Intelligent Malware Detection and Behavioral Analysis Through Machine Learning

V.Sravanthi
Sr.Asst. Professor, Dept. of CSE, GCET, Hyderabad.
vsravanthi,cse@gcet.edu.in

Abstract— **mainly the challenging thing at present over the internet is malware attacks. The impact of malicious software is becoming the problem day by day. Malicious type viruses are designed to interruption or for damaging computers, networks as well as other components which are linked to it. Different types of virus shares various behavioural format identifying their originality and cause. Behavioural mechanisms were caused either by fixed manner or it change with time to time which is termed with the help of ML Mechanisms. This concept shows a general brief regarding the procedures as well as mechanisms for identifying as well as estimating the virus.**

Keywords — Malware, **Virus, Static Analysis, Dynamic Analysis, Worms, Spyware.**

I. Introduction

Software that “intentionally fulfils the damaging intent of an attacker” is called a malicious software program or malware. The term malware is short for of malevolent programming, as the name proposes malwares are planned to hurt PCs and PC clients by taking data, adulterating documents or simply by doing wicked exercises to irritate clients. In spite of the critical improvement of network safety instruments and their persistent advancement, malware is still among the best dangers in the internet. Malware investigation applies strategies from a few distinct fields, for example, program examination and organization examination, for the investigation of noxious to create a more profound comprehension on a few perspectives, including their conduct and how they advance after some time. security scientists use malware location frameworks to identify malware. A significant objective of malware examination is to catch extra properties to be utilized to further develop safety efforts and make avoidance as hard as could really be expected. The paper is defined as follows: In the next section, under material and method, we explain types of Malware, Malware Detection techniques, Malware Analysis techniques, Section III explain regarding the removal of malware from devices. Finally, conclusion of survey.

II. Literature work

According to Guillermo Suarez-Tangil at all, [1] This article inspects the issue of malware in savvy gadgets and ongoing advancement made in recognition strategies. Here the creator originally introduced a definite examination weather virus was advanced at course with last a huge time period for most famous stages. We distinguish displayed practices, sought after objectives, contamination and appropriation procedures, and so forth and give various models through contextual analyses of the most important examples. We next review, arrange and examine endeavours made on recognizing malware along dubious programming like gray ware, focusing over more than 25very important methods proposed somewhere at ranging from 2010 and 2013. At ends extricated from this review, the creator at long last gives helpful conversation on open examination issues and regions where we accept the lot of task was required.

According to Heng yin et al, [2] pernicious data access and handling conduct is the essential quality of various malware classifications breaking clients' protection which incorporates key loggers, secret word criminals, network sniffers, covertness secondary passages, spyware, and root kits that isolates these malignant applications from harmless programming. In light of this issue, the creator proposed the Panorama to distinguish just as to investigate malware by catching this principal attribute. In the analyses, Panorama effectively identified all the malware tests and had not very many bogus up-sides. Moreover, by utilizing Google Desktop as contextual analysis, the creators in this work was precisely catch its data access and handling conduct, and we can affirm that it sends back touchy data to distant servers in specific settings. We accept that a framework, for example, Panorama will offer key help to code investigators and malware analysts by empowering them to rapidly understand the conduct and internal activities of an obscure example.

According to shabtai et al, [3] Here, the author introduced another conduct-based anomaly recognition framework for identifying significant deviations in a versatile application's organization conduct. The principle objective of the work is to secure mobile phones clients and cell framework organizations from vindictive applications by distinguishing proof of malevolent assaults or disguising applications introduced on a cell phone and ID of republished famous applications infused with a noxious code like repackaging. All the more unmistakably, the creator did an endeavor to identify another sort of portable malware with self-refreshing capacities that were as of late found on the authority Google Android commercial center. Malware of this kind can't be recognized utilizing the standard marks approach or by applying normal static or dynamic examination strategies. The discovery is performed dependent on the application's organization traffic designs as it were. For every application, a model addressing its particular traffic design is adapted locally (i.e., on the gadget). Semi-regulated AI techniques were used for designing the ordinary personal conduct standards and for distinguishing deviations from the application's normal conduct. These techniques were executed and assessed on Android gadgets. Our proposed work demonstrates that

(1) Various applications have explicit organization traffic designs and certain application classifications can be recognized by their organization designs

(2) Different degrees of deviation from ordinary conduct can be distinguished precisely

(3) on account of self-refreshing malware, unique and contaminated forms of an application have unique and discernable organization traffic designs that by and large, can be recognized inside a couple of moments after the malware is executed while introducing extremely low bogus cautions rate (4) Local learning is doable and has a low presentation overhead on cell phones.

According to Laura Garcia et al, [4] the author In this paper, the author considers and describes malware focusing on iOS gadgets. To this respect, we concentrate on the highlights of iOS malware and arrange tests of 36 iOS malware families found somewhere starting from 2009 and 2015. We likewise show the system for iOS malware examination and give a nitty-gritty investigation of a malware test. Our discoveries prove that the ot of them are circulated out of true business sectors, target jailbroken iOS gadgets, and not very many endeavors any weakness.

According to Sajedul Talukder [5] In this paper, authors did a review in regards to the rundown of malware recognition and examination methods and apparatuses. Specifically, the various devices accessible for malware discovery, memory legal sciences, parcel investigation, scanners/sandboxes, figuring out, hacking, and site examination have been tossed light. Since large amount of the current reviews commonly focus on a particular subset of the model, this paper offers a top to bottom investigation of techniques to recognize and assess malware with a reasonable comprehension of area explicit examination.

According to Sajedul Talukder [6] the author had reviewed an outline of strategies and devices for identifying and breaking down the malware. Specifically, a light has been tossed on different apparatuses accessible for malware identification, memory legal sciences, parcel examination, scanners/sandboxes, picking apart, troubleshooting, and site investigation. Since the vast majority part cantered around a particular subset of the norm, this paper gives a careful investigation of devices for distinguishing and examining malware with a reasonable comprehension of area explicit examination.

According to Mohammad Wazid et al, [7] the authors firstly perform an investigation of different kinds of malware assaults and their manifestations. We additionally examine a few structures of the IoT climate alongside their applications. Then, a scientific categorization of safety conventions in the IoT climate is given. In addition, we direct a near report on different existing plans for malware discovery and avoidance in the IoT climate. At last, some future examination difficulties and headings of malware discovery in the IoT/IoMT climate are featured.

According to Andreas Moser [8] the main agenda was to know the stopping the fixed location. Here we have designed twofold confusion mechanisms that were natives that helps for giving the position in consistent into a register to such an extent that an examination product which cannot identify its value. According to Anshan Damodaran [9] In this research, the authors compared malware recognition strategies dependent on static, dynamic, and mixture investigation. In particular, we train Hidden Markov Models (HMMs) on both static and moving capacity which uses the subsequent identification rates over a generous number of malware families. We additionally think about crossbreed cases, where dynamic investigation is utilized in the preparation stage, with static procedures utilized in the identification position where as the other way around. In our tests, a completely unique position for the majority part yields the best discovery rates. We examine the ramifications of this examination for malware discovery dependent on crossbreed strategies.

According to Muzzamil [10] the following article shows the malware examination framework, fit for experiencing known avoidance techniques for malware. A clever procedure for the discovery of malware shifty conduct is introduced, which depends on estimating the deviation from the ordinary conduct malware. Assessments and examination show that this methodology is compelling against distinguishing the varieties in malware conduct. Additionally, countermeasures executed by the Analysis Evasion Malware Sandbox (AEMS) are viable for an enormous level of malware recognition.

According to P V Shijo [11] This article proposes an incorporated static and dynamic investigation technique to examinations and characterize an obscure executable record. The strategy utilizes AI in which known malware and harmless projects are utilized as preparing information. The element vector is chosen by dissecting the parallel code just as unique conduct. The proposed strategy uses benefits of fixed as well as moving examination consequently the proficiency and the arrangement result are improved. Our trial results shows an exactness of 95.8% utilizing static, 97.1% utilizing dynamic and 98.7% utilizing coordinated strategy. Contrasting and the independent dynamic and static strategies, our incorporated strategy gives better precision.

III. METHODOLOGY

Here we will observe an overview of malware types, malware detection methods, and analysis techniques.

A. *Malware Types*

Malware is software program this is given to the device without user knowledge. It having the capable to damage computer by compromising computer functions, stealing records, or evading access controls. Malwares exist in unique forms, they may be widely categorised in following classes: Virus, Worms, Trojan Horse, Root kit, Spyware, Adware, Ransom ware, Key loggers, Sniffers.

- **Virus:** A malicious software program that duplicates itself through injecting its code into different programs. It can't exist independently so it attaches with different files more exactly executable files and applications and because of its replicating features, it unfolds throughout the files and even computer systems via network.
- **Worms:** Worms are malicious portions of code that exist independently. Worms can also encrypt documents or send junk emails. Unlike viruses, worms carry themselves in their very own containers. As they can develop multiple copies of themselves, antivirus scanners can pick out those codes due to more than one existence.
- **Trojan horse**: Trojan horse behaves like a beneficial program however it has a damaging purpose. They do not reflect themselves however it is transferred to a

computer by internet interaction like downloading. Trojans permit attackers to get benefit access to the effective pc and extract user private data like password and banking details.

- **Rootkit:** Rootkit is a group consisting of harmful code that is programmed to access a computer system and permit different forms of malware to get into the device. It takes control of the working system such that it may conceal itself or can take secure environment for different malware to hide with inside the device.
- **Spyware**: Spyware is a software program that constantly spies on the user's activities. Which helps to steal a person's private data or keep the watch on the user's activities which is used to collect data approximately the users like web pages frequently visited and credit card details without their knowledge, then sends that data again to the attackers. Even companies with massive names like Google uses this mechanism for gather the desired data of their customers.
- **Adware:** Adware is pretty annoying most of the time because it performs advertisements on a personal computer without its permission and interrupts its current activity. The primary reason for adware is to get economic benefits. It does as much damage as other malware. Usually, adware is bundled with free downloaded software and programs like free playing games.
- **Ransom ware:** A dangerous software program that permits the hacker to lock the computer and restrict the victim access to critical information. Ransom ware encrypts the critical data at the infected computer or network then asks for price to raise the restriction
- **Key loggers:** this was a kind of spyware that is used to record keystrokes to steal passwords, credit card data along other crucial and sensitive figures. It is transferred in a computer when a few other malicious software programs are installed or any infected site is visited by user.
- **Sniffers:** Sniffers are a software program that takes a look at and record the network traffic. They examine different fields of packets along collect information for the guidance of the malware attack.

B. *Malware Detection Techniques*

Malware identification mechanisms were identified into two important categories that consist of behaviour-based and signature-based methods. Also, which were fixed as well as dynamic malware analysis that commonly carried out in locating malicious applications.

1. Signature-Based Malware Detection:

Signature-based detection is the most normally applied procedure in antivirus programming highlighting exact correlation. The majority of the available antivirus software program uses the signature-based method. This method extracts a completely unique signature from captured malware report and uses this signature to detect similar malware. A signature is a series of bytes or a file hash that may be used to identify unique malware. Therefore, this technique has small false positive rate. However, it isn't always tough for attackers to change malware signature to prevent being detected through antivirus software. Signature-based could be very powerful and rapid in detecting recognized malware, however, it's far incapable to capture new released malware.

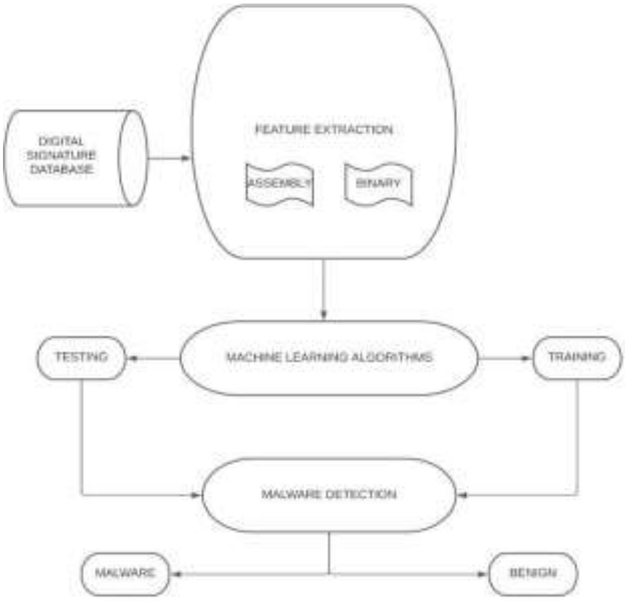


Figure 1: Malware Detection with Machine Learning Architecture

The signature-based technique relies upon enforcing fixed mechanisms to take out exceptional byte sequences called marks. The diagram defines the signature-based general technique for malware detection.

ADVANTAGES	DISADVANTAGES
Fast identification	Replicating data in the large database.
Identifying complete malware information	Failing to discover the polymorphic malware.
It is Easy to run	-

Table 1: Pros and Cons of Using ML for Malware Technique

Malware authors have created some other challenge for signature-based method by the usage of various strategies. This strategy consists of dead code, register allotment, instruction mixing as well as program manipulation.

2. *Behaviour-Based Malware Detection:*

Behaviour-based detection is also called Anomaly or Heuristic-based detection. It reviews the chosen behaviour-based approaches in data mining. the first advantage of the behaviour-based approach is that offers a superior mechanism how virus is made as well as implemented. These activities performed by malware throughout runtime are analysed during a training phase. after that, the file is labelled as a malicious or legitimate file during a testing phase based on a pattern extracted during the training test. within the conduct based malware approach, the dubious items are evaluated dependent on the exercises that they can't execute in the framework. Malevolent conduct is perceived utilizing a unique investigation that assesses pernicious goal by the article's code and construction. inside the conduct based identification, the API calls and get together choices are two principle methodologies for applying AI calculations. Behaviour-based typically relies upon data mining strategies in order to recognize the behaviours of running files, such strategies consist of SVM, NB, DT, along Random Forest. Unlike signature-based, behaviour-based technique is successful to come across each

Unknown malware and malware that makes use of obfuscation techniques.

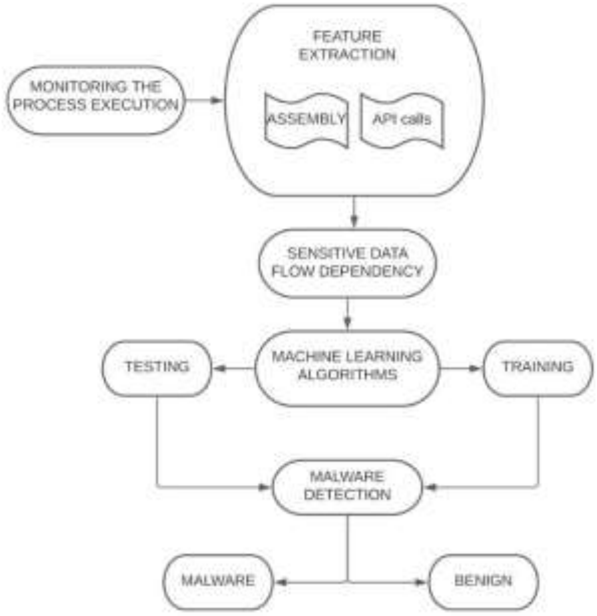

MONITORING THE PROCESS EXECUTION
FEATURE EXTRACTION
ASSEMBLY
API calls
SENSITIVE DATA FLOW DEPENDENCY
TESTING
MACHINE LEARNING ALGORITHMS
TRAINING
MALWARE DETECTION
MALWARE
BENIGN

Figure 2: Malware Detection Work Flow

However, the primary drawbacks of behaviour-based are a considerable false-positive rate and excessive tracking time.

ADVANTAGES	DISADVANTAGES
Data-flow dependency detector	Time Complexity
Identifying unconceived forms of malware attacks	Storage complexity for behavioural patterns.
Identifying the polymorphic malwares	-

Table 2: behaviour of various virus, worms.

Further, the reduction of hundreds of extracted features, compare similarities among them, and tracking malware activities are directly affecting the ability of detecting zero-day malware attacks.

C. *Malware Analysis techniques:*

Malware analysis is that the strategy of determining the cause and characteristics of a given malware sample like a bug, worm, virus. This method can be an important step in order for identification in designing good working techniques of malicious code. Mechanisms used in malware analysis can basically be broken into a couple of techniques which were fixed and moving. The static analysis tools commit to analyse a binary without very penalty the binary. Live analysis tools will study the response of a binary once it is been dead. Fixed and Moving techniques are delineating in detail in subsequent sections. automated malware analysis may even be an almost uncontrollable problem. it is simply not possible for one program to determine the correct behaviour of some other program. To come across malware, first, we should analyse, how malware carries out its feature and what's the motive behind malware development so that this kind of knowledge about malware makes it easy for the developers of malware detectors to implement the protective functionality. Malware analysis techniques were classified into 3 classes on the premise of time and approach to do the analysis.

1) *Static Malware Analysis:*

Static analysis is easy and quick but it could be useless towards sophisticated malware and can pass over the crucial behaviours of the malware.

- Virus Total: It is an online cloud anti-virus scanner. Virus Total examines the uploaded file in opposition to the known databases and anti-virus engines. If the uploaded file includes any malicious code or if the uploaded file is infected it's going to be detected. Not only files URLs also scanned to detect the virus.
- PEid: PEid check if a malware is packed or it is unpacked and also It can check compiler of malware and of course it is also used to unpack the malware.
- WinMD5: It was defined for virus fingerprinting. Malware fingerprinting is to generate the hash value of the malware which is unique to that malware load. The hash value is also used by anti-virus software for fast signature detection because it's only unique to the malware itself.
- Preview: This mechanism is used to view some administrative information about some malware things like its compilation date.
- Dependency walker: It basically shows us the important dynamically linked libraries known as DRLs and their important functions co by looking at all that we can actually give us an idea of the functionality of the malware.
- Bin Text: It is used for analysing usable strings. So rotating text helps us decipher the data into a usable string and can help us detect some hosts or network-based intruders like the IP address and the popup file.

2) *Dynamic Malware Analysis:*

To run the malware in a safe and contained environment, watch when it runs to determine the functionality of the malware. When software functionality is analysed and observed through execution, this is known as dynamic or behavioural analysis and can be done by tracking function calls, control flows, and also by analysing function instructions and parameters. Malicious code is executed over the unseen atmosphere in order to observe its behaviour and develop measures to counter this negative behaviour. Dynamic analysis was most effective than static analysis, because it uses this technique for running an infected software in a virtual machine for monitoring purposes and to test malware. The mechanisms which were defined for dynamic analysis are cloud sandbox, Reg shot, Process Explorer, Process Monitor, Patterns, Net cat.

- Cloud Sandbox: Multiple virtual machines are hosted in the cloud and upload the malware to the website, then the website sends the malware to the VMS network club and then run the malware, then this thing has artificial intelligence to get an exact data by the VMS to collect and present it. in the report and is then displayed.

- Regshot: Basically, Regshot works by taking two registry snapshots at different times, such as one before the malware runs and one after the malware runs, to compare number for registry entries that were well modified by analysing those registry changes.
- Process Explorer: The process explorer is able to monitor the complete techniques were are running in a virtual machine for each process, we can see the murexes created that process them and the DLLs that relevant strings were uploaded over process and certain processes are created or removed by malware when malware is executed.
- Process Monitor: It is able to monitor file system logs that are being modified and network activities that are displayed in real-time by various server processes on a computer. So when we run malware, we can use the process monitor to see step by step how the malware is running.
- Patterns: spoof the DNS response to a user-specified IP address, for example a local host address of 127.0.0.1 The purpose of separate DNS is to simulate a network on a single host in order to trick the malware into thinking it is a legitimate network trades malware to check your network's functionality.
- Netcat: Netcat is a tool for monitoring internal and external connections on a certain netcat port. We can tell if information was changing transferred remotely or not.

IV. Removing Malware from the Device

To remove the virus from the device you need to follow some steps.

Step 1: Disconnect from the internet.

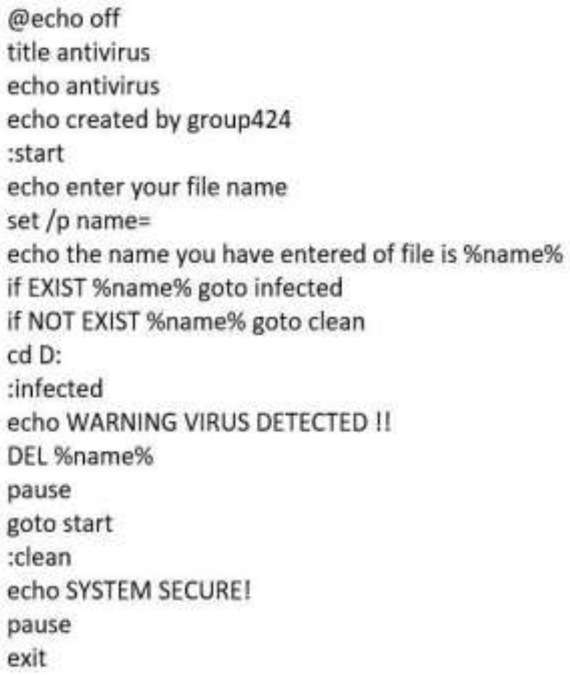
Step 2: Enter to the safe mode.

Step 3: Avoid Logging into accounts during Malware Removal.

Step 4: Check your activity monitor.

Step 5: Run a Malware scanner

The following code represents the malware scanner program.



```
@echo off
title antivirus
echo antivirus
echo created by group424
:start
echo enter your file name
set /p name=
echo the name you have entered of file is %name%
if EXIST %name% goto infected
if NOT EXIST %name% goto clean
cd D:
:infected
echo WARNING VIRUS DETECTED !!
DEL %name%
pause
goto start
:clean
echo SYSTEM SECURE!
pause
exit
```

Figure 3: Algorithm

The program scans the device for malicious code if the malware is present on the device. This file will be removed as soon as the malware is detected.

V. CONCLUSION

This survey paper presents a summary of virus identification as well as estimation techniques. We learned the basics of malware, malware detection techniques, and malware analysis techniques. In malware analysis techniques we have learned about fixed and movable malware analysis, moving virus techniques were the best way to analyze malware samples. In malware detection techniques, we've looked at signature-based malware detection and behavior-based malware detection. We used a batch file to remove malware from the device.

VI. REFERENCES

[1] G. Suarez-Tangil, J. E. Tapiador, P. Peris-Lopez and A. Ribagorda, "Evolution, Detection and

Analysis of Malware for Smart Devices," in IEEE Communications Surveys & Tutorials, vol. 16, no. 2, pp. 961-987, Second Quarter 2014, doi: 10.1109/SURV.2013.101613.00077.

[2] Heng Yin, Dawn Song, Manuel Egele, Christopher Kruegel, and Engin Kirda. 2007. Panorama: capturing system-wide information flow for malware detection and analysis. In Proceedings of the 14th ACM conference on Computer and communications security (CCS '07). Association for Computing Machinery, New York, NY, USA, 116–127. DOI:https://doi.org/10.1145/1315245.1315261.

[3] A. Shabtai, L. Tenenboim-Chekina, D. Mimran, L. Rokach, B. Shapira, Y. Elovici, Mobile malware detection through analysis of deviations in application network behavior, Computers & Security,Volume 43,2014, Pages 1-18,ISSN 0167-4048.

[4] L. García and R. J. Rodríguez, "A Peek under the Hood of iOS Malware," 2016 11th International Conference on Availability, Reliability and Security (ARES), 2016, pp. 590-598, doi: 10.1109/ARES.2016.15.

[5] Sajedul Talukder and Zahidur Talukder, "A Survey on Malware Detection and analysis tools",International Journal of Network Security & Its Applications (IJNSA) Vol. 12, No.2, March 2020.

[6] Sajedul Talukder "Tools and techniques for malware detection and analysis", International Journal of Network Security & Its Applications (IJNSA), Febuary,2020.

[7] M. Wazid, A. K. Das, J. J. P. C. Rodrigues, S. Shetty and Y. Park, "IoMT Malware Detection Approaches: Analysis and Research Challenges," in IEEE Access, vol. 7, pp. 182459-182476, 2019, doi: 10.1109/ACCESS.2019.2960412.

[8] A. Moser, C. Kruegel and E. Kirda, "Limits of Static Analysis for Malware Detection," Twenty-Third Annual Computer Security Applications Conference (ACSAC 2007), 2007, pp. 421-430, doi: 10.1109/ACSAC.2007.21.

[9] Damodaran, A., Troia, F.D., Visaggio, C.A. *et al.* A comparison of static, dynamic, and hybrid analysis for malware detection. *J Comput Virol Hack Tech* **13,** 1–12 (2017). https://doi.org/10.1007/s11416-015-0261-z.

[10] Muzzamil Noor, Haider Abbas, Waleed Bin Shahid, Countering cyber threats for industrial applications: An automated approach for malware evasion detection and analysis, Journal of Network and Computer Applications, Volume 103, 2018, Pages 249-261, ISSN1084-8045, https://doi.org/10.116/j.jnca.2017.10.004.

[11] P.V. Shijo, A. Salim, Integrated Static and Dynamic Analysis for Malware Detection, Procedia Computer Science, Volume 46, 2015, Pages 804-811, ISSN 1877-0509,https://doi.org/10.1016/j.procs.2015.02.149.