SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502

DETECTION OF MALWARE USING DEEP LEARNING AND MACHINE LEARNING ALGORITHMS

H. Shanthi
Assistant Professor, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India, ttshanthi876@gmail.com

M. Kalpana
U.G Student, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India

M. Pavani
U.G Student, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India

P. Nandini
U.G Student, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India

ABSTRACT—Malware has become a significant downside in computers resulting in a rise in malware attacks within the kind of malicious software system. Previous malware sleuthing techniques were supported Static and Dynamic analysis of malware. As these techniques are long recent malware use Machine Learning algorithms like Deep learning algorithms and polymorphic, metamorphic techniques to enhance malware detection. There's a requirement to mitigate bias and measure these strategies severally to make new increased strategies for effective zero-day malware detection. During this paper, we have a tendency to find and measure the malware mistreatment Machine Learning algorithms (MLAs) and deep learning architectures for detection, classification, and categorization of malware by providing datasets and validators all the machine learning and deep learning algorithms. Then we are going to determine the accuracy, precision, and prediction of each formula then compare them and find that malware is gift within the given dataset. Overall, this work proposes an efficient visual detection of malware employing an ascendable and hybrid deep learning framework for period of time deployments.

Therefore, the deep learning and machine learning algorithms for static, dynamic, and image process analysis approach could be a new technique for malware detection. A proof-of-concept model has been developed for example the effectiveness of the projected system.

KEYWORDS— Malware detection, Static and Dynamic Analysis, Machine Learning, Deep Learning, Image Processing

I. INTRODUCTION

In this world due to the advancement of technology, the daily life activities of personal lives

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502
has affected. Malware is software that is specially designed to damage or gain unauthorized access to a computer system. Examples of malware are spyware, backdoor, rootkit, Trojan, etc., To remove malware, you must be able to identify malicious actors quickly. This needs a constant network scanning. Once the threat is identified, we have to remove the malware from our network. Today's antivirus products are not enough to protect against advised cyber threats. Sufficient advanced malware protection requires multiple layers of safeguards along with a high-level network.

RESEARCH BACKGROUND

The first-ever Trojan horse created its look as "Morris worm" in 1988-1989 that is intended to sight the malware by finding a match with the virus definition information updated from time to time known as Signature-based malware detection. The key challenge of this can be that it is unable to sight zero-day malware as new malware use antivirus evasion techniques. Moreover, it needs a bigger time, throughout that an assaulter would attack the system.

NEED FOR STUDY

Machine Learning Algorithms (MLAs) are used to detect malware. The performance is biased by training the data. MLAs depend on future engineering, future selection, and future representation methods. It contains a set of features to create a separate plane and divides a separate family. The features are obtained based on static and dynamic analysis. These algorithms improve the performance as the no. of samples available for increasing the learning. Dynamic analysis is used at run time for the process of monitoring malware behavior. Malware which is detected based on dynamic analysis is more robust when compared to static analysis. Due to security and privacy concerns the available data is less for publishing. A new model named deep learning is invented to overcome the issues and drawbacks that came with machine learning algorithms. By using machine learning algorithms (MLA's) we get diminishing outputs to avoid this problem deep learning model is invented. Deep learning is introduced to improve cyber security. MLAs are scalable and use more data. So, to avoid these issues we need to develop the techniques.

II. RELATED WORK

The process of development of malware detection is divided into two stages: feature extraction and classification. They extracted features like API calls, permission, etc. And used SVC and F1-score for the malware detection model which has the accuracy of 94% and 96%-99% respectively.

Large Scale Identification of Malicious Singleton Files

Li, B., Roundy, K., Gates, C., & Vorobeychik, Y. (2017,March) [2], studied a dataset of billions of program binary files that appeared on 100 million computers over the course of twelve months, discovering that 94 percent of these files were talented on one machine. Though malware polymorphism is one cause for the large form of singleton files, more factors put together contribute to polymorphism, as long because the quantitative relation of benign to malicious singleton files is 80:1. The massive form of benign singletons makes it troublesome to faithfully establish the minority of malicious singletons. We've got a bent to gift a large-scale study of the properties, characteristics, and distribution of benign and malicious singleton files.

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502

We've got a bent to leverage the insights from this study to form a classifier based strictly on static choices to identify xcii of the remaining malicious singletons at a 1.4% false positive rate, despite important use of obfuscation and packing techniques by most malicious singleton files that we've got a bent to form no decide to deobfuscate. Finally, we've got a bent to demonstrate strength of our classifier to special classes of automated evasion attacks.

Measuring the value of cyber-crime

Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M. J., Levi, M., ... & Savage, S. (2013) [1] made a square measure very inefficient at fighting cybercrime; or to place it differently, cyber-crooks square measure like terrorists or metal thieves in this their activities impose disproportionate prices on society. A number of the explanations for this square measure well-known: Cyber-crimes square measure international and have robust externalities, whereas ancient crimes like felony and automobile stealing square measure native, and therefore the associated equilibrium have emerged when a few years of optimization. As for the lot of direct question of what ought to be done, our figures counsel that we must always pay less in anticipation of law-breaking (on antivirus, firewalls, etc.) and a lot of in response – that's, on the prosaic business of searching down cyber-criminals and throwing them in jail.

Convolutional Neural Network

CNN is used primarily for image recognition and processing, due to its ability to recognize patterns in images. It is used for image classification. It has multilayered neural networks and deep learning methods. It also has a specific structure called the convolution layer.

Malimg dataset

Vinayakumar Ravi, Mamoun Alazab, Soman Kp, Prabaharan Poornachandran (2019, April) [11]. From 25 various malware families, a Malimg dataset consists of 9,339 malware samples. The detailed dataset statistics are shown below.

No.	Family	Family Name	No. of variants
1	Dialer	Adialer.C	122
2	Backdoor	Agent.FYI	166
3	Worm	Allaple.A	2949
4	Worm	Allaple.L	1591
5	Trojan	Alueron.gen!J	198
6	Worm: AutoIT	Autorun.K	106
7	Trojan	C2Lop.P	146
8	Trojan	C2Lop.gen!G	200
9	Dialer	Dialplatform.B	177
10	Trojan Downloader	Dontovo.A	162

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502

11	Rogue	Fakerean	381
12	Dialer	Instantaccess	431
13	PWS	Lolyda.AA 1	213
14	PWS	Lolyda.AA 2	184
15	PWS	Lolyda.AA 3	123
16	PWS	Lolyda.AT	159
17	Trojan	Malex.gen!J	136
18	Trojan Downloader	Obfuscator.AD	142
19	Backdoor	Rbot!gen	158
20	Trojan	Skintrim.N	80
21	Trojan Downloader	Swizzor.gen!E	128
22	Trojan Downloader	Swizzor.gen!I	132
23	Worm	VB.AT	408
24	Trojan Downloader	Wintrim.BX	97
25	Worm	Yuner.A	800

Table 1: Statistics of the dataset

III. CLASSIFICATION OF MALWARE MODELS

The models of malwares are classified based on dynamic and static analysis.By using these models we can understand the different malwares presented. The models are,

1.Malware classification using static analysis.

2.Malware classification using dynamic analysis.

If malware contains big data we need to consider another technique named as Image processing techniques. It is used for good decision making.

Malware Classification using Static Analysis

The Static analysis permits US a chance to research the ASCII text file while not capital punishment the program. Static analysis is used to find the safety vulnerabilities, and it's additionally used to detect:

Performance problems

Non-compliance with standards

Use of out of date program constructs

It is performed throughout the continual Integer(CI) method to come up with a report for problems. The computer memory unit n-gram and string are the 2 most typically used ways for

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502
static malware detection.

Static analyzers are tools that assist you check your code while not very running your code. It will to pinpoint the precise vulnerable line within the code. Throughout a static Analysis, as you'll be able to trace the contaminated information from supply to sink, one can determine the purpose result's a vulnerability. During this analysis by victimization Windows-Static Brain-Droid (WSBD), we will value the performance of benchmarked models.

Malware Classification using Dynamic Analysis

In this analysis, the process of testing and evaluating a program while the software is running. The dynamic analysis is more robust to obfuscation methods as compared to static analysis. API calls were extracted and passed on to CNN for classification of dynamic analysis. In this dynamic analysis, the application of RNN, LSTM, and CNN was employed for the classification of malware. In dynamic analysis, the malware behavior is analyzed in a dynamic controlled environment.

Malware Classification using Image Processing

Malware attacks are increasing day by day. Many malware are existing from known malware, we need to avoid this malware to protect from attacking of new viruses. The malware variants are similar in structure, digital signal, and Image processing used for malware detection. This technique is used to convert the malware binaries into grayscale images, when we observe the malware family, it appears to be quite similar in layout and texture. The image processing technique is very fast when compared to static analysis and dynamic analysis. It contains high accuracy.

IV. SYSTEM ARCHITECTURE

Malware is detected by finding a match with the virus definition dataset that is updated from time to time. This technique is called as signature-based detection. However, once a brand-new malware variant is found, it could not notice it. However, this needs in deep domain data for the process of reverse engineering. To avoid this detection, hackers use polymorphism as Associate in Nursing obfuscation technique. As this method may be a resource-relative task, presented this as a 3 step method. Within the initiative, malware is unpacked. In step 2, the possible is disassembled. In step 3, the API decision is extracted. Step four involves API decision mapping and applied math feature analysis. Here it constitutes of machine learning techniques. Recently with the rise in malware attacks and obfuscated malware, several researches' area unit up machine learning algorithms for malware detection. Machine learning algorithms (MLAs) require feature engineering and have choice. Varied options are obtained through Static and Dynamic analysis. Static Analysis is done by capturing the data from executables while not running it. Dynamic Associate in Nursingalysis is finished by running the malware in an isolated setting. Dynamic Analysis is Associate in Nursing economical and long resolution for malware detection. However, deploying it might take an extended time to research the possible. Anti-malware programs usually use a hybrid of static and dynamic analysis. In recent times, Deep learning is improved a great deal. Here feature engineering isn't needed as a result of it'll learn.

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502

ML and DL ALGORITHMS

1. SVM Algorithm

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression issues. It uses the Kernel trick technique to transform the data and finds an optimal boundary between the outputs based on the transformations.

2. KNN Algorithm

K-Nearest Neighbour is also a supervised machine learning algorithm which assumes the similarity between the new data and available data and put the new data into the category that is similar to available categories. It stores all the present data and classifies a new data based on similarity. Used to solve classification and regression problems.

3. Naïve Bayes Algorithm

The Naïve Bayes algorithm is a technique based on the Bayes theorem with an assumption of independence among predictors.

4. CNN

Convolutional Neural Network (CNN) is picture process technique is used for image recognition and classification. DL algorithm therefore acknowledges objects in a picture by employing CNN.

5. LSTM

Long Short-Term Memory (LSTM) is used for classifying, processing, and making predictions based on time series data.

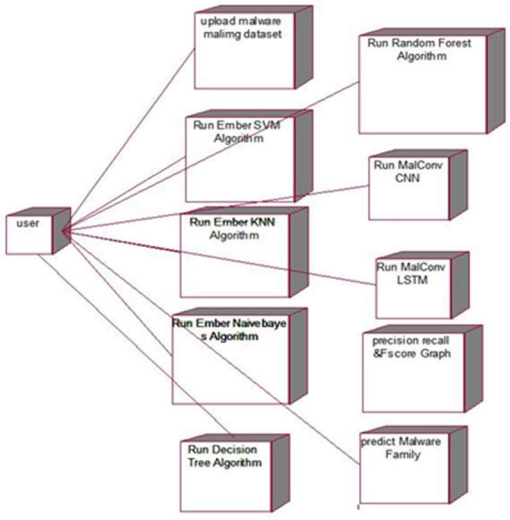
6. Decision Tree Algorithm

Decision Tree Algorithm which belongs to supervised MLAs family performs both classification and regression by forming a tree type flowchart and takes all the possible outcomes and comes to a conclusion.

7. Random Forest Algorithm

Random Forest is collection of decision trees used for large datasets which is also used to avoid classification and regression problems. Random forest also belongs to the family of supervised MLA. It is used when interpretability is a minor concern.

BLOCK DIAGRAM



SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502

Figure 1: Block diagram of proposed system design.

V.FINDINGS

In this section we will be finding out the accuracy of each and every ML and DL algorithm. We also discuss about the recall, precision, and FScore measures.

Accuracy is the measure of closeness of a calculated value to its actual value. Here we find 70 - 85% of accuracy in order to detect the malware using deep learning and machine learning.

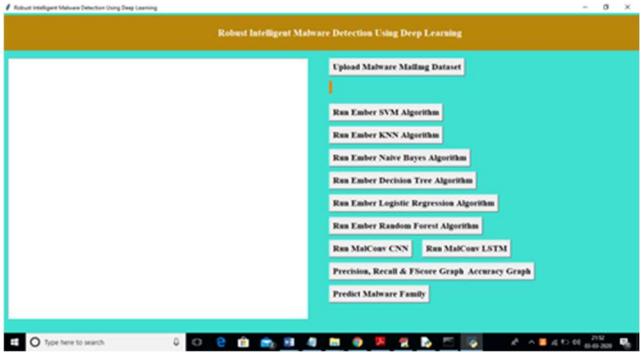
Precision refers to the degree to which a process will repeat the same value. Precision is the measure of quality.To detect the malware using deep learning we get the recall percentage as 70 - 85% .

Recall is the measure of how good a model is correctly predicting positive classes. Recall is the measure of quantity. He we obtain the recall percentage for detecting malware is about 70 - 85%.

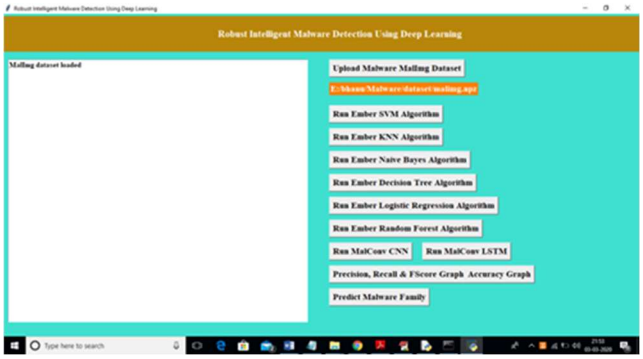
FScore/FMeasure measures te model's accuracy on a dataset. We get the Fmeasure of 70 - 85% for the detection of malware using deep learning and machine learning.

VI. RESULTS

Below screenshots shows the implementation of malware detection using deep learning algorithms.

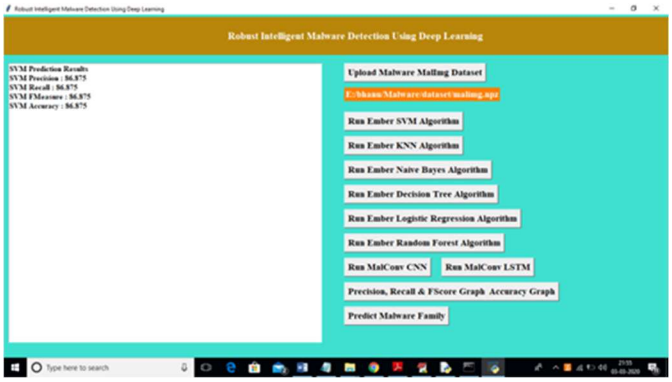


In above screen click on 'Upload Malware MalImg dataset' button to upload dataset.

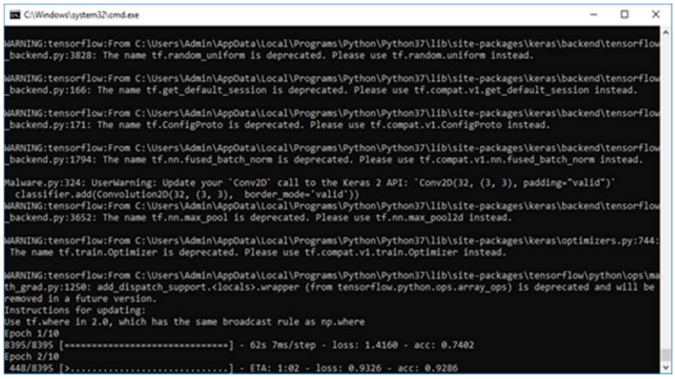


Now click on 'Run Ember SVM algorithm' button to read malware dataset and generate train and test model and then apply SVM algorithm to calculate its prediction accuracy, FSCORE, Precision and Recall. If algorithm performance is good then its accuracy, precision or recall values will be closer to 100.

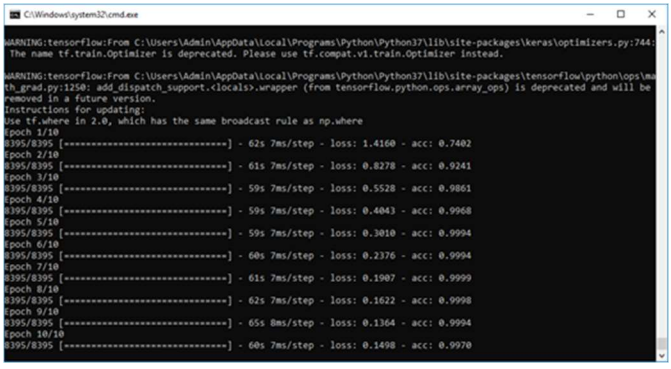
SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502



In above screen we got SVM precision, recall and FSCORE. Now click on 'Run Ember KNN Algorithm' button to get its performance. Similarly, click on all the remaining algorithms. Then we get precision, recall and FSCORE.



In above console it will take 10 epochs iteration and for each iteration it calculate accuracy for 8395 malware data. So you need to wait till all 10 epochs completed then u will get its performance details.



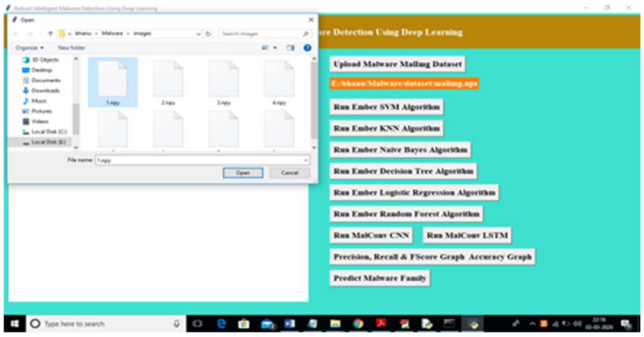
In above screen we can see CNN complete all 10 epochs and after that we will get accuracy details in the above main screen. So you need wait till all 10 epochs completed, then you will be able to see the performance details of CNN algorithm.

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502

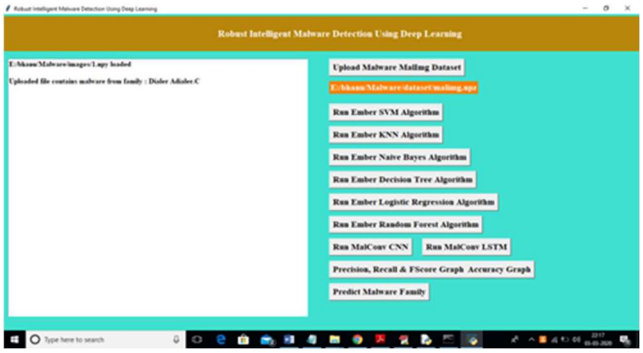
In above screen we can see precision graph for all algorithms and CNN get better performance. In above graph x-axis represents algorithm name and y-axis represents precision value and now close above graph to get recall graph.

Now click on accuracy button to get accuracy graph.

Now click on 'Predict Malware Family' button and upload binary file to get or predict class of malware



In above graph I am uploading one binary file called 1.npy and below is the malware prediction of that file.



In above screen we can see uploaded test file contains 'Dialer Adialer.C' malware attack.

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502
Similarly u can upload other files and predict class.

VII .CONCLUSION

In this paper, we have a tendency to project a replacement framework for police investigation malware underneath ever-changing environments. The projected framework is predicated on a denoising auto-encoder, that permits United States of America to extract sturdy and helpful options to reinforce the detection accuracy underneath ever-changing environments. Specifically, the denoising auto-encoder is employed as a building block throughout the coaching of deep neural networks. The projected model is employed to be told a way to reconstruct malware when applying noise thereon. This can be helpful to extract options that area unit sturdy against anti-analysis techniques and unstable environments. Our model was enforced employing a real-world dataset. The results demonstrate the effectiveness of the projected framework compared to the progressive malware detection ways.

REFERENCES

[1] Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M. J., Levi, M., ... & Savage, S. (2013). Measuring the cost of cybercrime. In The economics of information security and privacy (pp. 265-300). Springer, Berlin, Heidelberg.

[2] Li, B., Roundy, K., Gates, C., & Vorobeychik, Y. (2017, March). LargeScale Identification of Malicious Singleton Files. In Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy (pp. 227-238). ACM.

[3] Alazab, M., Venkataraman, S., & Watters, P. (2010, July). Towards understanding malware behaviour by the extraction of API calls. In 2010 Second Cybercrime and Trustworthy Computing Workshop (pp. 52-59). IEEE.

[4] Tang, M., Alazab, M., & Luo, Y. (2017). Big data for cybersecurity: vulnerability disclosure trends and dependencies. IEEE Transactions on Big Data.

[5] Alazab, M., Venkatraman, S., Watters, P., & Alazab, M. (2011, December). Zero-day malware detection based on supervised learning algorithms of API call signatures. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 171-182). Australian Computer Society, Inc..

[6] Alazab, M., Venkatraman, S., Watters, P., Alazab, M., & Alazab, A. (2011, January). Cybercrime: the case of obfuscated malware. In 7th ICGS3/4th e-Democracy Joint Conferences 2011: Proceedings of the International Conference in Global Security, Safety and Sustainability/International Conference on e-Democracy (pp. 1-8). [Springer].

[7] Alazab, M. (2015). Profiling and classifying the behavior of malicious codes. Journal of Systems and Software, 100, 91-102.

[8] Huda, S., Abawajy, J., Alazab, M., Abdollalihian, M., Islam, R., & Yearwood, J. (2016). Hybrids of support vector machine wrapper and filter based framework for malware detection. Future Generation Computer Systems, 55, 376-390.

[9] Raff, E., Sylvester, J., & Nicholas, C. (2017, November). Learning the PE Header, Malware Detection with Minimal Domain Knowledge. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (pp. 121-132). ACM.

[10] Rossow, C., Dietrich, C. J., Grier, C., Kreibich, C., Paxson, V., Pohlmann, N., & Van

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6877502
Steen, M. (2012, May). Prudent practices for designing malware experiments: Status quo and outlook. In Security and Privacy (SP), 2012 IEEE Symposium on (pp. 65-79). IEEE.
[11]Vinayakumar Ravi, Mamoun Alazab, Soman Kp, Prabaharan Poornachandran (2019, April). Malware Analysis using machine learning and deep learning architectures IEEE Access PP(99):1-1.