Republic of Iraq
Ministry of Higher Education
and Scientific Research
University of Diyala
College of Science

Detection of Malware under Android Mobile Application

A Thesis
Submitted to the Department of Computer Science\ College of Science \ University of Diyala as a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Saja Ibraheem Hani Ismail

Supervised By

Prof. Naji M. Sahib

2020 A.D. 1441 A.H.

يَا أَيُّهَا الَّذِينَ آمَنُوا إِذَا قِيلَ لَكُمْ تَفَسَّحُوا فِي الْمَجَالِسِ فَافْسَحُوا يَفْسَحِ اللَّهُ لَكُمْ ۖ وَإِذَا قِيلَ انْشُزُوا فَانْشُزُوا يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنْكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ وَاللَّهُ بِمَا تَعْمَلُونَ خَبِيرٌ ﴿١١﴾

سورة المجادلة آية ﴿١١﴾

Dedication

To those who taught me how to stand firmly on the ground

Dear father

To the source of love, altruism and generosity

My Mother

To the closest people to myself
My faithful husband

To my soul, my eyesight, and my heartbeat
My children

I present to you a summary of my scientific efforts

Saja Ibraheem

Acknowledgment

At the end of my thesis, I am pleased to thank and show my gratitude to those who have helped me to achieve my study whether through assisting to provide resources or with guidance and presenting me advice.

First, I would like to thank my honorable prof. Naji Muttar Sahib who supervised my thesis and offered me many scientific advices and guidelines; to him I give all of my appreciation.

In addition, I want to thank my respected professors in Department of Computer Science with whom I completed a very important stage of education within the academic journey through their scientific experiences and helping us to complete it.

Also, I would love to thank prof. Azhar Hasan Nussief who provided me with solid resources, so I give my gratitude to her.

In addition, I want to thank my family who supported me and stood by me during my study journey.

Saja Ibraheem

Supervisors' Certification

I certify that this thesis entitled **"Detection of Malware under Android Mobile Application"** was prepared by **"Saja Ibraheem Hani"** under my supervisions at the University of Diyala collage of Science Department of Computer Science, as a partial fulfillment of the requirements needed to award the degree of Master of Science in Computer Science.

(Supervisor)
Name: Prof. Naji M. Sahib

Signature:
Date: / / 2020

Approved by University of Diyala\ College of Science\ Department of Computer Science.

Signature:
Name: Asst. Prof. Dr. Taha Mohammed Hasan
Date:

Head of Computer Science Department

(Linguistic Certification)

I certify that this research entitled "**Detection of Malware under Android Mobile Application**" was prepared by "**Saja Ibraheem Hani**" and was reviewed linguistically. Its language was amended meet the style of English language.

Signature:
Name:
Date:

Scientific Amendment

I certify that the thesis entitled "**Detection of Malware under Android Mobile Application"** was prepared by **"Saja Ibraheem Hani"** has been evaluated scientifically; therefore, it is suitable for debate by examining committee.

Signature :
Name :
Date : **/ / 2020**

Abstract

Smartphones have become essential in our daily life. It also can do a lot of work and can browse the Internet, and download many applications for each device, through the available store. As a result, the number of malware applications downloaded also increases.

This malware carries out various activities behind the scenes; Such as confidentiality, breach of privacy, loss of confidentiality, system breakdown, theft of sensitive information, etc.

Many types of research and studies that proposed different techniques to detect malicious programs, but they contained weak points, which are illustrated by efficiency, speed, and lack of comprehensiveness.

In this thesis, a proposed system is developing implementing to detect malware in smartphones, and contains two parts:

In the first part, access control is initially started upon the system launch. The user authentication algorithm adopts the user's permission to detect a threat factor after applying a user's permission policy by improving the method with which the user's activities are extracted. While, **in the second part,** anomaly detection technology begins to extract the important features that play an effective role in detecting malicious code.

The proposed system has been tested by using a hybrid genetic algorithm, and the SVM data has been registered an accuracy of (0.9282).

The experimental results indicate that the proposed system has a high average accuracy rate compared to other existing methods where it (0.8848) average accuracy using PNN, while the average accuracy is (0.8835) and (0.8715) with SVM and K-NN respectively.

Table of Contents

List of Figures

List of Table

List of Abbreviations

Abbreviations	Meaning
API	Application Programming Interface
APK	Android Application Package
AV	Anti-Virus
GA	Genetic Algorithm
IOS	IPhone Operating System
JVM	Java Virtual Machine
K-NN	K-Nearest Neighbors
PNN	Probabilistic Neural Networks
RF	Random Forests
SVM	Support Vector Machine
PE	Portable Executable
DEX	Dalvik Executable
XML	Extensible markup Language
HTML	Hypertext Markup Language
SQL	Structured Query Language

Chapter One

Introduction

Chapter one

Introduction

1.1 Overview

In the past few years, it was clear that smartphone users have increased exponentially. Besides, the operating systems for smartphones are Symbian, iPhone Operating System (IOS), Android, and Blackberry. The smartphone is viewed as a portable Personal Computer system, PCs, as they have all the functionalities of a desktop PC integrated into them. Just as there are hackers/attackers releasing malware for PCs, there are attackers who are now targeting smartphones. The main reason for this is that mobile security is still in its initial stages and the lack of user awareness regarding how their devices can be undermined by using if they are not careful enough. Google's is open-source operating systems. Android is among the most popular smartphone operating systems. Android is a Linux-based operating system that also includes key applications and middleware. In order to fully benefit from and explore the functionality of Android, Google allows third-party developers to create applications and release them to the Android [1].

A recent work indicates the number of mobile applications is increased extremely which also increases malware application as shown in Figure (1.1) [2].

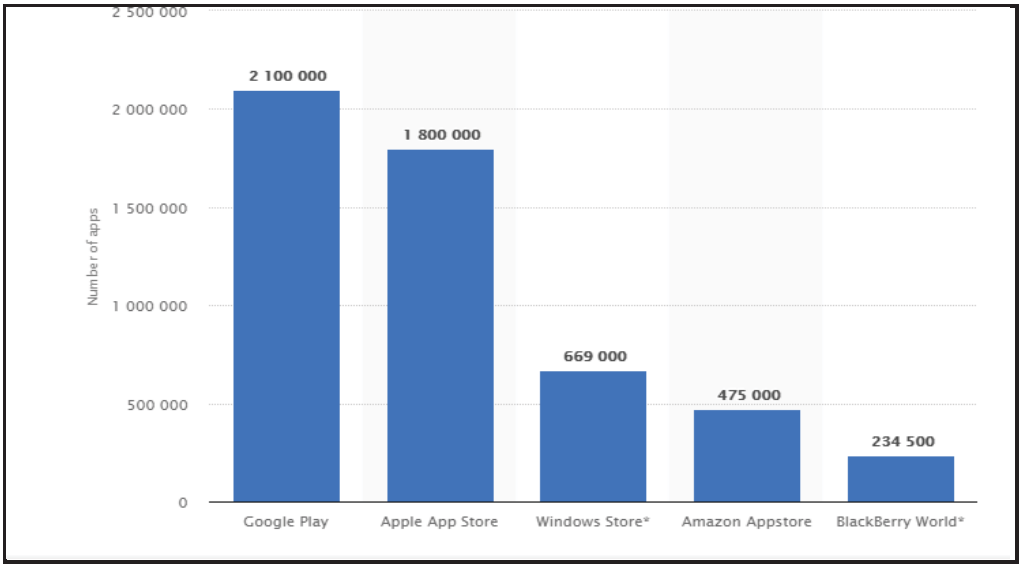

Figure (1.1): Rate of mobile Application downloaded [2]

Malware produced as malicious content that damages the stand-alone computer or networked computers damaged by its harmful effects. It can come in many forms; Spyware, Worm, Trojan or Virus, etc. But whatsoever the form is; its function is the same that is to harm the computers. This intrusive virus can be executable code or no- executable code. Many malware types have the ability to multiply so strict precautions are needed [3].

There are various types of malicious software with different structures, properties, and effects. They also vary in the intensity of the threats they pose [4].

There are different types of malware, the most common one is the virus (an infected code), which after execution multiplies itself and infects other files. It also adds malicious code to other files in order to attack more vigorously. Although viruses are hosted and controlled by a third party, worms create all the damage without being controlled by anyone. They can propagate themselves by infecting the other files. Another side

we have that worms are simply standalone malicious software, which creates the same damage as viruses [5].

Another most common type of malware is adware. It is supposed to be used solely for advertisement and generating revenue. However, nowadays adware has been combined with spyware, it keeps track of user's activities. Addition, Spyware attacks to steal user's sensitive information. On the other hand, adware is simply the popping up of ads on websites or applications. Normally this comes up with free software. Intruders are making use of it and transform adware into spyware stealing user's information and misusing it [6].

Spyware is one of the dangerous malwares, which keeps an evil eye on the user's activities by recording those keystrokes and personal information. Furthermore, personal information can vary from login credentials to sensitive bank account details. Not only stealing users' information, but also spyware can intrude into the user' s computer in order to change some software' s security or privacy settings or browser' setting making the networks public. The other type of malware is the bot which is an automatic malicious code that intrudes exclusively network of computers. The bots are casually used for positive purposes, but now malicious harmful attacks have been imposed on them. Denial-of-service-attack is an attack on the host computer that transfers its virus to all the networked computers, which is an output of bots. Also, this is a malicious code, which works automatically. Spambots are one of its types, which spam the internet with malicious websites. It is difficult to get rid of such malware but not impossible [7]. Ransomware is also a type of malware, which takes over the hard drive of the user, and the user has to pay some ransom to regain access. It is a crucial kind of malware, which restricts user access to its own computer. Furthermore, it spreads through the downloaded file or any vulnerability in the network service. Trojan horse

is a trick to the users; it presents itself as an authentic file which the user can download. Afterward, the attacker gets access to the infected computer remotely. Now the Trojan can add more malware to the infected computer, control the security configurations, monitor user key logs, steal sensitive information, etc. Such infected computers can be used in botnets [8].

The user authentication algorithm of user permission is dependent in order to detect a threat actor after applying the user permission to approach via the improvement of the user activities extraction method.

In addition, to adapt the unknown sessions and use the rule-based on integration with the attacks, the heuristic sequence will be used. A very important aspect is the identification of the classification algorithm of the most reliable detection accuracy. Therefore, the strongest classifiers are identified through the evaluation of the activeness and detection accuracy of all machine-learning algorithms. Modifying the default input values can enhance the efficiency and accuracy of a classifier. However, enabling an equivalent comparison between the classifiers dictated the implementation of the classifiers with their default input values. Many studies have been emerged to discover and treat malicious programs based on the artificial intelligence algorithm by many researches, such as K-Nearest Neighbors (k-NN), Naive Bayes, Random Forests (RF), Support Vector Machines (SVM), and Genetic algorithm [9].

The k-NN algorithm belongs to the family of methods known as instance-based methods. These methods are based on the principle that observations (instances) within a dataset are usually placed close to other observations that have similar attributes; this method selects the closest observations from the dataset in such a way to minimize the distance.

In machine learning, random forest (RF) is already widely used in bioinformatics the best available methods and superior to most methods

in common use. As the name suggests, RF combines many classification trees to produce more accurate classifications. By-products of the RF calculations include measures of variable importance and measures of similarity of data points that may be used for clustering, multidimensional scaling, graphical representation, and missing value imputation.

The machine learning techniques are Support Vector Machine (SVM) which is used for binary classification. It is a very general technique that can be applied in a wide variety of situations. Also, it has special characteristics that are used to implement efficient parallel algorithms in terms of time and memory. One characteristic is that the solution to the classification problem is obtained by only a few samples called Support Vectors (SV) that determine the maximum margin separating hyperplane. Another characteristic of SVM is to perform the nonlinear mapping without knowing the mapping function using predefined functions called kernels for calculating the inner product of mapping functions [10].

Genetic algorithm (GA) is a class of stochastic global search techniques based on biological evolution principles; several have applied the genetic algorithm to geophysical optimization problems such as seismic attributes. The genetic algorithm represents parameters as an encoded binary string and works with the binary strings to minimize the cost, while the other works with the continuous parameters themselves to minimize cost [11].

1.2 Related work

Several researchers have shown their interest in the detection of malware in a smartphone, the following are some of the published works that are relevant to the current work:

- **Lu, et al., 2013 [12]:** compared Bayesian method alone and Bayesian method combined with Chi Square feature selection method results are compared to evaluate the performance of the two ML algorithms. The study concluded that Bayesian method with Chi Squared yielded an accuracy of 89% while Bayesian method alone yielded 80%.
- **Nuray Baltaci, et al., 2014 [13]:** The main purpose of the study is to investigate the contribution of other application market metadata to the detection of malicious applications in addition to requested permissions. Hence, the information of applications presented on the official market when a user wants to download them was used as the feature set for training supervised classification algorithms.
- **Kurniawan, et al., 2015 [14]**: used Logger, a default application which is inbuilt in Android was used to extract the sum of Internet traffic, percentage of battery used and battery temperature for every minute. This information collected as set of features and is fed into weka, an open source learning library for testing and training with Naive Bayes, J48 decision tree and Random Forest algorithms. The author concluded that Random Forest has high accuracy of 85.6% with these features and proposes other features that can be combined with existing system to improve the accuracy.
- **Weng, et al. [15]:** in their work, published in 2017, propounded a model that classifies using machine learning techniques such as SVM, the nearest neighbors, by extracting 11 deferent static features. This model can be used in the management of large application markets. A correct classification rate of 99% for the malicious set and 82% for the benign set was obtained working with 100,000 benign and 8,000 malicious application set.

- **Mohsen Kakavand, et al.,2018 [16]:** this work involves in static analysis of applications, which checks for the presence and frequency of keywords in Android application' manifest file and drives the static feature sets from a 400- application dataset to produce better malware detection results. The classification performance of the ML algorithms is measured in terms of accuracy and true positive rate and interpreted to determine which algorithm is more applicable for the Android malware detection. The experimental results for a dataset of real malware and benign applications indicate the average accuracy rate of 79.80% and 80.50% with average true positive rate of over 67% and 80% using SVM and KNN, respectively.

- **Matthew Leeds, et al., 2017 [17]:** Malware is a current threat facing Android users. As users have come to depend on these devices for communication and information, it is essential to make sure they are secure. Therefore, developing and testing new sophisticated malware detection techniques must be a priority. This paper compared two prominent features used to detect Android malware, permissions and system calls, and applied machine learning to both. The results showed that permissions data was better at detecting malware than system call data. An average classification accuracy rate of 80% was achieved when using permissions data to determine malicious activity on Android devices. Therefore, it is a reliable way to detect malware.
- **Michal Kedziora, et al., 2018[18]:** In this study, an overview of Android malware analysis was presented, and a unique set of features was chosen that was later used in the study of malware classification. Five classification algorithms (Random Forest, SVM, K-NN, Nave Bayes, and Logistic Regression) and three attribute

selection algorithms were examined in order to choose those that would provide the most effective malware detection.

- **H al-kaaf1, et al., 2019 [19]:** In this study, proposed feature selection methods to identify clean and malicious applications based on selecting a set combination of permission patterns using different classification algorithms such as sequential minimal optimization (SMO), decision Tree (J48), and Naive Bayes. The experimental results show that sequential minimal optimization (SMO) combining with the SymmetricalUncertAttributeEval method achieved the highest accuracy rate of 0.88, with the lowest false positive rate of 0.085 and the highest precision of 0.910. And the findings prove that feature selection methods enhanced the result of classification.

1.3 Problem Statement

Develop a new method for more classifications performance serve to protect Mobil Application malware.

Applications on smartphones must take Care when Downloading, Due to that, many malicious attacks target them. The majority of operating systems in the smartphone business are operating using the Android OS. However, around 97% of mobile malware targets Android phones. In the Therefore, these incidents motivated us to study mobile application security, especially in Android because the viruses pose risk to the applications as well as the operator.

1.4 Aim of Thesis

The current study aims to find whether or not Google Play, Google's app, and Android's official application market, metadata of Android

applications assist in explaining the malicious behaviors when joined with user's permissions analysis.

1.5 Contribution

Using a hybrid system to detect malware based on static and dynamic approaches. Hence, this contribution will provide protection for the Android mobile based-on access control and anomaly detection.

1.6 Thesis Outline

Beside this chapter, the remaining parts of this thesis include the following chapters:

Chapter Two: Theoretical background

The start with an overview of the Android system architecture and describe the implementation design of Android. Also, discuss an overview of the core components which are found in android applications and it's included in the concepts. That concept is used in this thesis, where the methods used in the malware app such as classifying android apps, SVM, and Genetic Algorithm.

Chapter Three: The propose Detection Malware System

In this chapter, the discussed the proposed system for the Authentication process and the check authentication method with Algorithms SVM as well as the Genetic algorithm.

Chapter Four: Implementation and Experimental Results

This chapter involves studies and results, which are obtained from the system running as well as the performance measures of the results of the test, and comparisons.

Chapter Five: **Conclusion and Future Work**

In this chapter, the present a list of conclusions from the results of the presented work and some suggestions for future works.

Chapter Two

Theoretical Background