MALWARE DETECTION BY USING TWO-LAYER STACKING SVM CLASSIFIER

RAJA NURUL AINI BINTI RAJA MOHD ANUAR

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS LAPORAN

JUDUL: MOBILE MALWARE DETECTION USING TWO-LAYER STACKING SVM CLASSIFIER

SESI PENGAJIAN: [2020 / 2021]

Saya RAJA NURUL AINI BINTI RAJA MOHD ANUAR mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan unituk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. * Sila tandakan (✓)

_____	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_____	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan)
_____	TIDAK TERHAD	

(TANDATANGAN PELAJAR)

(TANDATANGAN PENYELIA)

Alamat tetap: Batu 17, Dusun Tua
Seberang 43200, Hulu Langat, Selangor

Ts Nor Azman Bin Mat Ariff

Tarikh: 12 SEPTEMBER 2021

Tarikh: 12 SEPTEMBER 2021

CATATAN: * Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

MOBILE MALWARE DETECTION USING TWO-LAYER STACKING SVM CLASSIFIER

RAJA NURUL AINI BINTI RAJA MOHD ANUAR

This report is submitted in partial fulfillment of the requirements for the Bachelor of [Computer Science (Computer Network)] with Honours.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2021

DECLARATION

I hereby declare that this project report entitled

MOBILE MALWARE DETECTION BY USING TWO-LAYER STACKING SVM CLASSIFIER

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT: _____ Date : 12 SEPTEMBER 2021

RAJA NURUL AINI BINTI RAJA MOHD ANUAR

I hereby declare that I have read this project report and found this project report is sufficient in term of the scope and quality for the award of Bachelor of [Computer Science (Software Development)] with Honours.

SUPERVISOR : _____ Date: 12 SEPTEMBER 2021

TS. NOR AZMAN MAT ARIFF

DEDICATION

Special dedicated to my beloved parents and fried who have encouraged, guided and inspired me throughout this journey of education

To my helpful lecturer, thank you for the guidance from the beginning until the end of this final year project

ACKNOWLEDGEMENTS

All praise to Allah SWT for giving me the strength to finish my PSM, I manage to complete this Projek Sarjana Muda 1 (PSM1) and Projek Sarjana Muda 2 (PSM2). I would like to thank Ts. Nor Azman Bin Mat Ariff for giving assistant to complete this project successfully. Without his guide, this report and project cannot be completed.

Special thanks to my family for the continuous support throughout this journey from day to day. I would like to thanks to all my friend for their time, concern efforts, and always encourage me when preparing this report. With all the help from involves parties, I manage to complete my report and project which is Mobile Malware Detection Using Machine Learning.

Last but not least, thanks to Universiti Teknikal Malaysia Melaka (UTeM) for the opportunity given.

ABSTRACT

Mobile device usage increased with new high technology that attracted the attacker to launch the attack, such as mobile malware. Mobile malware is malicious that is loaded into the device system and causes damage. Malware has become more prevalent in recent years. Nowadays, there are many techniques available to detect this attack. In this research, N-gram with opcode sequence dataset was used, focusing on mobile malware detection model. The dataset undergoes feature selection which is Information Gain and Chi-square, to reduce irrelevant data and redundant data. Then, the classifier used in this project is Support Vector Machine (SVM) to develop a model. The developed model will test and verify the accuracy with the test set. The project is giving hope to produce a system that can detect mobile malware attacks.

ABSTRAK

Penggunaan peranti mudah alih meningkat dengan teknologi tinggi baru yang menarik penyerang untuk melancarkan serangan, seperti perisian hasad mudah alih. Perisian hasad mudah alih berniat jahat yang dimuat ke dalam sistem peranti dan menyebabkan kerosakan. Perisian hasad semakin berleluasa dalam beberapa tahun kebelakangan ini. Pada masa kini, terdapat banyak teknik yang ada untuk mengesan serangan ini. Dalam penyelidikan ini, N-gram dengan dataset urutan opcode digunakan, dengan fokus pada model pengesanan malware mudah alih. Set data menjalani pemilihan ciri iaitu *Information Gain* dan *Chi-Square*, untuk mengurangkan data yang tidak relevan dan data yang berlebihan. Kemudian, pengklasifikasi yang digunakan dalam projek ini adalah Mesin Sokongan Vektor (SVM) untuk membangunkan model. Model yang dibangunkan akan diuji dan mengesahkan ketepatannya dengan set ujian. Projek ini memberi harapan untuk menghasilkan sistem yang dapat mengesan serangan perisian hasad mudah alih.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

LIST OF ABBREVIATIONS

FYP	**-**	**Final Year Project**
IT	**-**	**Information Technology**
OS	**-**	**Operating System**
ML	**-**	**Machine Learning**
AI	**-**	**Artificial Intelligence**
IG	**-**	**Information Gain**
CS	**-**	**Chi-Square**
NLP	**-**	**Natural Language Process**

LIST OF ATTACHMENTS

CHAPTER 1: INTRODUCTION

1.1 Introduction

Nowadays, mobile device usage is increasing and it is expanded with new advanced technology every year. The continuous development of mobile hardware and technology has created a highly connected world. Mobile devices are used in many fields such as business, healthcare, education, and, Information Technology (IT). This transformation of mobile devices is attractive to the attacker to attempt to steal the information from the devices or system. Malware is malicious software that loaded onto mobile devices and causes damage then destroys it. The more malware attacks are introducing every year and it needs to be overcome before it widespread. This project focuses on mobile malware attack detection using a machine learning algorithm technique. This technique was chosen because it has good performance to track the malware and able to get high accuracy in classification.

1.2 Problem Statement

The mobile device becomes a place where intruders try their malicious code to fulfill their intention. There have many researchers proposed to prevent the mobile malware attack by introducing techniques that can protect the devices. Malware developers think it is easy to transfer attackers on mobile devices because mobile applications downloaded from Apple App Stores, Google Play, and websites usually have no time to analyze them (Gyamfi & Owusu, 2019). The current malware can do anything, their threat level becomes bigger and keeps growing. The variation of this malware and the exposed vulnerabilities require new and improved methods for detection.

There are plenty of attack situations in which an attacker can compromise a user's information by exploiting operating system's vulnerabilities. The growth of malware has led to different mobile antiviruses, firewalls, and encoding products, which have been released by various security vendors such as F-Secure, Kaspersky Lab, McAfee, and Symantec. The number of F-Secure mobile malware in 2008 was 401, while McAfee numbered 457. Mobile malware is malicious software that has certain mobile device targets. It first appeared for Symbian Operating System (OS) in 2004 and has now grown exponentially with smartphones in popularity (Gyamfi & Owusu, 2019). Figure 1.1 shows the mobile malware growth from 2010 up to 2018.

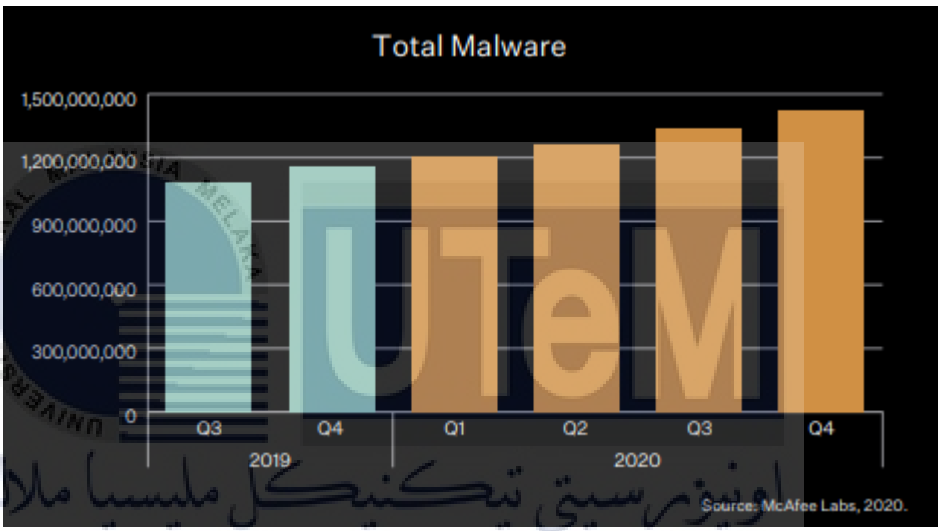

Figure 1.1: Total Malware by Year
(McAfee 2020)

COVID-19 pandemic Movement Control Order (MCO) is another alarming issue with mobile malware attacks. According to McAfee (2020), McAfee Lab's average volume of malware threats was 588 per minute, up from 169 per-minute threats in the third quarter of 2020 (40%). The fourth quarter was 648 per minute threats, an increase of 60 per minute threats (10%). Mobile malware grew 118% by SMS Reg driven from Q3 to Q4 of 2020. The best solution to the problem is the detection of mobile malware. The problem was defined as in table 1.1.

Table 1.1: Problem Statement

PS	Problem Statement
PS1	An increasing number of users connect to the Internet has been driven by the tremendous growth in mobile and smart device usage, it is hard to discover effective and fast technique in differentiate between benign and malware. Another issue is to identify the suitable machine learning technique to use in this project to process the dataset.

1.3 Project Question (PQ)

In this project, there are three Project Question (PQ) needed answers made out from the problem statement. Table 1.2 shows the summary of the project question based on the problem statement.

Table 1.2: Summary of Project Question

PS	PQ	Project Question
PS1	PQ1	What is malware in mobile devices?
	PQ2	How mobile malware classified?
	PQ3	Is the machine learning technique give a result model accurately?

1.4 Project Objective (PO)

Project Objective (PO) is a goal for the expected outcome of the research. To fulfill the goals of the project, the problem statement and project question must be assigned. Three objectives in this research to achieve and the relationship between PS, PQ, and PO for this project as shown in table 1.3.

Table 1.3: Summary of Project Objective

PS	PQ	PO	Project Objective
	PQ1	PO1	To study the behavior of malware.
PS1	PQ2, PQ3	PO2	To implement machine learning classifier to detect malware attack.
	PQ1, PQ2, PQ3	PO3	To test and verify the accuracy of machine learning to detect the malware.

1.5 Project Scope

The scope of this project is related to the following criteria:

I. The dataset only focuses on mobile malware.

II. This project focus on machine learning method.

III. The effectiveness of the proposed model is measured using accuracy.

1.6 Project Contribution

The result of this project is depending on the project implementation to detect the mobile malware attacks and select effective features that will give high accuracy. The following is a brief description of this project's contribution:

I. Identification of different mobile malware attacks based on their taxonomy.

II. Propose a model that can detect malware attacks on the mobile device.

III. Propose to check the accuracy of mobile malware attack detection and test and validate the developed program.

1.7 Thesis Organization

This report divided to seven chapter namely Chapter 1: Introduction, Chapter 2: Literature Review, Chapter 3: Project Methodology, Chapter 4: Analysis and Design, Chapter 5: Implementation, Chapter 6: Testing and Validation and Chapter 7: Project Conclusion.

1.7.1 Chapter 1: Introduction

This chapter act as the guideline to identifying the important things prior to the project. It will explain what and how this project attempted to accomplish, as well as its relevance.

1.7.2 Chapter 2: Literature Review

This chapter consists the related studies and early task from previous researcher about this project and critical analysis of this chapter.

1.7.3 Chapter 3: Project Methodology

Chapter 3 focuses on planning works that must answer the objectives of this project that has been mention previously. This chapter includes the procedures throughout this project.

1.7.4 Chapter 4: Analysis and Design

This section is related to the procedure that correlate with the experiment of this project. The experiment needs to be analyzed the effectiveness by referring to the literature review and design the project clearly to reach the goals.

1.7.5 Chapter 5: Implementation

Chapter 5 is the main part that the procedure from the previous chapter needs to be carried out. The experiment will be processed in this section until has the result and it will be gathered the result to be compared with another method.

1.7.6 Chapter 6: Discussion

This chapter explains the result and analyzed them where the findings are answers the objectives.

1.7.7 Chapter 7: Project Conclusion

This chapter will summarize the project, state the contribution, and highlight the constraints that were encountered throughout the project. This chapter will also describe how to improve the project in the future.