International Journal of Innovative Science and Modern Engineering (IJISME)
ISSN: 2319-6386 (Online), Volume-6 Issue-6, April, 2020

Malware Detection using Deep Learning Methods

Nourin N.S, Sulphikar A

Abstract: ***Rapid development of the internet leads the malware to become one of the most significant threads nowadays. Malware, is any kind of program or file which would adversely affect the computer users in a harmful way. Malware exist in different forms which includes worms, viruses in computer, Trojan horses, etc. These malicious contents can degrade the overall performance of the system. It includes activities like stealing, encrypting or deleting sensitive data, etc. without the consent of the user. Malware detection is a milestone in the field of computer security. For detecting malware many methods have been evolved. Researchers are mainly concentrated in malware identification methods based on machine learning. Malware can be detected in two ways. They are static approach and dynamic approach. This paper mainly deals with the current challenges faced by malware detection methods and also explores a categorized new method in machine learning. The methods discussed here are combined static and dynamic approach, random forest, Bayes classification. This work will help in cyber security area and also which will help the researchers to do efficient researches.***

Keywords: Computer Security, Dynamic Analysis, Machine Learning, Malware Detection, Static Analysis.

I. INTRODUCTION

Recently, it was found that various kinds of malware are kept increasing by expeditious use of internet. Different kinds of malware are Worm, Virus, Trojan-horse, etc. The goal behind developing malware is to collect the personal information without the knowledge of users. According to previous studies, thousands of new malwares are created every single day. The nature of malwares being complex makes it difficult to be detected using the traditional detection techniques like signature-based detection and behavior-based detection. Thus, the methods for detection and classification of malware need to be improved to do the required prevention mechanism. Different kinds of software provide wealth resources to users which results in danger. As a result, malware identification is one of critical issue faced in computer security. Many machine learning methods has been used for efficient detection of malwares. Malware is a program that makes frameworks which will accomplish something that an assailant needs it to do. Malware designers will redesign the old code with the new components instead of preparing the new codes for the malware generation. The analysis of these malwares can be partitioned mainly into two classes that include static analysis methods and dynamic analysis methods.

Revised Manuscript Received on April 22, 2019.
* Correspondence Author
Nourin N.S*, Department of Computer Science and Engineering, LBSITW, Trivandrum, India.
Sulphikar A, Department of Computer Science and Engineering, LBSITW, Trivandrum, India.

Existence of zero-day malwares made the commercial antivirus vendors to offer 100% protection for their personal system. If signatures are used to catch the malicious code, then the method is termed as signature based. They have high detection ratio, but at the same time will be vulnerable in some situations. The detection of zero- day is difficult due to the use of new malware signatures. To overcome this limitation of signature-based detection technique the behavior-based detection was developed. On the other hand, behavior-based detection identifies newly created malware whose signatures are not known. The limitation of this method is that it takes more space which will affect the system performance. Data mining approaches like Bayesian methods, tree-methods, etc. are used to detect malwares which will automatically and accurately classify the malicious executables. Improvement on automatic malware detection is required which helps in the tremendous growth in number of different malwares. And to deal with the above problem, machine learning techniques where used as a suitable solution.

II. RELATED WORKS

Nowadays, the internet usage is dramatically increased in all sectors. Thereby the key and major threats in internet is malicious software package, named as a malware. The malware or malicious contents which are mainly designed by attacker's that has the full authorization to vary their code as they wish and it will propagate through internet. As of now static and dynamic methods are replaced by machine learning techniques which are mostly used for the malware classifications. In Shijo and Salim et al. [1] discusses a combined approach for malware detection. Where they are combining the features of both static and dynamic analysis This approach helps the unknown executable files to be analyzed and classified. With the help of known malware and benign dataset classification can be efficiently done by machine learning approach. By integrating static and dynamic method, which utilize the benefits of both the methods and thereby enhanced the classification accuracy to 98.7%. By this they have proven that the combined features of both those methods gave high detection accuracy. In Badwik and Bagdi et al. [2] illustrated a probabilistic discriminative model. Here uses a logistic regression technique for the identification of malware in android applications. The source code which is decompiled will produce accurate detection result. There are some limitations which are reflected by the static technique.

The dead code will result in unreliable feature extraction and representation.

In Salehi et al. [3] proposed a scalable identification method for high complexity malware identification. Here classification is carried out by means of features like Application Programming Interface (API) calls and a combined product of API names and their arguments. Thus, malware which cannot quickly identified by signature method can easily detected by using this method. To make analysis time efficient feature selection techniques are applied. One of the machine learning technique like Random forest (RF) algorithm shows the best accurate rate of 98.4% when compared with other techniques for the classification. In Yerima et al. [4] illustrated a method for detecting malwares that are spreading in smartphones. Here uses a machine learning approach based on Bayesian classification which uncover all unknown malwares through static analysis. Large dataset containing malwares samples are considered for this method thereby improved the high accuracy detection rate. The empirical result shows this effective approach is one of the best solutions for detecting unknown malwares in the android.

Lee and Lue et al. [5] presented a multi stages method to analyze the malware behavior by code review and live testing. In the first step, the malware sample is extracted from the security repository and then categorized into different malware categories (Trojan, worm, viruses, etc.). In the second step, the malware runs on the device emulator platform to analyze the malware interaction between the device and the user. Therefore, the enhanced isolation of the data and the access to the data in the systemalsohelps.

III. MALWARE ANALYSIS APPROACHES

In recent years the malware detection has become one of the main concern areas of research. To deal with the growing amount of malware, several techniques have been proposed. That includes static analysis approach, dynamic analysis approach and machine learning methods.

Static Analysis for Malware Detection

As the increase in count of malicious code, it becomes necessary to take security steps against those malware attacks. One of the traditional methods used for these detections are spyware scanners. This type of defensive method can be easily eluded by simple code transformations. To overcome this limitation static analysis technique are came up and it provides better detection of malware when compared with traditional methods.

Static type of analysis usually looks at portable executable (PE) files without having to run them in an administrative environment. One of the common methods that comes under this analysis is signature-based. This method uses the basis of specific manually designed features. It always checks against the signature pattern when a new malware detects. If the patterns are similar then it will be considered as malware. Signatures will not lay hold of new features which are extracted from newly released malwares.

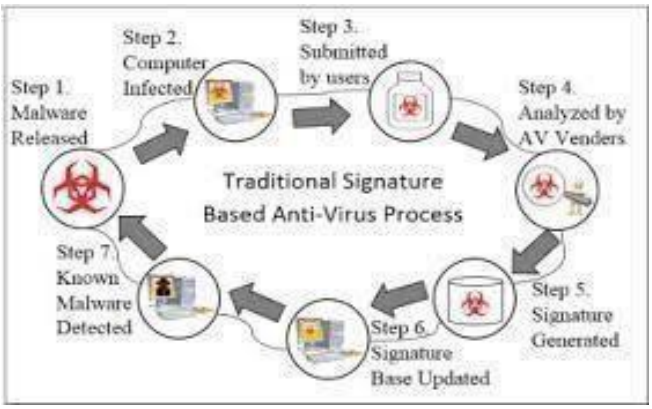

Fig.1. Steps in signature-based malware detection

Signature-based detection method steps are depicted in Figure 1. When a new malware is released it will be analyzed and the signatures corresponding to the malware is generated. Later on, it will be cross checked with the signatures in the database where all known detected malware signatures reside.

In static analysis the patterns which are used for the detection are extracted like [6]:

i. **Windows Application interfaces (APIs):** which are used as the communication calls with operating system. API calls disclose the behavior of programs and this is used for detection process.

ii. **Control Flow Graph (CFG):** illustrates control flow of a program. In PEs it extracts the structure of program.

iii. **Opcodes:** which identifies the operation which is to carried out by CPU and feature is extracted by measuring the similarity between sequences.

iv. **Strings:** holds the semantic information which are inevitable that helps in detection.

Analysis done Dynamically for Malware Detection

Dynamic analysis technique is also known as behavior-based analysis. In this method files will run in a supervised environment and detect the malware affected files. The supervised environments are Virtual Machine, emulator, debuggers, simulators. When compared with static analysis method, this is an efficient approach to analyses the affected file without disassembling it. In the case of dynamic analysis, it can able to detect the unknown malware too. Features that are extracted from the files helps to train the model for the accurate detection, the features include [6]:

i. **API calls** which helps to reveal malware behavior.

ii. **File system, Windows registry, network**

iii. **APIs sequence**

iv. **User API, native API:** these features are used in open source emulators that gives automated malware analysis.

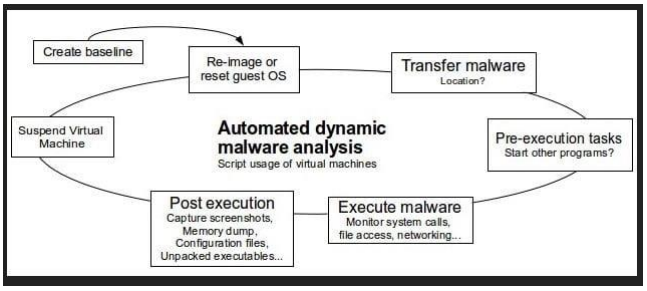

Fig.2. Process, Monitor and Explore the Dynamic Analysis

The steps which are handling in dynamic analysis in a virtual environment are shown in Figure 2. One of the main problems faced by dynamic analysis are time vigorous and resource consumption.

Machine Learning Based Malware Detection

Machine learning for malware detection mainly consist of two steps:

i. The first one is extracting features from the binary files

ii. Secondly, it mainly deals with the learning phase which trains a classification model using features that are extracted from the first phase.

Due to the tremendous increase in the number of malware family, machine learning methods have been used to detect unknown malware samples. Ancient anti-virus software becomes ineffective and expensive. Schultz el at. [7] first proposed a data mining framework to find new malware precisely and automatically. The samples in their data sets were automatically identified and used patterns which identify a set of new malicious binaries. Comparing their identification methods with the traditional signature method, this method is currently double identification of rates for new malicious enforcement.

Kong and Yoon et al. [8] proposed a new classification method that uses a combination of Principal Component Analysis and Probabilistic K-Nearest Neighbor (PKN) methods. That combination gives a good result to the classification of malware. Here the k-fold cross validation method is used in comparison with the K-nearest neighbor (KNN) method. This makes PKNN a better method compared to KNN. With PKNN, classifications that are heavy to perform by KNN can be easily classified.

Kolter and Maloof [9] proposed an efficient method by using more than 255 million unique n-grams as feature for training the model. They have observed various inductive methods, that including naive Bayes, decision trees, SVM, and boosting. Finally, found that boosted decision trees outperformed other inductive methods. Boosted decision trees attain a true-positive rate of 0.98 and a desired false-positive rate of 0.05 for 291 malicious executables. If TPR is higher and FPR is lower in percentage then that model is considered as the best one.

Some of researches have attempted to combine both behavior and static features. They expect a good result by integrating both behavior and static features. A hybrid malware detection tool known as OPEM, has been proposed by Santos et al. [10]. This tool is based on machine learning algorithm. It makes use of some features derived from both static and behavior analysis of malicious code. With the help of SVM classifier 96.60% detection rate is obtained. Their experiment proved that the better performance can be acquired with this hybrid method.

The above-mentioned malware identification based on machine learning has achieved a comparatively good result. We all know that; the malware will grow continuously year by year. For controlling these increment necessary steps should be taken. By this we conclude that there are many different methods which helps in malware identification. From those the most efficient and best techniques are used in the machine learning area. As we know, the malware identification becomes a crucial role in computer security field.

IV. DATA SET USED

Data set traditionally used for detection of the malware were in the binary file format which was too difficult to analyze and visualize due to its high computational cost. Thus, malware files were converted to images for easy detection and classification. Malware developers change small portions of the original code which in turn produce new malware variants. Images can identify these small differences and yet retain the global structure. Nowadays, image textures used for classification provide more complex features in terms of obfuscation techniques.

V. PROGRESS IN MALWARE DETECTION

The information needed to write and execute customized malware samples has actually decreased significantly, though the fact that malware is more complex. The main reason for this is the availability of automated tools for creating malware. These tools allow less experienced attackers to create and customize malware programs to perform cybercrime. And furthermore, many samples of various malware have to go through anti-malware detection. Instead of writing new malware from scratch, malware creators often use the most profitable way to use existing samples.

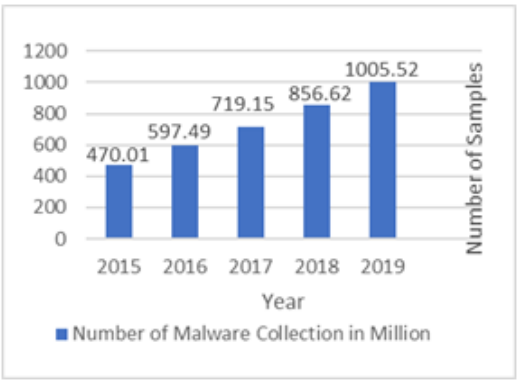

Fig.3. The increment of malware samples [11]

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

Therefore, such versions of malware models have evolved into a stream-line process. Since malware versions are often created automatically and faster, malware makers can replace older malware. Latest malware is the latest feature of the malware mutation process. Based on data provided by the AV-Test Institute, Figure 3 shows the increasing trend of malware models from 2015–2019. This proves that the number of malware samples has increased and the number of new malware samples collected has reached hundreds of millions. Then it gradually increased to 1005.52 million in 2019. Unfortunately, this trend is likely to continue. Malware continues to be one of the biggest security threats faced by Internet users and, therefore, the detection and remediation measures in terms of Internet security will be more limited.

VI. RESULTS

From the related works mentioned, malware has been detected using many different ways. In the below Table 1 shows the methods and their corresponding accuracy in the area of malware detection. Which shows that combined method applied in [1] gives more accuracy when compared with individual accuracy of those techniques. That means it integrate both advantages in the two static and dynamic methods.

Tab. 1. Results in malware detection methods.

SL NO	METHOD	REMARKS
1	Combined Static and Dynamic Approach [1]	98.7%
2	Random Forest [3]	98.4%
3	Bayesian Classification [4]	97%

VII. CONCLUSION

In this paper, efficient methods used for malware detection are discussed. These include static analysis approach, dynamic analysis approach, and machine learning methods. Static analysis malware detections are traditionally used techniques but there are some limitations which leads to the development of the dynamic analysis malware detection. For more sophisticated methods machine learning methods were introduced. Due to the scalability, rapidity and flexibility of malware family machine learning methods for example Support Vector Machine, Decision Tree have been used recently to identify and classify not known malware samples. After going through all these methods, we can conclude that machine learning techniques are comparatively more suitable and effective in the field of malware detection. This gives a better detection rate and also will cope with the new upcoming malware samples. This paper helps the researchers to do efficient researches in the area of malware detection area and thereby using it in the field of cyber security.

REFERENCES

1. Shijo, P.V. & Salim, A. (2015). Integrated Static and Dynamic Analysis for Malware Detection. Procedia Computer Science. 46. 804-811. 10.1016/j.procs.2015.02.149.
2. Nisha Badwaik, Prof. Vijay Bagdi. “A SURVEY ON SUPERVISED METHOD FOR DETECTION OF MALWARE”. In International Journal Of Engineering Science And Research Technology.ISSN (Online): 2277-9655 Impact Factor: 4.116, Vol. 5, Issue 6, June 2016.
3. Z. Salehi, , M. Ghiasi and A. Sami, " A miner for malware detection based on API function calls and their arguments," The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), Shiraz, Fars, 2012, pp. 563-568. doi:10.1109/AISP.2012.6313810.
4. S. Y. Yerima, S. Sezer,G. McWilliams.“Analysisof Bayesian Classification Based Approaches for Android Malware Detection” IET Information Security, Volume 8, Issue 1, January 2014, p. 2536, Print ISSN 1751-8709, Online ISSN 1751-8717.
5. Minzheng, pratick p.c. Lee, and John C.S Lue,”ADAM:An Automatic & extension platform tostresstestandroid antivirus system”, Dept of CSE.J.
6. Sihwail, Rami & Omar, Khairuddin & Zainol Ariffin, Khairul Akram. (2018). A survey on malware analysis techniques: static, dynamic, hybrid and memory analysis. 8. 1662. 10.18517/ijaseit.8.4-2.6827.
7. M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo, “Data mining methods for detection of new malicious executables,” in Proceedings of the IEEE Symposium on Security and Privacy (S & P), pp. 38–49, May 2001.
8. Kang, Seungjun & Yoon, Ji. (2015). Probabilistic K-nearest neighbor classifier for detection of malware in android mobile. Journal of the Korea Institute of Information Security and Cryptology. 25. 817-827. 10.13089/JKIISC.2015.25.4.817.
9. J. Z. Kolter And M. A. Maloof, “Learning to Detect and Classify Malicious Executables In The Wild,” Journal Of Machine Learning Research, Vol. 7, Pp. 2721–2744, 2004.
   Santos, Y. K. Penya, J. Devesa, and P. G. Bringas, “N-grams-based file signatures for malware detection,” in Proceedings of the ICEIS 2009 - 11th International Confeenceon Enterprise Information Systems, pp. 317–320, May 2009.
10. https://www.av- test.org/en/statistics/malware/

AUTHORS PROFILE

Nourin N.S is pursuing (4th Semester) Master’s Degree in Computer Science and Engineering from LBS Institute of Technology for Women, Kerala, India affiliated under Kerala Technical University.

Sulphikar A currently, he is working as an Associate Professor in Computer Science and Engineering, LBS Institute of Technology for Women, Trivandrum, India, affiliated under Kerala Technical University