Chapter One

Introduction

1.1 Background of the Problem

The traditional approaches of malicious Android apps detection are various. At the same time, it is not deniable that each of which has proven to be struggling with number of limitations that render them ineffective in the face of the continuous evolving of cyber-attacks and hacking mechanisms. Traditional approaches of detection either memorize a list of the attacks they should block/allow, or identify a normal behavior, which is used as a threshold for comparisons against other behaviors being investigated. In either cases, attacks can slip away from the detection systems when they take advantage of limitations of these traditional methods. [1][2]

In a Machine Learning approach that is adopted to help in malicious apps detection, the algorithm or the model learns rather than memorize; something that is useful for the task of identifying hidden patterns and associations. Identifying those hidden patterns is difficult without the use of sophisticated models built upon algorithms that are based on advanced probabilities, math and/or statistics. For those patterns are major contributors to insight forming and decision making. [3]

Addressing the Android detection problem with Machine Learning is not a straightforward and easy task itself. It involves number of considerations that need to be regarded.

One must observe the Android app having in mind the data factor, and thinks of ways that represent the behavior of the Android in a dataset that is efficient enough for the ML model to start its working with, in addition to the dilemma of which model to choose. Moreover, with the increasing number of Android applications attacks, the detection problem tends to have different natures and thus different considerations for each problem.

This research tries to present an analysis framework that guides security practitioners and ML analysts who are aiming to combine both fields to tackle the detection problem of malicious Android applications.

1.2 Statement of the Problem

The questions that the research is trying to answer are:

1. How Android security detection problem can be addressed with machine learning?
2. Do machine learning methods themselves differ from each other? Or are they all similar in the ways they operate?
3. And whether the machine learning approaches exceed the traditional detection methods.
4. And in what ways are the machine-learning based detection approaches are limited?

1.3 Objectives of the Research

The research aims to evaluate and compare between the different machine learning methodologies that are applied to detect malicious Android applications. It also focuses on teaching the common practices that are followed to handle a detection problem from a data science perspective, starting from the very beginning of how data regarding Android applications are obtained, extracted, and analyzed, until how to decide which machine learning algorithm to choose among the many available alternatives.

The study also explores the opinions of Android developers and mobile application security practitioners on whether they would accept an intelligent-based approach that separate between malicious from a benign Android app, over the traditional state-of-the art methodologies.

1.4 Scope of the Research

The study has specified two scopes: first, from the Android security side, the focus was on the malware analysis approaches that are based on static analysis. Approaches based on dynamic analysis were not handled in a detailing manner, but a comparison between static and dynamic analysis was briefly mentioned.

From the machine learning side, the scope focuses on the supervised-learning methods, particularly the classification algorithms. For that reason, unsupervised and semi-supervised algorithms are not included.

1.5 Significance of the Research

The significance of this study is showed in many points. First, the limitations of the traditional approaches have pressed the need to seek alternative approaches that adopt a level of intelligence, instead of relying on rule-based unintelligent frameworks.

Second, it signifies on how the intelligent solutions themselves are applied differently, and on what ways they are similar or different from each other, and if they overlap.

There is an ongoing debate regarding this topic, between the mobile-application security practitioners and machine-learning practitioners, and among the ML community as well. Casting the light into the similarities and differences would help in future in a better understanding and implementing of machine learning methodologies in the area of Android security.

1.6 Structure of the Thesis

The thesis is divided into five chapters. The first chapter is the introductory, which speak of the main problem and the scope and significant of the research. Chapter two is the background, which explains the necessary Android prerequisite knowledge that is needed in order to understand the following chapter, chapter three, which compares between the different machine learning algorithms that were being applied to tackle the Android security problem.

Chapter four handles the practical part of the research, at which a sample dataset was being experimented with using 5 different algorithms within two different approaches. The thesis is concluded with chapter five, which illustrates the limitation of the study, in addition to recommendations and final thoughts for future researches in the area of machine learning applications in Android security.