DESIGN AND ANALYSIS OF ROBUST MACHINE LEARNING IN THE CONTEXT OF COMPUTER SECURITY

Thesis Abstract

Manaar Alam (Roll No.: 16CS91R07)
Supervisor: Prof. Debdeep Mukhopadhyay
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

The rapid advancements and increasing popularity of modern internet-connected devices have triggered their overwhelming adoption in almost every aspect of contemporary life and work. The enormity of data-oriented user interactions in the connected world ecosystem also creates an alluring and coveted threat surface for an adversary to compromise the security and privacy of millions of consumers and jeopardize the economy of numerous business organizations. The existence of such threats makes a safe and secure operation within the ecosystem more demanding. In this thesis, we first resort to the benefits of unparalleled real-world decision-making performance of Machine Learning (ML) / Deep Learning (DL) algorithms to develop security-critical applications for safe and secure operations in a connected world ecosystem. However, the growing success of such ML/DL algorithms also attracted the attention of numerous adversaries for exploiting different vulnerabilities in these applications. The malicious intentions of these adversaries pose a serious concern over their security and integrity, especially in security-critical applications. Hence, in the second part of the thesis, we aim to develop countermeasures to uphold the integrity of ML/DL applications against such serious threats and also preserve their compelling efficiency.

An adversary can employ a multitude of malicious activities in the connected world ecosystem like injecting malware programs, encrypting data using ransomware, stealing the secret key of a cryptographic computation using shared resources, and fault injection. In the first part of the thesis, first, we develop a dynamic malware detection approach using a lightweight statistical hypothesis testing methodology guided by gradient-based backpropagation, especially useful for embedded platforms. Next, we develop a ransomware detection approach advocating the benefits of LSTM-based autoencoders by analyzing the hardware footprints of unknown programs, which are hard to formalize and difficult to manipulate for an adversary to conceal its behavior. Next, we develop a generalized approach to contemplate a large class of microarchitectural side-channel attacks by analyzing multiple hardware resources simultaneously using the benefits of various ML algorithms. Then, we develop an inherently fault-tolerant cryptographic primitive instead of having any add-on redundancy-based countermeasures leveraging the fault-tolerance property of Neural Network architectures.

The success of ML/DL-based applications in the connected world ecosystem has been severely hindered with the advent of adversarial attacks, where an adversary intends to mislead an ML/DL classifier from its correct classification by manipulating a legitimate input with a visually imperceptible noise. In the second part of the thesis, first, we propose an ensemble-based detection of adversarial attacks by introducing a transformation in the input domain that promotes lower-level, less important features to develop an ensemble of models with diverse decision boundaries. Next, we extend the notion of ensemble-based detection of adversarial attacks by proposing a loss function that promotes misalignments of gradients responsible for

adversarial attacks among all the models within an ensemble and that does not require any modification in the input dataset.

Keywords: Malware, Ransomware, Micro-architectural Side-Channel Attacks, Fault Attacks, Adversarial Attacks, Deep Learning, Countermeasures