Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 72-81 (2020)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Intrusion Detection in Cyber Security: Role of Machine Learning and Data Mining in Cyber Security

Gillala Rekha[1], Shaveta Malik[2], Amit Kumar Tyagi[3,*], Meghna Manoj Nair[3]

[1]*Koneru Lakshmaiah Education Foundation, Department of Computer Science and Engineering, Hyderabad, India – 522502*

[2]*Terna Engineering College, Department of CSE, Navi Mumbai, Maharashtra, India.*

[3]*Vellore Institute of Technology, School of Computer Science and Engineering, Chennai Campus, Chennai, 600127, Tamilnadu, India.*

ARTICLE INFO

Article history:
Received: 10 August, 2019
Accepted: 04 March, 2020
Online: 03 May, 2020

Keywords:
Cyber Security
Intrusion Detection System
Machine Learning
Data Mining

ABSTRACT

In recent years, cyber security has been received interest from several research communities with respect to Intrusion Detection System (IDS). Cyber security is "a fast-growing field demanding a great deal of attention because of remarkable progresses in social networks, cloud and web technologies, online banking, mobile environment, smart grid, etc." An IDS is a software that monitors a single or a network of computers from malicious activities (attacks). Detecting an intrusion or prevention (due to increase the usage of internet), is becoming a critical issue. In past, several techniques have been proposed to overcome or detect intrusion in a network. But most of the techniques (used now days in detecting IDS) are not able to overcome this problem (in efficient manner).Together this, Machine Learning (ML) also has been adopted in various applications (due to providing good accuracy results (in respective domain)). Hence, this work discusses "How machine learning anddata mining can be used to detect IDS in a network" in near future.ML use efficient methods like classification, regression, etc., with efficient results like high detection rates, lower false alarm rates and less communication costs. This work also provides a detail comparison with metrics in table 1-3 (with their performance/ algorithms/ dataset or metrics used).

1. Introduction

Cyber security involves the practice of preventing the exposure of computers, programs, etc. from attacks, unauthorized usage, modifications, destructions, etc. It's a common practice to find every Cyber Security system to have a firewall, antivirus techniques and Intrusion Detection System (IDS). IDS are a crucial component as they help in spotting any undesirable and unwanted changes in the system [1]. Intruders are mainly categorized as External Intrusions/Intruders (i.e., attack by the people who don't belong to the organization) and Internal Intrusions/Intruders (i.e., attack by the people from within the same establishment). However, cyber analytics can be separated on the following bases: i) on the basis of misuse or signatures ii) on the basis of anomalous encryptions iii) on the basis of hybrid nature.

The first form of classification is created to represent attacks following an ordered pattern to spot and prevent a similar attack in the further years along with the detection of famous attacks (though they become hard to use in the case of naïve outbreaks). It is to be pointed out that this method can't be used for the identification of novel (or zero day) catastrophes. The second classification (i.e., based on anomaly) replicates the behavioristic approach by developing an activity profile, hence differentiating the ambiguity from the normal attitude. This method can be used for the detection of novel-attacks and hence are deeply encouraged. Furthermore, it customizes the normal activity routine for every instance, ensuring that the intruders are unable to comprehend which of the activities can be performed incognito. But just like how every coin has two sides, this technique too has its own disadvantage – it is likely for False Alarm Rates (FARs). The last categorization involves the combination of the first two methods – misuse and anomaly detection. They are mainly implemented to raise the rate of detection of common attacks and reduce the False Positive (FP) rate for the minor attacks. IDS's can also be divided based on network or host. An IDS which depends on the network identifies attacks by keeping an eye on the traffic through the network devices. A host-based IDS screens all processes and file activities related to the software with a host.

[*] Corresponding Author: Amit Kumar Tyagi, amitkrtyagi025@gmail.com

https://dx.doi.org/10.25046/aj050310

a) Host-based IDS (HIDS): It mainly focuses on analyzing the internal functioning of a computing system. It might detect activities like which program is trying to access which particular resource and are there any attempts on illegitimate access. For example, a word processor which spontaneously alters the system password database.

b) Network-based IDS (NIDS): It focuses on analyzing and filtering the traffic among network device. It's commonly found that intrusions occur as ambiguous patters. These are mainly caused by the attacks launched by external intruders who wish to access the network to gamble the network and destroy it.

Hence, the article is organized into a number of sections. Section 2 discusses several classifications (like signature and anomaly) with respect to cyber security. Further, section 3 discusses several cyber data sets available for making a comparison and later on the significance of machine and data mining in detection of intrusion detection in cyber security/ applications (in near future) has been discussed in section 4. Further Section 5 discusses "how machine learning and artificial intelligence can be more useful for cyber security professionals for detecting vulnerabilities or preventing attacks". Finally, this work is concluded with some future enhancements in brief in section 6.

2. Cyber Security's Classifications

The three types of Intrusion detection in support of cyber security are [2]-[4]: *Misuse-based or Signature based, Anomaly-based, and Hybrid.* Here, each one can be discussed in detail as:

2.1. Misuse-Based or Signature Based

There are multiple ways to replicate an attack. The attack can be a pattern, or a signature used to identify the deviation. They are bound to detect a majority or most of the common attack techniques. However, they come to be of little use in the case of minor or unidentified attack patterns. These systems try to spot and differentiate on the principle of "bad" behavior. The prime obstacle to overcome is on how to create a signature that combines all the varieties of a consistent attack. A plethora of Machine Learning methodologies have been put into use for the detection of misuse in these systems. These detections prove to be useful to identify the outbreaks on networks by associating the routine activities with that of the expected actions of an intruder.

In [5], the author proposed a framework to identify and classify network activities based on Artificial Neural Network (ANN). The data sources are based on various formats, i.e., limited, incomplete, and nonlinear in nature. They implemented data detection that utilizes the analytical strengths of neural networks. A multi-layer classification prototype using MLP is used to detect the misuse by developing the architecture containing four fully connected layers. The neural architecture consists of 9 nodes as input and 2 output nodes. The data pre-processing were conducted at three different levels includes a) Protocol Identifier (PID) – the rules and regulations pertaining to an event (TCP = 0, UDP = 1, ICMP = 2, and Unknown = 3) b) Source Port c) Destination d) Source Address (IP address corresponding to a source) e) Destination Address (IP address of a destination) f) ICMP type (like echo requests, null, etc.) g) ICMP Code h) Raw Data Length (length of the data packets) i) Raw Data.

The neural network model was trained using a back-propagation algorithm for 10,000 iterations of the selected training data. Out of 9.462 records, 1000 were randomly selected for testing and the remaining was used to train the system. The neural network model required 26.13 hours to complete. The results reveal that on training data the root mean square error is 0.058298 and on Test data root mean square error is 0.069929. Finally, an accuracy of 93% can be considered based on RMS, where each data packet was classified as either a normal or an attack set.

In [6], the authors proposed Online Analytical Processing (OLAP) Mining and Classification based IDS, (OMC-IDS). OMC_IDS handle any intrusion detection data using historical data analysis from heterogeneous sources and summarization them by filtering the data by removing the irrelevant data. Apart, a data cube is constructed and integrate OLAP techniques. They applied association rule mining to extract the interesting patterns and classify each connection as normal or any attack. They proposed association rules to find the correlation between TCP/IP parameters and the types of attack on DARPA 1998 data set. They generated rules and less constraint is retained. After the rules are generated, a C4.5 classifier is applied for new connection records. The experiments were carried out on DARPA19985 dataset. The training data and test data are generated in the first seven weeks and in the next two weeks respectively. The results show that total of detection rates as 99%, 97%, 86% and 74%, respectively. The main drawback of association rule mining is that the generated rules may express correlation, but the approach is promising for attack signature building.

Further in [7], authors proposed an algorithm to use the existing signature data and find the signature of the related attack in less time. They compared their approach with algorithm based on Apriori called Signature Apriori (SA) and found that it takes less processing time. Such algorithms can be used to generate new signatures, i.e., used into misuse detection systems such as Snort. The proposed method finds newly attack signature based on the known signature. Scan Reduction method is also used for the reduction of time consumed for scanning of databases. This method involves the determination of a new attacking signature in an efficient way when compared to the Signature Apriori algorithm. Authors have implemented the data mining approach to complement the signature discovery in IDS based on network [8]. This not only generates signatures for the detection of misuses dependent on transfer protocols, but also for those based on content of traffic. The Signature Apriori (SA) is based on the typical association rules algorithm – Apriori algorithm [8]. The experiments have two parts to it: a) Speed testing of SA algorithm b) Accuracy testing of the signatures being mined. This evolves 70% support and the time consumed is extremely less (one is less than 50111s the other is 330 ms). On the whole, the techniques which are applied to tackle the cyber-attacks have been active predominantly as they emphasize on screening the traffic in the network, identification of anomalies and traffic sequences of cyber-attack. Apart from this, the misuse detection can be enforced for the detection of these outbreaks prior to them actually being a part of the attack. Some authors have spotted the command and control traffic (C2C) in Internet Relay Chat using the technique of machine learning to adhere to the botnet existence, for which TCP level data sets have been put into use. Wireless traffic sniffers were used extensively to gather complete TCP/IP headers from around 18 locations around the campus. This was divided into two major stages: (i) The initial stage involved the distinction between IRC and non-IRC traffic, (ii) after which, there was a distinction between botnet and real IRC traffic. For the initial stage, the comparison of performance is done between J48, naive Bayes, and Bayesian network classifiers to identify IRC and non-IRC traffic damages by attaining an excellent overall classification accuracy. Only the naïve Bayes classifiers were capable of achieving reduced false negative rate.

The naive Bayes classifiers accurately classified 35 out of the 38 botnet IRC (which flows correctly and achievesFalse Negative Rate (FNR) of 7.89%) [9]. In Stage (ii), by applying classification they accurately labelling IRC traffic as botnet and non-botnet were more challenging. In [10], author proposed an adaptive intrusion detection system which is considered as a framework for detecting intrusion detection using Naïve Bayesian network. The DARPA KDD99 dataset with 38 attacks are used to find the new intrusion signature like DoS, r21,u2r and probe.The dataset consists of 9 features in the inference network such as Protocol type, Service, Land, Wrong fragment, Numerous failed login, Logged in, Root shell, Is guest login. In the first stage, a junction tree inference technique is used to identify the normal or attack data with performance detection rate 87.68% on normal and 88.64% on intrusion. In the second stage, the dataset is classified into 4 classes: DoS, Probing, R2L and U2R.The performance determine a detection rate of 88.64% for DoS, 99.15% for Probing, 20.88% for R2L, 6.66% for U2R and 66.51% for other classes.

In [11], authors used reliable signatures generated based on supervised clustering algorithm and updating them in real-time using unsupervised clustering technique. The signature updating is done to change attack methods while retaining the signatures useful information. They used a simple density-based clustering algorithm, called Simple Logfile Clustering Tool (SLCT) to create clusters of regular and anomaly traffic. The study made use of a new user stricture, M, in SLCT which mentions the percentage of fixed attributes to be spotted out of all the attributes that a potential cluster is expected to have. If the value of M equates to 0, it then allows the formation of clusters irrespective of the number of fixed attributes. By equating the value of M to greater values they recapitulate the intruder ones, thus classifying the original data. This is inferred to with the help of parameter M as SLCT attack. Both the clustering techniques are implemented for the detection of normal or attack traffic and for identification of the usual traffic in a supervised manner accordingly. In [11], the author treated anomalous centroid of cluster as a signature. The experiments are carried out using KDD data sets using different attack percentages (0%, 1%, 5%, 10%, 25%, 50%, and 80%) and the author reported impressive results without prior knowledge of any attacks in the KDD datasets. Further, Kruegel at el. [12] installed an intrusion detection signature using clustering algorithms to derive decision tree for intrusion detection. It was a placement with Snort. With the help of a decision tree, we are able to choose the features which highly distinguish the characteristics of the rule set, permitting parallel evaluation for every unique feature. It provides a better performance with respect to Snort. In [12], the author make use of the tcpdump files as the necessary dataset for the ten days of test data when considering the evaluation of 1999 DARPA intrusion detection. On comparing and contrasting the rate of processing of Snort and the decision tree for the above data, it was observed that real performance gain vary drastically depending on the basis of the comprehended traffic. 103% was found to be the maximum speed, while 5% turned out to be the minimum. The decision trees performed better as they result in an average speed of 40.3%. The second task was also conducted with increased number of protocols right from 150 up to 1581. The results proved that the approach of the decision tree works efficiently, especially with respect to large rule sets. This approach notifies that the clustering action based on decision trees will definitely reduce the operating time, thus enhancing the processing speed. Furthermore, it portrays a generic solution to many of the other IDSs like host and network-based, and firewall and packet filters.

Zhang et al. [13] study proposed a complete intrusion detection framework containing a detector used for signature-based attack prediction and a database to identify outlier. All the anomaly patterns identified by the system or user either manually or automatically are stored in the database. Because of the extremely quick nature of its implementation, it’s often used as an online solution. Gharibian et al. [14] has put forth a comparative study with the help of probabilistic and futuristic ML methods and processes for detection of intruders and their malicious acts namely, Naïve Bayes and Gaussian along with those of Decision Tree and Random Forests. A lot of the training data sets which have been constructed from KDD99 are being deployed for effective functioning today and each of the methods have been used for categories of attack like DoS, Probe, R2L and U2R with a proper analytical study of their results. Normalization used in the formation of these datasets, complementing the argument that the features in KDD are not similar to those of the others and they possess high variance scales. The executional capability of Decision Trees (DT) and Random Forests (RF) portray valid results and operations in the identification of DoS. On the contrary, Gaussian and Naïve Bayes results shows much better in few of the varied attack domains like Probe, R2L and U2R. Based on the results, the author stated that the probabilistic techniques are more robustness in nature than predictive techniques for intrusion detection.

Mukkamala et al. [15] considered the performance of ANN, SVM and Multivariate Adaptive Regression Splines (MARS) and proved that ensembles of ANNs, SVMs and MARS is of top priority for individualized perspectives for the detection of these attacks with respect to precision of division. The five class classification experiments were performed on 11,982 records. They applied 3 classification algorithms like SVMs, MARS and ANNs. The ensemble of SVMs, MARS and ANNs approach out performs with accuracies of 99.71% for Normal, 99.85% for Probe or Scan, 99.97% for DoS, 76% for U2R, and 100% for R2L are reported respectively. The accuracy of four classes are 99% using SVM, RP, SCG, OSS algorithms and the accuracy on the U2R class is much less with 76%. In this paper [16] the author used genetic algorithms to generate simple rules for network traffic.

These rules are used to differentiate normal network connections from anomalous connections and these anomalous connections refer to events with probability of intrusions. Abraham et al. [17] applied genetic programming algorithms such as Linear Genetic Programming (LGP), Multi Expression Programming (MEP) and Gene Expression Programming (GEP) in attack classification. In [18], Hansen et al. used GP with homologous crossover for performing intrusion detection. Arnes et al. [19] proposed a novel approach to network risk assessment. The approach considers the risk level of a network as the composition of the risks of individual hosts. It is probabilistic and uses Hidden Markov models (HMMs) to represent the likelihood of transitions between security states. They tightly integrate the risk assessment tool with an existing framework for distributed, large scale intrusion detection, and apply the results of the risk analysis to prioritize the alerts generated by the intrusion detection sensors.

An HMM is denoted by (P, Q, Π). Lee et al. [20] developed a systematic framework using data mining techniques for automated IDS.In [21] the author trained Naïve Bayes classifier on KDD 1999. The data is partitioned into training set and test set and the data was grouped into four attacks (1. probe or scan, 2. DoS, 3.U2R, and 4. R2L). The author stated an accuracy of 96%, 99%, 90% and 90% for the respective attacks. Hu et al. [22] proposed a framework for malicious transactions. An cyber-attack detection model is needed as a prerequisite for fast damage recovery. The framework employed a sequential mining algorithm for finding the dependencies in database and presented as classification rules. The data captured from database logs including (Tname) transaction name, (TID) transaction ID, begin and end time, etc. They applied the framework for identifying U2R attacks as part of cyber security. The result presented 91% of TP (True Positive) rate and 29% of FP (False Positive) rate.

In [23], the author presented an IDS model with high accuracy and efficiency using machine learning algorithms including K-means, Support Vector Machine (SVM). They also employed feature reduction methods to eliminate the unwanted features. Table 1 shows the algorithm, data set, metric used for misuse-based intrusion detection.

2.2. Anomaly and Hybrid Detection

Lippmann et al. [24] proposed an IDS system on transcripts of telnet sessions. The combination of training data and new keywords were used to find the common attacks using neural network model. The system achieves 80% of high detection rate. Palagiri et al. [25] proposed a model for learning the normal traffic patterns from TCP/IP port. They applied preprocessing techniques then perform clustering on normal traffic and final trained using Artificial Neural Network (ANN). The study reported a 100% normal behavior.

Apiletti et al. [26] proposed NETMINE framework which classifies the traffic data using data mining/ machine learning techniques. The framework performs data stream processing, refinement analysis by using general association rule extraction for profile data, anomaly detection, and identifying recurrent patterns.

Intrusion Detection Systems (IDS) mainly intend towards protection of computerized systems and helps in spotting vulnerabilities and other attack exposures. A novice structural outline which has its' roots based on data mining methods have been put forth [27] for the creation of an IDS. This framework proposes Association Based Classification (BC) which is dependent on rules linked to fuzzy logic for the development of classifiers and this helps in categorization of normal and un-normal records. Compatibility threshold is the central parameter in this application. The approximate value for this depends on the ROC curve of the system which is produced by carrying out lots of tests on datasets, with varied threshold values. Therefore, 0.06 becomes the compatibility threshold which is to be dealt with in the detection of anomalous behavior. The FP error produced can be reduced to the level of that of misuse detection situation and there's a huge decrease in the detection rate of existing attacks. In the case of unforeseen intrusions, the ambiguous case outshines the misuse perspective, and this is the key advantage of anomaly-based approaches.

Luo et al. [28] has combined the association rule along with the frequency episodes with that of fuzzy logic to determine the sequence in the data. This produces short and flexible variations for intrusion detection as a lot of quantifying features come into play. To ensure that data instances don't outshine the contribution of that of the others, normalization is carried out before retrieving the fuzzy association rules. The required simulations have been conducted by customized programs and the results have proved the necessity of fuzzy rules and its' frequency occurrences in intrusion detection. Kruege et al. [29] implemented an intrusion detection system for identifying attacks against Operating System (OS), they analyzed OS calls to detect attacks against daemon applications and set uid programs. Also implemented on machines running with Linux or Solaris with individual system calls. A feature vector is represented which captures information specific to each system call such as the system call number, its return code, and its arguments. They applied Bayesian network to classify events during open and executive OS calls.

Table 1: The algorithm, data set, metric used for misuse-based intrusion detection.

Paper Citation	Algorithm Used	Data Set Used	Metric Used
[5]	Artificial Neural Network	RealSecure network monitor (Internet Security Systems)	Accuracy
[6]	OMC-IDS (OLAP and Association rule mining)	DARPA 1998	Accuracy
[7]	Signature Apriori (SA)	Signature based data	Accuracy
[8]	Apriori algorithm	SigSniffer architecture	Accuracy
[9]	J48, Naïve Bayes and Bayesian network	Dartmouth's wireless campus network (TCP level)	Accuracy
[10]	Bayesian network	DARPA KDD	Accuracy
[11]	Density-based clustering algorithm (SLCT)	KDD	Accuracy
[12]	Decision Tree	DARPA	Accuracy
[13]	Random Forest	KDD	Accuracy
[14]	Random Forest (Predictive techniques)	KDD	Accuracy
[15]	ANN, SVM and MARS	DARPA	Accuracy
[16]	Genetic algorithms	DARPA	Accuracy
[17]	Genetic algorithms	DARPA	Accuracy
[18]	Genetic algorithms	KDD	Accuracy
[19]	Hidden Markov Network	KDD	Accuracy
[20]	RIPPER	DARPA	Accuracy
[21]	Naïve Bayes	KDD	Accuracy
[22]	Apriori algorithm	Sequence patterns of log files from database are examined to find database intrusions.	Performance
[23]	Ant Colony Optimization (ACO)	KDD	Accuracy

Table 2: The algorithm, data set, metric used for anomaly and hybrid-based intrusion detection.

Paper Citation	Algorithm Used	Data Set Used	Metric Used
[24]	Artificial Neural Network (ANN)	Transcripts of telnet sessions	Accuracy and False alarm
[25]	Artificial Neural Network	DARPA	-------
[26]	NETMINE framework	Network capture tools are used to capture the network traffic packets and it was developed at Politecnico di Torino	Support

[27]	Fuzzy Association Based Classification (ABC)	KDD	Accuracy and FP rate
[28]	Fuzzy Logic	Tcpdump	Accuracy
[29]	Bayesian network	DARPA	Accuracy and False Alarm Rate (FAR)
[30]	Naïve Bayes algorithm	DARPA	------
[31]	sequence matching algorithms	User command level (shell commands)	Accuracy and False Alarm Rate (FAR)
[32]	EXPOSURE (C4.5 Decision Tree algorithm)	DSN	Accuracy and False Alarm Rate (FAR)
[33]	EXPOSURE (C4.5 Decision Tree algorithm)	Real-World Network	Accuracy and False Alarm Rate (FAR)
[34]	Genetic algorithms	KDD	Accuracy and False Alarm Rate (FAR)
[35]	Genetic Programming	DARPA	ROC (Receiver's Operating Curve) and False Alarm Rate (FAR)
[36]	Hidden Markov Network	KDD	(False Positive) FP rate and (False Negative) FN rate
[37]	RIPPER	DARPA	(False Alarm Rate) FAR
[38]	Bayesian network	KDD	Accuracy and False Alarm Rate (FAR)
[39]	Apriori algorithm	DARPA	Support
[40]	Robust Support Vector Machines	DARPA	Accuracy and False Alarm Rate (FAR)
[41]	Support Vector Machine	NetFlow data (Flame tool)	Accuracy and False Positive (FP) rate
[42]	Self-Organizing Feature Map (SOFM), Genetic Algorithms (GA), and Support Vector Machine (SVM)	DARPA 1999	Accuracy, (False Positive) FP rate and (False Negative) FN rate

The DARPA 1999 data set is used to excite the OS kernel by TCP/IP packets. These features are fed to Bayesian network model and if the output is close to zero it indicates normal or anomaly state.

In [30], the author proposed alert correlation method based on naïve bayes algorithm. 2000 DARPA dataset with their intrusion objective are used to train Bayesian network. In [31], the author proposed a model for differentiating masquerader's users from real users. The study stated a detection rate as high as 80.3% and a false positive rate as low as 15.3%. Table 2 shows the algorithm, data set, metric used for anomaly and hybrid-based intrusion detection. Now, next section will discuss availability of cyber security dataset (in current) globally.

Bilge et al. [32] introduced EXPOSURE, a system that employs large-scale, passive DNS analysis techniques to detect domains that are involved in malicious activity. Bilge et al. [33] presented

DISCLOSURE, a large-scale, wide-area botnet detection system that incorporates a combination of novel techniques to overcome the challenges imposed by the use of NetFlow data. In [34] the author broadly demonstrates how information of the network connection can be replicated as genes and how the parameters in GA can be define in this respect. Lu et al. [35] presented a rule evolution approach based on Genetic Programming (GP) for detecting novel attacks on networks. Joshi et al. [36] classify the TCP network traffic as an attack or normal using HMM and to build an anomaly detection system. Fan et al. [37] proposed an algorithm to generate artificial anomalies to coerce the inductive learner into discovering an accurate boundary between known classes of normal connections and known intrusions, and anomalies. Amor et al. [38] uses a simple form of a Bayesian network that can be considered a Nave Bayes classifier in intrusion detection. Li et al. [39] applied AprioriAll, an algorithm for mining frequent sequential pattern in Data mining field, to discovery multistage attack behavior patterns. Hu et al. [40] presented a new approach, based on Robust Support Vector Machines (RSVMs) for anomaly detection. Wanger et al. [41] proposed an approach for evaluating Netflow records by referring to a method of temporal aggregation applied to Machine Learning techniques. In paper [42], they proposed a new SVM approach, named Enhanced SVM, which combines soft-margin SVM and one class SVM methods

3. Cyber-Security Datasets

Data plays an important role for ML and DM models. Today data is new oil for digital world (or for industries), i.e., based on collecting data, competitors can launch affordable services in market. For example, based on collecting requirements/ demands of particular things in an area, companies can shift towards to sell their product in that specified area/ region. The necessary elements for the efficient conduction of research related to cyber security includes the right choice of data and its' proper utilization. To comprehend the ML and DM algorithms, put forth by a number of authors, requires a better understanding of data sets. We can achieve cyber security of data with the help of different gatherings like Win Dump or Wireshark tool to acquire the network data packets. It can also be done using the current public datasets.

a. **DARPA**: DARPA (Defense Advanced Research Projects Agency) intrusion detection datasets was collected and published by the Cyber Systems and Technology Group MIT/LL (Massachusetts Institute of Technology Lincoln Laboratory. The data was generated using network simulation and compiled based on TCP/IP network data. The datasets can be downloaded from the website and it primarily includes: DARPA 1998, 1999, 2000. DARPA 1998 consists of data collected for 9 weeks, which includes training data (seven weeks) and of test data (two weeks). Similarly, DARPA 1999 consists of data collection for five weeks wherein training data is for three weeks and the last two weeks is test data. DARPA 2000 includes scenario-specific datasets. Table 3 lists the complete basic features of TCP connection.

b. **KDD 1999 cup datasets:** The most popular and widely used datasets for intrusion detection are KDD 1999 datasets created by KDD cup challenge. This dataset is based on DARPA 1998 dataset with 4 million records. The KDD 1999 datasets consist of normal and 22 attacks categorized into 5 main components. Dos (Denial of Service attacks), R2L (Root to Local attacks), Probe (Probing attacks), U2R (User to Root attack) and normal. There exist 41 number of attributes containing features related to basic, content and traffic.

Table 3: List of the Complete Basic Features of TCPconnection

Basic Features	Type	Represented	Description
Duration	Continuous	Integer	Time duration of connection
Protocol, type	Symbolic	Nominal	Type of the protocol (TCP, UDP and ICMP)
Service	Symbolic	Nominal	HTTP, Telnet, FTP, SMTP and others
Flag	Symbolic	Nominal	Connection status
Src bytes	Continuous	Integer	Number of bytes sent per connection
Dst bytes	Continuous	Integer	Number of bytes received per connection
Land	Symbolic	Binary	Value=1 if port numbers and src/ dst IP address are same
Wrong fragment	Continuous	Integer	Total of bad checksum packets
Urgent	Continuous	Integer	Sum of urgent packets

Hence, this section discusses current cyber security datasets in detail. Now next section will discuss a brief introduction of data mining and machine learning and necessary uses in detecting vulnerabilities or intrusion over cyber – network (cyber space).

4. Introduction to Data Mining (DL) and Machine Learning (ML) for Cyber Security

The terms Machine Learning (ML), Data Mining (DM), and Knowledge Discovery in Databases (KDD) are often used interchangeably. As per research, KDD process is represented as whole and deals with extracting valuable, earlier unknown knowledge/information from data. Fayyad et al. [43], has clearly mentioned and explained the process of DM as a specific step in KDD which handles the implementation of algorithms for retrieval of sequences from data. It can hence, be observed that they possess common characteristics between ML and DM. The steps involved in KDD process are as follows: data selection, data cleaning and pre-processing, data transformation, application of DM algorithms, result interpretation/ evaluation. DM is one step among all and used for extracting patterns from data by applying algorithms. It's to be pointed out that there is a plethora of publications [e.g., Cross Industry Standard Process for Data Mining (CRISP-DM) [44] along with industry participants who consider the process DM.

These two terms are commonly discussed together and are applied interchangeably. According to Arthur Samuel Creator of Machine Learning (ML) defined "ML as a field of study that makes the computers to learn by itself without being explicitly programmed". The machine learning algorithms mainly focus on classification and prediction techniques. The ML algorithms learn from the training/ past data and finds the insights for future/unknown conditions. The various classification algorithms in general applied to cyber security are discussed as below.

- Decision Trees

Decision trees are the important and popular techniques used for classification. A decision tree is nothing but a simple flowchart similar to that of the structure of a tree which has every internal node denoting a test with respect to an attribute such that each branch indicates the outcome of the test and each leaf node acquired a class label.ID3 (Iterative Dichotomiser) is a decision tree algorithm which was developed by Ross Quinlan. He then represented the successor of ID3 – C4.5 which has turned out to be a benchmark for comprehending algorithms

- C4.5 Algorithm

This model forms its basis from ID3 algorithm along with additional characteristics to acknowledge the issues faced by that of ID3. It's considered to have a greedy approach and it is said to possess a top-down recursive divide and conquer method.Given a data samples S, C4.5 applies divide and conquer algorithm for tree generation and the process is stated as follows:

a) If S is small or all the data samples in S belong to the same class, then the leaf node is labeled with the most frequent class in S.
b) Or else, the process of selecting attributed is made use of to control the criterion of the splitting process. The criterion for the process of splitting indicates which attribute is to be tested at node S by identifying the most efficient way to distinguish the tuples into separate classes.

The process continues recursively to form a decision tree.

- Naive Bayes Algorithm:

The Naive Bayes algorithm (NB) employs a simplified version of Bayesian learning method. It involves statistical classifiers. The probabilities of membership can be determined with the help of these classifiers and it has its foundation on Bayes theorem. It's assumed that the effect of a feature value of a given class doesn't depend on the values from other features and is called conditional independence. One of the most efficient, robust and best methods to prevent noisy data is by making use of Naïve – Bayes classifiers. The highlight feature being that it calls for only a small amount of training data to approximate the strictures needed for categorization.

- K-Nearest-Neighbor

K-Nearest-Neighbor (k-NN) is a classification which is one of the simplest and fundamental ones, working well even in the presence of little or absolutely no prior knowledge regarding the data distribution and it's based on the process of learning by equivalence. 'm' dimensional numerical attributes are used for describing the training samples with each sample replicating a certain point in the m-dimensional space. Hence, we can see that all the points are stored in an m-dimensional pattern space. In the case of an unknown data sample, a k-nearest neighbor classifier checks out the pattern space for the k training data modules which are quite close to that of an unknown sample. 'Closeness' refers to Euclidean distance. The new and unknown sample is designated with the most common class from it is nearest k neighbors.

- Support Vector Machine

It mainly plots the input vector into a space of very high dimensions and helps in the construction of a hyper plane. The hyper plane has the capacity to separate the data points into different classes. A great level of distinction is obtained by hyper planes which has the greatest distance to the closest training data point of any class which is called as the functional margin. It's observed that with increase in margin, there's a lower generalization error for the classifier. The hyperplane is a decision boundary for the two classes. In reality, the persistence of a decision boundary ensures the detection of a misclassification which is created by a particular method. Classification, regression, and other jobs are implemented with SVM.

- Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

RIPPER, is a generic methodology used for effectively applying separate-and -conquer rule learning. It helps in increasing the precision of protocols by replacing or re-enforcing the individual norms. Reduce Error Pruning was implemented to create the rule and the created rules are often restricted to a smaller number. It ensures the pruning of each rule right after the creation and removal of data samples. Reduced error pruning facilitates the handling of huge training sets, thus improving the precision. The below mentioned steps are carried out: Spot the characters/ features from the training data and identifies the split of all attributes essential for categorization (i.e., feature/dimensionality reduction). Comprehend models using the training data and use the trained model to segregate the unknown data. In the initial stage pf training, each feature with a corresponding class is acquired by using suitable algorithms from the training set. The perspectives of ML/DM are mainly categorized into three classes supervised, unsupervised and semi-supervised. The different machine learning and data mining methods applied for cyber security is mentioned in Table 1-2.

5. Role of Machine Learning and Artificial Intelligence towards Cyber Security

Today cyber security has put everything on risk, due to attracting billions of online users over internet and storage of data over internet (at cloud side). Everyday every country is facing critical attacks by enemy nations on their computer labs, systems or network, which can create a situation of third world war. Till today, we are detecting cyber attackers or hackers through human work-force, for that we require a huge number of skilled workforce to look over or prevent against any cyber threats. But in near future, there is a possibility that intrusion or vulnerabilities detection can be done by using machine learning and artificial intelligence. Also, it will provide several benefits to society and avoid the problem of weaker security, lower efficiency, leaking of personal information by Internet of Things, increasing vulnerabilities on cyber and physical space or cyber physical systems. Note that recently many critical attacks have been measured by several countries on their nuclear programs/ sites [45]-[48] by their enemy nations. On other side, Artificial Intelligence (AI) will reduce required workforce (requirement of cyber security professionals), speed of detection of intrusion, etc. AI can help in living life longer and better through its emerging innovations. Such benefits of AI are listed in following ways.

- Handling huge volumes of security data
- Picking out threat needles in cyber haystacks
- Acceleration of detection and response times
- Keeping up in the Artificial Intelligence arms race
- Breathing space for human cyber security teams.

Hence, data mining, machine learning and artificial intelligence are necessary components for 21stcenturygeneration. So, we will see the tremendous uses of Machine learning, Artificial intelligence in next 20-30 years, which will do many/ everyday task and will serve humanity better and better.

6. Conclusion and Future Enhancements

In the recent/ several decades, several attacks have been measured/ noticed. Due to this reason, cyber security and intrusion detection has been coined in this smart era. Due to enormous internet usage (in the past decade), the vulnerabilities of network security (in a network) need to be overcome. Overcoming such issue has become an important issue today. In general terms, Intrusion detection system is used to identify the flaws in the system such as unauthorized access and unusual attacks over the secured networks. Hence, to solve this issue, several authors had discussed many studies. In that, we found that (from literature, refer section 2 and 3) machine learning can be more useful in solving these issues/such problems using regression, prediction, and classification techniques. In this smart era, we have large amount of data (generated from internet/ web browsing) and shortage of talented employees in cyber-security domain/area. So, Machine Learning is the only solution to provide efficient results in minimum time. Hence, in order to understand importance of ML techniques for solving the IDS problems, which focus on the design of the single, hybrid and ensemble classifier models (with discussing several algorithms, used datasets). This work also discussed "How Machine earning, and data mining can be useful in identifying/ detecting intrusion, in section 4"?

Hence, we found that uses of different classifier/ ML techniques in IDS a promising study in cyber security and artificial intelligence. It will make attraction of young scientists from research communities for a long time. For future work, this work has identified some valid points which are: removal of data redundancy and irrelevant features for the training phase (have important role in system performance), i.e., consideration of best feature selection algorithm will play an important role in the classification techniques in near future. Also, multiple or different selection of algorithms for featured selection will provide best possible solutions in various scenarios/ intrusion detection in a network. Last, but not the least, cyber security and intrusion detection systems works well and shows a better performance with ensemble classification algorithms when compared to single classification algorithms.

Authors' Contributions

Gillala Rekha drafted this manuscript, whereas Shaveta Malik and Meghna Manoj Nair have put this article's content in correct order. In last, Amit Kumar Tyagi has approved this manuscript.

Acknowledgement

This research is funded by the Anumit Academy's Research and Innovation Network (AARIN), India. The author would like to thank AARIN India, an education foundation body and a research network for supporting the project through its financial assistance.

Conflict of interest

The authors declare that they do not have any conflict of interest with respect to publication of this research work.

Scope of the Work

This work has been written through collecting articles from several international journals like ACM, IEEE, Springer, Wiley, etc. This work will be useful for future researchers who are working towards computer vision/ the use of machinelearning or artificial intelligence towards cyber security.

References

[1] S. Mukkamala, A. Sung, A. Abraham, Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools, Vemuri, V. Rao, Enhancing Computer Security with Smart Technology. (Auerbach, 2006) (2005) 125–163.

[2] A. Sundaram, An introduction to intrusion detectionCrossroads 2 (4) (1996) 3–7.

[3] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM computing surveys (CSUR) 41 (3) (2009) 15.

[4] B.-C. Park, Y. J. Won, M.-S. Kim, J. W. Hong, Towards automated application signature generation for traffic identification, in: Network Operations and Management Symposium, 2008. NOMS 2008. IEEE, IEEE, 2008, pp. 160–167.

[5] J. Cannady, Artificial neural networks for misuse detection, in: National information systems security conference, Vol. 26, Baltimore, 1998.

[6] H. Brahmi, I. Brahmi, S. B. Yahia, Omc-ids: at the cross-roads of OLAP mining and intrusion detection, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2012, pp. 13–24.

[7] H. Zhengbing, L. Zhitang, W. Junqi, A Novel Network Intrusion Detection System (NIDS) based on signatures search of data mining, in: Proceedings of the 1stinternational Conference on Forensic Applications and Techniques in Telecommunications, information, and Multimedia and Workshop, ICST, 2008, p. 45.

[8] H. Han, X.-L. Lu, L.-Y. Ren, Using data mining to discover signatures in network-based intrusion detection, in: Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on, Vol. 1, IEEE, 2002, pp. 13–17.

[9] L. Carl, et al., Using machine learning techniques to identify botnet traffic, in: Local Computer Networks, Proceedings 2006 31stIEEE Conference on. IEEE, 2006.

[10] F. Jemili, M. Zaghdoud, M. B. Ahmed, A framework for an adaptive intrusion detection system using bayesian network, in: Intelligence and Security Informatics, 2007 IEEE, IEEE, 2007, pp. 66–70.

[11] G. R. Hendry, S. J. Yang, Intrusion signature creation via clustering anomalies, in: Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2008, Vol. 6973, International Society for Optics and Photonics, 2008, p. 69730C.

[12] C. Kruegel, T. Toth, Using decision trees to improve signature-based intrusion detection, in: International Workshop on Recent Advances in Intrusion Detection, Springer, 2003, pp. 173–191.

[13] J. Zhang, M. Zulkernine, A. Haque, Random-forests-based network intrusion detection systems, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38 (5) (2008) 649–659.

[14] F. Gharibian, A. A. Ghorbani, Comparative study of supervised machine learning techniques for intrusion detection, in: Communication Networks and Services Research, 2007. CNSR'07. Fifth Annual Conference on, IEEE, 2007, pp. 350–358.

[15] S. Mukkamala, A. H. Sung, A. Abraham, Intrusion detection using an ensemble of intelligent paradigms, Journal of network and computer applications 28 (2) (2005) 167–182.

[16] W. Li, Using genetic algorithm for network intrusion detection, Proceedings of the United States Department of Energy Cyber Security Group 1 (2004) 1–8.

[17] A. Abraham, C. Grosan, C. Martin-Vide, Evolutionary design of intrusion detection programs., IJ Network Security 4 (3) (2007) 328–339.

[18] J. V. Hansen, P. B. Lowry, R. D. Meservy, D. M. McDonald, Genetic programming for prevention of cyber-terrorism through dynamic and evolving intrusion detection, Decision Support Systems 43 (4) (2007) 1362–1374.

[19] A. ˚Arnes, F. Valeur, G. Vigna, R. A. Kemmerer, Using hidden markov models to evaluate the risks of intrusions, in: International Workshop on Recent Advances in Intrusion Detection, Springer, 2006, pp. 145–164.

[20] W. Lee, S. J. Stolfo, K. W. Mok, A data mining framework for building intrusion detection models, in: Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on, IEEE, 1999, pp. 120–132.

[21] M. Panda, M. R. Patra, Network intrusion detection using naive bayes, International journal of computer science and network security 7 (12) (2007) 258–263.

[22] Y. Hu, B. Panda, A data mining approach for database intrusion detection, in: Proceedings of the 2004 ACM symposium on Applied computing, ACM, 2004, pp. 711–716.

[23] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, K. Dai, An efficient intrusion detection system based on support vector machines and gradually feature removal method, Expert Systems with Applications 39 (1) (2012) 424–430.

[24] R. P. Lippmann, R. K. Cunningham, Improving intrusion detection performance using keyword selection and neural networks, Computer networks 34 (4) (2000) 597–603.

[25] C. Palagiri, Network-based intrusion detection using neural networks, department of Computer Science Rensselaer Polytechnic Institute Troy, New York (2002) 12180–3590.

[26] D. Apiletti, E. Baralis, T. Cerquitelli, V. DElia, Characterizing network traffic by means of the netmine framework, Computer Networks 53 (6) (2009) 774–789.

[27] A. Tajbakhsh, M. Rahmati, A. Mirzaei, Intrusion detection using fuzzy association rules, Applied Soft Computing 9 (2) (2009) 462–469.

[28] J. Luo, S. M. Bridges, Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection, International Journal of Intelligent Systems 15 (8) (2000) 687–703.

[29] C. Kruegel, D. Mutz, W. Robertson, F. Valeur, Bayesian event classification for intrusion detection, in: Computer Security Applications Conference, 2003. Proceedings. 19th Annual, IEEE, 2003, pp. 14–23.

[30] S. Benferhat, T. Kenaza, A. Mokhtari, A naive bayes approach for detecting coordinated attacks, in: Computer Software and Applications, 2008. COMPSAC'08. 32nd Annual IEEE International, IEEE, 2008, pp. 704–709.

[31] K. Sequeira, M. Zaki, Admit: anomaly-based data mining for intrusions, in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2002, pp. 386–395.

[32] L. Bilge, E. Kirda, C. Kruegel, M. Balduzzi, Exposure: Finding malicious domains using passive dnsanalysis., in: Ndss, 2011.

[33] L. Bilge, D. Balzarotti, W. Robertson, E. Kirda, C. Kruegel, Disclosure: detecting botnet command and control servers through large-scale netflow analysis, in: Proceedings of the 28thAnnual Computer Security Applications Conference, ACM, 2012, pp. 129–138.

[34] M. S. A. Khan, Rule based network intrusion detection using genetic algorithm, International Journal of Computer Applications 18 (8) (2011) 26–29.

[35] W. Lu, I. Traore, Detecting new forms of network intrusion using genetic programming, Computational intelligence 20 (3) (2004) 475–494.

[36] S. S. Joshi, V. V. Phoha, Investigating hidden markov models capabilities in anomaly detection, in: Proceedings of the 43rd annual Southeast regional conference-Volume 1, ACM, 2005, pp. 98–103.

[37] W. Fan, M. Miller, S. Stolfo, W. Lee, P. Chan, Using artificial anomalies to detect unknown and known network intrusions, Knowledge and Information Systems 6 (5) (2004) 507–527.

[38] N. B. Amor, S. Benferhat, Z. Elouedi, Naive bayesvs decision trees in intrusion detection systems, in: Proceedings of the 2004 ACM symposium on Applied computing, ACM, 2004, pp. 420–424.

[39] Z. Li, A. Zhang, J. Lei, L. Wang, Real-time correlation of network security alerts, in: e-Business Engineering, 2007. ICEBE 2007. IEEE International Conference on, IEEE, 2007, pp. 73–80.

[40] W. Hu, Y. Liao, V. R. Vemuri, Robust support vector machines for anomaly detection in computer security., in: ICMLA, 2003, pp. 168–174.

[41] C. Wagner, J. Francois, T. Engel, et al., Machine learning approach for ip-flow record anomaly detection, in: International Conference on Research in Networking, Springer, 2011, pp. 28–39.

[42] T. Shon, J. Moon, A hybrid machine learning approach to network anomaly detection, Information Sciences 177 (18) (2007) 3799–3821.

[43] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM 39 (11) (1996) 27–34.

[44] C. Shearer, The crisp-dm model: the new blueprint for data mining, Journal of data warehousing 5 (4) (2000) 13–22.

[45] Tyagi, Amit Kumar, Building a Smart and Sustainable Environment using Internet of Things (February 22, 2019). Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur - India, February 26-28, 2019. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3356500

[46] Tyagi. Amit Kumar, Cyber Physical Systems (CPSs)- Opportunities and challenges for improving cyber security, International Journal of Computer Applications, 2016,137 (14).

[47] Sravanthi Reddy, M. Shamila, Amit Kumar Tyagi, Cyber Physical Systems: The Role of Machine Learning and Cyber Security in Present and Future, Computer Reviews Journal, PURKH, Vol. 5 (2019).

[48] Meghna Manoj Nair, Amit KumarTyagi, RichaGoyal, Medical Cyber Physical Systems and Its Issues, Procedia Computer Science Volume 165, 2019, Pages 647-65.