A MOBILE MALWARE DETECTION FRAMEWORK BASED ON ENSEMBLE CLASSIFIER OF MULTIPLE N-GRAM OPCODE PROBABILITY OUTPUT

NOOR AZLEEN BINTI ANUAR

MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION TECHNOLOGY

2023

Faculty of Information and Communication Technology

A MOBILE MALWARE DETECTION FRAMEWORK BASED ON ENSEMBLE CLASSIFIER OF MULTIPLE N-GRAM OPCODE PROBABILITY OUTPUT

NOOR AZLEEN BINTI ANUAR

Master of Science in Information and Communication Technology

2023

A MOBILE MALWARE DETECTION FRAMEWORK BASED ON ENSEMBLE CLASSIFIER OF MULTIPLE N-GRAM OPCODE PROBABILITY OUTPUT

NOOR AZLEEN BINTI ANUAR

A thesis submitted
in fulfilment of the requirements for the degree of Master of Science in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2023

DECLARATION

I declare that this thesis entitled "A Mobile Malware Detection Framework based on Ensemble Classifier of Multiple N-Gram Opcode Probability Output" is the result of my own research work except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

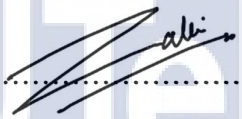
Signature : .....

Name : Noor Azleen Binti Anuar

Date : 18th August 2023

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature : 

Supervisor Name : Ts. Dr. Mohd Zaki Bin Mas'ud

Date : 18th August 2023

DEDICATION

This thesis is dedicated with

Deepest love and affections to my beloved parents,

Anuar bin Jabbar and Rudziah binti Adam

Brother and sisters

Syafiq, Syaziela, and Amirah

Their love, patience, guidance, wisdom and strength

Have inspired me throughout these years in

Universiti Teknikal Malaysia Melaka

To be the best that I can be.

ABSTRACT

The advancement of mobile devices nowadays lets users do varieties of activities including surfing the internet, online banking transactions, engaging in social networking and hence increasing the usage of mobile devices. This scenario opens the possibility for cybercriminals to launch a mobile malware attack towards users. The complexity of detecting mobile malware also contributes to the possibility of mobile malware remaining dormant in the application store which can expose users to being tricked into installing the infected programs. Current mobile malware detection methods such as Static analysis and signature-based detection can address these issues, but it can be very difficult to detect zero-day or obfuscated code because it relies on a unique signature. Meanwhile, Dynamic analysis and anomaly-based detection can curb the problem, yet it can result in a relatively high rate of false alerts. In addition, a single model classifier is not strong enough to produce a good detection result. Based on this reason, this research intends to enhance the current Mobile Malware Detection Framework using multiple N-Gram opcode probability output and weighted ensemble to enhance the accuracy, TPR, and FPR. The aim of this research is to identify the features of malicious activity from mobile malware application through static analysis. The features obtained were used in formulating and evaluating the enhanced MMD Framework. The generation of N-Gram opcode sequence represents the malicious features and feature selection method is used to search for optimum features. Additionally, the weighted ensemble method is introduced to combine several probability outputs from multiple classification models. Particle Swarm Optimization is used in searching for optimum weight to be used together with the probability output to improve mobile malware detection. In conclusion, the proposed MMD Framework had shown an enhanced performance with an accuracy of 96.55%, TPR of 99.10%, and FPR of 0.90%. Based on the encouraging results, future studies could explore the possibility of using a dynamic analysis detection approach and applying n-gram to features other than opcode sequence. Ultimately, other datasets and other mobile malware variants should also be explored in future.

RANGKA KERJA PENGESANAN PERISIAN HASAD MUDAH ALIH BERDASARKAN PENGELAS ENSEMBEL BAGI OUTPUT KEBARANGKALIAN OPKOD N-GRAM BERBILANG

ABSTRAK

Kemajuan peranti mudah alih pada masa kini membolehkan pengguna melakukan pelbagai aktiviti termasuk melayari internet, transaksi perbankan dalam talian, terlibat dalam rangkaian sosial dan lain-lain. Walau bagaimanapun, senario ini membuka kemungkinan penjenayah siber melancarkan serangan perisian hasad mudah alih terhadap pengguna. Kerumitan dalam mengesan perisian hasad mudah alih juga menyumbang kepada kemungkinan perisian hasad mudah alih kekal tidak aktif di kedai aplikasi yang boleh mendedahkan pengguna untuk ditipu untuk memasang program yang dijangkiti. Kaedah pengesanan perisian hasad mudah alih semasa seperti analisis statik dan pengesanan berasaskan tandatangan boleh menangani isu-isu ini, tetapi ia boleh menjadi sangat sukar untuk mengesan kod sifar hari atau obfuscated kerana ia bergantung pada tandatangan yang unik. Sementara itu, analisis dinamik dan pengesanan berasaskan anomali dapat membendung masalah, namun ia boleh mengakibatkan kadar amaran palsu yang agak tinggi. Di samping itu, pengelas model tunggal tidak cukup kuat untuk menghasilkan hasil pengesanan yang baik. Berdasarkan sebab ini, penyelidikan ini berhasrat untuk meningkatkan kaedah Pengesanan Perisian Hasad Mudah Alih semasa menggunakan pelbagai ciri urutan opkod N-Gram dan kaedah klasifikasi ensembel berwajaran dari segi ketepatan pengesanannya, Kadar Positif Sebenar dan Kadar Positif Palsu. Matlamat penyelidikan ini adalah untuk mengenal pasti ciri aktiviti berniat jahat daripada aplikasi perisian hasad mudah alih melalui analisis statik. Ciri-ciri yang diperolehi digunakan dalam merumus dan menilai Rangka Kerja MMD yang dipertingkatkan. Penjanaan jujukan opkod N-Gram mewakili ciri berniat jahat dan kaedah pemilihan ciri digunakan untuk mencari ciri optimum. Selain itu, kaedah ensembel berwajaran diperkenalkan untuk menggabungkan beberapa keluaran kebarangkalian daripada pelbagai model klasifikasi. Pengoptimuman Kawanan Zarah digunakan dalam mencari berat optimum untuk digunakan bersama-sama dengan output kebarangkalian untuk meningkatkan pengesanan perisian hasad mudah alih. Kesimpulannya, Rangka Kerja MMD yang dicadangkan telah menunjukkan prestasi yang dipertingkatkan dengan ketepatan sebanyak 96.55%, TPR 99.10%, dan FPR 0.90%. Berdasarkan keputusan yang menggalakkan, kajian masa depan boleh meneroka kemungkinan menggunakan pendekatan pengesanan analisis dinamik dan menggunakan n-gram pada ciri selain daripada urutan opcode. Akhirnya, set data lain dan varian perisian hasad mudah alih yang lain juga harus diterokai pada masa hadapan.

ACKNOWLEDGMENT

In the name of Allah, the Most Gracious and the Most Merciful

First and foremost, All Praise to Allah SWT the Almighty for the strength and blessing in completing my thesis and Masters journey. A sincere gratitude I'll give to my supervisor, Ts. Dr. Mohd Zaki bin Mas'ud, whose expertise, understanding, patience, and guidance had helped me enhance my graduate experience. I am so grateful for his priceless effort in helping me whenever I find difficulties in completing my task and giving comments for improvement of this thesis.

Special thanks to my co-supervisor; Ts. Dr. Nazrulazhar bin Bahaman, my mentor; Ts. Nor Azman bin Mat Ariff for their support and aid in making my Masters journey a success. My appreciation also goes to Universiti Teknikal Malaysia Melaka (UTeM) for the opportunity given.

Last but not least, from the bottom of my heart a highest gratitude to my family for their love and support. Especially to my father, Anuar bin Jabbar and my mother, Rudziah binti Adam for their encouragements and blessings. Finally, to those who indirectly contributed to this research, your kindness has inspired me to embark on this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AB	-	Anomaly Based
AMD	-	Android Malware Dataset
API	-	Application Program Interface
APK	-	Android Package
AUC	-	Area Under Curve
BOW	-	Bag of Words
C&C	-	Command and Control
CPU	-	Central Processing Unit
CS	-	Chi-Square
DVM	-	Dalvik Virtual Machine
EA	-	Evaluation Algorithm
FN	-	False Negative
FP	-	False Positive
FPR	-	False Positive Rate
GA	-	Genetic Algorithm
GB	-	Gigabyte
GPS	-	Global Positioning System
GSS	-	Greedy Search Strategy
HIS	-	Hybrid Intelligent System
HTTP	-	Hypertext Transfer Protocol
ICCID	-	Integrated Circuit Card ID

IDS	-	Intrusion Detection System
IG	-	Information Gain
IMEI	-	International Mobile Station Equipment Identity
IMSI	-	International Mobile Subscriber Identity
IP	-	Internet Protocol
IPC	-	Inter Process Communication
kNN	-	K-Nearest Neighbour
MCT	-	Multiple Classifier Technique
MMD	-	Mobile Malware Detection
OS	-	Operating System
PO	-	Probability Output
POC	-	Proof of Concept
PSO	-	Particle Swam Optimization
RAM	-	Random Access Memory
RC	-	Research Contribution
RO	-	Research Objective
RP	-	Research Problem
RQ	-	Research Question
SB	-	Signature Base
SBS	-	Sequential Backward Search
SEO	-	Search Engine Optimization
SFS	-	Sequential Forward Search
SMS	-	Short Message Services
SU	-	Symmetrical Uncertainty
SVM	-	Support Vector Machine
TN	-	True Negative
TNR	-	True Negative Rate

TP	-	True Positive
TPR	-	True Positive Rate
UID	-	Unique Identifier

LIST OF SYMBOLS

f_{ngram}	-	Normalize occurrence of n-gram opcode sequence frequency
f_c	-	Normal n-gram frequency
min(f)	-	The minimum value of the attribute
max(f)	-	The maximum value of attribute
S	-	The scaling factor for the output range
T	-	The translation factor for the output range
w_i	-	Weight Assigned
h_i	-	Probability Output for each N-Gram Classifier
h(x)	-	Weighted Combination of Probability Output of N-Gram Classifiers

LIST OF PUBLICATIONS

1. Noor Azleen Anuar, Mohd Zaki Mas'ud, Nazrulazhar Bahaman, and Nor Azman Mat Ariff., 2020. Mobile Malware Behavior through Opcode Analysis. *International Journal of Communication Networks and Information Security*, 12(3), pp. 345–354.

2. Noor Azleen Anuar, Mohd Zaki Mas'ud, Nazrulazhar Bahaman, and Nor Azman Mat Ariff., 2020. Analysis of Machine Learning Classifier in Android Malware Detection Through Opcode. *2020 IEEE Conference on Application, Information and Network Security (AINS)*, pp. 7–11. https://doi.org/10.1109/AINS50155.2020.9315060

CHAPTER 1

INTRODUCTION

1.1 Background

With the great evolution of technology and communication, all living souls are obligated to use their mobile devices on a regular basis. Mobile gadgets enable people to communicate with one another from anywhere in the globe. Nonetheless, the presence of mobile devices provided enough opportunities for attackers to exploit them by introducing malware into the devices without the users' awareness (Varna and Visalakshi, 2020). Due to the difficulty in detecting mobile malware, they now have the capacity to go unnoticed in the app store (Tenenboim-Chekina et al., 2013), causing users to be deceived into installing the infected applications. Once the software is installed on the victims' devices, malicious malware penetration and replication begin, which can create major difficulties if left undetected.

Malware, often known as malicious software, is well-known for its tendency to disrupt computer software and hardware. Attackers frequently exploit it to take advantage of accessible resources and for other cybercriminal purposes such as stealing users' data, passwords, and credit card numbers. This assault might occur when victims open an infected email or download malicious software. When a malicious link or attachment received in an email is clicked, malware may be installed without the user's knowledge, and a ransomware attack may happen, causing the entire system to freeze. This might

potentially expose sensitive information, financial and corporate information. In addition, the attacker could obtain and extract the victim's login credentials and account information.

According to the I Threat Evolution in Q1 2020 issued by Kaspersky, (Kaspersky, 2022) the discovery of latest mobile malware variations had decreased in the year 2021. In Quarter 1, Quarter 2, and Quarter 3 of 2020, the number of attacks were 14,446,496, followed by 14,203,865, and 16,440,099. The highest number of attacks were in the Quarter 4 of 2020 with 18,085,657 of attacks. Following that, it can be seen that there was a decrease in the year 2021 with 15,239,031 number of attacks in Quarter 1 of 2021. Next, in Quarter 2, Quarter 3, and Quarter 4 of 2021 were 14,465,670, followed by 9,599,322, and 6,931,266 number of attacks. Although it can be seen that there was a slight decrease in the mobile malware threats, the number of attacks were still high. Figure 1.1 shows the trends of malware attacks on mobile devices.

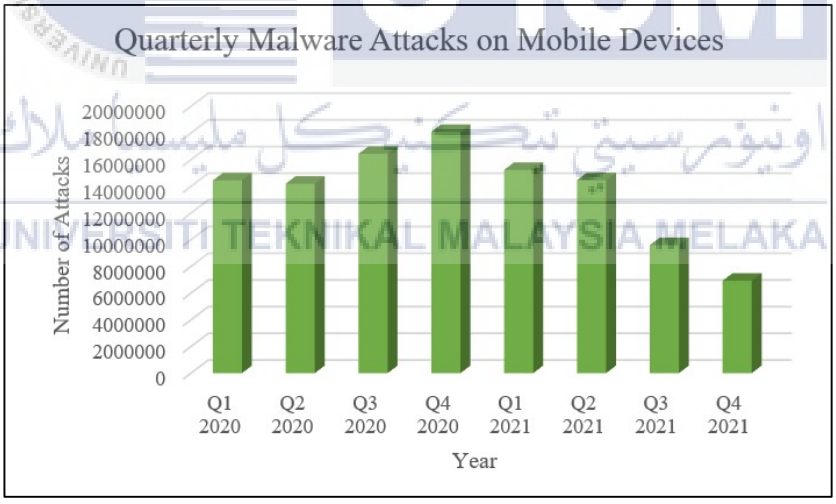


Figure 1.1: Mobile Malware Attack Trends (Kaspersky, 2022)

As reported by McAfee Labs (McAfee, 2017), several evasion techniques were used by mobile malware in their journey to detour into Android protection mechanisms. The most common evasion techniques are anti-sandbox, anti-security tools, code injection, anti-monitoring, and anti-debugging. Anti-sandbox can be defined as a method used to

isolate programs that is being executed in order to steer away from any unwanted applications to the system. Next, anti-security tools refer to a method used to bypass any detection done using security devices, or programs. On the other hand, code injection exists in the form of additional unwanted code or instructions to the binary in order to distract the disassembly view or waste the analyst's time. Anti-monitoring acts as a monitoring agent in order to prevent any reverse engineering from happening. Finally, yet importantly, anti-debugging is a program used so that the reverse engineering processing time will increase. All of the evasion techniques mentioned had made it harder to identify mobile malware. Figure 1.2 below shows the percentage of each evasion technique that had been reported in the year 2017.

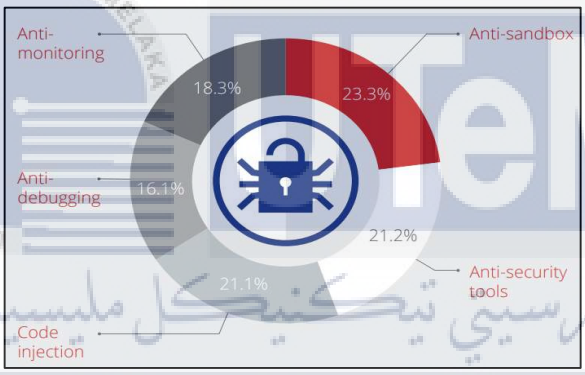


Figure 1.2: Evasion Techniques (McAfee, 2017)

1.2 Research Problems

Various analysis technique were implemented in order to solve the rising mobile malware threats. The two type of malware analysis are the static and dynamic analysis. Other than that, there are two types of classifier method in mobile malware detection that is single classifier and multiple classifier.

In static analysis, the code of an application is extracted and analyzed without having to be executed (Medvet and Mercaldo, 2016).The features of each application are