Chapter 1
Machine Learning for Malware Analysis:
Methods, Challenges, and Future Directions

Krishna Yadav
National Institute of Technology, Kurukshetra, India

Aarushi Sethi
National Institute of Technology, Kurukshetra, India

Mavneet Kaur
National Institute of Technology, Kurukshetra, India

Dragan Perakovic
https://orcid.org/0000-0002-0476-9373
University of Zagreb, Croatia

ABSTRACT

Companies and organizations are collecting all sorts of data ranging from nominal feedback like customer reviews to highly classified data like medical records. With data being such a critical aspect of most of the operations around us, cybercriminals are looking for an opportunity to misuse this information. One such device that cybercriminals use to further their malicious intent is malware. Over the years, these cybercriminals have become immensely powerful using the knowledge of previous attacks. Hence, malware analysis and methods to troubleshoot the problems arising due to malware attacks is the need of the hour. Over time, different new approaches have been developed to defend malware. However, in recent times, machine learning-based malware analysis has gained popularity. The capacity to detect possible future malware by learning from existing malware patterns makes this method very popular. In this chapter, the authors have introduced different malware and the machine learning-based approach that has been developed in recent times to mitigate malware.

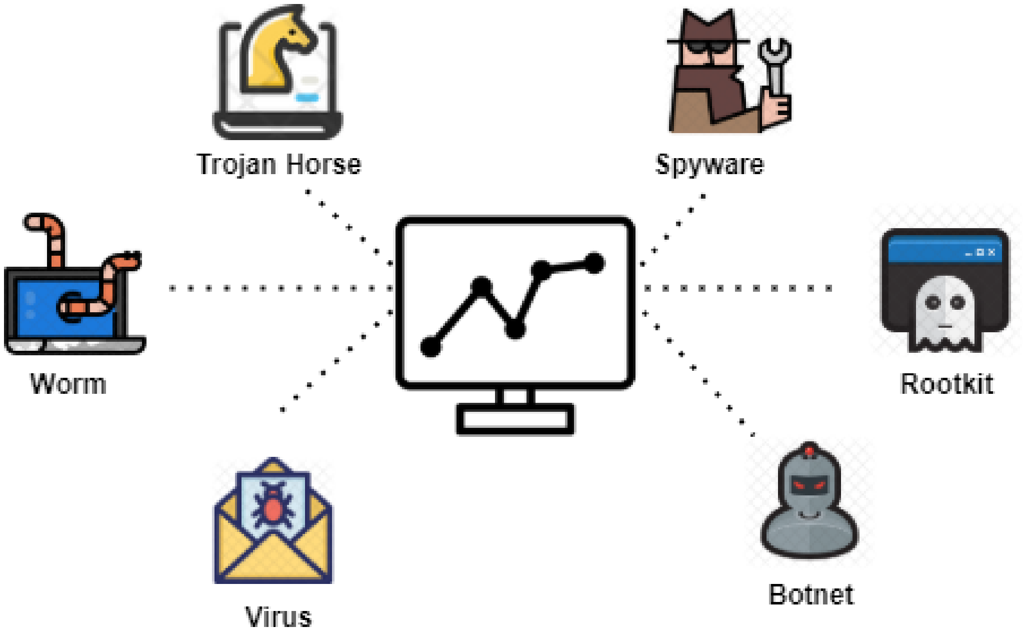
DOI: 10.4018/978-1-7998-7789-9.ch001

INTRODUCTION

Malware or malicious software is an umbrella term that comprises all the software that is designed with the purpose of corrupting or harming any program, device, service, or network. Once malware penetrates through the network and gets access to the files, it may infect or corrupt the data, steal the data or even use it for identity thefts. With IoT rapidly becoming a reality, more and more devices are being connected, which means that if one device in a system gets infected (Sharmeen, Shaila, 2019), all the other devices are at risk of being infected as well. The trend of the number of major cyberattacks over the years is pretty sporadic. However, with the introduction of machine learning models, there has been quite a downfall for the past two years (2019 and 2020), with the drop in the number of malware attacks going down by 43.3%. These machine learning models range from probabilistic to decision tree-based to deep neural networks. Over the years, numerous types of malware have been created that work in different ways to harm your files. Knowing what type of malware has infected your network is extremely helpful in order to find suitable techniques for patching.

There are various kinds of malware available these days. Figure 1 gives a pictorial representation of different types of malware. One of them is a computer virus (UCCI, Daniele, 2019). Computer viruses are one of the most common malware. They usually come attached to a file. Once the file is opened, it corrupts the system by transferring from one programme to another. It can transfer through programmes, computers, and even networks. Computer viruses can spread by email, text message attachments, audio files, etc. The next category of malware is worms. Unlike viruses, worms do not require any action by the victim; they replicate themselves by finding loopholes in the security of the software or operating system. The main purpose of worms is to destroy the files that it has access to and make them unusable by using certain encryption or corruption techniques.

Figure 1. Types of malware


Trojan Horse
Spyware
Worm
Rootkit
Virus
Botnet