A Review on Malware and Malware Detection Techniques

Dr.P.Sujatha*[1], S.Sivasankari[2], R.Devi[3]

Department of Information Technology, School of Computing Sciences, VISTAS

1 sujinagi@gmail.com

2 sankarivs96@gmail.com

3 devi.scs@velsuniv.ac.in

***Abstract*— A computer virus may be a serious threat. The foremost recent reports emphasize that making malicious software system damages the pc system and a few of them cover the connected system within the network or web affiliation. Researchers and manufactures have taken much effort to supply anti-malware systems with effective detection strategies to secure the computers. Signature-based and Heuristic-based detection are the fundamental approach to spot the malware. In this paper, an in-depth review has been conducted on this state of affairs of malware infection and also the work done to boost anti-malware detection techniques.**

***Keywords*— Malware Detection, Signature Based Detection, Behavior Malware Detection, Heuristic-based detection**

I. INTRODUCTION

Presently the utilization of web is the most integral part of our life. At this point web program downloads different kind of PC software [1]. One downside of utilizing web is numerous PC frame work to assaults and get contaminated malware. The diverse names for malware are malevolent code, malignant program, pernicious executable and more. There are various types of malware such as Virus, Trojanhorse, Spyware, Scareware, Adware or Tropdoor and so on. A malware location framework is a system used to decide if a program has noxious goal or not [2]. The malware locator is utilized as a tool to defend the malware. It takes two sources of information, i) signature or social parameters of code, ii) program under inspection, employ its discovery techniques to detect whether the program has malware. The main purpose of this review paper is investigating the different forms of malware analysis and detection techniques. The detection of malware is an area of major concern not only to the research community but also to the general public. This paper gives a detailed study of various malware classification and detection techniques using data mining.

II. REVIEW OF LITERATURE

Galal et al. [22] proposed a behavior-based features model that defines malicious activitiesshown by malware example. To remove the proposed model, the authors initially perform dynamic examination on a generally late malware dataset inside a controlled virtual environment and catch hints of API calls conjured by malware examples. The traces are then generalized into high-level features refer to as actions. The proposed method is assessed using some renowned classification methods such as random forests, decision tree and SVM. The experimental results demonstrate that the classifiers attain high precision and acceptable outcome in the identifying the malware variants.

Sheen et al. [23] have considered Android-based malware for examination and an adaptable recognition component is planned to use multi-feature collaborative decision fusion (MCDF). The particular features of a malicious record like the consent-based features and the API call-based features are considered keeping in mind the end goal to give a superior discovery via preparing a gathering of classifiers and combining their choices utilizing collective approach in view of likelihood hypothesis. The execution of the proposed model is assessed on a gathering of Android-based malware including diverse malware families and the resultexhibit that the presented approach gives a superior execution than best in class troupe plans accessible.

Ming et al. [24] have presented a substitution attacks to cover comparable practices by harming behavior-based specifications. The key procedure for the attacks is to supplant a system call dependence graph to its semantically identical variations so that the comparable malware tests confidential unique family end up being characteristic. Accordingly, malware investigators need to put more endeavors into reconsidering the similar samples which may have been examined sometime recently. They distill common attacking strategies by mining more than 5200 malware tests' behavior specifications and execute a compiler-level model to automate replacement attacks. By assessing on the real malicious examples, the effectiveness of the proposed method to obstruct several behavior-based malware analysis tasks, such as clustering and malware comparison. Finally, they discussed likely countermeasures to support current malware protection.

Eskandari et al. [25] presented a novel hybrid approach, HDM-Analyzer, is shown which takes points of interest of dynamic and static investigation techniques for rising pace while ensuring the precision at a sensible level. HDM-Analyzer can anticipate the dominant part of basic leadership focuses on using the factual data which is gathered by component examination; along these lines, they have no performance overhead. The basic commitment of this paper is taking precision preferred standpoint of the element investigation and solidifying it into static examination keeping in mind the end goal to enlarge the accuracy of static investigation. Honestly, the execution overhead has been continued in learning stage; henceforth, it does not force on highlight extraction stage which is performed in examining operation. The exploratory result represents that HDM-Analyzer achieves better general exactness and time versatile quality than static and element investigation strategies.

III. MALWARE CLASSIFICATION

The classification of malware is a troublesome method. Software package that enables unauthorized user management of the system is malicious malware comes in numerous forms and classes [3]. The malicious code "any code added, changed, or removed from a software system in order to intentionally cause harm or subvert the intended function of the system" which describes malware as "a generic term that encompasses viruses, trojans, spywares and different intrusive code". Some of the canonical malwares are viruses, worms, and Trojan horses.

A. Virus

A computer virus propagates from one PC to a different PC by inserting the code into another program. A virus is a code that replicates by inserting itself into different programs. A program that a virus has inserted itself into is infected, and is cited as virus's host. A virus needs an existing host program in order to cause harm existing host. For instance, in order to get into an automatic data processing system, a virus could attach itself to some software package utility like a data processing application [5]. Launching the data processing application might then activate the virus which will duplicate itself and disable malware detectors enabled on the system.

B. Worm

It is a self –replicating program from one computer to a different by transmitting to copy of personal information via a network without user permission. A computer worm replicates itself by executing its own code independent of any other program. The primary distinction between a virus and a worm is that a worm does not need a host to cause harm. Another distinction between viruses and worms is their propagation model [6]. In general, viruses try to unfold through programs/files on a single computer. However, worms unfold via network connections with the goal of infecting as several computer systems connected to the network as possible.

C. Trojanhorse

A Trojan horse is an appearing to be something user information to detect the financial data, password, date of birth, mobile number etc. A Trojan horse is malware embedded by its designer in an application or system. The application or system appears to perform some useful function (e.g., give the local weather), but is performing some unauthorized action (e.g., capturing the user's keystrokes and sending this information to a malicious host). Trojan horses are typically associated with accessing and sending unauthorized information from its host [7]. Such Trojan horses can be classified as spyware as well. Malware embedded by its designer is not limited by this kind of malicious activity. The embedded malware could also be a time bomb.

D. Spyware:

Spyware is software that installed on system without user's permission which monitors and gathers personal information and then user send the information back to attacker to install the software [8].

IV. MALWARE ANALYSIS TECHNIQUE

Malware analysis is necessary to develop effective malware detection technique. It is the process of analysing the purpose and functionality of a malware, so the goal of malware analysis is to understand how a specific piece of malware works so that defence can be built to protect the organization's network [9]. There are three types of malware analysis which has the same goal of explaining, how malware works and their effects on the system. The performance of analysis differs by the tools, time and skills required.

A. Static analysis:

It is also called as code analysis. It is the process of analyzing the program by examining it i.e. software code of malware is observed to gain the knowledge of how malware's functions work [11]. In this technique reverse engineering is performed by

using disassemble tool, decompile tool, debugger, source code analyzer tools such as IDA Pro and Ollydbg in order to understand structure of malware.

B. Dynamic analysis:

It is also known as behavioral analysis. Analysis of infected file during its execution is known as dynamic analysis. Infected files are analyzed in simulated atmosphere like a virtual machine, simulator, emulator, sandbox etc. After that malware researchers use SysAnalyzer, Process Explorer, ProcMon, RegShot, and other tools to identify the general behavior of file [12]. In dynamic analysis the file is detected after executing it in real environment, during execution of file its system interaction, its behavior and effect on the machine are monitored.

C. Hybrid Analysis

This technique is proposed to overcome the limitations of static and dynamic analysis techniques [13]. It firstly analyses the signature specification of any malware code & then combines it with the other behavioral parameters for enhancement of complete malware analysis. Due to this approach hybrid analysis overcomes the limitations of both static and dynamic analysis.

V. THE MALWARE DETECTION TECHNIQUES

As described in the introduction, a malware detector is the implementation of some malware detection technique(s). The malware detector makes an attempt to assist shield the system by detecting malicious behavior which may or may not reside on the same system it is trying to protect [14]. The malware detector performs its protection through the manifested malware detection techniques. A Malware detector uses two inputs. They are i) Knowledge of the malicious behavior and ii) Program under inspection.

i) Knowledge of the malicious behavior: The detector has the knowledge about the distinction about typical behavior and malicious behavior.

ii) Program under inspection: The detector has the program that must be identified under surveillance.

iii) Once the malware detector has the knowledge of what is considered malicious behavior (normal behavior) and the program under inspection, it can employ its detection technique to decide if the program is malicious or benign. The malware detection techniques are utilized to identify the malware and keep the computer system from being infected, shielding it from potential information loss and system compromise.

iv) As a result of the creating malware in the innovation, the information of obscure malware protection is a principal subject in the malware recognition as per the machine learning strategies[15]. The machine learning strategies are categorized into supervised and unsupervised classes. Malware detection approaches are classified into two main categories that include behavior-based and signature-based methods.

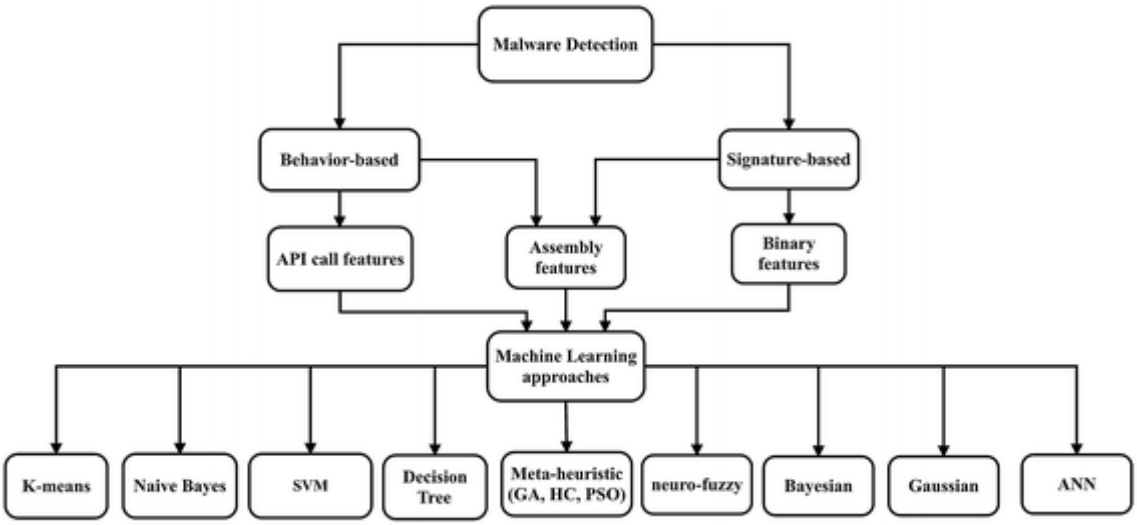

Fig. 1 Taxonomy of malware detection approaches

In the above Fig. 1, the malware detection taxonomy is outlined depending on machine learning approaches. As indicated by this figure, the API calls features, assembly features, and binary features are existing methodology for malware detection method. These features use machine learning methods for predicting and recognizing malicious files.

A. Signature Based Detection:

Recently, signature-based detection is the most commonly used procedure in antivirus programming highlighting exact correlation. The signature-based system discovers interruptions using a predefined list of known assaults. This technique requires steady overhauling of the predefined signature database to identify malware in the versatile application. Within the malware structure, existing malicious objects have attributes that can be used to produce a unique digital signature. The anti-malware provider utilizes the meta-heuristic algorithms that can examine effectively the malicious object to control its signature [16]. After identifying the malicious object, the detected signature is added to the current database as the recognized malware. The database sources incorporate huge number of the different signatures that classify malicious objects. In the signature-based malware detection, there are some different characteristics including fast identification, easy to run, and broadly accessible.

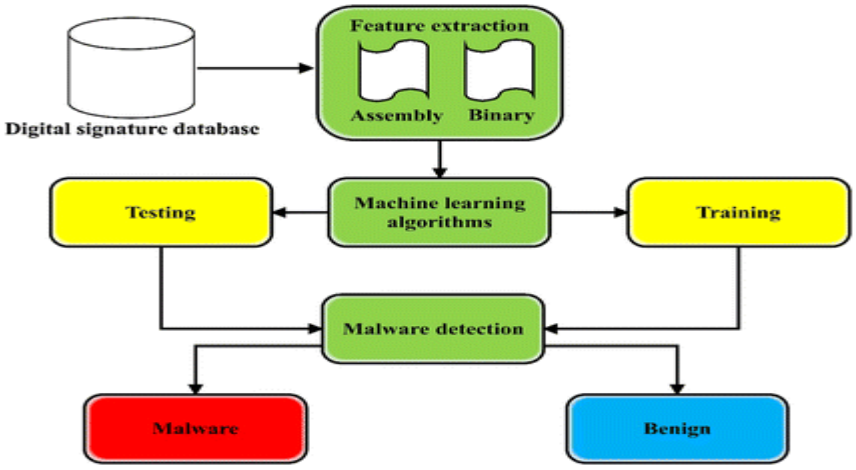

Fig. 2 The signature-based malware detection framework

B. Behavior-Based Detection

In the behavior-based malware approach, the suspicious objects are evaluated based on their activities that they cannot execute in system. Efforts to achieve activities that are obviously irregular or informal would indicate the suspicious object is malicious, or at least apprehensive. A malicious behavior is known using a dynamic analysis that assesses malicious intent by the object's code and structure [17]. In the behavior-based detection, the API calls and assembly features are two main principle for applying machine learning algorithms. Figure 2 describes a standard behavior-based malware detection approach using data mining algorithms.

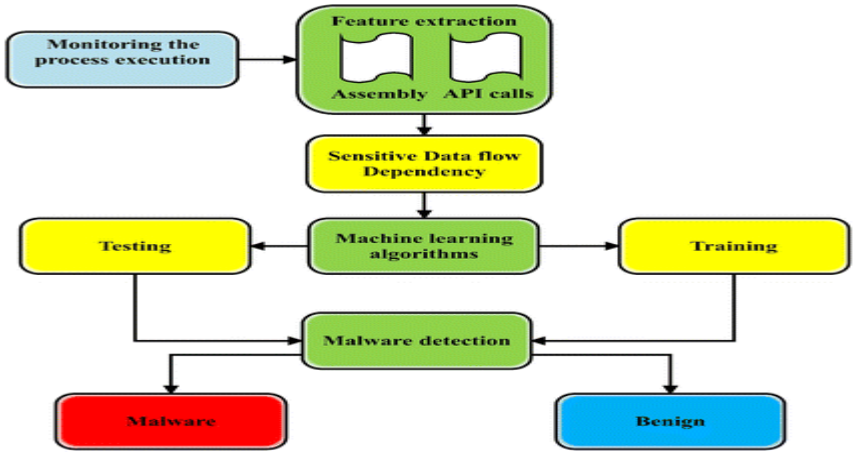

Fig. 3 The behavior-based malware detection framework

VI. Analysis of Various Malware Detection Approaches

The malware identification approaches were compared and analyzed according to different essential factors such as classification approaches, data analysis methods, the number of the used dataset, accuracy factor and case study analysis. The advantage and disadvantage of each method were deliberated in the malware detection methods. In the malware analysis stage, the most methods are proposed for the android smart phones. Likewise, using meta-heuristic algorithms in malware detection analysis can accelerate and enhance the execution time and the overall accuracy of the data mining process. The Table I depicts the analysis of various signature based approaches.

TABLE I
ANALYSIS OF VARIOUS SIGNATURE BASED APPROACH

S.No	Case study	Classification approach	Data analysis method	Used dataset	Total dataset	Accuracy %
1.	Polymorphic Malware Detection	K-means	Dynamic	ClamAV, VirusTotal,	2876	99
2.	Android malware detection	SVM	Dynamic	Google play store	5494	94
3.	Graph malware detection	Graph-SVM	Dynamic	Windows DLL calls	6671	88
4.	Droid malware detection	SVM	Dynamic	Windows API library	7000	98
5.	API malware detection	Naive Bayes and Decision Tree—SVM	Dynamic	Google play store	7000	95
6.	N-grams malware detection	SVM	Dynamic	Google play store	658	97
7.	Smartphone malware detection	K-means—artificial immune system	Hybrid	Android malware database XVNA	1300	89.8
8.	Frequent pattern mining	Minimal contrast frequent subgraphs	Static	Several websites	2083	92
9.	Service-Oriented mobile malware detection	Naive Bayes and Decision Tree	Hybrid	Key Laboratory of Network Security, Fujian Normal University	3000	97.3
10.	Multi-objective evolutionary detection	Multi-objective evolutionary by GA	Static	Viruseshair and VirusTotal websites	9383	95.15

The following Table II gives the analysis of various behaviour based approaches.

TABLE II
ANALYSIS OF VARIOUS BEHAVIOR BASED APPROACH

S.No	Case Study	Classification Approach	Data Analysis Method	Used Dataset	Total dataset	Accuracy Percentage
1.	Deep learning malware detection	DeepAM	Dynamic	Windows API calls in Comodo Cloud Security Center	2000	98
2.	Graph mining in	Graph search	Dynamic	Windows sandbox	6994	96

	malware detection			malware		
3.	AMAL: automated malware analysis	Decision trees	Dynamic	Random sample from internal user and external customers such as antivirus companies	2086	98
4.	Deep Packet Inspection for malware	BoostedJ48, J48, Naïve Bayesian and SVM	Dynamic	Wireless and Secure Networks Research Lab	4560	99
5.	Android malware characterization and detection	Deep belief networks	Hybrid	Google play and android Malware genome project	1860	96.76
6.	Objective Oriented malware	Multiple association rules	Hybrid	Several websites	8000	97.2
7.	Android based malware	J48, SVM, IBk, NaiveBayes	Static	Google play and android Malware services	2000	98.91
8.	Behavioral Malware	Regression, SVM, J48	Dynamic	Web data commons library in VirusSign and VXHeaven	7000	98.3
9.	Malicious code based on API	Decision tree, SVM and random forest	Dynamic	API hooking library in VirusSign	2000	96.89
10.	Deep-learning malware detection	Naive Bayes, PART, Logistic Regression, SVM and MLP	Hybrid	Google play, virus share	11,000	95.05

VII. CONCLUSION

Malware is a crucial warning to client's computer system in terms of stealing private data, corrupting or damage security system. This survey paper presents few existing technologies utilized by security researchers to handle these threats. It describes static, dynamic and hybrid malware examination methods, their comparative study, existing traditional malware identification techniques and their advantages-disadvantages. According to the comparative study advanced malware detection technique i.e. data mining and machine learning method overcome the drawbacks of existing malware detection techniques.

REFERENCES

[1] Souri A, Norouzi M, Asghari P (2017) An analytical automated refnement approach for structural modeling largescale codes using reverse engineering. Int J InfTechnol 9:329–333.

[2] Hashemi H, Azmoodeh A, Hamzeh A, Hashemi S (2017) Graph embedding as a new approach for unknown malware detection. J ComputVirol Hacking Tech 13:153–166.

[3] Souri A, Asghari P, Rezaei R (2017) Software as a service-based CRM providers in the cloud computing: challenges and technical issues. J ServSci Res 9:219–237

[4] Safarkhanlou A, Souri A, Norouzi M, Sardroud SEH (2015) Formalizing and verifcation of an antivirus protection service using model checking. Procedia ComputSci 57:1324–1331.

[5] Malhotra R, Jangra R (2017) Prediction & assessment of change prone classes using statistical & machine learning techniques. J Inf Process Syst 13(4):778–804

[6] Chowdhury M, Rahman A, Islam R (2018) Malware analysis and detection using data mining and machine learning classifcation. In: Abawajy J, Choo K-KR, Islam R (eds) International conference on applications and techniques in cyber security and intelligence: applications and techniques in cyber security and intelligence. Springer International Publishing, Cham, pp 266–274
[7] Sun M, Li X, Lui JC, Ma RT, Liang Z (2017) Monet: a user-oriented behavior-based malware variants detection system for android. IEEE Trans Inf Forensics Secur 12:1103–1112
[8] Boukhtouta A, Mokhov SA, Lakhdari N-E, Debbabi M, Paquet J (2016) Network malware classifcation comparison using DPI and fow packet headers. J ComputVirol Hacking Tech 12:69–100.
[9] Dali Z, Hao J, Ying Y, Wu D, Weiyi C (2017) DeepFlow: deep learning-based malware detection by mining Android application for abnormal usage of sensitive data. In: 2017 IEEE symposium on computers and communications (ISCC), pp 438–443
[10] Wu S, Wang P, Li X, Zhang Y (2016) Effective detection of android malware based on the usage of data flow APIs and machine learning. InfSoftwTechnol 75:17–25.
[11] Mao W, Cai Z, Towsley D, Feng Q, Guan X (2017) Security importance assessment for system objects and malware detection. ComputSecur 68:47–68.
[12] Galal HS, Mahdy YB, Atiea MA (2016) Behavior-based features model for malware detection. J ComputVirol Hacking Tech 12:59–67.
[13] Norouzi M, Souri A, SamadZamini M (2016) A data mining classification approach for behavioral malware detection. J ComputNetwCommun 2016:9
[14] Bhattacharya A, Goswami RT (2017) Comparative analysis of different feature ranking techniques in data miningbased android malware detection. In: Satapathy SC, Bhateja V, Udgata SK, Pattnaik PK (eds) Proceedings of the 5th international conference on frontiers in intelligent computing: theory and applications: FICTA 2016, Volume 1. Springer Singapore, Singapore, pp 39–49
[15] Wuechner T, Cislak A, Ochoa M, Pretschner A (2017) Leveraging compression-based graph mining for behavior-based malware detection. IEEE Trans Dependable SecurComput.
[16] Sun L, Li Z, Yan Q, Srisa-an W, Pan Y (2016) SigPID: significant permission identification for android malware detection. In: 2016 11th international conference on malicious and unwanted software (MALWARE), pp 1–8
[17] Palumbo P, Sayfullina L, Komashinskiy D, Eirola E, Karhunen J (2017) A pragmatic android malware detection procedure. ComputSecur 70:689–701
[18] Wu B, Lu T, Zheng K, Zhang D, Lin X (2014) Smartphone malware detection model based on artificial immune system. China Commun 11:86–92
[19] Fan Y, Ye Y, Chen L (2016) Malicious sequential pattern mining for automatic malware detection. Expert SystAppl 52:16–25
[20] Hellal A, Romdhane LB (2016) Minimal contrast frequent pattern mining for malware detection. ComputSecur 62:19–32.
[21] Santos I, Brezo F, Ugarte-Pedrero X, Bringas PG (2013) Opcode sequences as representation of executables for datamining-based unknown malware detection. InfSci 231:64–82
[22] Galal HS, Mahdy YB, Atiea MA (2016) Behavior-based features model for malware detection. J ComputVirol Hacking Tech 12:59–67
[23] Sheen S, Anitha R, Natarajan V (2015) Android based malware detection using a multifeature collaborative decision fusion approach. Neurocomputing 151(Part 2):905–912.
[24] Ming J, Xin Z, Lan P, Wu D, Liu P, Mao B (2016) Impeding behavior-based malware analysis via replacement attacks to malware specifcations. J ComputVirol Hacking Tech
[25] Eskandari M, Khorshidpour Z, Hashemi S (2013) HDM-Analyser: a hybrid analysis approach based on data mining techniques for malware detection. J ComputVirol Hacking Tech 9:77–93.