A Review of Machine Learning Methodologies for Malware Detection in Cloud Infrastructure

Dr.D.Praveena[1],G.Kamalesh[2],T.Kamalesh[3],D.Lokesh[4]

[1]Assistant Professor, Department of Information Technology, R.M.D Engineering College

[2,3,4]Student, Department of Information Technology, R.M.D Engineering College

Abstract—The Internet continues to spread itself over the globe, providing a great opportunity for various threats which are growing on a daily basis. Current static detection techniques only detect known malicious attacks and they also require frequent updates to signature-based databases. To reduce this work, systems are proposed for network intrusion detection systems capable of analyzing contents of the network by means of machine learning techniques to analyze and classify the malicious contents. Various machine learning algorithms are used for developing a Network Intrusion(Malware) Detection System. The review intends to provide an exhaustive survey of the currently proposed machine learning based intrusion detection systems in order to assist Network Intrusion Detection System developers to gain a better intuition.

Keywords—Static signature-based detection techniques; Network intrusion detection systems; Machine learning

I. INTRODUCTION

Usage of the Internet has been on the rise ever since its inception. Along with its boons, the internet has given rise to many vices. This has led to an increase in the number of attacks. These attacks may affect individuals as well as organizations.

The Equifax attack that had taken place in July 2017 leaked personal and financial information of almost 150 million people [17]. Such breaches continued in 2018. During January 2018, the Unique Identification Authority of India data breach, a massive 1.1 billion Aadhar data records of Indian citizens were exposed to the buyers for a mere sum of INR 500 (Approximately $7). In March 2018, a political data firm named Cambridge Analytica got hold of personal information of 87 million Facebook users. In June 2018, it was revealed that information of 120 million users was disclosed by ‘Nametests.com’. Similarly, in June 2018, email addresses of 92 million MyHeritage customers were leaked. MyHeritage is a platform which stores genetic data of people [18].

The vulnerability of victims of such attacks has drastically increased. This can be exemplified by the fact that the approximate total cost of cyber attacks in 2014 was $445 billion. It increased to $600 billion in 2017. It is projected that this cost will escalate to $6 trillion per year by 2021 [17]. As per the ‘Global State of Information Security Survey 2018’, 87% of all CEOs are devoting their resources to improve their security against cyber attacks]. Their belief is that it is the best way to obtain the customers’ trust.

Based on the aforementioned data it is evident that the extent of these attacks is vast and their impact on those who are targeted is severe. New types of attacks are being commenced frequently. Thus the need for pre-emptive Intrusion Detection Systems (IDS) is increasing as static detection techniques fail to provide security against new attacks.

II. INTRUSION DETECTION SYSTEM

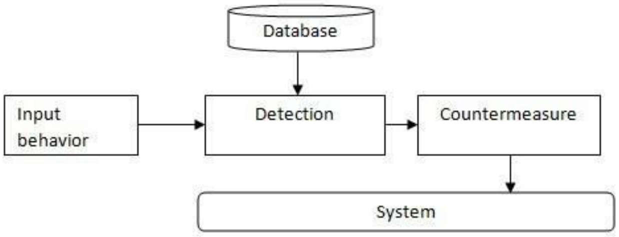

Fig. 1. Block Diagram of IDS

An IDS is a hardware or software application that monitors network traffic data on a system or a network. An IDS typically reports any policy violations or security breaches. A block diagram of a typical IDS is as shown in Figure 1 :

An intrusion detection system has a static database of identified malicious behavior. The input (i.e. the network traffic or system behavior) is compared with the entries from this database. If the input is malicious, the severity of the threat is detected and a proper countermeasure is used. The countermeasures range from simple notifications to blocking the activity which is suspected to be a threat. The most prevalent types of IDS are Host-based and Network based.

III. NETWORK INTRUSION DETECTION SYSTEM

A Network Intrusion Detection System (NIDS) is used to keep track of and provide analysis of internet traffic on the subnet. A NIDS reads all incoming data and looks out for suspicious behavior. The system reacts to such behavior based on the seriousness of the threat .

IV. Machine Learning

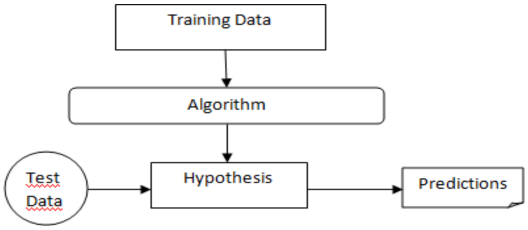

Fig. 2. The Framework of Machine Learning Process

Machine learning is a class of algorithms that allows software applications to become more precise in estimating outcomes without being explicitly programmed [13]. The algorithm applied to any data is jointly called a model. A machine learning algorithm learns from experience 'E' concerning some class of tasks 'T' and performance measure 'P' if its Performance at task 'T' enhances with experience 'E' [14]. A generic framework of Machine Learning process is shown in Fig. 2 [12]. A machine learning problem can be classified into:

A. *Supervised learning*

Supervised Learning is training a model with a dataset which also contains the correct answer for a prediction called as a label.

B. *Unsupervised Learning*

Unsupervised learning is training a model without using labels.

C. *Semi-supervised Learning*

Along with the above mentioned two categories, there is yet another field called Semi-Supervised learning, which contains datasets with a few labeled data points in addition to predominantly unlabelled data.

Machine Learning is used for Network Intrusion Detection to make the process dynamic as opposed to the current static detection techniques being used.

V. Reviewed Machine Learning Methodologies for Cloud Malware Detection

A few Machine Learning methodologies that are currently being used for Network Intrusion Detection are:

A. *Support Vector Machine*

Support Vector Machine (SVM) is an algorithm that rests upon the notion of decision planes known as decision boundaries. These decision planes assist SVM to classify data into their respective categories. Finding the optimal decision boundary is the main objective of SVM. These boundaries are constructed in the multidimensional space to achieve the optimal result in the case of non-linear data. This is a major advantage of SVM over a simple linear classifier. Margin holds the key for the correctness of classification of a new data point. The margin is the distance between nearest data point, also called as 'Support Vector', and the decision boundary.

Mathematical representation of SVM is given as follows:

$$\mathbf{w}^*, t^* = \underset{\mathbf{w},t}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||^2$$

B. *Algorithm proposed (Md Nasimuzzaman Chowdhury et. al.)*

The method proposed by Md Nasimuzzaman Chowdhury and Ken Ferens, Mike Ferens [1] begins with an arbitrary selection of 3 features at a time is done in the training samples. This combination is then fed to the SVM. This gives SVM the power to detect any odd activity from internet traffic data. The total number of features (N) from the dataset were identified and arbitrary combination of 'n' features was done(n belongs to N). SVM was applied to these training samples. Total S data samples were selected. SVM parameters such as Gamma, coefficient theta, nu etc. were selected. The T-train dataset is the training dataset; it contains n*S data samples. Testing dataset(T-test) was created using n*M data samples. T-train is used to train SVM. Testing the performance of SVM is done by using the T-test data set. Detection accuracy, FPR, FNR and total time taken by system defines the performance. Steps 2 and 3 were repeated until the highest detection accuracy and lowest FPR and FNR was achieved.

C. *Min-Max K-means clustering*

Another study by Mohsen et. al. have put forth the Min-Max K-means clustering [2] for intrusion detection. The suggested algorithm attempts to minimize the maximum internal variance of clusters instead of minimizing the sum of internal variance as that of the K-means algorithm. Every cluster has some weight and higher weights are assigned to the cluster with larger internal variance. Experimentation shows that Min-Max K-means is used to solve the initialization problem of K-means algorithm, as compared to clustering algorithms such as K-means++ [7] and pifs K-means. Min-Max K-means displayed 81% detection rate as compared to75% obtained by the K-means algorithm. False Positive Rate is improved from 14% to 9% for the Min-Max K-means algorithm. It is concluded that the Min-Max K-means clustering has a higher detection rate than the K-means clustering algorithm.

D. *Intelligent Intrusion Detection System*

The 'Intelligent Intrusion Detection Process' proposed by Jiaqi Li, Zhifeng Zhao and Rongpeng Li [3] consists of two phases. The first phase consists of using a Random Forest algorithm to obtain a subset of features by weighing their importance. The second phase includes a

'Hybrid Clustering-Based Adaboost' which acts as a classifier based on the subset of features as the input.

The 'Hybrid Clustering-Based Adaboost' is performed in two stages. The first stage consists of using the unsupervised clustering algorithm 'k-means++' [7] to create two preliminary clusters of malicious and benign activities. The 'k-means++' algorithm is preferred over the regular k-means algorithm in order to choose preliminary clustering centers which are as far away from each other as possible. The clusters are further classified into four types of anomaly clusters using AdaBoost [8]. AdaBoost is an ensemble classifier which consists of multiple smaller classifiers trained on the same data. The weight of every data point is the same at the beginning; however, if the example is misclassified in the previous classifier then the weight is increased, and conversely, if the example is correctly classified, then the weight is decreased, the same is followed in successive iterations.

E. *Artificial Neural Network*

Artificial Neural Networks (ANN) is a machine learning methodology inspired by the human nervous system. A single processing unit of an ANN is known as a perceptron. A perceptron receives weighted input along with a fixed bias value and generates an output. The mathematical representation of a perceptron is as follows:

$$\sum_{i=0}^{n} w^i x^i + b$$

where *w* is the weight vector, *x* is the input and *b* is bias value. A typical neural network consists of three types of layers - An input layer which gets real values from data points (Network dump files), hidden layers to process inputs and an output layer which provides an actual prediction. The design of the Artificial Neural Network for the Network Intrusion Detection System proposed by Alex Shenfield, David Day and Aladdin Ayesh [4] is as shown in Fig. 3.

The ANN consisted of 1000 input neurons to accommodate for 1000 bytes of contiguous data, 2 hidden layers consisting of 30 neurons each and the output layer consisting of 2 neurons. The data fed to the neural network was converted from byte data to integer data. 10-fold cross-validation was used to generate optimal results. The ANN was implemented using MATLAB (2016b edition) Neural Network Toolbox.

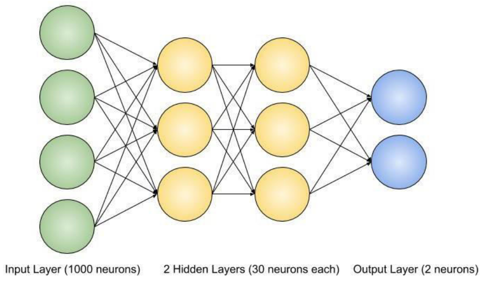

Fig. 3. Design of Proposed ANN

F. *Back Propagation Neural Network*

V. Jaiganesh et. al. have proposed a Back-propagation algorithm (BPA) [5] for intrusion detection. It is stated that BPA learns by examples that are used to find attacks. The algorithm is given examples of the function to be performed by the Neural Network. The parameter weights are changed during every iteration and the weights are provided as an input to the next iteration. The aim is to calculate attacks by the correct use of the back-propagation algorithm. The idea used behind the algorithm is to first apply for a forward pass through the neural network by supplying random weights (between -1 to +1) in order to get some output. The initial output obtained may not be precise as the weights are random. After obtaining the initial output, the following back-propagation strategy was applied.

VI. Datasets

A. *KDD Cup 1999*

The KDD Cup 1999 [16] data set was developed for 'The Third International Knowledge Discovery and Data Mining Tools Competition' which was held synchronously with 'The Fifth International Conference on Knowledge Discovery and Data Mining'. The complete dataset has 4 million data points and 42 features.

B. *UNSW-NB15*

Cyber Range lab of Australian Centre for Cyber Security (ACCS) has designed the UNSW NB 15 dataset [9]. It is a network-based dataset that has over two and a half a million records. Nine categories of attacks are present in this dataset, namely Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. 49 features were extracted with a class label, using the 12 algorithms of Argus and Bro-IDS tools. To generate the artificial network traffic, IXIA PerfectStorm Tool was used. This network data is a perfect combination of normal activities and attacks occurring simultaneously. 100 GBs of raw traffic was generated and captured with the help of Tcpdump tool.

C. *Custom Dataset*

The custom dataset used by Alex Shenfield et.al. [4] for the development of their NIDS consisted of normal traffic data added to malicious entries from the Exploit database.

VII. Observations

After the study of various machine learning techniques for developing a NIDS, the observations made can be summarized in Table 1. It was observed that four out of the six methodologies studied, showed high accuracy and all of the six techniques displayed a relatively low false positive rate. The algorithm proposed by Md Nasimuzzaman Chowdhury

TABLE I. OBSERVATION TABLE

Sr.No.	Algorithm	Accuracy (%)	FPR (%)	Dataset
1	SVM	88.03	4.2	UNSW-NB15
2	Algorithm proposed	98.76	0.09	UNSW-NB15
3	Min-Max K-means clustering	81	9%	KDD Cup 1999
4	Intelligent Intrusion Detection process	92.62	0.54	KDD Cup 1999
5	ANN	98.2	<2	Custom
6	Back Propagation Neural Network	78.15	low	KDD Cup 1999

et.al. [1] however, was the most promising with the accuracy of 98.76% and an extremely low false positive rate of 0.09%. The accuracy was obtained due to feature annealing [1] which focused on training the neural network based on the most influential features of the data. It was implemented on the UNSW-NB15 dataset. The ANN designed by Alex Shenfield, David Day and Aladdin Ayesh [4] also had a very high detection accuracy of 98.2% and a false positive rate of less than 2%. It was made possible due to the optimized selection of neurons per layer in the ANN design. The techniques which needed improvement were observed to be the Min-Max K-means clustering [2] and the Back Propagation Neural Network with relatively low detection accuracies of 81% and 78.15% respectively.

VIII. CONCLUSION

Based on the observations made it can be affirmed that the power of machine learning harnessed properly could have great potential in Network Intrusion Detection. The best methods were observed to be the algorithm proposed by Md Nasimuzzaman Chowdhury et. al. [1] and the Artificial Neural Network designed by Alex Shenfield, David Day and Aladdin Ayesh [4]. The best accuracies were observed on the machine learning models trained using the UNSW-NB15 dataset. Modern Machine Learning techniques using standardized and optimized machine learning libraries could be used to develop Network Intrusion Detection Systems with higher accuracies and low false positive rates. The accuracies for Network Intrusion Detection could be improved by optimized feature selection, optimizing learning algorithms by creating multiple weak classifiers for determining whether the network access is malicious or benign, normalization of data and optimizing the neural network design by modifying the architecture of the neural network and using regularization to prevent overfitting.

REFERENCES

[1] Md Nasimuzzaman Chowdhury & Ken Ferens, Mike Ferens (2016). “Network Intrusion Detection Using Machine Learning”.2016 Int’l Conf. Security and Management, SAM’16.

[2] Eslamnezhad, Mohsen & Varjani, A. (2014). “Intrusion detection based on MinMax K-means clustering”. 2014 7th International Symposium on Telecommunications, IST 2014. 804-808.

[3] Li, Jiaqi, Zhifeng Zhao and Rongpeng Li. “A Machine Learning Based Intrusion Detection System for Software Defined 5G Network.” CoRR

[4] Shenfield, Alex & Day, David & Ayesh, Aladdin. (2018). “Intelligent intrusion detection systems using artificial neural networks”. ICT Express. 4.

[5] Jaiganesh, V & Sumathi, P & Mangayarkarasi, S. (2013). “An analysis of intrusion detection system using back propagation neural network”. 2013 International Conference on Information Communication and Embedded Systems, ICICES 2013. 232-236.

[6] Ghosal, Amrita & Halder, Subir. (2017). “A survey on energy efficient intrusion detection in wireless sensor networks”. Journal of Ambient Intelligence and Smart Environments. 9. 239-261. 10.3233/AIS-170426.

[7] Arthur, David & Vassilvitskii, Sergei. (2007). “K-Means++: The Advantages of Careful Seeding”. Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms. 8.

[8] Wu P., Zhao H. (2011) “Some Analysis and Research of the AdaBoost Algorithm”. In: Chen R. (eds) Intelligent Computing and Information Science. ICICIS 2011. Communications in Computer and Information Science, vol 134. Springer, Berlin, Heidelberg.

[9] Moustafa, Nour & Slay, Jill. (2015).”UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)”.

[10] http://www.statsoft.com/textbook/support-vector-machines/.

[11] https://www.datacamp.com/community/tutorials/support-vevtor- machines-r/.

[12] https://www.ritchieng.com/one-variable-linear-regression/.

[13] https://searchenterpriseai.techtarget.com/definition/machine-learning- ML.

[14] https://machinelearningmastery.com/what-is-machine-learning/

[15] https://www.techopedia.com/definition/12941/network-based-intrusion- detection-system-nids.

[16] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[17] https://investingnews.com/daily/tech-investing/cybersecurity- investing/why-is-cybersecurity-important/

[18] https://blog.barkly.com/biggest-data-breaches-2018-so-far/.

[19] Flach, P. (2012). “Machine Learning – The Art and Science of Algorithms that Make Sense of Data”. Cambridge: Cambridge University Press.212.