Machine learning approach for detection of fileless cryptocurrency mining malware

W B T Handaya[1], M N Yusoff[2], A Jantan[2]

[1] Department of Informatics, Universitas Atma Jaya Yogyakarta, Jalan Babarsari 44, Daerah Istimewa Yogyakarta, Indonesia

[2] School of Computer Sciences, Universiti Sains Malaysia, 11800 Gelugor, Penang, Malaysia

E-mail: wilfridus.bambang@uajy.ac.id

Abstract. Cybercrime is the highest threat to every private company and government agency in the world. Using synergistic threats to attack provides many success alternatives that lead to the same goal, which is to take over the network and carry out illegal mining activities using CPU resources from the victim's computer. One of the main motives for the success of this criminal business is its relatively low cost and high return of investment. Using the infection chain method in carrying out cryptocurrency mining malware attacks with fileless techniques involves loading malicious code into system memory. Monero (XMR) is by far the highest popular cryptocurrency among threat actor installing mining malware because it comes with full anonymity and resistance to an application-specific circuit mining (ASIC). This work proposes a better method for classifying conventional malware and cryptocurrency mining malware. On the other hand, grouping specific of suitable features extracted from the sources of EMBER dataset shown as malware and need to categorize as a cryptocurrency mining malware. The proposed approach is defining a better algorithm for enhancing accuracy and efficiency for cryptocurrency mining malware detection.

1. Introduction

1.1. Research background

Malicious software designed by malware creator to attack the target and compromise the data. Currently, the malware metamorphosed and enhanced its ability to use a combination of vectors and threats in hijacking the system. Based on the variety of attack scenarios chosen by the malware creator, from the spread over the network, launching organized Denial of Service (DoS) attacks to the server, and running cryptocurrency mining. Although the malware detection problem has been discussed in numerous papers, it proposed that better solutions are needed, and more research is encouraged to mitigate the threat. The conventional malware types employ standard techniques to attack and avoid detection.

The variant of techniques that avoid signature detection already occurs, such as code obfuscation. The latest method used by malware creators is fileless malware [1], which shows the benign behavior of software by manipulating the Windows Management Instrumentation (WMI) and Power Shell. The fileless attack [2] utilizes a task automation and configuration management framework from Microsoft Windows Operating System, dealing with scripting language processing, as well as optimizing the use

of the command-line shell to carry out specialized functions on the target system. In this research, there is a need to build an approach to enhance the fileless cryptocurrency malware detection and classification models to identify the variations of cryptocurrency mining malware, based on the specific features extraction result from common malware dataset.

1.2. Motivation

Cybercrime activities have a fatal damage impact in all sectors, so that the increasing number of cyber threats currently makes companies, both private and government give more attention to efforts to maintain security, authenticity, and the availability of data and information from users. Cybercrime, on the other hand, also makes the trend of security and hardware and shopping spending increasingly high, because nowadays the better level of knowledge and awareness of the executive level towards data from consumers is getting better. According to published research by Chen L *et al*. [3], more than 1 million malware attacks recorded on the Internet network every day while malware examples counted in Q3 2017 to reach until 57.6 million records.

This notice has not yet comprehensively covered the number of malware attacks in the first semester in the year 2019, further than 430,000 sole users were attacked by financial threat. Continuity and consistency of malware attacks change the techniques, which no longer use conventional methods to enter and control the target but using synergistic threat techniques. Based on previous research, this research idea wants to compare the most accurate supervised learning algorithm [4], which succeed in the detection of the common malware and use those the same algorithms for enhancing the detection model of cryptocurrency mining malware.

1.3. Problems

Cryptocurrency has changed the newest trend in the cyber world, and no time wasted in manipulating its features to earn a rapid income. The campaigns specify that attackers goal a comprehensive range of sources, from a data center, personal computers, mobile devices [5], and IoT devices. They are resumed, as any vulnerable device that can provide CPU cycles. Cybercriminals have also increasingly turned to browser-based cryptocurrency mining recently has it become a common issue, due to the gradual growth in cryptocurrency as well as the launch in 2017 of mining services like Coinhive. Attackers also exploit this method in compromised websites, where even the website owner is ignorant of the activities from the malicious script that successively and demanding the victim's CPU resources.

Cryptocurrency mining is exploited via various approaches by hackers to achieve an income. There is malicious malware that drops the cryptocurrency mining function as its payloads. These payloads are released and resident into the victim computer and executed to exploit the CPU cycle to mine [6], and loading the malicious code to the victim system's memory as a part of involvement from the infection chain in fileless attack.

Dargahi [7] from the previous research, notes that less residual of pieces of evidence from the script-based malware make them have a stealthier possibility for attackers pointing the victims. The left behind of footprint as an evidence, representing an infection is the manifestation of a malicious file, modified the WMI service, and runs complete access to Windows operating system jobs such as the Component Object Model (COM) objects and Windows Management Instrumentation (WMI), as a mentioned before by Hendler *et al* [8]. Fileless attacks allow attackers to hide malware in memory and may hide as a part of a legitimate system process to avoid the detection system [9].

Cryptocurrencies are the first and most established application of blockchain technologies. Monero (XMR), as one of the kinds of cryptocurrency, is favored by criminal actors because anonymous and untraceable. Monero based on the CryptoNight hash proof-of-work algorithm, which comes from the CryptoNote protocol [10]. The CryptoNote protocol has significant algorithmic differences about Blockchain obfuscation. The ring one-time ring signatures anonymize the sender's address of a transaction. Besides, the mining process against Monero does not depend on specialized architectures such as GPU. Based on the CryptoNight hash proof-of-work algorithm [11] and comes with ASIC

(Application-Specific Integrated Circuit)-resistant cryptocurrency, with hard-forks in order to keep this coin available for mining via CPUs and GPUs.

1.4. Research scopes

The scope of this research is accurately carried out on the creation of an approach where machine learning algorithms are used to detect and evaluate malware attacks that use the fileless method to perform its primary functions as cryptocurrency-mining. This type of malware will focus on the impacts to use a computer or CPU resources from target computers that have successfully infected, then run a script tasked with carrying out Monero mining [12]. For the operating system chosen as an environment of research is avoiding antivirus detection or the malware detection system that installed on the user's computer through fileless attack techniques especially on Microsoft Windows operating systems, wherein it can hide malware in memory and as part of a legal process system to avoid detection systems.

Machine learning has a variety of approaches that take a solution rather than a single method, and it makes autonomous decisions. Previous research by Le *et al*. [13] stated that the Machine Learning algorithms could detect unknown malware if data sets can be trained to analyze micro behaviors of malware attacks. On the other hand, the Machine Learning algorithms can use to develop a framework to analyze scripts with the highest stage of cybersecurity solution with the aim of a zero-attack day. Machine Learning algorithms used in this study is a combination of the k-Nearest Neighbors (kNN) [14], SVM [15], and random forest [16], wherefrom the results published research is that each of these algorithms has been used to provide high accuracy in recognizing the characteristics of malware [17]. For the diagram to visualize the research scope, shows in Figure 1.

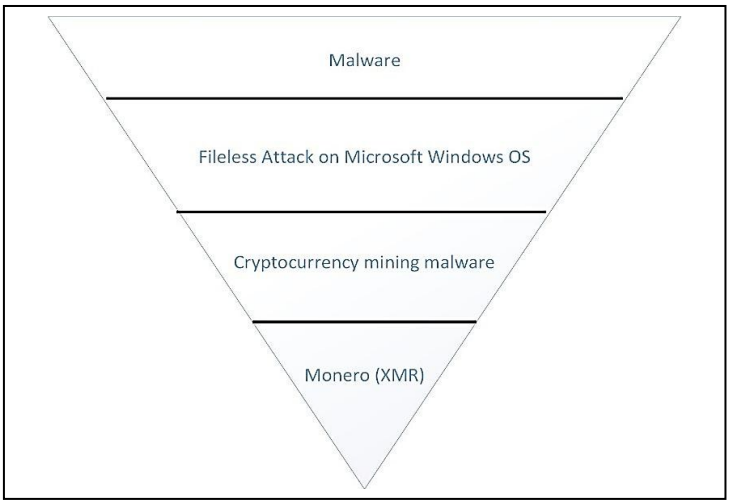

Figure 1. The research scopes diagram.

2. Literature review

Various trends in cyber-attacks have no longer directed personal computer users but have shifted to companies' systems in all countries. The results of research and investigations, as outlined in the report per semester from McAfee, state clearly that malicious software deployment techniques using the built-in features of the Microsoft Windows operating system known as PowerShell recorded in statistics that move up significantly to 267% in the Q4 period of 2017.

The malware attack, malicious code, or malware is a program that can intentionally and unexpectedly interfere with the regular operation of a computer system [18]. Usually, malware has been designed to get financial benefits or additional pandemic. The anticipation of future attacks is undoubtedly an integral part; it can be added capabilities by adopting various techniques and other knowledge outside of computer security, such as artificial intelligence and behavior analysis techniques [19]. Whereas for

more details in Indonesia, based on the annual statistics report in 2018 from the ID-SIRTII (Indonesia Security Incident Response Team on Internet Infrastructure), specifically regarding malware attack trends in more detail are taken from monthly reports, for example, from the start in January 2018, reported for 2,937,420 incidents. Moreover, gradually increase until 10,468,704, in the last month of 2018.

2.1. Blockchain

In order for a digital system to work, the precise function of these systems should be ensured and subsequently maintained. Given the continual acceptance and enhance of Cryptocurrency transactions due to use in digital transfer, it is irresistible that the consequence years would witness a display of transactions that would cause by the using of Monero (XMR) like Blockchain protocols [20]. A comprehensive study was conducted by collecting actual data, from the preparation stage to the extraction process of information in the digital payment system used. With its advantages, Blockchain remains an innovative technology and areas of anxieties that can enhance to manage perfect productivity [21].

2.2. Phases of malware analysis

Applying the machine learning algorithms are necessary to access a dataset that contains a massive number of examples to be analyzed [22]. The malware analysis and extraction methods can mostly characterize into two categories: (i) based on features drawn from an unpacked static version of the executable file without performing the analyzed executable files [23]. Furthermore, (ii) based on dynamic features or behavior features found through the execution of the executable files.

Static Analysis, also named static code analysis [24], is the process of debugging the application without performing the full function inside. The information generated regarding functions and other technical indicators helps create digital signatures of malicious code. Notes in the form of technical indicators collected by static analysis include file names, MD5 checksums or hashes, file types, file sizes, and introduction codes by commonly used antivirus detection devices. When doing static analysis, various tools and techniques are used to gather as much information about malware as possible. Alam S *et al*. [11] offered a method to detect malicious behavior in executables using static analysis with detecting obfuscation patterns in malware by showing good accuracy.

Damodaran, A *et al*. [25] compare malware detection methods based on static, dynamic, and hybrid analysis on both static and dynamic feature sets then compare the result from the detection rates over an outstanding amount of malware samples. Choudary *et al*. [26] prove that dynamic analysis shows the execution process of executable files to understand its behavior, which is results achieved shows high accuracy and indicating that this method can be improved further by using larger sample space. Bai *et al*. [27] projected a malware detection approach by mining format information of portable executable (PE) files and described experiments conducted against recent Win32 malware. However, Stiborek, J *et al.* [28] said that a malware creator already triggered the development of evasion methods such as system-call injection attacks, shadow attacks, or sandbox detection.

2.3. Cryptojacking

The code behind crypto-jacking malware is relatively simple, and it can deliver via phishing campaigns, malware advertising, compromised websites, or software downloads. While some attackers have been known to spin up CPUs to one hundred percent capacity brazenly, those campaigns do not last long because they can cause irreversible damage to the device, and a broken system does not provide any benefit to malicious miners. It is why those with severe networks of hijacked machines are modifying directives to systems: running an unknown operation, and it makes the computing process on the CPU working and sometimes taking over all the resources that the computer executes the cryptocurrency mining scripts.

2.4. Dataset

Applying the machine learning algorithms are necessary to access a dataset that contains a large number of samples to be analyzed [22]. The researcher can use the EMBER dataset, with more than 1 million samples of sha256 hashes from PE files that scanned in 2018, including 900K training samples and 200K test samples. The visualization of the EMBER dataset shows in Figure 2.

Figure 2. The visualization of the EMBER dataset.

Proposed research [29, 30] suggests that deep neural networks achieve comparably and well suited to address the problem of malware detection using static portable executable (PE) features and performed better than any other classical machine learning classifiers. Raff in previous research used the EMBER dataset to test of generalization over time [31].

3. Methodology

The design of this system must be composed of some distinct components of software to function correctly as a system and be informed by the collection. These separate parts should function as part of the malware analysis platform; this separation and abstraction also allow for concurrency of task processing. One of the core design principles adhered to is the abstraction and modularity of separate parts of the system, that while functioning effectively as constituent modules and objects, should be able to work together as a cohesive whole. These separate parts should function as part of the malware analysis platform; this separation and abstraction also allow for concurrency of task processing. A workflow of the proposed malware detection system architecture shows in Figure 3.

From a perspective of machine learning, malware detection can see as a classification problem: unknown types of malware should be grouped into some clusters based on specific algorithm-identified properties. Shijo, P *et al*. [32] mentioned a combined static and dynamic analysis method to analyze and classify an unknown executable file. The method uses a support vector machine and a random forest machine learning algorithm and combines with the customized feature vector for analyzing the code from binary.

The diagram in Figure 4 shows the details of the classification model with a supervised learning algorithm used to enhance the detection of cryptocurrency mining malware are support vector machines, k-Nearest Neighbors (kNN) [30], and random forest [33]. From a perspective of machine learning, malware detection can see as a classification [34] or clustering problem [35]: unknown types of malware should be grouped into several clusters based on specific algorithm-identified properties. On the other hand, taking trained a model in a broad set of malicious and benign files, this problem can be reduced to classification. This issue can only be reduced to the classification of known malware families with a limited number of classes, of which the malware sample certainly belongs, the right class is more accessible to identify, and the result would be more precise than with clustering algorithms.

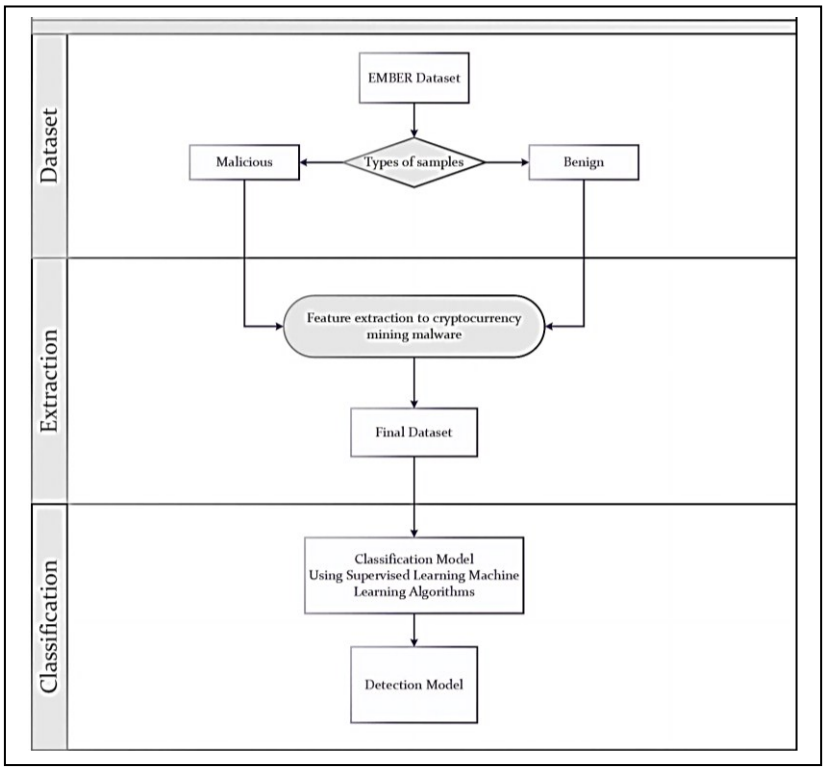

Figure 3. The workflow of a proposed malware detection.

Those three supervised learning algorithms would be compared whether the results are equally useful and accurate for the common malware and when applied to the detection of cryptocurrency mining malware.

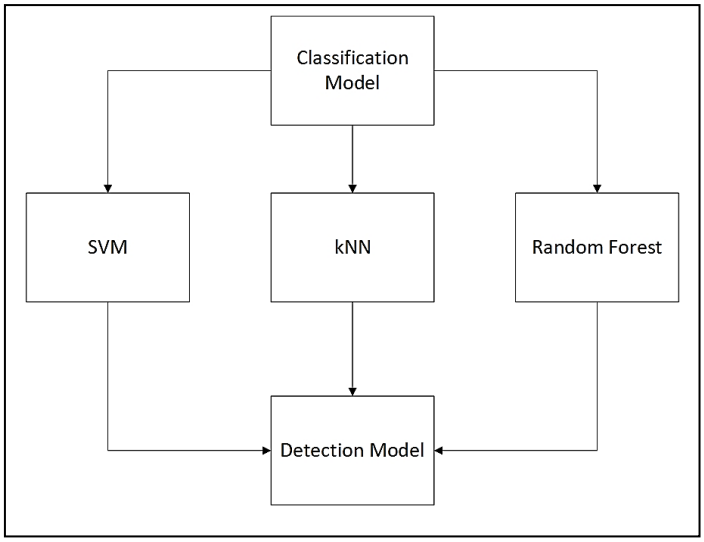

Figure 4. Details of a classification model.

4. Research contributions

- A better method for carrying out identification and classification stages to determine the type of conventional malware and malware that infiltrates the target system through fileless attacks.
- A technique to extract the suitable features that show the type of malware, and explicitly it will be declared in a dataset that records the distinctive characteristics of fileless cryptocurrency mining malware. This study proposes an analysis of the EMBER dataset using Elastic outlier detection functionality, with use an ensemble of four well-formed outlier detection techniques: distance to kNN, the average distance to kNN, local outlier factor, and local density-based outlier factor.
- A better approach for providing high accuracy results to detect the fileless cryptocurrency mining malware which running for focus mining on Monero (XMR). The results from this experimental phase would be compared with the previous research that uses the same classifier algorithm to detect the common malware types.

5. Conclusions

This paper focuses on cryptocurrency mining malware using fileless attack techniques, the latest techniques used by the malware creator to evade the antivirus mechanisms, which use legitimate tools built-in from the Microsoft Windows operating system like Powershell and Windows Management Instrumentation (WMI). On the other hand, the researcher requiring a piece of knowledge to develop the new extraction method of the specific features from the source of common malware types datasets and used to identify the cryptocurrency mining malware correctly. However, in a workable application, the method proposed in this preliminary research needs to have further enhancements in order to enhance its performance to detect the cryptocurrency mining malware with efficient and accurate.

6. References

[1] Bulazel A and Yener B 2017 *Proceedings of the 1st Reversing and Offensive-oriented Trends Symposium on - ROOTS* 1-21
[2] Souri A and Hosseini R 2018 *Human-centric Computing and Information Sciences* **8**
[3] Chen L, Sultana S and Sahita R 2018 *IEEE Security and Privacy Workshops (SPW)* 109-115
[4] F.Y O, J.E.T A, O A, J. O H, O O and J A 2017 *International Journal of Computer Trends and Technology* **48** 128-138
[5] Soviany S, Scheianu A, Suciu G, Vulpe A, Fratu O and Istrate C 2018 *IEEE/IFIP International Conference on Embedded and Ubiquitous Computing* p 14-21
[6] Pastrana S S-T and Guillermo 2019 *Proceedings of the Internet Measurement Conference* 73-86
[7] Dargahi T, Dehghantanha A, Bahrami P N, Conti M, Bianchi G and Benedetto L 2019 *Journal of Computer Virology and Hacking Techniques*
[8] Hendler D, Kels S and Rubin A 2018 ASIACCS '18: *Proceedings of 2018 on Asia Conference on Computer and Communications Security* p 187-197
[9] St'Astna J and Tomasek M 2018 2017 *IEEE 14th International Scientific Conference on Informatics* p 406-411
[10] Möser M, Soska K, Heilman E, Lee K, Heffan H, Srivastava S, Hogan K, Hennessey J, Miller A, Narayanan A and Christin N 2018 *Proceedings on Privacy Enhancing Technologies* **3** 143–163
[11] Alam S, Horspool R N, Traore I and Sogukpinar I 2015 *Computers & Security* **48** 212-233
[12] Carlin D, O'Kane P, Sezer S and Burgess J 2018 *Proceedings of the 2018 International Conference on Privacy, Security, and Trust*
[13] Le Q, Boydell O, Mac Namee B and Scanlon M 2018 *Digital Investigation* **26** S118-S126
[14] Kruczkowski M N-S, Ewa 2014 *Journal of Telecommunications and Information Technology* **4** 24-33
[15] Alazab M, Huda S, Abawajy J, Islam R, Yearwood J, Venkatraman S and Broadhurst R 2014 *Journal of Networks* **9** 2878-2891

[16] Xiaofeng L, Xiao Z, Fangshuo J, Shengwei Y and Jing S 2018 *Procedia Computer Science* **129** 248-256
[17] Zhang K, Li C, Wang Y, Zhu X and Wang H 2017 *Procedia Computer Science* **108** 1682-1691
[18] Sabar N R, Yi X and Song A 2018 *IEEE Access* **6** 10421-10431
[19] Canzanese R M, Spiros, Kam, Moshe 2015 *2015 IEEE International Conference on Software Quality, Reliability and Security* 119-124
[20] Ciaian P, Rajcaniova M and Kancs d A 2018 *Journal of International Financial Markets, Institutions & Money* **52** 173-195
[21] Peng, Yaohao, Albuquerque P H M, Camboim de Sá J M, Padula A J A and Montenegro M R 2018 *Expert Systems With Applications* **97** 177–192
[22] Azab A, Layton R, Alazab M and Oliver J 2014 *Fifth Cybercrime and Trustworthy Computing Conference* **44**-53
[23] Gandotra E, Bansal D and Sofat S 2014 *Journal of Information Security* **05** 56-64
[24] Wagner M, Rind A, Thür N, and Aigner W 2017 *Computers & Security* **67** 1-15
[25] Damodaran A, Troia F D, Visaggio C A, Austin T H and Stamp M 2015 *Journal of Computer Virology and Hacking Techniques* **13** 1-12
[26] Choudhary S P and Vidyarthi M D 2015 *Procedia Computer Science* **54** 265-270
[27] Bai J, Wang J and Zou G 2014 *The Scientific World Journal* 2014
[28] Stiborek J, Pevný T, and Rehák M 2018 *Computers & Security* **74** 221-239
[29] Vinayakumar R and Soman K P 2018 *ICT Express* **4** 255-258
[30] Vinayakumar R, Alazab M, Soman K P, Poornachandran P and Venkatraman S 2019 *IEEE Access* **7** 46717-46738
[31] Raff E, Zak R, Cox R, Sylvester J, Yacci P, Ward R, Tracy A, McLean M and Nicholas C 2016 *Journal of Computer Virology and Hacking Techniques* **14** 1-20
[32] Shijo P V and Salim A 2015 *Procedia Computer Science* **46** 804-811
[33] Shiva Darshan S L and Jaidhar C D 2018 *Procedia Computer Science* **125** 346-356
[34] Banin S and Dyrkolbotn G O 2018 *Digital Investigation* **26** S107-S117
[35] Mohaisen A, Alrawi O and Mohaisen M 2015 *Computers & Security* **52** 251-266

Acknowledgments
The author acknowledges Universitas Atma Jaya Yogyakarta and Yayasan Slamet Rijadi Yogyakarta for supporting the author's study at Universiti Sains Malaysia. The author also acknowledges the members in the Security & Forensics Research Group, Universiti Sains Malaysia, for their helpful discussion and suggestion. This work is supported by USM Short-term Grant No.304/PKOMP/6315237.