IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 1, January 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001076

A Machine Learning Framework for Domain Generation Algorithm (DGA) Based Malware Detection

Juweriya Solkar [1], Priyanka Bandagale [2], Daniya Solkar [3]

B.E Student, Department of Information Technology, Finolex Academy of Management & Technology, Ratnagiri, Maharashtra, India[1]

Assistant Prof. Department of Information Technology, Finolex Academy of Management & Technology, Ratnagiri, Maharashtra, India[2]

B.E Student, Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India[3]

ABSTRACT - Most of the time an intruder makes use of the Command and Control Server to exploit the transmission. To perform an attack, the attacker deploys a Domain Generation Algorithm (DGA) which creates various network locations. The existing precaution measures like blacklisting are incapable to handle the threat. In this paper, we put forward a machine learning framework to recognize and detect domain threats to reduce probable attacks. Using the data gathered from one year's span we will be classifying the DGA domains with the help of a deep learning model. In the first place, the normal domains are filtered out from the DGA domains, and later on, clustering is used to spot out the algorithms that produce DGA domains. A Deep Neural Network (DNN) is used to boost the capability of enormous datasets.

KEYWORDS: Feature Extraction, Random Forest GPU, Histogram, LBP (Local Binary Pattern), Nearest Neighbour Classifier, Pattern Matching.

I. INTRODUCTION

The evolution of technology has created ease in data availability but on the other side, there is a constant struggle to maintain the confidentiality of the data being shared and processed online. As the technology enhances so does the techniques of the attacker to invade the data systems. Data is the most crucial part of an organization. And on the second priority comes the network, as all the data and important reports are shared online through a network. So, to maintain network security is a weighty concern. There is a tedious anti-malware software list that is developed to cease the attacks. But most of the software fail to provide the amount of security required. Techniques such as static string matching, network communication whitelisting, hashing schemes, etc are used by software to fight network attacks. But these solutions are too lucid for the attacker to crack.

The Command and Control server is most often used for DDoS attacks. It is used to send controls and commands to the zombie computers. The zombie computers are under complete control of the attacker and they have numerous solitary IP addresses. The attacker may create software that abducts passwords or other confidential data which can lead to a huge mishap. To avoid the happening of unfavorable situations, a proper precautionary measure is very much important.

If the user gets mail from a domain that is categorized as spam, the mail will go into the spam bin directly which will deteriorate the attacker's plot. This is where the Domain Generation Algorithm comes into the picture. The Domain Generation Algorithm alters the domain from time to time. Using this the attacker's server will be prevented from blacklisting. Now the user has the delusion that he is browsing on a safe domain but he is not. So, whatever data that the user uploads or shares on the network is accessible to the attacker and the network is no more private.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001076

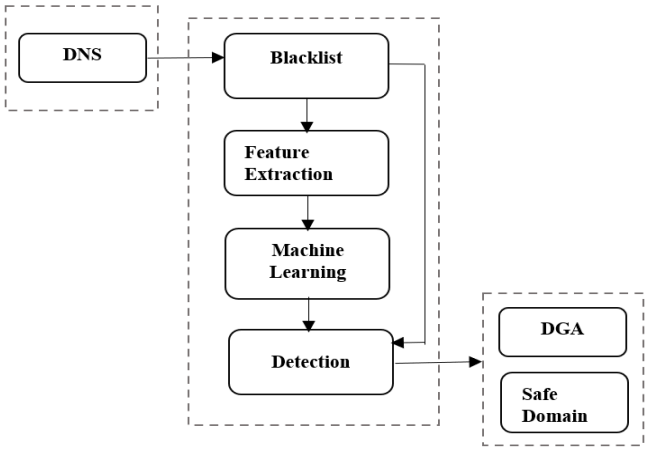
II. LITERATURE SURVEY

A Machine Learning Framework for Domain Generation Algorithm-Based Malware Detection was developed in 2019. This framework made use of the DBSCAN algorithm. The domains which were procured from classification, only those were used for clustering. The disservice caused here was that the system could classify only the known domains and it couldn't recognize other domains that were unknown and decide whether if the domain is safe or unsafe. Another model named Learning and Classification of Malware Behaviour was initiated which was capable to identify the occurrence of malware. It performed and identified 70% more efficiently than anti-virus software could. But even this approach could not give quite satisfying results as it worked based on the information attained from surveillance of the network.

An SDN based framework for guaranteeing security and performance in information-centric cloud networks was proposed in 2017. The framework was able to progressively cipher the routing path to ensure the safety and pursuance of the network being used. But the approach got demoted as the queries which didn't match the knowledge base were stored in the product backlog.

III. PROPOSED FRAMEWORK

In the proffered framework we will not be only recognizing the blacklisted domains but we will be also identifying the malicious domains. Here python is used for programming and we have an SQL Database. The entire process from the beginning to the end can be illustrated in a diagram as shown below:



IV. STEPS INVOLVED

1.1 Blacklisting:

Domain names are the most important thing in the process as it plays the role of input for all the techniques we will be implementing. It is the core of our operations. In this framework, we will be maintaining a blacklist. This list will consist of the domain names which have been recognized as harmful in previous times. When a domain name is given as an input to the system the blacklist will be checked thoroughly and if the name is to be found present there then, the feature extraction step and the machine learning step will be skipped. And if the domain name is new then all the further methodswill be applied to it and based on that the prediction will be done. The list will have pre-stored domain names that are gathered precedingly.

1.2 Feature Extraction:

In this step primarily, the length of the input will be calculated that is, the number of characters. The second step will involve counting the ratio of meaningful words. It comprises detecting if the words in the input make any sense, do they form a meaningful word or are they just random. The accuracy of this will be calculated in numbers. And then the third step involves counting the percentage of numerical in the input value. It will count the number of times a

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001076

numerical exists in the input string and then output the value of the calculated data. The fourth step is to count the pronounceability score. It will detect if words in the input are pronounceable. Again, the output of this will be given in percentage and it will output the percentage of pronounceable string from the total string. The fifth step comprises calculating which is the longest string which forms a meaningful sentence or a word from the input and based on that it gives percentage values. It measures the minimum number of single-character edits between a current domain and its previous domain in a stream of DNS queries received by the server. The Levenshtein distance is calculated based on a domain and its predecessor. To sum up, the exact work of feature extraction consider an example of a Facebook URL, A normal Facebook URL must look like https:facebook.u433=False-123.com. If we refer to the given example, we can notice that the URL consists of pronounceable and meaningful words, also the number count is reasonable. It is observed that fake domains often have unusual words that do not form a meaningful word and have a conspicuous count of numbers. All these minute details most often go unnoticed by the user.

1.3 Machine Learning:

After the data gained by feature extraction, we move on to the next process which involves applying approaches and techniques to generate an output, based on which detection of the DNS will be done.

Decision Tree: This step will involve decision making. A decision is made based on the fulfilment of certain conditions. The conditions will be pre-defined and the decision to be made will depend upon the input provided.

Artificial Neural Network: ANN is a part of the supervised learning technique. ANN's are made to function like a human brain. It decides how a human would have. It takes into consideration many factors while deciding.

Support Vector Machines: SVM does the work of classification and regression. SVM falls under the category of supervised learning techniques. It will perform the task of analysing the dataset. Here we will be using the Scikit-learn library of Python.

Multiple Logistic Regression: Here, regression is applied to data and, classification is done. Multiple Logistic Regression gives probable value as output and it is not a single value.

Naïve Bayes Classifier: This classifier implements the Naïve Bayes Theorem. It will analyse the data and look for connecting factors between features.

Random Forest: It comprises a large number of decision trees. It does the work of classification, regression, and value prediction. After applying all these machine learning techniques, the accuracy will be calculated. And using the DBSCAN algorithm clustering will be performed. The domain name will be clustered based on conditions specified in the algorithm.

Time Series Prediction: The historic data from previous incidents will be used to predict the outcomes in the upcoming times. It will take into consideration the output of each step of feature extraction and based on observations of the gathered data will predict values.

1.4 Detection:

As a concluding step, it will decide based on the previous processes the final category of the domain. It will decide if the domain is safe or is malicious.

V. APPLICATION

a) For checking the security of an online banking website**.**
b) In organizations where data is shared via a single network.
c) Verifying online shopping websites.

VI. CONCLUSION

The proffered framework will do the work one step ahead as that of anti-virus software. It not only classifies but also identifies the domains. Implementation of this framework will reduce the number of domain attacks. It will help maintain security when browsing on the internet and keep intact the confidentiality of data.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001076

REFERENCES

[1] Yi Li, KaikiXiong, Tommy Chin, Chengbin Hu. Institute of Electrical and Electronics Engineers, 2019.
[2] Konrad Rieck, Thorsten Holz, Carsten Willems, Patrick Dussel, Pavel Laskov. Kluwer Academic Publishers, Dordrecht (2002)
[3] C. Khancome, V.Boonjing, and P.Chanvarasuth. Tenth International Conference on Information Technology: New Generations(ITNG). IEE, 2013.
[4]U.Ghosh, P.Chatterjee, D.Tosh, S. Shetty,K. Xiong, and C. Kamhou. 11th IEEE International Conference on Cloud Computing (IEEE Cloud), 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details