods, the MAS-IDS reduced the processing time up to 70%, while improving the detection accuracy.

G. Wang et al. [21] proposed FC-ANN, based on ANN and fuzzy clustering, to solve the problem and help IDS achieve higher detection rate, less false positive rate and stronger stability. The procedure of FC-ANN is initially, fuzzy clustering technique, used to generate heterogeneous training set which is divided in to several homogenous subsets. ANN module learns the pattern of every subset. ANN employs classic feed- forwarded neural networks trained with back propagation algorithm. Finally a meta-learner and fuzzy aggregation module aggregate ANN’s result and reduce the detection errors. KDD CUP 1999 dataset used to demonstrate the effectiveness of new approach especially for low-frequent attacks, i.e., R2L and U2R attacks in terms of detection stability and precision. It is unable to handle the noisy data and also it is complex in large dataset.

S.S. Sindhu et al. [22] proposed the crucial part of light weight IDS depending on preprocessing of network data, identifying important features in the design of efficient learning algorithm that classify normal and anomalous patterns. The goal of this paper is to initially remove redundant instances that cause the learning algorithm to be unbiased, identifying suitable subset of features by employing a wrapper based feature selection algorithm, and finally realizing proposed IDS with neurotree to achieve better detection accuracy. The combination of GA and neurotree algorithm is known as wrapper approach. An extensive experimental evaluation of the proposed approach compared with a family of six decision tree classifiers namely Decision Stump, C4.5, Naive Baye’s Tree, Random Forest, Random Tree and Representative Tree model. The performance metrics of proposed system are TP rate, FP rate, Precision, Recall, and F-measure are better even when the dataset is presented with different number of classes.

Reda M. Elbasiony et al.[23] proposed hybrid framework, the anomaly part is improved by replacing the k – means algorithm with another one called weighted k-means algorithm, moreover, it uses a proposed method in choosing the anomalous clusters by injecting known attacks into uncertain connections data. Our approaches are evaluated over the Knowledge Discovery and Data Mining (KDD’99) datasets. The results show that the hybrid framework achieves detection rates and false positive rates better than the other earlier proposed techniques.

Amreen Sultana and M.A.Jabbar [24] proposed intelligent network intrusion detection system using Average one dependence estimators (AODE) algorithm for the detection of different types of attacks. In order to evaluate the performance of our proposed system, we conducted experiments on NSL-KDD data set. Experimental results prove that accuracy, DR and MCC for four types of attacks are increased by our proposed method. Empirical results show that proposed model compared with naive bayes generates low false alarm rate and high detection rate.

M.S.M. Pozi et al. [25] identified that missing rare attacks can be defined as anomalous rare attacks and has hardly been solved in IDS proposed new classifier to improve the anomalous attacks detection rate based on support vector machine (SVM) and genetic programming (GP).Based on the experimental results, our classifier, GPSVM, managed to get higher detection rate on the anomalous rare attacks, without significant reduction on the overall accuracy, because GPSVM optimization task is to ensure that accuracy is balanced between classes without reducing the generalization property of SVM. The proposed classification algorithm, GPSVM, is meant to improve the detection rate of anomalous rare attacks and produce a more balanced classification accuracy on NSL-KDD dataset without a need to perform any reduction technique such as resampling and feature selection.

Ahmed I. Saleh et al. [26] designed a Hybrid IDS (HIDS) that can be successfully employed in a real time manner and suitable for resolving the multi-class classification problem. HIDS relies on a Naïve Base feature selection (NBFS) technique, which is used to reduce the dimensionality of sample data. Outliers are noisy input samples that can lead to high rate of misclassification. Optimized Support Vector Machines (OSVM) used for rejecting outliers. After outlier rejection, HIDS can successfully detect attacks through applying a Prioritized K-Nearest Neighbors (PKNN) classifier. Hence, HIDS is a triple edged strategy as it has three main contributions, which are: (i) NBFS, which has been employed for dimensionality reduction, (ii) OSVM, which is applied for outlier rejection, and (iii) PKNN, which is used for detecting input attacks. HIDS has been compared against recent techniques using three well-known intrusion detection datasets: KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets. HIDS has the ability to quickly detect attacks and accordingly can be employed for real time intrusion detection. OSVM and PKNN HIDS performed high detection rates specifically for the attacks which are rare such as R2L and U2R. PKNN is also suitable for resolving the multi-label classification problem.

F. Amiri et al. [27] proposed Feature Selection method in order to improve the performance of existing classifiers by excluding non-related features. Furthermore, an improved Partial Least Squares Support Vector Machine called PLSSVM has been introduced. A linear and non linear measure for the feature selection within pre-processing phase has been considered in this work. PLSSVM performed well in classifying normal and probe attacks records. In this work, the effect of changing feature goodness measure and evaluation function has been investigated by linear correlation- based feature selection (LCFS), forward feature selection (FFSA) and modified mutual information feature selection algorithms (MMIFS). Experiments on KDDcup99 dataset demonstrate that feature selection algorithms can greatly improve the classification accuracy. In contrast, PLSSVM missed a big number of dynamic attacks such as DoS and U2R attacks that behave quite similar to the normal behavior.

B. M. Aslahi-Shahri et.al [28] proposed a hybrid method of support vector machine (SVM) and genetic algorithm (GA). The proposed hybrid algorithm reduced the number of features from 45 to 10. Using GA the features are categorized into three priorities, as the highest important is the first priority and lowest is the third priority. The feature distribution is done in a way that 4 features are went in the first priority, 4 features in the second, and 2 features in the third priority. The proposed algorithm could show an outstanding true-positive value as well as low false-positive value.

Yassine Maleh, et.al [29] proposed hybrid, and light weight intrusion detection system for sensor networks. For anomaly detection uses based on support vector machine (SVM) algorithm and a set of signature rules to detect malicious behaviors and provide global lightweight IDS. The combination of these two techniques provide simulation results, those that can detect abnormal events efficiently and has low false alarm rate with high detection rate.

Jabez J et.al [30] proposed a new approach called outlier detection where, the anomaly dataset is measured by the Neighborhood Outlier Factor (NOF). This approach is used for improving the performance of Intrusion Detection system from big dataset with distributed storage environment. This technique tested with the KDD datasets that are received from real world. This approach takes less execution time and storage to test the dataset. The performances of this approach well and could significantly detect almost all anomaly data in the computer network.

R. A. R. Ashfaq et.al [31] proposed a novel fuzziness based semi-supervised learning approach for unlabeled samples, assisted with supervised learning algorithm. The unlabeled samples with their predicted labels are categorized according to the magnitude of fuzziness. A single hidden layer feed-forward neural network (SLFN) used for train data, and unlabeled samples are categorized into low, mid, and high fuzziness categories. Neural network with random weights (NNRw) is used for better learning performance. The classifier retained after incorporating each category separately into the original training set. This method used on the NSL-KDD dataset and unlabeled samples belonging to low and high fuzziness groups make major contributions to improve the classifier's performance, compared to existing classifiers such as naïve bayes, support vector machine, random forests.

Observations of literature survey:

In this area, let us discuss the different approaches utilized by different authors in their research work. From the literature review discussed above, it is concluded that most of the researchers used combination of clusters and classifiers. Some of the researchers used feature extraction methods to extract discriminative features. These features help to classify efficiently. Recently many researchers used hybrid methods as combination of clustering, classification and feature extraction. Very few number of researchers used ensemble classifiers as intrusion detection. The researchers used KDD CUP 99 and NSL-KDD datasets for evaluation of designed techniques.

VI. CONCLUSION

The intrusion detection system using Machine learning and Deep learning methods has received much attention for network security. Machine Learning algorithms are unable to detect zero-day attacks, and low frequent attacks such as R2L, and U2R. Data set always contain a huge number of features where most of these are redundant or irrelevant. Employing feature reduction method is an essential factor to reduce the computational cost and to increase the classifier performance. Feature selection and feature extraction are having advantages and disadvantages, which makes it hard to choose a single method to implement.

It is there by recommended to use feature extraction followed by feature selection as a hybrid approach to increase the accuracy of intrusion detection. Deep learning algorithms are recommended to attain more accuracy compared to Machine Learning algorithms.

REFERENCE

[1]. Guide to Intrusion Detection and Prevention Systems (IDPS), National Institute of Standards and Technology, Gaithersburg

[2]. D. E. Denning, "An intrusion-detection model", IEEE Trans. Softw. Eng., vol. 13, no. 2, pp. 222–232, Feb, 1987.

[3]. C. Guo, Y.-J. Zhou, Y. Ping, S.-S. Luo, Y.-P. Lai, and Z.-K. Zhang, "Efficient intrusion detection using representative instances," Computers & Security, vol. 39, pp. 255–267, 2013.

[4]. F. Amiri , M. M. Rezaei Yousefi , CaroLucas , A.Shakery and NasserYazdani, "Mutual information-based feature selection for intrusion detection systems" Journal of Network and Computer Applications 34 , 1184–1199, 2011.

[5]. S. Soni1 , P. Sharma, "Review of Hybrid Intrusion Detection System", International Journal of Computer Sciences and Engineering, Vol.-6, Issue-6, pp 1100-1104, 2018.

[6]. Uzair Bashir and Manzoor Chachoo, "Intrusion Detection and Prevention System: Challenges & Opportunities" IEEE ,pp. 806-809, 2014.

[7]. J. R. Quinlan, "Introduction of Decision Trees", Machine Learning vol. 1.

[8]. Xiao-Bai Li, "A scalable decision tree system and its application in pattern recognition and intrusion detections", Decision Support Systems 41 ,pp 112–130, 2005.

[9]. Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin and Wei-Yang Lin, "Intrusion detection by machine learning: A review", Expert Systems with Applications, vol.36,11994–12000, 2009.

[10]. Whitley, Darrell, "A Genetic Algorithm Tutorial." Statistics and Computing vol 4, 65-85, 1994.

[11]. Meng Jianliang, Shang Haikun and Bian Ling, "The Application on Intrusion Detection Based on K means Cluster Algorithm", International Forum on Information Technology and Applications, IEEE , 2009.

[12]. Shengfeng Tian, Chuanhuan Yin, and Shaomin Mu, "High-Order Markov Kernels for Network Intrusion Detection", Springer-Verlag Berlin Heidelberg, pp. 184 –191, 2006.

[13]. Ping yi , Yue Wu , "A New Intrusion Detection System Based on KNN Classification Algorithm in Wireless Sensor Network", Journal of Electrical and Computer Engineering, 2014.

[14]. C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[15]. C. Amali Pushpam , J. Gnana Jayanthi," A Review on effect of SVM in Intrusion Detection System", International Journal of Computer Sciences and Engineering, Vol. 6, Issue.12,pp.471-474, 2018.

[16]. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications (CISDA'09), IEEE Press, Piscataway, NJ, USA, p. 53-58, 2009.

[17]. L.Dhanabal, Dr. S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, June, pp 446-452, 2015.

[18]. Hossein Shirazi, "Anomaly intrusion detection system using information theory, K-NN and KMC algorithms", Aus tralian Journal of Bas ic and Applied Sciences, Vol. 3, pp- 2581-2597, 2009.

[19]. Ahmad, I., Abdullah, A., Alghamdi, A., & Hussain, M, "Optimized intrusion detection mechanism using soft computing techniques". *Telecommunication Systems, Vol.* 52(4), 2187–2195, 2013.

[20]. W.L. Al-Yaseen, Zulaiha Ali Othman, and Mohd Zakree Ahmad Nazri, "Hybrid Modified-Means with C4. 5 for Intrusion Detection Systems in Multiagent Systems", The Scientific World Journal, 2015.

[21]. G. Wang, J. Hao, J. Ma, and L. Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering," Expert systems with applications, vol. 37, no. 9, pp. 6225-6232, 2010.

[22]. S.S. Sindhu, S.G. Sivatha, and A. Kannan, "Decision tree based light weight intrusion detection using a wrapper approach," Expert Systems with applications, vol. 39, no. 1, pp. 129-141, 2012.

[23]. Reda M. Elbasiony , Elsayed A. Sallam , Tarek E. Eltobely , And Mahmoud M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means" Ain Shams Engineering Journal , 753–762, 2013.

[24]. 23 A. Sultana, and M.A. Jabbar, "Intelligent network intrusion detection system using data mining techniques," In the Proceedings of 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 329-333, 2016.

[25]. M.S.M. Pozi, M.N. Sulaiman, N. Mustapha, T. Perumal, "Improving anomalous rare attack detection rate for intrusion detection system using support vector machine and genetic programming," Neural Processing Letters, vol. 44, no. 2, pp. 279-290, 2016.

[26]. Ahmed I. Saleh , FatmaM. Talaat, LabibM. Labib, "A hybrid intrusion detection system (HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers", Artif Intell Rev, 2017.

[27]. Oladeji Patrick Akomolafe and Adeleke Ifeoluwa Adegboyega, "An Improved KNN Classifier for Anomaly Intrusion Detection System Using Cluster Optimization", International Journal of Computer Science and Telecommunications (IJCST) Vol. 8, Issue 2, 2017.

[28]. B. M. Aslahi-Shahri, R. Rahmani, M. Chizari, A. Maralani, M. Eslami, M. J. Golkar, and A. Ebrahimi, "A hybrid method consisting of GA and SVM for intrusion detection system", The Natural Computing Applications Forum 27 ,1669–1676, 2016.

[29]. Yassine Maleh, Abdellah Ezzati, Youssef Qasmaoui, Mohamed Mbida, 2015 ,"A Global Hybrid Intrusion Detection System for Wireless Sensor Networks", Procedia Computer Science vol. 52, pp 1047 – 1052.

[30]. Jabez J , Dr.B.Muthukumar, "Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach", Procedia Computer Science, vol. 48 , 338 – 346, 2015.

[31]. R. A. R. Ashfaq, X. Wang, J. Z. Huang , H. Abbas , and Yu-Lin He, " Fuzziness based semi-supervised learning approach for intrusion detection system", Information Sciences, vol.378 , pp 484–497, 2016.

Authors Profile

Mrs. P Anitha have completed M.Tech (Computer Science and Engineering) from Gayatri Vidya Parishad College of Engineering, Visakhapatnam. She is currently working as a Assistant Professor in Department of Computer Science and Engineering, at Gayatri Vidya Parishad College for Degree and P.G Courses (A) since 2017. Her main research work focuses on Artificial Intelligence and Network security based education. She has 7 years of teaching experience.

Mr. D. Rajesh have completed M.Tech (Software Engineering) from CVSR Engineering College, Hyderabad. He is currently working as a Assistant Professor in Department of Computer Science and Engineering, at Rise Krishna Sai Prakasam Group of Institutions, Ongole, since 2010. His main research work focuses on Artificial Intelligence and Data-mining based education. He has 10 years of teaching experience.

Mrs. K. Venkata Ratnam have completed M.Tech (Computer Science and Engineering) from Avanthi Institute of Technologies, Visakhapatnam. She is currently working as a Assistant Professor in Department of Computer Science and Engineering, at Gayatri Vidya Parishad College for Degree and P.G Courses (A) since 2016. Her main research work focuses on Artificial Intelligence and Data-mining based education. He has 5 years of teaching experience.