A Survey on The Accuracy of Machine Learning Techniques for Intrusion and Anomaly Detection on Public Data Sets.

Rizal Tjut Adek
Department of Informatics,
Universitas Malikussaleh
Bukit Indah, Aceh Utara,Indonesia
Rizal@unimal.ac.id

Munirul Ula
Department of Information System,
Universitas Malikussaleh
Bukit Indah, Aceh Utara,Indonesia
Munirulula@unimal.ac.id

Abstract - Machine learning (ML) is growing popularity due to their ability to solve the problem in many areas. In digital world including information security, some intrusion detection systems (IDS) are being upgraded with Machine Learning elements for improving the performance of the system. It is known that is very limited real data set available for information security (IS) research. Therefore, many IS researches relies on the public data set. However public data set have many limitations. The aim of this paper is to analyze the accuracy and performance of the Machine Learning in intrusion detection system and to highlight some recommendation for future research. This study involves an academic papers systematic literature review on intrusion detection related to the application of machine learning methods using public data set. This paper elaborates the used of Machine Learning algorithms in intrusion detection system, highlighting the accuracy and the limitations of the methods for detecting attackers. The goal of this research is to provide an academic base for future research in the adoption of machine learning methods for IDS.

Keywords: machine learning, intrusion detection, anomaly activity detection, information security, publics data set

I. INTRODUCTION

In modern world, with the growing threat of information security, researches are focusing on machine learning and its ability to identify, stop and respond to sophisticated cyber-attacks [1]. Machine learning can be leveraged in various domains of Information security to provide analytical based approaches for attack detection and response. Security officer may benefit from detection and analysis tools based on machine learning methods. However, the accuracy of these methods are still in question mark[2].

The aim of this paper is to address security officers regarding the accuracy (performance) and the limitations of the Machine Learning in detecting attacker and to highlight some recommendation and solution. This study is based on an extensive systematics review of the literatures on academic papers in information security field by considering one specific application typically oriented to Artificial Intelligence (AI) and machine learning methods using public data set.

This systematical review paper is intended for researchers who demand to begin research in machine learning for IDS. To get comprehensive systematic review results, this paper conducted a search query on the Google search engine. The paper used for further analysis are the result of the first 50 google page regardless of the year of publication, as well as the type of paper. A search was also conducted on two popular online literature databases in the field of computer science, namely IEEE Explore and the Association for Computing Machinery (ACM) digital libraries from 2017 to 2020.

The systematic literature review conducts the process of identifying, evaluating and interpreting the results. This process started by searching papers published in the library database of journals with specific keywords for search, starting on January 20, 2019 until May 30, 2020. These keywords are "(machine AND learning AND (model OR algorithm)) AND information security OR intrusion detection".

From the initial search results, 3567 are obtained, then the first inclusion filtering was conducted, namely the criteria I1. I1filter; the select papers must be in the form of journals, magazines. As the results, there are 518 paper titles pass the filter. Then filtering I2 is conducted by quick reading of abstracts which requires describing methods or models of algorithms or applied examples of applications. From this screening produced 123 papers that were determined to pass to the next filtering. Then I3 papers were screened which had to be full in English, and then produced 110 papers, then continued filtering I4; papers must be fully accessible without limitation, which resulted in 79 papers. The last filtering of E2 is done by reading the paper comprehensively. In papers that are not directly related to

978-1-7281-9792-0/20/$31.00 ©2020 IEEE

intrusion detection and using public data set will be ignored here, and in this final process produced 23 papers.

This study is started by presenting an original taxonomy of machine learning approaches. Then, the Machine Learning algorithms applied to intrusion detection, Highlighting the accuracy and the limitations of the methods for detecting attackers. The goal of this research is to evaluate the maturity of these machine learning methods and to identify its limitations in the adoption of these methods for IDS. The conclusions from the analysis of the schematics review of literatures will provide recommendations for the application for machine learning methods in IDS.

This paper is structured as follows; In section 2, the research method used in this study will be discussed. This study uses a systematic literature review (SLR) approach in summarizing the result. The SLR is a most well-known method for gaining better understanding about the available literatures related to machine learning research in the area of information security. SLR is a secondary study for collecting and analysing previous research related to primary research. SLR is expected to help in finding solutions in previous research. The knowledge can be used to develop further research. To get comprehensive results, the publications search was carried out on the Google search engine by applying certain query strings without limitation of publication year. In addition, a search was also carried out on the two largest research databases in the field of computer science namely IEEE Xplore and ACM by limiting paper year of publication from 2017 to 2020. In the part III. Survey results will be displayed relevant results from the study of literatures analysing in this section. In analysis section, this paper will make an analysis of the survey results and the recommendations. The last section, Conclusions, this paper will provide an opinion of the conclusions from the overall results of this study and research suggestions that can be done in the future.

II. Public Data Sets

It is known that is very limited real data set available for information security (IS) research. Therefore, many IS researches relies on the public data set. The most common public data set used in IS researches are DARPA 1998 and KDD1999 [3], [4]. The DARPA 1998 set belong to the Lincoln Laboratory (MIT). A simulation network was built and data were compiled based on TCP/IP network data, Solaris Basic Security Module log data, and Solaris file system dumps for user and root. Effectively, the assembled data set was composed of network and operating system (OS) data. The data were collected for 9 weeks, with the first 7 assigned as the training set and last 2 assigned as the testing set. Attack simulations were organized during the training and testing weeks.

Similarly, the DARPA 1999 data set was collected for a total of 5 weeks, with the first 3 assigned as the training set and the last 2 assigned as the testing set. This data set had substantially more attack types than the DARPA 1998 data set. In both collections, the data sets were processed and curated to be used in the experiments.

Other famous data set is the KDD 1999 [5] created for the KDD Cup challenge in 1999. The KDD 1999 look like NetFlow data with simulated attack. The DARPA 1998 has Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L), and Probe or Scan. DARPA 1999 added a new attack type one where the attacker attempts to exfiltrate special files that have to remain on the victim computer.

The KDD 1999 data set has 4 million records of normal and attack traffic [6] and known to have some serious limitations such as creating the network and attack data because of privacy concerns, an unknown number of dropped packets caused by traffic overflow, and vague attack definitions. Using KDD199 as data set, will causing bias, because it has 78% redundant records in its training data, and 75% in test data. Therefore, we can conclude that by using this data set as the training and testing data, will lead to unrealistic accuracies result.

The type of attacks in Public Data Set are including [5];

- Denial of Service Attacks (DoS) - DoS attack makes a computer too busy or too full to serve real networking requests.
- Remote to User Attacks (R2L) - R2L attack is in which an attacker sends packets to a machine that he has no access right to expose the vulnerabilities of the devices.
- User to Root Attacks (U2R) is an exploitation attacks begins with a normal user account on the system and tries to abuse system vulnerabilities to obtain super user privileges.
- Probing attacks is by scanning the system to identify vulnerabilities that can be exploited later in order to compromise the system.

III. Previous Research

Until recently, there are many literature reporting research works related to the deployment of machine learning methods for intrusion and anomaly detection systems. The previous related studies provided in Table 1 are very recent research related to our research work.

A previous research conducted by Zarpelao et al. [8] developing a comparative study for the method and algorithms used in intrusion detection in digital devices. They classified the detection method, IDS placement technique, and security issues. Other research by Milenkoski et al. [9] elaborated the most common practices for detecting anomaly and intruder. They analyse the existing approaches related to each of the evaluation standard parameters, namely, workloads, metrics, and technique. Other research studies [10], [15], [16], [17] focus on the application of machine learning methods in intrusion and anomaly detection. However, these works do not take account the datasets used in their researches.

TABLE I. PREVIOUS STUDIES ON INFORMATION SECURITY INTRUSION detection.

Author	Method	Public Data sets	Year
Folino et al. [7]	ML	Yes	2016
Zarpelao et al. [8]	ML	No	2017
Aburomman and Reaz [9]	ML	Yes	2017
Xin et al. [10]	ML	Yes	2018
Ring et al. [11]	Other	Yes	2019
Loukas et al. [12]	Other	Yes	2019
da Costa et al. [13]	Other	Yes	2019
Chaabouni et al. [14]	ML	Yes	2019
Berman et al. [15]	ML	Yes	2019
Mahdavifar et al. [16]	ML	Yes	2019
Sultana et al. [17]	ML	No	2019
Our Study	ML	Yes	/

VI. SURVEY RESULTS

In this literature review survey, there are some machine learning method have been widely used in information security application.

A. *Hidden Markov Model (HMM)*

HMM model has been used widely in information security areas [18] . As the name suggests the Markov model has a hidden chain that cannot be directly observed, Figure 1. shows a dotted line as a curtain that prevents it from being directly monitored.

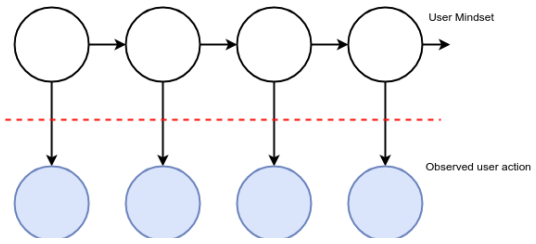

FIGURE 1. HIDDEN MARKOV MODEL

By using HMM, there are several advantages, one of them is because probabilistic characteristics of HMM, so we can obtain useful information in the process. Applications of HMM are widely used in computer science, and for example in information security have been applied for detecting intruder [18] [19] . At the beginning of its emergence, anomaly can be recognized well through a distinctive pattern in the intruder activities, then in its journey the intruder is morphed to avoid detection from protection system [20] . In addition, HMM is also used in an intrusion detection system (IDS) [21] [22] .

B. *Support Vector Machines (SVM)*

SVM is one of the popular algorithms in machine learning. With a simple concept, but because of its characteristics, it is very difficult to be understood internally. From a comprehensive point of view, SVM has the characteristic of trying to provide maximum class distances to data using the term hyperplane, by maximizing distance between classes, as illustrated in Figure 2. Kernel tricks are applied to SVM to increase distance between classes without giving high computational burden [24] [24] .

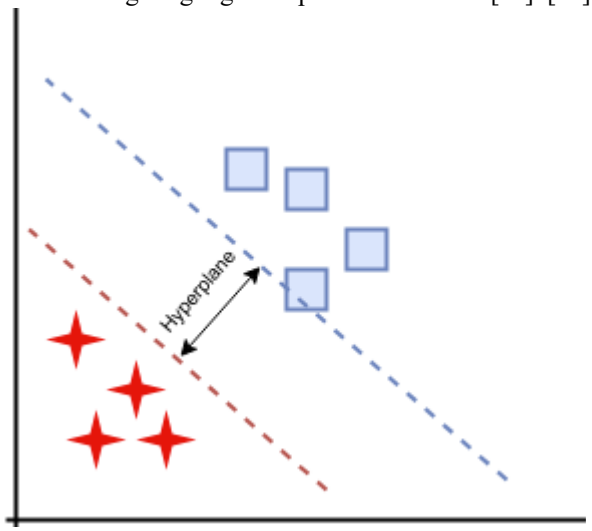

FIGURE 2. ILLUSTRATION OF SVM MODELS BY MAXIMIZING DISTANCE

SVM has become one of the best models for detecting intruder in cyber space [25] [26]. The SVM algorithms can detect the activities of the intruder activities [27] [28] [29] [30]. SVM also can be applied in analysing spam images in emails [31] [32] and also analysing spam based on text [33], analysis of attacks on network [34].

C. *Clustering*

In conducting data grouping, sometimes the data to be grouped does not yet have a special label. However, the researchers want that the machine can group the data based on the similarity characteristics. In the process the clustering technique explores data which seeks to gain insight into the anomalous data values. At the beginning of clustering research, there are some older papers that have discussed it [35] [36] . K-Means is one of the popular clustering methods. K-means principle is to do an iterative process to calculate the mean value of each edge, then group them according to the closest similar value. K value define the grouping. The clustering model has been extensively researched and applied, specifically focusing on the information security field, for example in the detection, analysis intruder. In malware clustering experiments have been carried out by several studies [37] [38] [39] . Various types of clustering have also been applied to various other information security issues including spam detection [40] dan, and network attacks [41] , and have been proven to detect intruders in the

network [40] , botnet detection in network activity [28] , as well as various privacy violation issues [29] as well as many other applications.

Other clustering algorithms such as DBSCAN carry out clustering by focusing to the connectivity of densities [29]. DBSCAN is an algorithm with the aim of separating high-density samples from low density samples. This algorithm is superior to its ability to detect outliers / noise but the clusters that are formed depend on the input values provided. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering method that builds areas based on connected densities. DBSCAN is a type of partition clustering where areas of high density are considered clusters while those whose density is low or are not joined to a cluster are considered noise [29].

In this algorithm several terms are known as follows [30]:
1. Core: The centre point in the cluster is based on the density where there are a number of points that must be within Eps (radius or threshold value), minimum points in the cluster user-defined.
2. Border: The point that becomes the boundary in the region of the central point (core).
3. Noise: A point that cannot be reached by the core and is not a border.
4. Density is reached directly: A point is said to be directly reached if the point is connected directly with the centre point (core).
5. Affordable density: A point is said to be an affordable point if the point is connected indirectly to the centre point (core)

The advantages of this algorithm include [30]: it does not require to know the number of clusters; it can find arbitrary shaped clusters and it is able to overcome the noise. DBSCAN generally starts with a random starting point. Then find all the surrounding points within the EPS starting point distance. If the number of points around is greater than or equal to minimum points then a cluster is formed. If the number of points is less than minimum points, it is marked as noise.

D. Bayesian Network

The Bayesian Network (BN) method is one of the Probabilistic Graphical Models (PGM) which is built from probabilistic theory and graph theory. BN is a Directed Acrylic Graph (DAG) and is equipped with a Conditional Probability distribution Table (CPT) for each node. Each node represents a domain variable and each arrow between nodes represents a probabilistic [36]. In general, BN can be used to calculate the probability of a node by assigning values to other related nodes. The steps to build a Bayesian Network to detect anomaly is based on classifications made to data. First ± the input data is processed and entered into the pre-processing process, namely case folding and tokenizing. Case folding process is the process of changing all the letters in a document / sentence into lowercase letters. Next to enter the tokenizing process, namely the process of breaking down sentences into single words is done by scanning sentences using white space separators such as spaces, tabs, and newlines. After the case folding and tokenizing process is carried out, then enter the process of calculating the probability value using the Bayesian network method. The last process is comparing the probability value of the data included in the category of normal or anomaly [37].

E. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) is one method in the science of Artificial Intelligence in the form of a mathematical model to mimic the workings of the human brain. ANN is a non-linear algorithm that has been widely used to solve pattern recognition problems [39]. In recognizing a pattern, ANN can make generalizations so that it can recognize patterns that have either been trained or that have never been trained. The use of ANN techniques in intrusion detection is a promising research because it is an efficient way to improve the performance of IDS based on misuse detection and anomaly detection [40]. For IDS based on misuse detection, ANN can be used to generalize several signatures, while for IDS based on anomaly detection, ANN can be used to increase the level of pattern recognition of an attack to reduce false-alarms.

Artificial Neural Networks (ANN) is a method of how computers can learn and recognize something [39]. This is a representation of biological neural networks in human brain. In biological neural networks, there is a very wide network, which consists of interconnected neurons. There are three important components that each neuron has, namely dendrites, soma, and axons. Dendrites and axons are responsible for conveying information in the form of electrical impulses from one neuron to another. Soma, or cell body, will add up information delivered by dendrites through a sypnatic gap. The more information added by Soma, the greater the information from a neuron, which means that the intelligence of people increases. In ANN, the communication process between nodes, also takes place every time there is an impulse from one node to another node. Each node is connected to another node through a layer. To implement the addition of information carried out by SOMA on biological neural networks, at ANN, each layer has a certain weight, which will also always be added together. These weights will be used by the nodes to solve a problem. In biological tissues, these weights can be analogous to cations in the chemical processes that occur in the syptatic gap [40].

F. Convolutional Neural Network (CNN)

One of the new machine learning methods is the Convolutional Neural Network (CNN). CNNs were first developed for computer vision application [41]. CNN contains three layers which are Convolution layer, Pooling layer, and Fully connected layer. CNNs use on 2-dimensional data in the analysis, therefore, the input data should be in be in matrices form. The convolutional layer is used for extracting local features in the matrices data; the pooling layer is responsible for dimension reduction to enhance the feature generalizability and the connected layer is similar to the traditional neural network portion and is used to output the desired result as shown in Figure 3 [42].

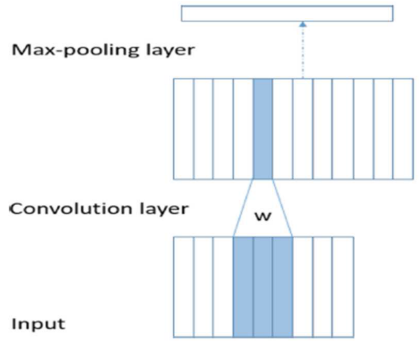

FIGURE. 3. THE CNN PRINCIPLE

G. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is a machine learning method designed for analyzing sequential data. The RNNs have been widely used in natural language processing. RNN has connections between nodes form a directed graph along a temporal sequence [43]. RNNs also could use its internal state memory to process sequences of inputs. This characteristic allows RNN to show temporal dynamic behavior of the data. This feature also makes RNN different from feedforward neural networks. The structure of an RNN is shown in Figure 4.

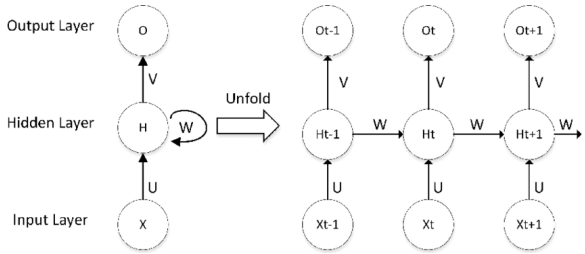

Figure 4. The structure of an RNN

V. THE ACCURACY OF THE MACHINE LEARNING METHODS

A. Artificial Neural Networks (ANN)

For detecting anomaly, Lippmann and Cunningham [39] proposed a system that uses keyword selection and ANN method. The method achieves 80% detection with roughly 1 false alarm per day. This false alarm rate represents a two orders of magnitude improvement from the baseline system with the same detection accuracy.

Other work by Bivens et al. [40] deploy a complete IDS that use a pre-processing stage, clustering the normal traffic, normalization, an ANN training stage, and an ANN decision stage. Bivens et al. [28] reported successfully predicting 100% of the normal behaviour. Their overall method is promising, even though there are some attacks were not detected and the FAR for some attacks reached 76%.

B. Bayesian Network

Jemili et al. [44] developed an IDS framework using Bayesian network classifiers. Their method used nine features of the KDD 1999 data in the inference network. In the detecting intruder and anomaly activity stage, the normal or attack decision have 88% accuracy on the normal and 89% accuracy on attack categories. They also reported the accuracy for denial of service (DoS) is 89%, unauthorized access from a remote machine (R2L) is 99%, the unauthorized access to local super-user privileges (U2R) is 21%, probing attack is 7%, and 66% for other classes of attack. The study suggests the low performance of the R2L and U2R categories is because the number of training instances is much lower than for the other categories.

Other research by Kruegel et al. [45] used other public data set, the DARPA 1999, for stimulate the OS kernel by TCP/IP packets. Because the detection threshold is used to control the False Alarm Rate (FAR), the system is flexible and can make self-adjustments against too many false alarms; 75% accuracy, 0.2% False Alarm Rate (FAR) and 100% accuracy, and 0.1% false alarm rate (FAR) are achieved by using different threshold values.

Benferhat et al. [46] used Bayesian network to detect denial of service (DoS) intrusion detector. The study setup has two different scenarios extracted from the DARPA 2000 data set; but unfortunately, the research does not report any numerical results.

C. Clustering

In their study, Blowers and Williams [44] use a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering method to group normal versus anomalous network packets. The KDD data set is pre-processed to select features using a correlation analysis. A 10% attack to no-attack ratio is set during pre-processing of the data. The reported performance is 98% for attack or no-attack detection.

D. Decision Trees

Kruegel and Toth [45] research work try replaced Snort detection engine with decision trees method. They used 10 days DARPA 1999 Tcpdump test data to evaluate the intrusion and anomaly detection. The summarise the result by comparing Snort performance and the decision-tree method. This research experiment also conducted by increasing the number of rules in Snort 2.0 from 150 to 1581. As the result, the performance of the Decision trees method depending on the type of traffic; the maximum speed-up was 105%, the average 40.3%, and the minimum 5%.

E. Hidden Markov Models

Ariu et al. [46] conducted a research for detecting the attacks on XSS and SQL-Injection web applications. HMMs are used to extract attack signatures. The study reported 50% of the discovered vulnerabilities in 2009 affected web applications. In the experiment section, Ariu's study also used the DARPA 1999 data set as well as some other HTTP data sets. In most of the experiments, the mean Area Under the ROC Curve (AUC) of 0.915 to 0.976 is achieved, for an FP rate ranging from 10E-4 to

10E-1. For Fault Positive rates higher than 10E-3. For smaller Faults Positive rates, the percentage of detected attacks decreases but still remains higher than 70% for a FP rate of 10E-4.

Joshi and Phoha [50] also used HMMs algorithms for detecting intruders. Their research applied an HMM method with five stages and six observation symbols per stage. The KDD 1999 data set was used, and 5 out of 41 features were chosen for modelling. The accuracy was 79%; the remaining 21% is refer to as a Faults Positive rate (i.e., classifying anomaly as normal) and an FN rate (i.e., classifying normal as an attack). The researchers suggested to increase the number of features used for significantly improve the accuracy.

F. Naïve Bayes

Panda and Patra [48] used the Naïve Bayes classifier and used the public data set, KDD 1999 for training and testing. The data were grouped into four attack types namely denial of service (DoS), unauthorized access from a remote machine (R2L), unauthorized access to local super-user privileges (U2R), and probing. The accuracy achieved are 96%, 99%, 90%, and 90%, respectively, on these categories with 3% cumulative False Alarm Rate (FAR).

Amor et al. [52] develop a simple form of Naïve Bayes classifier and utilizing the KDD 1999 data set. The data group in to three categories. They reported the accuracy as 97% for Normal, 96% for DOS attack, 9% for local remote machine attack, 12% for user privilege attack, and 88% for probing and scan attack. The paper does not have information about false alarm in the research, however, by 97% normal is reported, the False Alarm Rate (FAR) can be expected to be less than 3%. The accuracy for detecting intruder and anomaly activity experiment is reported as 98% and 89% for the Normal and Abnormal categories, respectively.

G. Support Vector Machine

In the work by Li et al. [53], an SVM classifier with an Radial Basis Function (RBF) kernel was used to classify the KDD 1999 data set into predefined categories (denial of service (DoS), unauthorized access from a remote machine (R2L), unauthorized access to local super-user privileges (U2R), and Normal). The study reported performance as overall 98% accuracy with unknown variance. The lowest accuracy of 53% was for the U2R category.

Amiri et al. [54] research used a least-squared SVM to have a faster system to train on large KDD public data sets. Only 19 features in the KDD data set from 41 were used. To predict the attack type, five classifiers were built for each category. In this manner, a cost is associated with each category and the final classification was determined. The accuracy reported are 99% on the DoS, Probe or Scan, R2L, and Normal classes and as 93% on the U2R class with 99% confidence interval.

Hu et al. [55] used the robust support vector machine (RSVM), with DARPA 1998 public data set to pre-process training and testing data. The work reported a good classification accuracy in the presence of noise (such as some mislabelling of the training data set) and reported 75% accuracy with no false alarms and 100% accuracy with a 3% False Alarm Rate (FAR).

Shon and Moon [56] study used the DARPA 1999 data set for intrusion detection. In the experiment, they used the subset of DARPA 1999 that consisted of 1% to 1.5% attacks and 98.5% to 99% normal traffic to have more realistic, the real world like data. The paper reported the accuracy of the Enhanced SVMs is 87.74%, with 10.20% Fault Positive rate, and a 27.27% Fault Negative rate. Those results were considerably better than those from the one-class SVMs, but not as good as soft-margin SVMs. However, the Enhanced SVM can detect novel attack patterns, whereas a soft-margin SVM cannot.

H. Convolutional Neural Network (CNN)

Safaa et al. [43] proposed a CNN-based intrusion detection method. They conducted experiments on the KDD+ datasets. They constructed CNN and Inc-CNN to classify the attacks. The proposed CNN performed good, reaching accuracies of 86% for CNN and 89% for Inc-CNN.

I. Recurrent Neural Network (RNN)

RNNs have been used in detecting intrusions, all used the public data sets, KDD-1999. Kim et al. [57] used KDD-1999 with additional data. Moreover, Yin et al. [58] achieved 83.28% accuracy on the test data, and 68.55% on a harder subset of the test data. Research work by Safaa et al. [43] show that RNN method achieved 84,33% accuracy for Bi-LSTM and 78,98% for GRU.

VI. RECOMMENDATIONS

One of the most important factors related to the accuracy of IDSs is the type and level of the input data. As previously discussed, several studies used DARPA or KDD data sets because they are easy to obtain and contain network-level data (either tcpdump or NetFlow) as well as OS-level data (e.g., network logs, security logs, kernel system calls). The biggest gap seen is the availability of the labelled data, and definitely a worthwhile investment would be to collect data and label some of it. Using this new data set, significant advances could be made to Machine Learning techniques in information security and breakthroughs could be possible. Otherwise, the best possible available data set right now is the KDD 1999 corrected data set. (However, being 15 years old, this data set does not have examples of all the new attacks that have occurred in the last 15 years.)

The second factor related to the accuracy of the IDSs is the type of Machine Learning algorithms employed and the overall system design. For detecting intruder and anomaly activity, a clustering method, Table 2 summarize the accuracy of the machine learning methods using public data sets. From the table, ANN has 100% overall accuracy, however it contains 76% Fault Positive, Therefore, ANN is not a reliable method for IDS. CNN

also have a good accuracy (85%), it can be an option for future research.

TABLE 2: THE ACCURACY OF THE MACHINE LEARNING METHODS ON PUBLICS DATA SETS

Methods	Data Set	Accuracy					
		Overall	DOS	R2L	U2L	Probing	FAR
ANN	DARPA 1999	100%	No Data	No Data	No Data	No Data	76%
BN	KDD 1999	No Data	89%	99%	21%	7%	No Data
	DARPA 1999	75-100%	No Data	No Data	No Data	No Data	0.1-0.2%
Clustering	KDD 1999	98%	No Data	No Data	No Data	No Data	No Data
HMM	DARPA 1999	70%	No Data	No Data	No Data	No Data	No Data
	KDD 1999	79%	No Data	No Data	No Data	No Data	21%
Naïve Bayes	KDD 1999	No Data	96%,	99%,	90%,	90%,	3%
	KDD 1999	No Data	96%	9%	12%	88%	3%
SVM	KDD 1999	98%	No Data	No Data	52%	No Data	No Data
	KDD 1999	99%	99%	99%	93%	99%	No Data
	DARPA 1998	100%	No Data	No Data	No Data	No Data	3%
CNN	KDD+	86%	84%	21%	21%	71%	
Inc-CNN	KDD+	89%	70%	19%	22%	61%	
Bi-LSTM	KDD+	84%	72%	29%	24%	64%	
GRU	KDD+	79%	81%	36%	10%	56%	

Naïve Bayes also shows high accuracy for one research but it shows not promising result in other research work. One-class SVMs also have high accuracy and performance in detecting anomaly in traffic data. SVM can be one of the best choices of machine learning method for detecting anomaly and intruder in information system. With SVM, much can be learned by extracting association rules or sequential patterns from available normal traffic data. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) are the very accurate, easy to implement, less parameter or distribution dependent, and have high processing speeds.

VII. CONCLUSIONS

This research work elaborates the literature review of Machine Learning techniques used for information security application especially focus on the accuracy of the machine learning methods in intrusion and anomaly detection in public data sets. The question rises as follow: Which Machine learning methods have high accuracy for detecting anomaly activities? Among the machine learning algorithms, for anomaly detectors, one-class SVMs perform well and should be an option in future research.

The data availability is one of the most critical aspect of Machine Learning application for security intrusion detection. Its need a proper data for training and detecting, machine learning techniques cannot work without representative data, and it is difficult and time consuming to obtain such data sets. The major problems found in some research work are related to data set. Most of the application of the machine learning method used public data sets namely DARPA 1998, and KDD 1999. The best possible available data set right now is the KDD 1999 corrected data set. However, being 15 years old, this data set does not have examples of all the new attacks that have occurred in the last 15 years. Moreover, those public data are simulation data, not a real event data. It is very difficult to have a real data from a network or information system. It is highly needed to have a new data set. So that, several promising Machine Learning method could be used to develop models and compared, narrowing the list of Machine Learning effective for information security applications. Significant advances could be made to Machine Learning techniques in information security using this data set and breakthroughs could be possible.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, 2015.

[2] A. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, 2015.

[3] R. Lippmann, J. Haines, D. Fried, J. Korba, and K. Das, "The 1999 DARPA offline intrusion detection evaluation," Computer Network, vol. 34, pp. 579–595, 2000.

[4] R. Lippmann et al., "Evaluating intrusion detection systems: The 1998 DARPA offline intrusion detection evaluation," in Proc. IEEE DARPA Information Surviv. Conference Expo., 2000, pp. 12–26.

[5] S. JOURNAL Stolfo, KDD Cup 1999 Data Set, University of California Irvine, KDD repository [Online]. Available: http://kdd.ics.uci.edu, accessed on Jun. 2019.

[6] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A detailed analysis of the KDD Cup 1999 data set," in Proc. 2nd IEEE Symp. Computer Intelligent Security Defense Application, 2009, pp. 1–6.

[7] G. Folino, P. Sabatino. "Ensemble based collaborative and distributed intrusion detection systems: a survey". Journal Network Computer Application, 66 (2016), pp. 1-16

[8] B.B. Zarpelao, R.S. Miani, C.T. Kawakani, S.C. de Alvarenga. "A survey of intrusion detection in internet of things". Journal Network Computer Application, 84 (2017), pp. 25-37

[9] A.A. Aburomman, M.B.I. Reaz. "A survey of intrusion detection systems based on ensemble and hybrid classifiers Computer Security", 65 (2017), pp. 135-152

[10] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, et al. "Machine learning and deep learning methods for cybersecurity". IEEE Access, 6 (2018), pp. 35365-35381

[11] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, A. Hotho. "A survey of network-based intrusion detection data sets Computer Security". (2019)

[12] G. Loukas, E. Karapistoli, E. Panaousis, P. Sarigiannidis, A. Bezemskij, T. Vuong. "A taxonomy and survey of cyber-physical intrusion detection approaches for vehicles Ad Hoc Network"., 84 (2019), pp. 124-147

[13] K.A. da Costa, JOURNALP. Papa, C.O. Lisboa, R. Munoz, V.H.C. de Albuquerque. "Internet of things: a survey on machine learning-based intrusion detection approaches". Computer Network, 151 (2019), pp. 147-157

[14] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, P. Faruki, "Network intrusion detection for IoT security based on learning techniques". IEEE Commun. Survey. Tut. (2019)

[15] D.S. Berman, A.L. Buczak, JOURNALS. Chavis, C.L. Corbett, A survey of deep learning methods for cyber security Information, 10 (4) (2019), p. 122

[16] S. Mahdavifar, A.A. Ghorbani. "Application of deep learning to cybersecurity: a survey Neurocomputing" (2019)
[17] N. Sultana, N. Chilamkurti, W. Peng, R. Alhadad. "Survey on SDN based network intrusion detection system using machine learning approaches Peer-to-Peer Network Application", 12 (2) (2019), pp. 493-501
[18] M. Stamp, "A Revealing Introduction to Hidden Markov Models," no. October 2018, pp. 1–11, 2018.
[19] A. Kalbhor, T. H. Austin, E. Filiol, S. Josse, and M. Stamp, "Dueling hidden Markov models for virus analysis," JOURNAL Computer Virol. Hacking Tech., vol. 11, no. 2, pp. 103–118, May 2015.
[20] W. Wong and M. Stamp, "Hunting for metamorphic engines," JOURNAL Computer Virol., vol. 2, no. 3, pp. 211–229, Nov. 2006.
[21] T. Okamoto and Y. Ishida, "Framework of an Immunity-Based Anomaly Detection System for User Behavior," in Knowledge-Based Intelligent Information and Engineering Systems, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 821–829.
[22] R. Posadas, C. Mex-Perera, R. Monroy, and J. Nolazco-Flores, "Hybrid Method for Detecting Masqueraders Using Session Folding and Hidden Markov Models," Springer, Berlin, Heidelberg, 2006, pp. 622–631.
[23] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press, 2000.
[24] R. Berwick and V. Idiot, "An Idiot's guide to Support vector machines (SVMs)," 2003.
[15] I. Firdausi, C. lim, A. Erwin, and A. S. Nugroho, "Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection," in 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2010, pp. 201–203.
[26] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, "A Survey on Malware Detection Using Data Mining Techniques," ACM Computer Survey, vol. 50, no. 3, pp. 1–40, Jun. 2017.
[27] K. Rieck, P. Trinius, C. Willems, and T. Holz, "Automatic Analysis of Malware Behavior using Machine Learning," Journal of Computer Security, pp. 1–30, 2011.
[28] W. Hu and W. Hu, "Robust Support Vector Machines for Anomaly Detection," PROC. 2003 International Conference Machine Learning Application (ICMLA'03, pp. 23--24, 2003.
[29] S. Mukkamala, G. Janoski, and A. Sung, "Intrusion detection using neural networks and support vector machines," in Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), pp. 1702–1707.
[30] L. Khan, M. Awad, and B. Thuraisingham, "A new intrusion detection system using support vector machines and hierarchical clustering," VLDB Journal, vol. 16, no. 4, pp. 507–521, Aug. 2007.
[31] M. Kamble and C. Dule, "Image Spam Detection: A Review," 2017.
[32] A. Annadatha and M. Stamp, "Image spam analysis and detection," Journal Computer Virol. Hacking Tech., vol. 14, no. 1, pp. 39–52, Feb. 2018.
[33] H. Drucker, Donghui Wu, and V. N. Vapnik, "Support vector machines for spam categorization," IEEE Trans. Neural Networks, vol. 10, no. 5, pp. 1048–1054, 1999.
[34] T. Sohn, J. Seo, and J. Moon, "A Study on the Covert Channel Detection of TCP/IP Header Using Support Vector Machine," Springer, Berlin, Heidelberg, 2003, pp. 313–324.
[35] Kumar, "Cluster Analysis: Basic Concepts and Algorithms," Psychology.
[36] B. Mirkin, "Choosing the number of clusters," Wiley Interdisciplinary Rev. Data Min. Knowl. Discov., vol. 1, no. 3, pp. 252–260, 2011.
[37] D. Heckerman, "A Tutorial on Learning with Bayesian Networks". New York, NY, USA: Springer, 1998.
[38] F. V. Jensen, "Bayesian Networks and Decision Graphs". New York, NY, USA: Springer, 2001.
[39] R. P. Lippmann and R. K. Cunningham, "Improving intrusion detection performance using keyword selection and neural networks," Computer Network, vol. 34, pp. 597–603, 2000.
[40] A. Bivens, C. Palagiri, R. Smith, B. Szymanski, and M. Embrechts, "Network-based intrusion detection using neural networks," Intelligent Eng. Syst. Artificial Neural Network, vol. 12, no. 1, pp. 579–584, 2002.
[41] Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 806–813.
[42] Aastha Puri and Nidhi Sharma, "A Novel Technique for Intrusion Detection System for Network Security Using Hybrid SVMCART," International Journal of Engineering Development and Research, Volume 5, Issue 2, ISSN: 2321-9939, pp. 155-161.2017
[43] L. Safaa, Yassini K. E., and Hasnaoui. M. L., "A deep learning methods for intrusion detection systems based machine learning in MANET". SCA '19: Proceedings of the 4th International Conference on Smart City Applications, October 2019. https://doi.org/10.1145/3368756.3369021
[44] F. Jemili, M. Zaghdoud, and A. Ben, "A framework for an adaptive intrusion detection system using Bayesian network," in Proc. IEEE Intelligent Security Information, 2007, pp. 66–70.
[45] C. Kruegel, D. Mutz,W. Robertson, and F. Valeur, "Bayesian event classification for intrusion detection," in Proc. IEEE 19th Annual Computer Security Application Conference, 2003, pp. 14–23.
[46] S. Benferhat, T. Kenaza, and A. Mokhtari, "A Naïve Bayes approach for detecting coordinated attacks," in Proc. 32nd Annual IEEE International Computer Software Application Conference, 2008, pp. 704–709.
[47] M. Blowers and JOURNAL Williams, "Machine learning applied to cyber operations," in Network Science and Cybersecurity. New York, NY, USA: Springer, 2014, pp. 55–175.
[48] C. Kruegel and T. Toth, "Using decision trees to improve signature based intrusion detection," in Proc. 6th International Workshop Recent Adv. Intrusion Detect., West Lafayette, IN, USA, 2003, pp. 173–191.
[49] D. Ariu, R. Tronci, and G. Giacinto, "HMMPayl: An intrusion detection system based on hidden Markov models," Computer Security, vol. 30, no. 4, pp. 221–241, 2011.
[50] S. S. Joshi and V. V. Phoha, "Investigating hidden Markov models capabilities in anomaly detection," in Proc. ACM

43rd Annual Southeast Reg. Conference, 2005, vol. 1, pp. 98–103.

[51] M. Panda and M. R. Patra, "Network intrusion detection using Naïve Bayes," International JOURNAL Computer Science Network Security, vol. 7, no. 12, pp. 258–263, 2007.

[52] N. B. Amor, S. Benferhat, and Z. Elouedi, "Naïve Bayes vs. decision trees in intrusion detection systems," in Proc ACMSymp. Application Computer, 2004, pp. 420–424.

[53] Y. Li, JOURNAL Xia, S. Zhang, JOURNAL Yan, X. Ai, and K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," Expert Syst. Application, vol. 39, no. 1, pp. 424–430, 2012.

[54] F. Amiri, M. Mahdi, R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for IDSs," JournaL Network Computer Application, vol. 34, no. 4, pp. 1184–1199, 2011.

[55] W. J. Hu, Y. H. Liao, and V. R. Vemuri, "Robust support vector machines for anomaly detection in computer security," in Proc. 20th International Conference Machine Learning, 2003, pp. 282–289.

[56] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," Information Science, vol. 177, no. 18, pp. 3799–3821, Sep. 2007.

[57] Kim, J.; Kim, J.; Thu, H.L.T.; Kim, H. "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection". In Proceedings of the 2016 International Conference Platform Technology and Service (PlatCon), Jeju, Korea, 15–17 February 2016; pp. 1–5.

[58] Yin, C.L.; Zhu, Y.F.; Fei, J.L.; He, X.Z. "A deep learning approach for intrusion detection using recurrent neural networks". IEEE Access 2017, 5, 21954–21961