Chapter 10
Malware Analysis and Classification Using Machine Learning Models

Swadeep Swadeep
Vellore Institute of Technology, Chennai, India

Karmel Arockiasamy
Vellore Institute of Technology, Chennai, India

Karthika Perumal
Vellore Institute of Technology, Chennai, India

ABSTRACT

In modern times, it has become common practice for major corporations to utilize computers for storing data. Unfortunately, the frequency of malware attacks has increased, which facilitates unauthorized individuals' access to private information. Analyzing malware has become a critical task in safeguarding information systems against malicious attacks. Therefore, machine learning techniques have become an effective tool for automating investigations using static and dynamic analysis, combining malware with similar behavior into separate families based on proximity. Deep learning techniques improve the accuracy of malware variant detection and classification by building neural networks with more potentially different layers. This research aims to address this issue by training machine learning models using various algorithms on a dataset obtained by performing static and dynamic analysis on both malicious and benign samples. The resulting models were then combined to produce superior results compared to those obtained from a single model, which can be seen in the results.

1. INTRODUCTION

The current world heavily depends on computers and a microprocessor or a microcontroller can be found in all appliances today. This computing power is being used in all industries and is connected to the

DOI: 10.4018/978-1-6684-8531-6.ch010

world with the help of the internet. People are becoming more dependent on these systems to increase their productivity. But these vast amounts of systems also produce vast amounts of data and use vast amounts of computing resources. To gain access to these resources and the vast amounts of data, people try to infect the host computer with malicious software, also known as malware, which can be used for financial gains. Malware is created to cause harm to systems or networks and gain access to these systems even when they're not authorized.

2. LITERATURE SURVEY

In Akbar et al. (2021), various machine learning models have been used to classify APTs by converting the APT traces into a graph. Their work mainly focuses on identifying different tactics of Advanced Persistent Threats based on logs from executing MITRE ATT&CK framework, which is then reduced to make it noise-free, which also yields limited attack traces for identification. It was seen that SetConv, A New Approach for Learning from Imbalanced Data, gave the best accuracy metrics.

The main study of Rath et al. (2022) is to find the current trends and applications to achieve organization-level cyber security with the help of AI. This paper also talks about the various cyber threats such as viruses, worms, rootkits, botnets, etc. It also talks about the drawbacks of AI systems in cyber security such as the accumulation of data. They also had to gather large amounts of malicious codes, non-malicious codes, and various abnormalities to create a model which can be used at an organizational level. Using hypothesis testing, they concluded that the use of up-to-date preventive antivirus and anti-cyber protection software is a nice remedy for updated cybercrime and control. Cybenko & Hallman (2021) studies the various advantages that machine learning, game theory, and secure distributed computing offer for current technology such as IoT, unmanned autonomous vehicles, etc. They also look at adaptive cyber defense where the cyber defense techniques adapt to the changes of the attacker and the operating environment (such as reconfiguration). Here, the problem of improving performance while interacting with the real environment is studied with the help of distributed upper-confidence bound algorithm. Byzantine Fault tolerance and blockchain technologies are used in a distributed adaptive cyber-defensive system to make the systems robust under untrusted agent operations.

In Hota & Hota (2022), the trend of open banking is studied in recent years, and the various threats to open banking are discussed. Security in open banking is an important aspect, especially with the rise of open banking since the covid outbreak. According to their studies, 48% of the customers want the banks to provide product information based on their actions. It can be seen that data breaches and human error are the two primary risks associated with open banking. This paper also studies the risks faced in digital transactions and open banking. The evolution of advanced persistent threats increases the difficulty of detecting cyber-attack campaigns and hence Zou et al. (2020) studies various approaches to help detect such attacks. The common characteristics of APT attack campaigns are analyzed by studying past campaigns and looking at the impact they caused. Current methods to detect APTs are studied. It was seen that existing approaches were based on unsupervised connecting the dots through provenance tracking across multiple events that look safe singularly but malicious when together and have a drawback of dependence-explosion. A top-down approach can take advantage of known APTs and solve the problem of dependence explosion and an approach to this was modeled.

In Costales et al. (2020), a live attack on deep learning systems was proposed which patches the parameters of the model to achieve pre-defined malicious behavior on a specific set of inputs. The fea-