INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING & COMPUTER SYSTEMS (IJSECS)
ISSN: 2289-8522 e-ISSN: 2180-0650
VOL. 8, ISSUE 2, 25 – 32
DOI: https://doi.org/10.15282/ijsecs.8.2.2022.3.0100

ORIGINAL ARTICLE

Malware Detection System Using Cloud Sandbox, Machine Learning

Mohd Azuwan Efendy Mail[1], Mohd Faizal Ab Razak[2], and Munirah Ab Rahman[3]

[1]Department of Information Technology and Communication, Politeknik Mersing Johor, 86800 Johor, Malaysia.
[2]Faculty of Computing, Universiti Malaysia Pahang, 26600 Pahang, Malaysia.
[3]Department of Information Technology and Communication, Politeknik Mersing Johor, 86800 Johor, Malaysia.

ABSTRACT – Today's internet continues to move forward, and with it comes the development of many applications. Therefore, these applications are also directly accessible via the Internet, which makes it one of the important things these days. In addition to this, these applications are sometimes developed as software that can be installed on users computers, laptops and even smartphones, which often attracts many attackers to compromise their computers with malware that is unintentionally installed in the computer. Gadgets and even computer systems. computer background. Many solutions have been employed to detect if these malware are installed. This paper aims to evaluate and study the effectiveness of machine learning methods in detecting and classifying malware being installed. This paper employs heuristics and machine learning classifiers to identify malware attacks detected in each website or software application. The study compares 3 classifiers to find the best machine learning classifier for detecting malware attacks. Prove that the cloud sandbox can achieve a high detection accuracy of 99.8% true positive rate value when identifying malware attacks? Use website features. Results show that Cloud Sandbox is an effective classifier for detecting malware attacks.

ARTICLE HISTORY
Received: 17 Feb 2022
Revised: 26 May 2022
Accepted: 13 June 2022

KEYWORDS
Software
Machine Learning
Malware
Website
Classifiers
Cloud Sandbox

INTRODUCTION

Malware is short for malicious software and usually consists of code developed by attackers, cyber attackers that usually plan to cause massive damage to systems and data in order to gain unauthorized access to a particular network. Examples of malware include viruses, ransomware, and spyware. Basically, malware is any software that causes unintentional harm due to lack and flaws, this is known as a software bug [1]. Malware usually will cause a long-term bad effect that will affect the daily operation and thus will affect the security of the system and will easily be breached. It will also cause the computer to slow down which in bad condition will cause our computer to crash [2]. For example, from 2009 until 2018, the malware attack has been rising significantly until that year wherein 2018 the recorded cases were 812.67 million meanwhile the recorded cases in 2009 were 12.4 million only. Approximately, nearly 67.72 million cases were recorded each month in that particular year, 2018. Every year, malware attacks have increased nationally. Based on the figure above, the cases that were recorded show that from 2014, the cases reported increased rosily since, in 2014, the technology was up-to-date significantly. The technology has kept rising as there were many new technologies introduced during that particular year and afterward [3]. This shows how the cases were very serious as we were experiencing new technology almost every year [4], and that it will make each company that was experiencing these attacks, will give a bad reputation for the company revenues [5]. Figure 1 shows illustrate the statistics of malware attacks. Figure 1 shows the increase of malware attacks in the year 2009 until 2018. The increase of the new technology is also because there are a lot of malware types such as computer viruses, worms, Trojan horses, ransomware, and spyware. in the meantime, this malware can be created by using the payload, the well-known payload creator is TheFatRat, an exploiting tool that compiles malware that can be executed on many operating systems (OS), such as Linux, Windows, Mac, and even Android. TheFatRat software can easily get through almost all antivirus software since it is simple to build the backdoors and payloads [6]. Additionally, there's a requirement for utilitarian anti-malware answers for identifying malware assaults and controlling web dangers. There are a few anti-malware discoveries that have been executed by the past analyst such as PeStudio [7], Handle Programmer [8], and Handle Screen (ProcMon) [9]. These anti-malware arrangements have been fathoming malware assaults of late, the clients are still inclined to unused malware assaults as the unused technology keeps coming within the future. This happens since their exercises are not inactive for the assailant; the assailants need to be undetected as conceivable by changing their mode of action [10].

*CORRESPONDING AUTHOR | M.F.A. Razak | ✉ faizalrazak@ump.edu.my

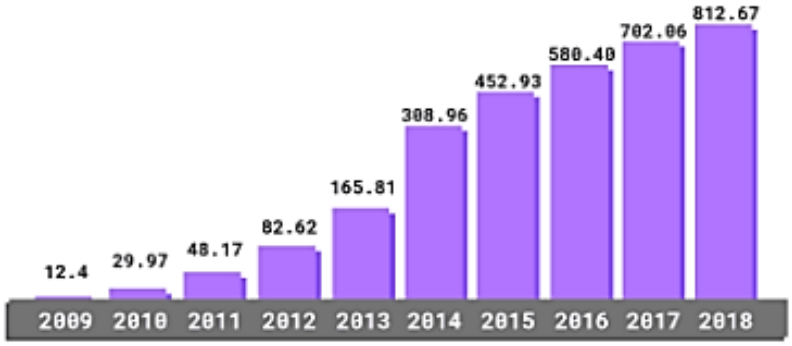

Figure 1. Statistic of malware attack

In spite of the fact that there are numerous sorts of existing frameworks to identify the assaults of malware, there. Indeed, in spite of the fact that appropriately connected innovation, besides security mindfulness are able to reduce the malware assaults, it is troublesome to apply in way of life [11]. For illustration, mail has its own protection approach from a malware assault but is incapable to ensure developing malware assault designs [11]. This can be since mail utilized existing malware assault designs, in this way making clients inclined to unused malware assaults. This is driven by the disclosure of machine learning classifiers to distinguish malware assaults.

Machine learning is counterfeit insights (AI) that apply an information mining approach to find obscure or existing highlights within the dataset [12]. The highlights will be identified on the off chance that the calculation was the malware or not. This paper will propose in case malware discovery will show up or not in the event that we use machine learning. It is additionally proposed that ready to identify the malware within the site approach. Within the result, the center of this paper propose is to distinguish the malware assault within the computer, the commitments of this propose are the taking after underneath:

1) The assessment of the ponder connected malware assault for the malevolent and compassionate test from the SoReL-20M dataset.
2) The proposed bit swarm optimization has been made strides for the optimization by utilizing ten times cross-validation for malware attack form.
3) The proposition from the forest has expanded precision in classifying the malware assaults on computer applications and frameworks.

The rest of the paper is organized as takes after. Fragment 2 looks at related works of the ask almost. Fragment 3 portrays the methodology which consolidates highlights optimization and common plan. Section 4 evaluates the ampleness of malware ambush revelation systems. At long last, Section 5 conclusion of this paper.

RELATED WORK

This part presents a simple diagram of malware attacks taken from the approach used in today's malware attacks. Currently, there are various types of malware attacks. It falls into six different types, the most common types being infections, worms, Trojan horses, spyware, adware, and ransomware. Infections are designed to refocus your computer by destroying information, reformatting your hard drive, or shutting down your system completely.[11]. Worms had been greater regularly now no longer unfold our laptop structures at that factor abusing the operating framework vulnerabilities [13]. The Trojan horse or "Trojan" infiltrates our framework, tricking us into downloading and presenting a program disguised as a log or a regular and secure program that will trick us into tracking it [14]. Spyware is often introduced into our computers without our information . The ability of spyware has been described as tracking our browsing trends and web movements by the week or even daily solstice [15]. Adware is commonly known as a powerful adware program that often brings unwanted advertisements to our computer screen when we use the computer and interface it with the organization [16]. In conclusion, the most well-known malware is Ransomware, the malware that has been generating the most cash because it has been set up and has been spreading out. Ransomware may be malware that holds our information and the programmer, as a rule, requests an installment to discharge the information that's been held sometime after the installment has been made [17].

Three types of approaches are used to attack malware, namely using anti-virus software tools, generic programming methods, and cloud-based online application approaches. Malware detection typically uses signature approaches and virus methods to protect against malicious attempts and anonymize software. Most anti-virus engines depend on regular expressions and patterns to classify malware by following the provided pattern. It's fairly easy for anti-virus software to update their databases to detect and prevent malware, since the functionality of the files that need updating to be used before it's used, this will almost certainly work. requires maximum human effort [18]. A generic programming approach

also had been applied before to avoid evasion attacks before. A report will be generated which will contain and cluster various malware files and group them by data analysis methods. But, these features require more complex implementation methods and surely will need a higher resource consumption where needed [19]. In this paper, we have proposed an online platform that will benefit us and protect us all from almost all types of malware attacks by using Cloud Sandbox as one of the tools to protect our computers from this malicious software. We have barely and previously taken the consideration that malware specifications that were collected from different samples can uplift resource cost and time with different accuracy [20]. In this work, we proposed one of the efficient online machine learning algorithms that gain its experience over time from samples files that we gained before.

RESEARCH METHOD

The malware detection framework consists of five components that have been clarified in detail in this section. The five components collect information, characterize the malware, prove it, test it, and finally, the results are compared. Figure 2 shows the composition and workflow of a malware detection framework in action.

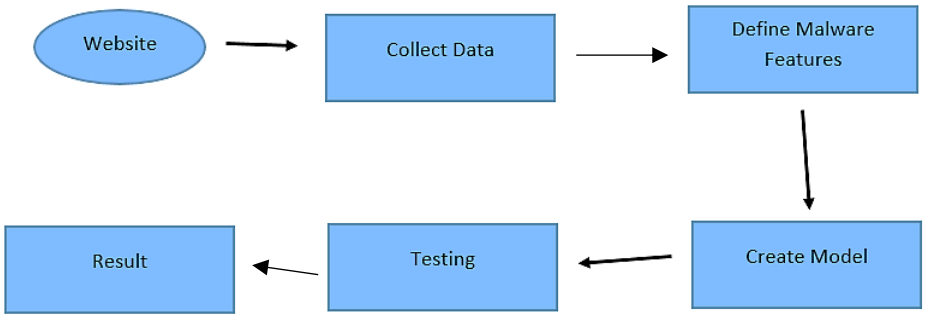

Figure 2. Component of the malware detection system

Data Collection

The first part of the execution segment is to discover, assemble, and collect the dataset as a dataset for use in malware detection frameworks. The data set phase is an essential and vital part of the organization to maintain efficiency and accuracy. There will be more understanding and clarification of the dataset on how the malware performs its activities and it appears that legitimate activities have occurred in the past. The testing then goes further with the dataset, which is then examined and then analyzed, the events used to predict and find long-term malware attacks.

All the highlights have been collected. More than 10 malicious site highlights were collected. The well-known administrators of VirusShare, MalwareBazaar, Hybrid Analysis, URLhaus, VirusBay, USB-IDS Dataset, SecRepo, SoReL20M, Digital Corpora,The Honeynet Project, InQuestLabs and Google Look are where the dataset was collected. The collected data set contains categorical values such as "Legal ", " Suspicious " and "Malware", which have been replaced with numerical values by substituting values "1", "0" and "1" instead of "Legal", "Suspicious", and "Malware" separately.

(5)

Machine Learning Approach

Machine learning can be a part of fabricated ideas (AI), and computer science focuses on using calculations and information to mimic how humans learn [21]. A machine learning approach is used in this mindset to ensure that IT customers can optimize and perfect the highlights of malware attacks through a covered optimization approach. included in this thinking approach of machine learning. This approach brings and delivers shorter time as well as simple test and preparation time . In this way, it streamlines IT customers to use and understand this malware localization framework.

IBM provides a comprehensive history of machine learning. One of its owners, Arthur Samuel, was forced to coin the term “machine learning” in his research (PDF, 481 KB) on female pleasures. Robert Nealey, the self-proclaimed ace of checkers, played the game on the IBM 7094 in 1962, and he switched to it. A few decades later, the advancement of this innovation in the field of controllability and processing will enable some of the innovations we will see today, like the YouTube suggestion engine or the car. independent driving [21].

Cloud Sandbox is machine learning to detect malware. Cloud Sandbox is to tests the software that can categorize "safe" or "unsafe" at the end of the process [22].

1. Malware can become more dangerous if not tested by Cloud Sandbox first. There are many links, applications, and outsourced downloads that could be endless if it is not damned early.
2. Cloud Sandbox can be used as a tool to attack and block malware before it enters the network.
3. The Cloud Sandbox collects sample behaviors by executing them in an isolated Windows environment.
4. Provide an additional layer of security to separate from the threat in online networks

Machine gaining knowledge is distinctly vital in the record sciences field with the making use of authentic strategies, algorithms. They are organized to shape categorization or forecasts. These bits of information Some time later, power preference is made in the programs and businesses, in a really perfect global affecting key improvement measurements. As huge records proceed to expand and expand, the call for the records Researchers will increase and require them to assist in the acknowledgment of commercial enterprise questions and, in this manner, the records to answer them. The reimbursement of the record researchers may be tall considering that they ought to get what records are saying [21].

Sandbox Version Member

When an application is enabled for sandbox, a Sandbox member needs to be created under the Version dimension. Version number under the Version Sandbox member was added when a sandbox was created with the name given by the creator of the sandbox [23]. For example:

Version

- Sandbox
  - Sandbox 1
  - Sandbox 2
  - Sandbox 3

The data in the sandbox was stored at the intersection of each member from the version.

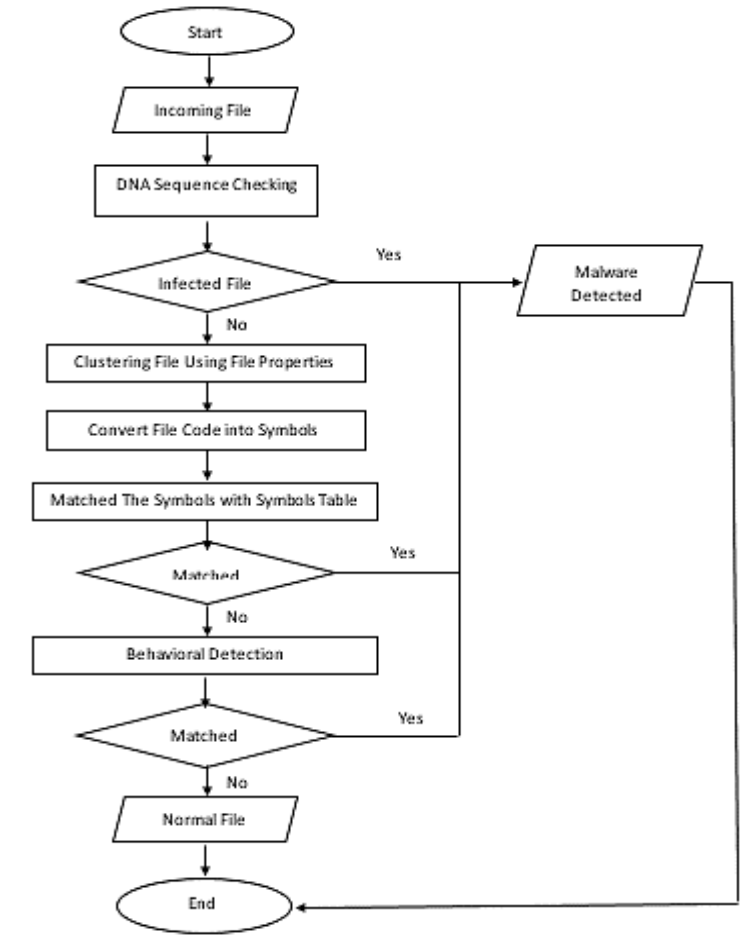

Figure 3. Workflow of the Cloud Sandbox operates

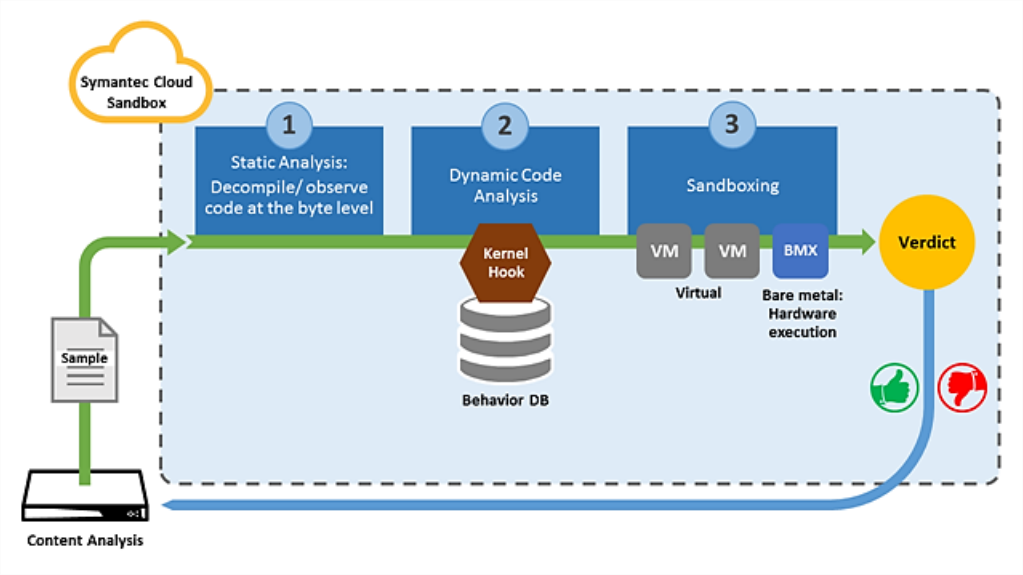

Figure 4. Animation version of explanation.

SANDBOX VERSION MEMBER

To create or modify member formulas, for the data to be calculated correctly in both the base and sandbox views, member formulas must refer to the intersection of the database member and the sandbox Version member. For example:

```
elseif( @ismbr ( @relative ( Sandboxes, 0 ) ) )

" Product revenue "->"ConsolidatedData " * " COGS % "->
                "ConsolidatedData ";
```

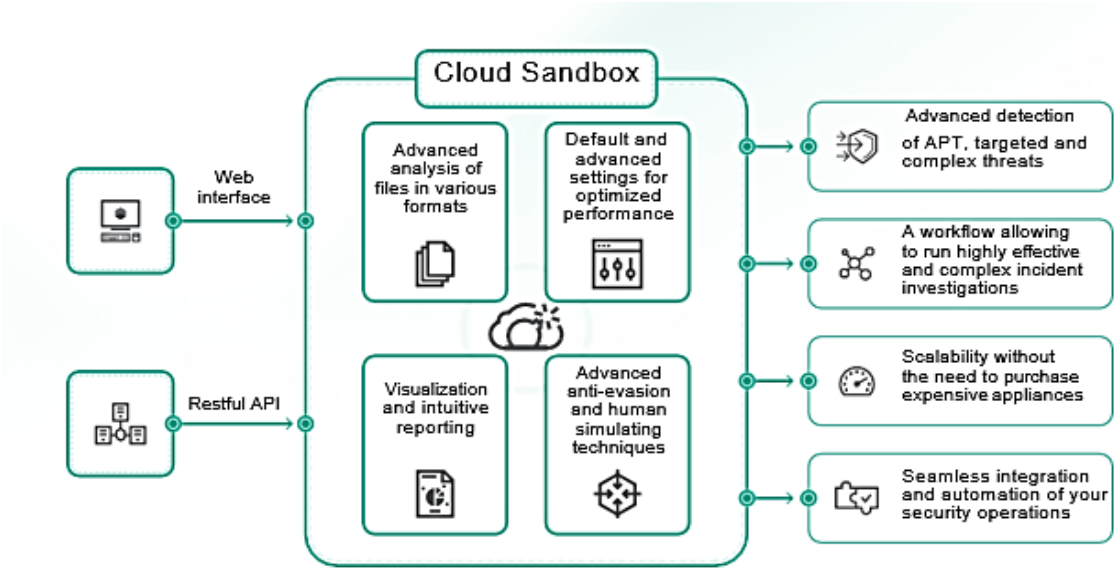

Figure 5. Cloud Sandbox Diagram.

Table 1. Malware website Feature

Malware feature	Description
Access_WiFi_State	Allow Wi-Fi networks information to be access by website
Get_Accounts	Allow the account list in the Accounts Service to be accessed.
SSLFinal_State	A small data file that binds with a cryptographic key to details for organization.
Request_URL	Outside objects contained inside a web page such as pictures and other media are stacked from another space to the same space.
DNSRecord	To mapping records that DNS server will tell around which IP address for each space is related
SFH	IS contain an purge page or appear “about: blank” is considered dubious since activity should be taken when the data is submitted.
Rootkit	Is an administrator-level access to the victim of the system.

EXPERIMENTAL RESULTS

This investigation uses a guided machine learning approach because of the named test information set (malware and common). Also, using machine learning can be a great activity to reduce errors. This exam will run three classifiers to track the results of the classifiers. The three classifiers are Random Forest (RF), J48, Naïve Bayers (NB). These weights used evaluation parameters such as TP rate, FP rate, Precision, ROC area and Accuracy. Table 2 appears after using machine learning for the three classifiers below.

No	Classifier	TP rate	FP rate	Precision	ROC Area	Accuracy
1	Random Forest(RF)	0.99	0.0	1	1.0	99.9
2	J-48	0.99	0.0	1	1.0	99.8
3	Naive Bayers	0.96	0.18	0.85	0.95	90.27

Table 2. Show the time comparison (in second) to build a model depends on the dataset size with the time increase.

Table 3. Table caption.

Classifier	Dataset Size 9000	Dataset Size 9600
Random Forest	1.17	1.18
J-48	0.29	0.27
Naive Bayers	0.1	0.11

CONCLUSION

This paper reports the primary endeavor and has displayed the execution of the proposed approach by investigating the Cloud Sandbox in recognizing malware assaults. The proposed approach executes the machine learning classifier and has accurately recognized and classified the sorts of malware assaults by utilizing important features that were included within the Cloud Sandbox. Within the tests, this paper considers connected genuine malware endeavors and program tests application dataset. The explore comes about because the proposed approach recorded high exactness in classifying the malware assaults. For future research, distinctive strategies of procedures should be considered to improve encourage the accuracy of distinguishing the sorts of malware assaults that are being utilized within the innovation approaches.

ACKNOWLEDGEMENT

This work was supported by Politeknik Mersing Johor and Universiti Malaysia Pahang, under the Grant Faculty of Computer Systems and Software Engineering (FSK1000), RDU1803163.

REFERENCES

[1] R. Sihwail, K. Omar and K. Z. Ariffin. "A survey on malware analysis techniques: Static, dynamic, hybrid and memory analysis." Int. J. Adv. Sci. Eng. Inf. Technol, 8(4-2), 1662-1671, 2018.

[2] A. V. Kabulov, I. K. Yarashov and M. T. Jo'Rayev. "Computer viruses and virus protection problems." Science and Education, 1(9), 179-184, 2020.

[3] J. Firch and L.L.C. PurpleSec, L. L. C. "Cyber Security Trends You Can't Ignore In 2021." PurpleSec. Available online: https://purplesec. us/cyber-security-trends-2021/(Accessed 07/04-2021). 10.

[4] K. Stansberry, J. Anderson and L. Rainie. "Leading concerns about the future of digital life." Pew Research Center, 28, 2019.

[5] R. G. Eccles, S. C. Newquist and R. Schatz. "Reputation and its risks." Harvard Business Review, 85(2), 104, 2007.

[6] F. A. Garba, K. I. Kunya, S. A. Ibrahim, A. B. Isa, K. M. Muhammad and N. N. Wali. "Evaluating the state of the art antivirus evasion tools on windows and android platform." In 2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf) (pp. 1-4), 2019.

[7] S. Talukder. "Tools and Techniques for Malware Detection and Analysis." *arXiv preprint arXiv:2002.06819*, 2020.

[8] P. Hacker. "Overview - Process Hacker." Accessed February 13, 2019.

[9] A. Mohanta and A. Saldanha, A. "Windows Internals." In Malware Analysis and Detection Engineering (pp. 123-162). Apress, Berkeley, CA, 2020.

[10] R. Safi and G. J. Browne. "Detecting Cybersecurity Threats: The Role of the Recency and Risk Compensating Effects." Information Systems Frontiers, 1-16, 2022.

[11] J. Jang-Jaccard and S. Nepal. "A survey of emerging threats in cybersecurity." Journal of Computer and System Sciences, 80(5), 973-993, 2014.

[12] I. H. Sarker. "Machine learning: Algorithms, real-world applications and research directions." SN Computer Science, 2(3), 1-21, 2021.

[13] N. Ochieng, W. Mwangi and I. Ateya. "Optimizing computer worm detection using ensembles." Security and Communication Networks, 2019.

[14] J. Kaur, J. "Taxonomy of malware: Virus, worms and trojan." Int. J. Res. Anal. Rev, 6(1), 192-196, 2019.

[15] R. Chatterjee, P. Doerfler, H. Orgad, S. Havron, J. Palmer, D. Freed and T. Ristenpart. "The spyware used in intimate partner violence." In 2018 IEEE Symposium on Security and Privacy (SP) (pp. 441-458), 2018.

[16] J. Gao, L. Li, P. Kong, T. F. Bissyandé and J. Klein. " Should you consider adware as malware in your study?." In 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER) (pp. 604-608), 2019.

[17] J. Fruhlinger. "Ransomware explained: How it works and how to remove it." CSO Online, 19. it.html#:%7E:text=GandCrab%20might%20be%20the%20most,payouts%20as%20of%20July%202019, 2020.

[18] Kanrar, S. (2019, March). A Novel Approach for Predicting Malware Attacks. A Novel Approach for Predicting Malware Attacks. Retrieved November 2021, from https://www.researchgate.net/publication/331790992_A_Novel_Approach_for_Predicting_the_Malware_Attacks

[19] U. Sivarajah, M. M. Kamal, Z. Irani and V. Weerakkod. "Critical analysis of Big Data challenges and analytical methods." Journal of business research, 70, 263-286, 2017.

[20] E. Cozzi, M. Graziano, Y. Fratantonio and D. Balzarotti. "Understanding linux malware." In 2018 IEEE symposium on security and privacy (SP) (pp. 161-175), 2018.

[21] M. Haenlein and A. Kaplan. "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence." California management review, 61(4), 5-14, 2019.

[22] E. M. Kandoussi, I. El Mir, M. Hanini and A. Haqiq. "Modeling a Sandbox Security Mechanism in Cloud Computing Environment using Bayesian Game." Journal of Information Assurance & Security, 13(1), 2018.

[23] F. Handrick da Costa, I. Medeiros, T. Menezes, J. V. da Silva, I. L. da Silva, R. Bonifácio and M. Ribeiro. "Exploring the Use of Static and Dynamic Analysis to Improve the Performance of the Mining Sandbox Approach for Android Malware Identification." arXiv e-prints, arXiv-2109, 2021.