A Review on Malware Detection Schemes Using Machine Learning Techniques

[1]Priya Sharma, [2]Jyoti Arora
[1]Research Scholar, [2]Assistant Professor
Department of Computer Science & Engineering, Desh Bhagat University,
Mandi Gobindgarh

***Abstract* - Malware is a one type of software which can harm the computer's operating system and it may also steal the personal information from the computer. Malware can be made by using any programming language by the programmer. It is very difficult to define a malware with a single term or a single name. A malware can be considered as a malicious software or malcode or it is also known as a malicious code .Malware do the bulk of the intrusive activities on a system and that spreads itself across the hosts in a network. Malware detection techniques can be classified into 2 categories - the static analysis techniques and the dynamic analysis techniques. The static techniques involve looking into the binaries directly or the reverse engineering. The code for patterns is the same. This paper attempts to provide a brief survey of all the work that has been done in the field of malware detection. All literatures have been properly reviewed and discussed for their merits and demerits.**

***Keywords* - Malware Detection, Machine Learning, Pattern Recognition, Signature based technique**

I. INTRODUCTION

Malware is defined as software performing actions intended by an attacker, mostly with malicious intentions of stealing information, identity or other resources in the computing systems. There are different types of malware like adware, bots, Trojan horses, viruses, bugs, rootkits, spyware and worms. The dynamic analysis techniques involve capturing the behavior of the malware sample by executing it in a sandboxed environment or by program analysis methods and then use that for extracting patterns for each family of virus.

Static Based Detection Technique and other detection techniques

Signature based detection technique - This technique is most popular in Malware detection technique. This technique is used to identify the behavior of virus and its family. Signature detector can made a list of signatures which can be used to identify the virus and this list need to update day by day with signature of new virus. The signature of new virus is not available in the list so, it cannot detect. Signature based technique is fast, simple and reliable. Polymorphic and metamorphic virus is not used as signature detector.

- **Metamorphic malware** is malicious software that is capable of changing its code. Metamorphic malware have virus mutation to the next level and each new variant create different signature which is difficult to recognize. Metamorphic malware is more difficult for anti-virus software to recognize but not impossible. It represents the next class of virus which can construct the completely new variant after reproduction. Metamorphic contain morphing engine [3]
- **Polymorphic malware**:-Polymorphic malware changes encrypted code by adding an additional component. The variants created by polymorphic malware continuously change. Polymorphic malware contain polymorphic engine in anywhere in the virus body.

Anomaly based detection Technique - In this technique, any abnormal activity is occur in the program it can be arise. By using this technique, any abnormal activity which is arising in the system can give alert message .It is more reliable technique as compared to another technique because it can be detecting new virus. In which the detector can learn the behavior of host and it can detect the zero day attackers. Zero day attackers are the unknown of previous malware detector.This technique can be occurred in two phases training phase and detection phase. In the training phase the detector try to learn its normal behavior. The main advantage of anomaly based detection is that it can detect the zero day attacks. This technique also have some limitations such as its high false alarm rate and complexity.

Emulation based detection Technique - In this technique, it can detect the behavior of malware and the sequence of malware. This technique is used to decrease the time of detection. It is used to detect the polymorphic malware as well as metamorphic malware.[4].

Heuristics based detection - It can be used with 2 approaches: static approach and dynamic approach .In static approach, it is used to find out the known pattern for matching if it can be present in program. In dynamic approach, it can open other executable files which can modify its contents. It can be promise to detect the unknown malware and which it can make system more vulnerable by taking the real malware as another [5].

Dynamic detection techniques - Automated, dynamic malware analysis systems work by monitoring a programs execution and generating an analysis report summarizing the behavior of the program. These analysis reports typically cover activities (e.g., what files were created), Windows registry activities (e.g., what registry values were set), network activities (e.g., what files were downloaded, what exploit were sent over the wire), process activities such as when a process was created or terminated and Windows service activities such as when the service was installed or started in the system etc. Several of them are publicly available on the Internet. The main thing to note about dynamic analysis systems is that they execute the binaryfor a limited amount of time.

II. RELATED WORK

AmanJantanet.al describes a new framework for malware behavior identification and its classification. Malware writers try to avoid detection by using several techniques such as polymorphic, metamorphic and also hiding technique. In order to overcome that issue, we proposed a new framework for malware behavior identification and classification that apply dynamic approach. This framework consists of two major processes such as behavior identification and malware classification. These two major processes will integrate together as interrelated process in our proposed framework. Result from this study is a new framework that able to identify and classify malware based on it behaviors**.** The main purpose of malware is to break the computer operation.
In this, paper describes to detect the malware by using signature based matching technique which is popular and reliable technique. Signature based matching technique have limitations which cannot provide appropriate time to learn and understand threat. For the reason is that the detection process is based on string matching without knowing malware goals and behavior. In this research paper the author proposed new framework for malware identification and classification. The proposed framework has three main objectives. First objective is to develop architecture for secure environment for behavior-based malware analysis. Secondly is to implement comprehensive approach to conduct malware analysis. The third objective is to classified malware into new possible group using adapted AI technique. This is truly host-based framework that was designed for the window environment that has high potential of malware attacks.
Schmidt et al. [10] proposed a host based malware detection system for android platform. It is a signature detection method which applies static approach on executables.Blasing et al. [11] aims to perform static and dynamic approaches on android applications.
In [5]a novel method has been developed to detect leaked sensitive phone-related information. Firstly, it decrypts Objective-C binary and generate control flow graph. Presence of leaks - paths arising from functions obtaining sensitive resources is checked in the graph.
Zmist is an advanced metamorphic virus that (shows or proves) a set of polymorphic andmetamorphic code writing skills which include entry-point hiding/blocking, randomly using an added polymorphic (change secret codes into readable messages)or, code combination and arrangement and code (combination of different things together that work as one unit).
Chouchane and Lakhotia[8] proposed using “engine signatures” to assist in detecting metamorphic malware. Basically, this technique evaluates collected forensic evidence from x86 code segments through a code scoring function. This score is a measure of how likely it is that the code has been generated by a known instruction substituting metamorphic engine.
Wagener et al. [8] proposed a flexible and automated approach to extract harmful programs or apps behavior by watching/ noticing/ celebrating/ obeying all the system function calls (did/done/completed) in a virtualized execution (surrounding conditions). (things that are almost the same as other things) and distances between harmful programs or apps behaviors are figured out/calculated which allows classifying harmful programs or apps behaviors. The main features of this approach reside in coupling a sequence matching-up related method to figure out/calculate (things that are almost the same as other things) and use/take advantage of the Hellinger distance to figure out/calculate connected distances. The classification process proposed by this work is using phylogenetic tree. However, this way of doing things still has a limitation due to the wrongly classified a few harmful programs or apps behavior.
Bayer et al. was used an (able to be made bigger or smaller) clustering approach to identify and group harmful programs or apps samples that show almost the same behavior [10]. This approach also (does/completes) energetic/changing analysis to get the execution traces of harmful programs or apps programs using automated tools. The execution traces are generalized into behavioral profiles, which describe/show the activity of a program in more fuzzy and unclear terms. Then the profiles serve as input to a (producing a lot with very little waste) clustering set of computer instructions that allows handling sample sets larger than previous approaches in term of harmful programs or apps behaviors.
Bergeron et al. proposed a new approach for the static detection of harmful programs or apps code in executable programs [10]. This approach carried out directly on binary code using (related to the meaning of words) analysis based on behavior of unknown harmful programs or apps. The reason for targeting binary executables is that the source code of those programs that need to detect evil and cruel code is often not available. The first (or most important) goal of the research is to describe in detail practical methods and tools with (related to ideas about how things work or why they happen) foundations for the static detection.
Ulrich et al. presented an approach to improve the (wasting very little while working or producing something) of energetic/changing harmful programs or apps analysis systems [11]. It is to overcome the huge number of new evil and cruel files now appears. It is due to changes of only a few harmful programs or apps programs. The proposed system avoids carefully studying harmful programs or apps binaries that simply make up/be equal to changed (in a bad way) events of already carefully studied polymorphic harmful programs or apps. It can extremely reduce the amount of time needed/demanded for carefully studying a set of harmful programs or apps programs. The limitation of this approach is due to the changes of the behavior after the analysis process that are caused by the limitation of energetic/changing analysis.
Tabish et al. proposed harmful programs or apps detection that applied data mining which is based on the analysis of byte level file content [12]. This way of doing things also designed to provide protection against first day launched harmful programs or

apps. This non-signature based way of doing things has the possible ability to detect (before that/before now) unknown and new launch harmful programs or apps. It does not memorize specific byte-sequences or string that appearing in the actual file content. Standard data mining set of computer instructions was used to classify the file content of every block as (usual/ commonly and regular/ healthy) or possibly evil and cruel by separate and label it as harmless or harmful programs or apps. The proposed way of doing things was tasted using six different file types such as doc, exe, jpg, mp3, .pdf and zip. Six different types of harmful programs or apps that consist of alternative, Trojan, virus, worm, constructor and miscellaneous was used as dataset.

III. CONCLUSION AND FUTURE SCOPE

All the literatures related to the field of malware detection has been discussed and analysed. It is found that there are different techniques like permission based strategy, signature based techniques, pattern based techniques which are being utilized in this field. The paper attempts to survey and analyze each techniques and a brief discussion about them is provided. In future, new algorithms such as machine learning techniques such as classification and prediction can be utilized in this field. Also the same technique can be applied in other fields such as android malware detection etc.

IV. REFERENCES

[1] NwokediIdika and Aditya P. Mathur "A Survey of Malware Detection Techniques" pp.1-47, February 2, 2007.
[2] Dennis Distler "Malware Anylasis an introduction" pp.1-64, December 14, 2007.
[3] Vinod.P, Laxmi.V.,Chauhan.G "MetamOrphicMalware Exploration Techniques Using MSA signatures"In: International Conference on Innovations in Information Technology (IIT),pp 232-237(2012).
[4] Kumar,N.D., Mishra.L., Charan.M.S., Kumar.B.D " The New Age Of Computer Virus and Their Detection"In: International Journal of Network Security & Its Applications (IJNSA), Vol.4, pp 79-96 ,(2012).
[5] Rafique,M.Z.,Chen,P.,Huygens,C., Joosen,W "Evolutionary Algorithms for Classification of Malware Families through Different Network Behaviors" Pp, 1-8, (2014).
[6] Imtithal A. Saeed,AliSelamatand Ali M. A. Abuagoub "A Survey on Malware and Malware Detection Systems" In:International Journal of Computer Applications,Vol.67,pp 25-31,(2013).
[7] Agrawal,H., Bahler,.L, Micallef,J., Snyder,S., Virodov,A.: Detection of Global, Metamorphic Malware Variants Using Control and Data Flow Analysis. Pp 1-6 ,IEEE(2013).
[8] Emad,S.A., Hashemi,.S. " A General Paradigm for Normalizing Metamorphic Malwares".In:10th International Conference on Frontiers of Information Technology. pp 349-353(2012).
[9] Cai,Y.L.,Ji,.D., Cai,.D.F. "A KNN Research Paper Classification Method Based on Shared Nearest Neighbo". Pp 336-340, (2010).
[10] Baoli,.L, Shiwen,.Y., Qin"An Improved k-Nearest Neighbor Algorithm for Text Categorization".In:20th International Conference on Computer Processing of Oriental Languages, pp 1-6 ,(2003).