INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

ROBUST INTELLIGENT MALWARE DETECTION USING DEEP LEARNING

[1] Mrs. D. SRAVANI , [2] J.MANIRAJ, [3] A.JAHNAVI REDDY, [4]MA.IMRAN

1. *Assistant Professor Department of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Rangareddy (TS).India.*

Email-: Damanapeta@gmail.com

[2,3,4.]B.Tech StudentstDepartment of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Rangareddy (TS).India.

Email-: [3.]manirajjakkoju@gmail.com[,2]akitijahnavi@gmail.com, [4.]mohammadabdulimran086@gmail.com

Abstract- Malicious software or malware continues to pose a major security concern in this digital age as computer users, corporations, and governments witness an exponential growth in malware attacks. Current malware detection solutions adopt Static and Dynamic analysis of malware signatures and behavior patterns that are time consuming and ineffective in identifying unknown malwares. Recent malwares use polymorphic, metamorphic and other evasive techniques to change the malware behaviors quickly and to generate large number of malwares. By using the advanced MLAs such as deep learning, the feature engineering phase can be completely avoided. The train and test splits of public and private datasets used in the experimental analysis are disjoint to each other's and collected in different timescales. In addition, we propose a novel image processing technique with optimal parameters for MLAs and deep learning architectures.

KEYWORDS: Malware, Machine Learning, Deep Learning, CNN, LSTM, Robust.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

1. INTRODUCTION

In this digital world of Industry 4.0, the rapid advancement of technologies has affected the daily activities in businesses as well as in personal lives. Internet of Things (IoT) and applications have led to the development of the modern concept of the information society. However, security concerns pose a major challenge in realising the benefits of this industrial revolution as cyber criminals attack individual PC's and networks for stealing confidential data for financial gains and causing denial of service to systems. Such attackers make use of malicious software or malware to cause serious threats and vulnerability of systems. A malware is a computer program with the purpose of causing harm to the operating system (OS). A malware gets different names such as adware, spyware, virus, worm, trojan, rootkit, backdoor, ransomware and command and control (C&C) bot, based on its purpose and behaviour. Detection and mitigation of malware is an evolving problem in the cyber security field. As researchers develop new techniques, malware authors improve their ability to evade detection.

2. LITERATURE SURVEY

Measuring the Cost of Cybercrime

In this paper we present what we believe to be the first systematic study of the costs of cybercrime. It was prepared in response to a request from the UK Ministry of Defence following scepticism that previous studies had hyped the problem. For each of the main categories of cybercrime we set out what is and is not known of the direct costs, indirect costs and defence costs – both to the UK and to the world as a whole. We distinguish carefully between traditional crimes that are now 'cyber' because they are conducted online (such as tax and welfare fraud); transitional crimes whose modus operandi has changed substantially as a result of the move online (such as credit card fraud); new crimes that owe their existence to the Internet; and what we might call platform crimes such as the provision of botnets which facilitate other crimes rather than being used to extract money from victims directly. As far as direct costs are concerned, we find that traditional offences such as tax and welfare fraud cost the typical citizen in the low hundreds of pounds/Euros/dollars a

year; transitional frauds cost a few pounds/Euros/dollars; while the new computer crimes cost in the tens of pence/cents. However, the indirect costs and defence costs are much higher for transitional and new crimes. For the former they may be roughly comparable to what the criminals earn, while for the latter they may be an order of magnitude more. As a striking example, the botnet behind a third of the spam sent in 2010 earned its owners around US$2.7m, while worldwide expenditures on spam prevention probably exceeded a billion dollars. We are extremely inefficient at fighting cybercrime; or to put it another way, cyber- crooks are like terrorists or metal thieves in that their activities impose disproportionate costs on society. Some of the reasons for this are well-known: cybercrimes are global and have strong externalities, while traditional crimes such as burglary and car theft are local, and the associated equilibria have emerged after many years of optimisation.

As for the more direct question of what should be done, our figures suggest that we should spend less in anticipation of cybercrime (on antivirus, firewalls, etc.) and more in response – that is, on the prosaic business of hunting down cyber-criminals and throwing them in jail.

Large-Scale Identification of Malicious Singleton Files.

We study a dataset of billions of program binary files that appeared on 100 million computers over the course of 12 months, discovering that 94% of these files were present on a single machine. Though malware polymorphism is one cause for the large number of singleton files, additional factors also contribute to polymorphism, given that the ratio of benign to malicious singleton files is 80:1. The huge number of benign singletons makes it challenging to reliably identify the minority of malicious singletons. We present a large-scale study of the properties, characteristics, and distribution of benign and malicious singleton files. We leverage the in-sights from this study to build a classifier based purely on static features to identify 92% of the remaining malicious singletons at a 1.4% percent false positive rate, despite heavy use of obfuscation and packing techniques by most malicious singleton files that we make no attempt to de-obfuscate. Finally, we demonstrate robustness of our classifier to important classes of automated evasion attack.

3. EXISTING SYSTEM:

Microsoft Defender for Endpoint uses a combination of machine learning and signature-based detection methods to detect and prevent malware infections

Darktrace: Darktrace is an AI-powered network security solution that uses machine learning algorithms to analyze network traffic and detect potential cyber threats.

DISADVANTAGES OF EXISTING SYSTEM:

They can only deal with known attacks.

Cannot detect unknown instances of malware.

4. PROPOSED SYSTEM:

In proposed a methodology to represent binaries into image representation. This can preserve the sequential information of byte codes and it is similar to.

The proposed method converts the byte code into byte streams and thereby this method is able to preserve the sequential order of binary code.

Various deep learning architectures such as CNN and bidirectional LSTM and combination of CNN and bidirectional LSTM architectures are evaluated with sampling and as well as without sampling techniques to handle the samples equally across all the classes.

ADVANTAGES OF PROPOSED SYSTEM:

Fast and reliable malware analysis.

It utilizes multilayer approach like ML.

5. MODULES:

USER

A new proposal of a scalable and hybrid framework, namely ScaleMalNet which facilitates to collect malware samples from different sources in a distributed way and to apply pre-processing in a distributed manner. The framework has the capability to process large number of malware samples both in real-time and on demand basis.

MALWARE CLASSIFICATION

Several security researchers have applied domain level knowledge of portable executables (PE) for static malware detection. At present, analysis of byte n-grams and strings are the two most commonly used methods for static malware detection without domain level knowledge. However, the ngram approach is computationally expensive and the performance is considerably very low . It is often difficult to apply domain level

knowledge to extract the necessary features when building a machine learning model to distinguish between the malware and benign files. This is due to the fact that the windows operating system does not consistently impose its own specifications and standards. Due to constantly changing specifications and standards from time to time, the malware detection system warrants revisions to meet future security requirements. To address this, has applied machine learning algorithms (MLAs) with the features obtained from parsed information of PE file. They adopted formatting of agnostic features such as raw byte histogram, byte entropy histogram which was taken from, and in addition employed string extraction.

DEEP NEURAL NETWORK (DNN)

A feed forward neural network (FFN) creates a directed graph in which a graph is composed of nodes and edges [16]. FFN passes information along edges from one node to another without formation of a cycle. Multi-layer perceptron (MLP) is a type of FFN that contains 3 or more layers, specifically one input layer, one or more hidden layer and an output layer in which each layer has many neurons, called as units in mathematical notation. The number of hidden layers is selected by following a hyper parameter tuning approach. The information is transformed from one layer to another layer in forward direction without considering the past values. Moreover, neurons in each layer are fully connected. An MLP with n hidden layers can be mathematically formulated as given below:

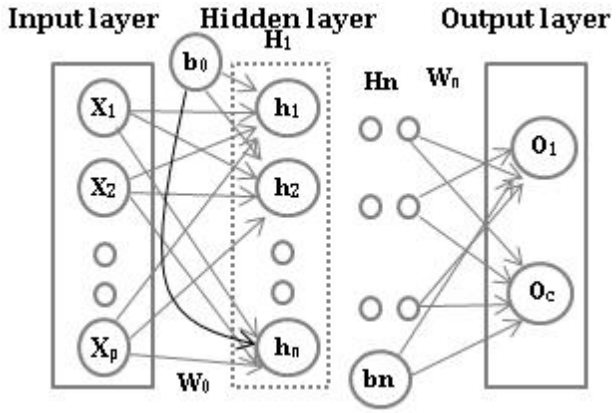

Fig 1: DNN MODEL

CONVOLUTIONAL NEURAL NETWORK (CNN):

Convolutional network or convolutional neural network or CNN is supplement to the classical feed forward network (FFN), primarily used in the field of image processing . It is where all connections and hidden layers and its units are not shown. Here, m denotes number of filters, ln denotes number of input features and p denotes reduced feature dimension, it depends on pooling length. In this work,

CNN network composed of convolution 1D layer, pooling 1D layer and fully connected layer. A CNN network can have more than one convolution 1D layer, pooling 1D layer and fully connected layer. In convolutional 1D layer, the filters slide over the 1D sequence data and extracts optimal features.

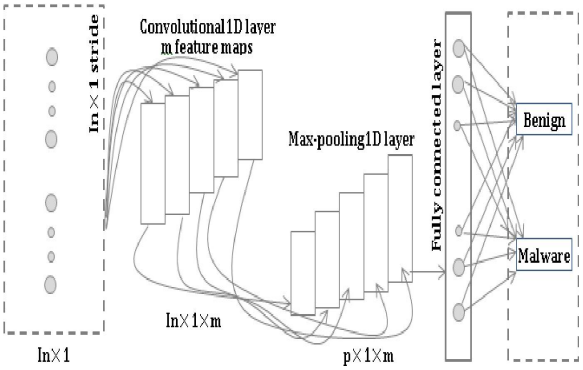

Fig 2: CNN ARCHITECTURE

6. RESULTS:

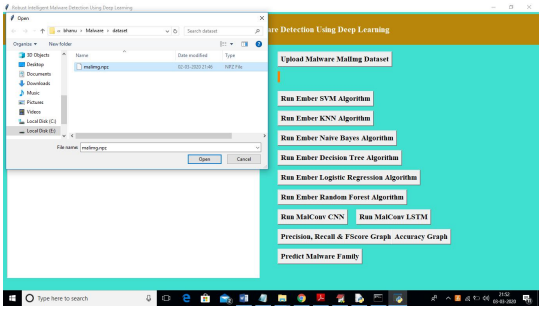

Fig 3: I am uploading 'malimg.npz' binary malware dataset and after uploading dataset will get below screen.

Fig 4: Now click on 'Predict Malware Family' button and upload binary file to get or predict class of malware.

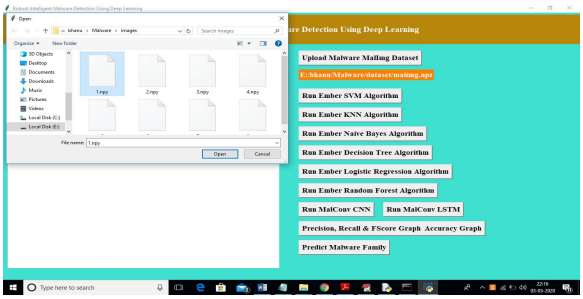

Fig 5: I am uploading one binary file called 1.npy and below is the malware prediction of that file.

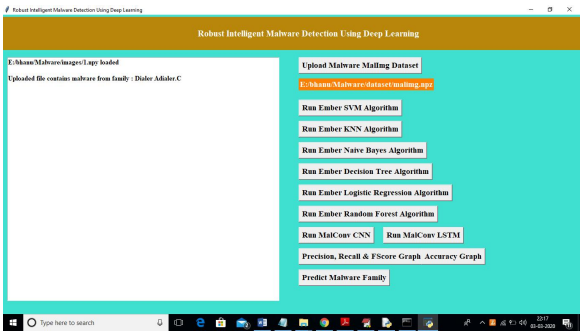

Fig 6: we can see uploaded test file contains 'Dialer Adialer.C' malware attack. Similarly u can upload other files and predict class.

7. CONCLUSION

This paper evaluated classical machine learning algorithms (MLAs) and deep learning architectures based on Static analysis, Dynamic analysis and image processing techniques for malware detection and designed a highly scalable framework called ScaleMalNet to detect, classify and categorize zero day malwares. This framework applies deep learning on the collected malwares from end user hosts and follows a two stage process for malware analysis. In the first stage, a hybrid of Static and Dynamic analysis was applied for malware classification. In the second stage, malwares were grouped into corresponding malware categories using image processing approaches.

8. 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). Large-Scale 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)..