Cyber Security: Threat Detection Model based on Machine learning Algorithm

Kushal Rashmikant Dalal
forkushaldalal@gmail.com
491 5th Avenue,Lyndhurst,NJ,07071

Mayur Rele
mayurrele@gmail.com
60 Bruan Place Apt D Clifton NJ 07012

Abstract— ***A threat can be anything that causes potential damage to the network system. These threats can turn out to be a attack to the system. Threat may occur in any forms like viruses, outright attack, and phishing attack from hackers to gain information. Such attacks put a user's system and also business system at risks. Cyber security aims at the protection of system from attacks like unauthorized network access, intrusions attacks etc. This paper presents a novel architecture model based on machine learning for the prediction of Cyber security malware that requires execution in a sandbox environment. In order to prevent the attackers from infiltrating the system Machine Learning approach is adapted.***

Keywords—Machine Learning, Sandbox

I. INTRODUCTION

IT infrastructure environments are growing complex day by day as more data is being produced. It is beyond the capacity of human brain monitoring, which fails to differentiate which data is standard and which one contains virus or malwares. Cyber security is that domain which finds difference between irregularities that still needs human expertise alongside with machine learning algorithms. It can be defined as the intersection of network security, computer security and information security. A cyber security system consists of two main parts a network and host security system. These two parts owns minimum of antivirus software, firewalls, and Intrusion Detection System (IDS). IDS help identify unauthorized use, alteration, duplication, and destruction of information systems [7].

Machine Learning(ML) is a data analysis method that automates building of an analytical model using algorithms that learn from data which can be easily automated, and find accuracy in the data without using explicit programming as to where look[8].

ML works on three phases. It first trains data, validate it and then test it through algorithms. In order to decide which the best model to work with is, the selection should be made on the basis of performance of model not on the accuracy of the data set given. Machine Learning works with following steps [9] :

1. Categorize the characteristics from training data (training model).
2. Classify split sets of attributes necessary for classification.
3. Model learning using training data (providing ML algorithm).
4. Using trained model to create classes of unknown data, and predict the result accurately.

The main advantage of ML algorithms is that it will learn and calculate based on practice and results. It means if today it is taking 1 day, tomorrow working time will be 20 hours, the next day it will spend 12 hours and so on. Machine Learning by "learning and predicting" calculates the automated tasks to the level of extent that human team cannot reach [10].

Several factors need to be measured while selecting the Machine learning algorithm. The factors take account of time intricacy, progressive up gradation ability, online/offline approach, and algorithm based on the detection rate of a system. Due to the diverse role of algorithm, it is quite difficult to make a choice of the most apt and well-organized algorithm [11].

Security Analytic System performance depends on two important factors i.e. class of input data and the implementation of Machine Learning (ML) algorithm. The ML algorithms scopes from supervised learning (e.g., Logistic Regression, Support Vector Machine, Naïve Bayes, and Decision Trees) to unsupervised learning algorithms(k-mean, clustering) and reinforcement learning .

II. LITERATURE REVIEW

In the paper [1] 'The promise of machine learning in cyber `Security' by authors J.B.Fraley, J.Kannady have mentioned some of the works have been carried out in the regards of securing the data from attackers. This paper will also helps to know the significant elementary part as well as finding solutions utilizing the system software. This paper is transpired from a regular lab scale which is termed as 'black

box' surrounding to an experimental platform. It has been utilized by numerous blooming commercialized agencies. Few of the applications synthesized are 'speech recognition, voice command, language processing and many more.

The cyber security and system learning is the best for managing high amount of data processing unit. Automated security devices are specifically based to offer signature based control. These applications specifically detect suspicious events and provides alert to data, network like Intrusion Prevention System/Intrusion Detection System. This paper reveals the higher pathway of approaching the system and their results in the security aspects.

Some of the reports such as from Macfee says that there are about 3 million newly synthesized vicious files generated every 60 minutes. The anticipated aim is that the system learning will be the main objective to eradicate the current attack of scenario.

To comprehend the exact situation in ongoing attack the exact numerical are mentioned in tabular column to know the state of the attacks. The e-mail, network, internet, malware in thousands. These numerical are in common for millions of users.

There are six steps of analyzing and overcoming the working process these operations are utilizes.
In the phase 'develop business understanding' step, here the actual reasons for the defects are taken into consideration in this meet are the answer for them are maintained through the system learning procedure.

In 'Analyse data and data dependencies' step, the group combines and to compare, analyse the data present with the situation and to bring out the out of the box answer for the problem. Globally there are 1200 alerts which are grouped in to 20 sections to ease the way of study.

The subsequent step is 'Engage subject matter experts' are indulged in analysis of the situation and in examining along with grouping of the factors. The accuracy of attacks along with proper numerical are listed for legible analysis in further steps. The SME's were in clear in-tension of bringing out transparent results in the scenario.

Followed by 'Paper data' in which the data will be fed to the system which will provide the proper solutions. To get an optimum result, the data which is fed must be analyzed, re-checked in an organized way. This step will consume around 85% of the total engineering work. The system will analyse for the gaps, in-completion, not in relation along with comprehension with the data present to clear the analysis.

Then comes the 'Develop model' which will be generated by the group from the previous step data fed to the set system. 'TensorFlow' is being used in this step for the optimizing the results. It has the capability to multitask by handling the system and also managing status in a parallel way.

Last step is the 'Evaluate model' here the models created with the data set fed to the system are analyzed and are evaluated with performing calibration. The stages are cross-examined to comprehend the performed status which is the major goal, later all are properly labelled.

The accuracy level was pointed to about 90% yield for the study. By knowing the results the time consumed from the traditional way was brought down by approximately 78% from the study. The stressed reported for every hour was brought down by 1/4th of the original problems reported.

In the paper[2] 'A machine learning approach to detecting senior data modification intrusion in WBANs' by A.Verner and D.Butvinik, In WBAN only little quality of the answers are produced, the best are code blue and alarm net. These are based on ECC and AEC type of security safety respectively. In this field it uses the similar method of enlisting the received data in an efficient format in many fields. In the methodology in this study they have made few assumptions which enables them to teach in a more understandable way. The keys used are 'Assumptions, method, Chosen ML features, chosen structure of negatively labelled vector' which are the key tools used in the study. The studying phase of SVM and also the execution time have taken longer period to complete in a way many of the vectors were combined to overcome the drawback of the time with a complete accuracy level. Still better modification is in need for this method.[2]

In the paper [3], 'Machine learning to detect anomalies in web log analysis' published by Q.Cao, Y.Qiao, and Z.Lyn..This paper segregates the whole system into 4 different parts are 'data pre-processing, decision free classifier, data extractor, Hidden Markov model' are the included models. It mainly relates in producing the solutions in a path of data. Clustering algorithm was the first ever anomaly network. This proposed method builds a proper model for regular files and then maintain them as the detector. This paper produced high level of optimum solution of about 93% and negative positive rate till 4%. The parameters were self-allotted to generate better results. The performed method was made in comparison with other models which shows better performance. Even with the better results the module was not assured with proper adaptable quality with regards.

In the paper [4], 'Evaluation of machine learning techniques for network intrusion detection' by M.Zaman and C.H.Lung, have mentioned that in the evaluation of machine learning for detection clustering types are mostly utilized. This platform is widely used in the research as the source data. The paper totally represented six separate methods of machine learning along with six ensemble methods which is used to gather the results for the attacks. But when compared with the ROC results, it was not satisfactory although it had opened a unique route of machine learning. The technique involved in tracking the traffic data which had involved true positive, false positive, true negative and false negative. The proposed method did not provide better

results hence further study and adaptable measure is a must in the future studies.

In the paper [5] 'A review of intrusion detection using anomaly based detection' by authors U.Kumari and U.Soni, have illustrated that the most important aspect of data security is data intrusion detection, in identifying the fraud, fake products, attacks. This intrusion is used to detect the attack cause or the root or the relation/bridge between the huge set back of attack. This key aspect is now been regularly used as the route to find the cause in all the fields both private, commercial fields. The unusual patterns can be easily identified for any terrorist activities as the data attack. This paper has improved the effectiveness of the system functionalities also with the secure feature. The huge drawback is that in securing the database the paper has left out [5].

In the paper [6], 'Security and privacy in machine learning' written by N. Papernot, P. McDaniel, A. Sinha, M. P. Wellman, have mentioned that the platform for data, security and privacy is an critical area which is the target for ransom, attacks etc. The paper brings out the model for reasoning of the threats by representing the diagrammatic networks like the 'physical domain, digital representation, machine learning'. By understanding the sensitive part of the privacy the study for learning the techniques are carried out. The spaces are further filled. Here the sensitivity of the deployed data facilities are focused and are studied. With partial success with the study still the generalization of sensitivity and the gap remains unclear.

III. ARCHITECHTURE MODEL FOR CYBER SECURITY USING SAND BOX

For machine learning algorithms to achieve a better performance it is necessary to run the malware in a sandbox to collect features from the malware which cannot be obtained statically. We present a novel architecture model based on machine learning for the prediction of malware that requires execution in a sandbox environment. In Figure 1 Network-based anti-malware gateway i.e. sandbox deployment led to widespread exposure of antivirus weaknesses resulting in a wave of next-generation endpoint security deployment as well as industry innovation.

The goal of machine learning based architecture is to make predictions on the unseen data and present a final model that performs the finest job defined by parameters like available historical data, the time spent on the architecture and finally the procedure.

Data from enterprise internet is collected for classification, i.e. the data is identified in two forms one is bad data and the other is, data for training set containing both positive and unlabelled data set. Problems which involve classification are considered as part of machine learning. In machine learning (ML) the computer gets the ability to learn from data. After the data is classified it is segregated into groups using clustering, on certain set of rules. This can be achieved by applying clustering models (Centroid, distribution and connectivity) [20].Clean data is moved directly to sand box as test data for further deep malware detection.

The primary requirement of training dataset is that it must be accurately labelled as it may predict misleading results and the future unknown sample of data. The training set containing one labelled and another unlabelled data set (positive data (labelled) as well as malicious sample (unlabelled)) [21] is applied to the training model. Now the rules are extracted from these trained data set. These rules are used to train the algorithm. Principally, this leads to the creation of classifier for new samples.

Here training data is prepared to train the model by applying it to various filters.

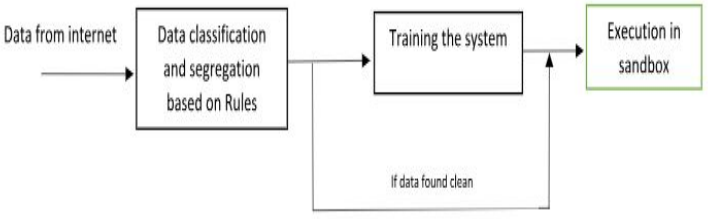

Figure.1 Threat Detection Model

The unknown sample of data is provided to the training model for its detection. The training set creates a model which understand the success rate of input data set. In fig 2 as below shows in detail the training model. It shows how it detects the malicious threats based on certain set of rules that are selected on the basis of trained set.

As shown in figure 2 below. The output from trained database model is in the form of decision or metric. The decision from user is taken in the form of true or false from user. If rules does not match (false), the sample data is send to training the database for adding new rules. After rule matches, that pure data is sent to database for storing.

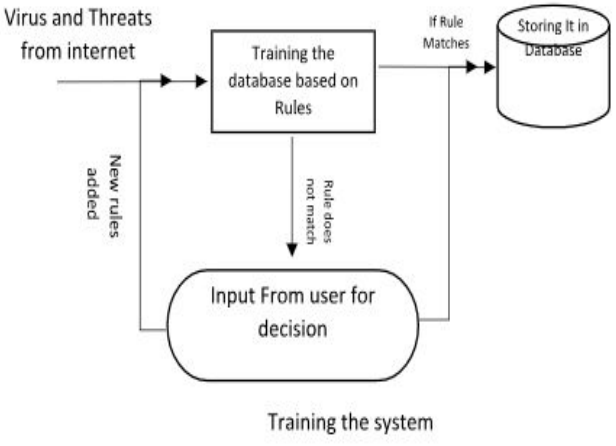

Figure. 2 Training the System

From database it will go to sandbox as in figure 1. Another powerful tool for advance threat detection is Sandbox. Sandboxing along with machine learning methods has emerged as a powerful cyber security tool. The time taken by the algorithm to train a database (i.e., training time) varies from algorithm to algorithm. Lots of research is needed for selecting the ML algorithm according to application. Area like if the input is continuous apply supervised machine learning methods and if discrete apply unsupervised learning methods. The selection of algorithm also depends upon the working mode (online or offline) of a security analytic system.

It is clear that to apply the machine learning algorithms to any problem, it is essential to represent the data in some form. For this purpose, Sandbox is used. The reports generated by the sandbox, describing the behavioural data of each sample, are pre-processed, and malware features are extracted from there [22].

Sandboxing engages in the capturing a document or executable file which is then opened within a secure virtual machine. In this controlled environment, in order to observe how the executing software behaves exactly, potential threats are run. Sandbox is used for executing untrusted, untested programs or code, possibly from unverified third parties, websites, suppliers, without risking any harm to the host machine in cyber-security system or any operating system.

An advanced sandboxing solution provides CPU-level threat protection which depends on the mishandling stage of the attack. This allows an organisation to be detected and blocked against advanced persistent threats (APT).

A software architect should take into consideration several things during the selection and integration of an optimized algorithm. The ML algorithm performs better for one type of security analytics (e.g., detecting DoS attack) may not perform well in another security analytics i.e detecting brute force attack).

The selection of an algorithm is tricky in a way that if it is giving performance, then it may degrade other system qualities like accuracy, complexity, and understand ability of the final result. For example, Researchers [24] compared SVM (Support Vector Machine) with Extreme Learning Machine (ELM) in terms of accuracy and performance. It is found out that SVM (Support Vector Machine) produced more accurate results but proved computationally expensive. On the other hand, ELM proved lighter but produced less accurate results. A practical trade-off should be recognized while selecting the algorithm among various system's qualities.

As it is said in paper[23] "While there is no silver bullet to solve the cybersecurity challenge, the key is to use layered defences along with machine learning threat detection capabilities for optimum results".

IV. EXAMPLES OF DIFFERENT ALGORITHMS EMPLOYED IN OTHER DETECTION APPLICATIONS

Cloud-based Threat Detector [25] the system has been implemented with two ML algorithms – K-mean and Naive Bayes. To explore the training time taken by both the algorithms to train a system, it is observed that with 500 GB training data, K-means takes around 60 seconds while Naive Bayes takes around 92 seconds to train a model.

● V APPLICATIONS OF APPLYING MACHINE LEARNING IN CYBER SECURITY

These are the threats that ML can save against [38]:-

1. Ransomware- malware that dont allows the user to access its personal files and ask for ransome payement for again accessing personal account.

2. Watering Hole- Hackers in this keep on tracking the websites on which users usually visits and access their identification is the concept of a watering hole.

3. Webshell - Webshell is a short code that allows the hacker to make modifications on server's web root directory. This means that full access to the database of the system is gained. If it is an e-commerce website, in order to collect credit card information of the customer who operates on everyday can access the database.

4. Spear Phishing- The Machine learning trained models can be used to detect whether the email is malicious or not by identifying the key features such as email headers, subsamples of body-data, punctuation patterns, etc.

VI. CONCLUSION

Sand boxing technique along with Machine Learning method is a dominant cyber security tool to prevent cyber attacks on the network system. A dominant cyber security tool that uses Machine Learning methods along with Sandbox is presented in this paper. It has proven to be a powerful method for advanced threat protection. In this propose method, the transactions will be on hold until the controller gets the confirmation from the sand box servers to execute on main servers. This technique proves to be an efficient and autonomous in advance threat detection systems.

REFERENCES

[1] J. Fraley and J. Cannady, "The promise of machine learning in cybersecurity", SoutheastCon 2017, 2017.

[2] A. Verner and D. Butvinik, "A Machine Learning Approach to Detecting Sensor Data Modification Intrusions in WBANs", *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.

[3] Q. Cao, Y. Qiao and Z. Lyu, "Machine learning to detect anomalies in web log analysis", *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 2017.

[4] [M. Zaman and C. Lung, "Evaluation of machine learning techniques for network intrusion detection", NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium, 2018.

[5] U. Kumari and U. Soni, "A review of intrusion detection using anomaly based detection", 2017 2nd International Conference on Communication and Electronics Systems (ICCES), 2017.

[6] N.Papernot, P.McDaniel, A.Sinha, M.P.Wellman, "Security and privacy in machine learning", Tutorial at IEEE WIFS, Rennes, France 2017.

[7] Mukkamala A, Sung A, Abraham A, "Cyber security challenges:Designing efficient intrusion detection systems andantivirus tools".

[8] R. Shanbhogue and B. Beena, "Survey of Data Mining (DM) and Machine Learning (ML) Methods on Cyber Security", *Indian Journal of Science and Technology*, vol. 10, no. 35, pp. 1-7, 2017.

[9] Congzheng Song, Thomas Ristenpart, Vitaly Shmatikov, "Machine Learning Models that Remember Too Much", 22 Sep 2017, **arXiv:1709.07886v1 [cs.CR].**

[10] Ullah, F., and Babar, "Architectural Tactics for Big Data Cybersecurity Analytic Systems: A Review", 9 Feb 2018, **arXiv:1802.03178v1 [cs.CR].**

[11] A. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153-1176, 2016.

[12] DeLiang Wang , "Unsupervised Learning - Foundations of Neural Computation" ,AI Magazine, vol. 22, pp. 101-102, September 2001.

[13] Hwanjo Yu,Jiong Yang, Jiawei Han, "Classifying Large Data Sets Using SVMs with Hierarchical Clusters," SIGKDD '03 Washington, DC, USA, ACM 1581137370/ 03/0008, 2003.

[14] Andrew Ng , CS229 Lecture notes, support vector machines,part- V, http://math480-s15-zarringhalam.wikispaces.umb.edu/file/view/SVMs_Ng_Stanford.pdf/538578768/SVMs_Ng_Stanford.pdf

[15] Jason Brownlee, "A Tour of Machine Learning Algorithms", 25 Nov 2013.

[16] T. Soni Madhulatha, "An Overview on Clustering Methods", *IOSR Journal of Engineering*, Apr 2012, Vol. 2(4) pp: 719-725, ISSN 2250-3021**, arXiv:1205.1117v1 [cs.DS].**

[17] Zoubin Ghahramani Sam Roweis, 'Probabilistic Models for Unsupervised Learning', *Gatsby Computational Neuroscience Unit University College London , NIPS Tutorials*, http://www.gatsby.ucl.ac.uk/, Dec 1999.

[18] G. N. Ramadevi, K. Usharani , "Study on Dimensionality Reduction Techniques and Applications", *Publications of Problems & Application in Engineering Research – Paper*, Vol 04, Special Issue01 , ISSN: 2230-8547; e-ISSN: 2230-8555, 2013.

[19] W. Lee, M. Cheon, C. Hyun and M. Park, "Best Basis Selection Method Using Learning Weights for Face Recognition", *Sensors*, vol. 13, no. 10, pp. 12830-12851, 2013.

[20] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", *Published in Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, IOS Press Amsterdam, The Netherlands, The Netherlands, pp. 3-24, ISBN: 978-1-58603-780-2, 2007.

[21] Anoop Kumar Jain and Satyam Maheswari, "Survey of Recent Clustering Techniques in Data Mining", *International Archive of Applied Sciences and Technology*, Volume 3 [2], ISSN: 0976-4828, pp. 68 - 75, June 2012.

[22] B. Liu, X. Li, W. Lee, & P. S. Yu, "Text Classification by Labeling Words. Retrieved", *American Association for Artificial Intelligence*, 2004.

[23] L. Vokorokos, A. Balaz and B Mados, "Application Security through Sandbox Virtualization", *Acta Polytechnica Hungarica*, Vol. 12, No. 1, 2015.

[24] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms", *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006.

[25] Chi Cheng, Wee Peng Tay and G. Huang, "Extreme learning machines for intrusion detection", *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012.

[26] N. Keegan, S. Ji, A. Chaudhary, C. Concolato, B. Yu and D. Jeong, "A survey of cloud-based network intrusion detection analysis", *Human-centric Computing and Information Sciences*, vol. 6, no. 1, 2016.

[27] A. Sallab, M. Abdou, E. Perot and S. Yogamani, "Deep Reinforcement Learning framework for Autonomous Driving", *Electronic Imaging*, vol. 2017, no. 19, pp. 70-76, 2017.

[1]
[2]