University Teknologi MARA

MALWARE DETECTION USING MACHINE LEARNING

NUR WAHIDA KAUSAR BINTI ABD RAHMAN
2018262934

Thesis submitted in fulfillment of the requirement for Bachelor of Computer Science (Hons.) Data Communication and Networking
Faculty of Computer and Mathematical Sciences

JULY 2021

ABSTRACT

After detecting malware, categorizing risky files is a crucial part of the malware investigation process. So far, a number of static and dynamic malware classification algorithms have been reported. This study shows how malware families may be classified using a deep learning-based malware detection (DLMD) strategy based on static methodologies. To categorize malware families, the proposed DLMD approach uses both byte and ASM files for feature engineering. Two distinct Deep Convolutional Neural Networks are used first to extract features from byte input (CNN). Then, utilizing a wrapper-based method and a Support Vector Machine (SVM) as a classifier, important and discriminative opcode features are discovered. The objective is to mix several feature spaces to produce a hybrid feature space that overcomes each feature space's shortcomings and thereby minimizes the likelihood of malware being undetected. Finally, a Multilayer Perceptron is trained to categorize all nine malware types using the hybrid feature space. The proposed DLMD approach, according to experimental results, gives a log-loss of 0.09 for ten independent runs. Furthermore, the suggested DLMD approach's performance is compared to that of other classifiers, demonstrating its efficacy in detecting malware.

ACKNOWLEDGEMENT

All praise is due to Allah S.W.T. for His strength and favour in the completion of this undertaking. Prof. Dr. Jasni binti Mohamad Zain, my supervisor, deserves special thanks for her advice and unwavering support in ensuring my success. Her bravery has provided me with vital feedback and has gone above and above to ensure that I have the proper materials and solutions for my project.

My gratitude and thanks to Dr. Fakariah Hani binti Mohd Ali, my CSP600 instructor, and Dr. Zolidah binti Kasiran, my CSP650 lecturer, for their support and assistance in completing my proposal. Thank you so much to all of my colleagues and friends for always encouraging me and cheering me up during my studies.

Last but not least, for their endless love, prayers and support, my heartfelt gratitude goes to mybeloved parents and also to my brothers and sisters. Special thanks to those around me who have always trusted in me until the very last moment.

TABLE OF CONTENT

CONTENT **PAGE**

CHAPTER 1

INTRODUCTION

1.1 Project Background

Malware is described as computer program that is designed to harm a computer, server, or network. Because early malware did not employ advanced cryptographic techniques, it was simple to detect and classify it by comparing specific code components. Malware categorization has become a difficult and time-consuming operation as a result of new polymorphism and metamorphism concepts such as obfuscation. Polymorphic malware uses an encryption mechanism that encrypts the code each time it iterates while keeping the encryption key safe, making it harder to detect. Metamorphic malware, on the other hand, encrypts the code and changes the encryption key every time it iterates, making it nearly impossible to detect. The total number of instances per day has increased significantly over time, rendering manual malware analysis impractical. The extensive usage of malware producers' obfuscation techniques, which implies that hazardous files from the same malware family (i.e., identical code and common origin) are updated and disguised on a regular basis, is one of the key causes for the enormous number of malware samples. As a result, a generic Machine Learning-based malware analysis is recognized as a viable approach capable of performing well on previously unknown samples. To detect and categorize malware, static and dynamic analysis are used during training in this example.

Static approaches, on the other hand, examine the malware's code (assembly or machine) without executing it. On the other hand, dynamic techniques keep track of the malware's actions while it runs. Each approach of analysis has its own set of disadvantages. For example, the vulnerability in the code cannot be detected in a precise area using dynamic analysis, but static techniques excel at this. Static analysis, on the other hand, offers the advantage of detecting malware before it is executed. Static analysis does not allow for the restoration of control of infected systems, while dynamic techniques do.

1.2 Problem Statement

Malware categorization is important in malware analysis because it allows researchers to understand how different types of malware infect computers, how dangerous they are, and how to defend against them. When malware is detected, a classification technique is used to classify it and assign it to the