UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

MSc Defence

Thursday October 13, 2022 at 2pm via Zoom

Bardia Esmaeili

Image-based and Graph-based Adversarial IoT Malware Detection and Classification

Chair: Dr. Fangju Wang
Advisor: Dr. Ali Dehghananha
Non-Advisory: Dr. Hadis Karimipour [SoE]
Non-Advisory: Dr. Neil Bruce

Abstract:

With the emergence of the Internet of Things (IoT) networks, a variety of physical devices can now send, receive, analyze and act upon data collected from their surrounding environment to automate various tasks and reduce human involvement. IoT networks have led to faster and more energy-efficient applications in industry and society. The increasing demand for IoT networks has been followed by a rise in malicious programs attacking IoT devices to exploit their data and sabotage their networks. As a result, researchers have investigated the malware software characteristics as well as algorithms that may assist in recognizing these harmful entities. Deep Learning (DL) and Machine Learning (ML) models have proven effective in classifying malware and have been widely adopted. However, these models are susceptible to a security-related vulnerability known as an adversarial attack, where the attacker can manipulate the input and deceive the model. An adversarial malware bypassing an ML or a DL-based model could have catastrophic consequences.

Therefore, in this study, we attempt to identify adversarial samples against DL-based IoT malware classifiers by utilizing adversarial detection models with image-based and graph-based input representations, widely adopted mediums for this task. The image-based detection model is optimized to generate new representations spatially close to each other
for conceptually similar inputs and considerably distant from one another for distinctive inputs. Then, an attempt to create an adversarial example could be detected if the recent inputs to the detector have been too similar. On the other hand, the graph-based detection model learns the distribution of non-adversarial data such that it can assign a high score to unseen adversarial inputs and identify them. We considerably improve the detection performance against adversarial examples on both mediums with 93.1% for the image-based adversarial detector and 98.96% for the graph-based adversarial detector.