ENHANCED MOBILE MALWARE DETECTION USING INTERSECTION ATTRIBUTES TECHNIQUE

NAJIAHTUL SYAFIQAH BINTI ISMAIL

DOCTOR OF PHILOSOPHY

2022

Faculty of Information and Communication Technology

ENHANCED MOBILE MALWARE DETECTION USING INTERSECTION ATTRIBUTES TECHNIQUE

Najiahtul Syafiqah binti Ismail

Doctor of Philosophy

2022

ENHANCED MOBILE MALWARE DETECTION USING INTERSECTION ATTRIBUTES TECHNIQUE

NAJIAHTUL SYAFIQAH BINTI ISMAIL

A thesis submitted
in fulfillment of the requirements for the degree of
Doctor of Philosophy

Faculty Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2022

DECLARATION

I declare that this thesis entitled “Enhanced Mobile Malware Detection Using Intersection Attributes Technique” is the result of my own research 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 : NAJIAHTUL SYAFIQAH BINTI ISMAIL

Date : 2/8/2022

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 Doctor of Philosophy.

Signature : .....

Supervisor Name : ASSOC PROF Ts. DR ROBIAH BINTI YUSOF

Date : 2/8/2022

DEDICATION

This thesis is dedicated with

Deepest love and affections to my beloved parents,

Haji Ismail Bin Salim (1954-2005), and Hajah Samidah Binti Ab Samad (1957-2021)

who instil the importance of hard work, patience and having higher education in life

For their love, patience, guidance, wisdom, strength, and financial support

And

My Brother and sisters

Naczathul Faeza, Amir Ikhwan and Nadiathul Raihana

For their continuous support in all aspects

Have kept me going throughout these years in my journey.

ABSTRACT

The user-friendly interface, easy to use, and have many alternatives for the applications market make Android one of the most popular smartphone operating systems in this 21st century. In a mobile application, permission is one of the essential elements to protect user's personal information and privacy. Permission-based detection has been used widely but is deemed insufficient because it still suffers from high false alarm rates due to the permission-based issue. The current detection technique generates high false alarm rates, making the detection technique less effective in detecting the permission-based attack. Therefore, this research aimed to improve permission-based detection by integrating permission attributes with intent. However, integrating multiple attributes will increase the number of attributes used in mobile malware detection and affect the false alarm rate. Thus, the optimal size of attributes was developed to reduce the high false alarm rates generated by the Mobile Malware Detection System. Hence, this research introduces an Intersection Attribute Technique to reduce the number of attributes generated and improve the quality of attributes selected in the attribute selection process. The proposed technique with the Venn Diagram concept determined the correlation between attributes in the same process during Pre-processing phase before undergoing the Correlation Feature Selection process. Support Vector Machine was used to classify the applications. A comparative analysis has been performed using the proposed approach and three other approaches. The dataset used in this research is from New Brunswick Repository. The result indicates the Intersection Attribute Technique can reduce the number of attributes generated by 18, accuracy 96.67% and the false positive rate is 0.04%. In conclusion, the Enhanced Mobile Malware Detection with Intersection Attribute Technique can classify benign and malicious mobile applications more accurately and minimize false alarms.

PENAMBAHBAIKAN PENGESANAN PERISIAN HASAD PERANTI MUDAH ALIH MENGGUNAKAN TEKNIK PERTINDIHAN ATRIBUT

ABSTRAK

Antara muka mesra pengguna, mudah untuk digunakan, dan mempunyai banyak alternatif untuk pasaran aplikasi membuat Android salah satu sistem operasi telefon pintar yang paling popular di abad ke-21 ini. Dalam aplikasi mudah alih, keizinan adalah salah satu elemen penting untuk melindungi maklumat peribadi pengguna dan privasi. Pengesanan berasaskan keizinan telah digunakan secara meluas tetapi dianggap tidak mencukupi kerana ia masih mengalami kadar penggera palsu yang tinggi. Teknik pengesanan semasa menjana kadar penggera palsu yang tinggi, menjadikan teknik pengesanan kurang berkesan dalam mengesan serangan berasaskan keizinan. Oleh itu, kajian ini bertujuan untuk meningkatkan pengesanan berasaskan keizinan dengan mengintegrasikan sifat-sifat keizinan dengan niat. Walau bagaimanapun, mengintegrasikan pelbagai atribut akan meningkatkan bilangan atribut yang digunakan dalam pengesanan perisian hasad mudah alih dan menjejaskan kadar penggera palsu. Oleh itu, sifat atribut yang optimum telah dibangunkan untuk mengurangkan kadar penggera palsu yang tinggi yang dihasilkan oleh sistem pengesanan perisian hasad mudah alih. Oleh itu, kajian ini memperkenalkan teknik pertindihan atribut untuk mengurangkan bilangan atribut yang dihasilkan dan meningkatkan kualiti atribut yang dipilih dalam proses pemilihan atribut. Teknik yang dicadangkan dengan konsep Rajah Venn menentukan korelasi antara atribut dalam proses yang sama semasa fasa pra-pemprosesan sebelum menjalani proses pemilihan ciri korelasi. Mesin Vektor Sokongan digunakan untuk mengklasifikasikan aplikasi. Analisis perbandingan telah dilakukan menggunakan pendekatan yang dicadangkan dan tiga pendekatan lain. Dataset yang digunakan dalam penyelidikan ini adalah dari repositori New Brunswick. Hasilnya menunjukkan teknik pertindihan atribut dapat mengurangkan bilangan atribut yang dihasilkan kepada 18, ketepatan 96.67% dan kadar positif palsu adalah 0.04%. Kesimpulannya, Penambahbaikan Pengesanan Perisian Hasad Peranti Mudah Alih Menggunakan Teknik Pertindihan Atribut dapat mengklasifikasikan aplikasi mudah alih yang berniat jahat dengan lebih tepat dan meminimumkan penggera palsu.

ACKNOWLEDGEMENTS

In the Name of Allah, the Most Gracious, the Most Merciful

Alhamdulillah, all praises to Allah for His strength and blessing in completing this thesis. I want to express my gratitude to my respectful supervisor, Professor Madya Dr. Robiah Binti Yusof, and my co-supervisor, Professor Madya Dr. Mohd Faizal Abdollah, for their continuous support, advice, patience, and immense knowledge. Their guidance helped me all the time of my years to make my PhD. journey a success.

My deepest appreciation and dedication to my late father, Haji Ismail bin Salim and my late mother, Hajjah Samidah Binti Ab Samad, and my siblings for their continuous support, encouragement, and blessing

Finally, the appreciation goes to my friends for the stimulating discussions, the sleepless nights, and all the fun we had together.

TABLE OF CONTENTS

	PAGE
DECLARATION	
APPROVAL	
DEDICATION	
ABSTRACT	**i**
ABSTRAK	**ii**
ACKNOWLEDGEMENTS	**iii**
TABLE OF CONTENTS	**iv**
LIST OF TABLES	**viii**
LIST OF FIGURES	**x**
LIST OF APPENDICES	**xiii**
LIST OF ABBREVIATIONS	**xiv**
LIST OF SYMBOLS	**xvi**
LIST OF PUBLICATIONS	**xvii**
CHAPTER	
1. INTRODUCTION	**1**
1.1 Introduction	1
1.2 Research problem	3
1.3 Research question	6
1.4 Research aims and objectives	7
1.5 Research scope	8
1.6 Thesis framework	9
1.7 Research contribution	13
1.8 Thesis organization	14
1.9 Summary	17
2. LITERATURE REVIEW	**18**
2.1 Introduction	18
2.2 Mobile malware	20
2.2.1 Mobile malware history/evolution	28
2.2.2 General mobile malware behaviour	31
2.2.3 Permission based issues in mobile malware	36
2.3 Permission based issues in mobile malware	39
2.4 Mobile malware detection	46
2.4.1 Mobile malware detection technique	50
2.4.1.1 Signature-based	50
2.4.1.2 Anomaly-based	51
2.4.1.3 Specification-based	52
2.4.1.4 Proposed detection technique	53
2.4.2 Mobile malware analysis	54
2.4.2.1 Static analysis	54
2.4.2.2 Dynamic analysis	58
2.4.2.3 Static or dynamic	61
2.4.2.4 Proposed mobile malware analysis	62
2.4.3 Mobile application attribute	62
2.4.3.1 Application attribute: Permission	64

2.4.3.2 Application attribute: Intent 67
2.4.3.3 Application attribute: API call 68
2.4.3.4 Application attribute: Network traffic 69
2.4.3.4 Application attribute: System call 70
2.4.3.6 Discussion on Attributes 70
2.4.3.7 Proposed Attributes 72
2.4.4 Discussion of mobile malware detection 72
2.4.5 Data pre-processing 74
2.4.5.1 Proposed data pre-processing 76
2.4.6 Attribute selection 76
2.4.6.1 Proposed attribute selection 79
2.5 Mobile malware behaviour 79
2.5.1 Mobile malware behaviour through application permission 80
2.6 Machine learning in mobile malware detection 84
2.6.1 Classification 86
2.6.1.1 SVM (Support Vector Machine) 88
2.6.1.2 Naïve Bayes 89
2.6.1.3 Proposed Classifier 91
2.7 Evaluation and validation 90
2.8 Mobile malware dataset 92
2.9 Summary 95

3. METHODOLOGY 97
3.1 Introduction 97
3.2 Chapter objective 97
3.3 Chapter outline 98
3.4 Research design 99
3.5 Research approach 99
3.6 Research framework 100
3.7 Research process 102
3.7.1 Experimental approach 104
3.7.1.1 Experimental approach 105
3.8 Design framework 107
3.8.1 Data pre-processing 107
3.8.2 Attribution selection evaluation 109
3.9 Summary 112

4. MOBILE MALWARE BEHAVIOUR THROUGH INTENT AND PERMISSION INTERSECTION ATTRIBUTES 114
4.1 Introduction 114
4.2 Chapter objective 114
4.3 Chapter outline 115
4.4 Mobile malware behaviour experiment approach overview 116
4.5 Mobile malware behaviour analysis through permission and intent attributes 117
4.5.1 Permission attribute 117
4.5.2 Intent attribute 118
4.5.3 Analysis of permission and intent 120
4.5.4 Permission frequency used in mobile application 120
4.5.5 Intent frequency used in applications 123

5. MOBILE MALWARE BEHAVIOUR THROUGH INTENT AND PERMISSION INTERSECTION ATTRIBUTES **114**
5.1 Introduction 114
5.2 Chapter objective 114
5.3 Chapter outline 115
5.4 Mobile malware behaviour experiment approach overview 116
5.5 Mobile malware behaviour analysis through permission and intent attributes 117
5.5.1 Permission attribute 117
5.5.2 Intent attribute 118
5.5.3 Analysis of permission and intent 120
5.5.4 Permission frequency used in mobile application 120
5.5.5 Intent frequency used in applications 123

6. MOBILE MALWARE BEHAVIOUR THROUGH INTENT AND PERMISSION INTERSECTION ATTRIBUTES **114**
6.1 Introduction 114
6.2 Chapter objective 114
6.3 Chapter outline 115
6.4 Mobile malware behaviour experiment approach overview 116
6.5 Mobile malware behaviour analysis through permission and intent attributes 117
6.5.1 Permission attribute 117
6.5.2 Intent attribute 118
6.5.3 Analysis of permission and intent 120
6.5.4 Permission frequency used in mobile application 120
6.5.5 Intent frequency used in applications 123
6.5.6 Permission and intent attributes in malware detection 125
6.5.7 Correlation between permission (same family) and intent (same family) 127
6.5.8 Correlation between permission and intent (different family) 128
6.5.9 Mobile malware behaviour classification based on permission and intent attribute. 129
6.5.9.1 Data theft 130
6.5.9.2 Gain-privilege access 131
6.5.9.3 Short Messaging System (SMS) 132
6.5.9.4 Network and Wi-Fi 133
6.5.9.5 Phone 133
6.6 Intersection attribute technique 134
6.6.1 Venn Diagram concept and intersection technique in mobile malware detection 135
6.6.2 Data collection 137
6.6.3 Attribute vector 139
6.6.4 Attribute tuning 140
6.6.5 Modifying attribute value 140
6.6.6 Removing extra attribute 141
6.6.7 Evaluation before and after intersection attribute process 142
6.7 Attribute selection process 142
6.7.1 Attribute selection evaluation and analysis 143

6.8 Intersection attribute technique with correlation feature 145
6.9 Summary 146

7. EVALUATION AND VALIDATION 147
7.1 Introduction 147
7.2 Chapter objective 147
7.3 Chapter outline 148
7.4 Dataset 149
7.5 Intersection attributes of permission and intent evaluation 151
7.5.1 Result and findings for intersection attributes of permission and intent evaluation 151
7.5.2 Result and findings for intersection attributes of permission and intent evaluation with without intersection attributes 157
7.6 Summary 163

8. CONCLUSION AND FUTURE WORK 165
8.1 Introduction 165
8.2 Research review 166
8.3 Research contribution 169
8.4 Intersection attribute technique 171
8.5 Research limitations 171
8.6 Recommendation and future work 172
8.7 Conclusion 173

REFERENCES 173
APPENDIX 208

LIST OF TABLES

TABLE	TITLE	PAGE
1.1	Summary of research problems	5
1.2	Summary of research question	7
1.3	Summary of research objectives	8
1.4	Summary of task list in research methodology and possible output	13
1.5	Summary of research contribution	14
2.1	Definition of malware	26
2.2	No Android malware attack	29
2.3	List of reference	33
2.4	Type of general mobile malware behavior	36
2.5	References of permission based attacks from 2005 to 2019	39
2.6	Previous works on mobile malware detection	47
2.7	Comparisons type of malware detection	52
2.8	Definitions of static analysis	54
2.9	Definitions of dynamic analysis	58
2.10	Comparisons of Static and Dynamic Analysis	61
2.11	Description of attribute type	63
2.12	Dangerous type permission provided by Android	64
2.13	Advantages and limitations of mobile malware detection category	71
2.14	Related research on data pre-processing step	73
2.15	Related works on machine learning and attribute selection	78

2.16	Analysis of mobile malware behaviour based on application permissions	80
2.17	Mobile malware behaviour classification through permission	83
2.18	Past research using machine learning technique	85
2.19	Comparisons of classification and clustering	86
2.20	The differences of classifier	91
2.21	Mobile malware dataset comparison	93
3.1	Previous research on permission and intent attribute selection	111
4.1	List of permission used by malware applications	122
4.2	List of intent use by malware applications	124
4.3	Algorithm 1 Sequence of attribute tuning value	140
4.4	Algorithm 2 Removing extra attributes	142
4.5	Evaluation of intersection attribute result	144
5.1	Details of each dataset	149
5.2	Evaluation and validation process summary	150
5.3	Intersection attribute technique evaluation with different authors	151
5.4	Summary of classification accuracy percentage	156
5.5	ANOVA with single factor result	156
5.6	Intersection attribute technique evaluation with non-intersection attribute	157
5.7	Summary of classification accuracy percentage	162

LIST OF FIGURES

FIGURE	TITLE	PAGE
1.1	Operational Framework: Literature review phase	10
1.2	Operational Framework: Analysis, Design, Implementation Test and Validation	11
1.3	Thesis outline	15
2.1	Chapter 2 outline	19
2.2	Number of smartphone subscriptions worldwide from 2016 to 2026 (in billions) (Source: Statista (O'Dea, 2021))	21
2.3	Statistics of detected malicious installation applications (Securelist, 2021)	22
2.4	Mobile Operating System Market Share Worldwide - May 2021Android (StatCounter, 2021)	23
2.5	Mobile users attacked in January 2019 until December 2020 (Securelist, 2021)	24
2.6	Types of mobile malware	34
2.7	Framework of data pre-processing	75
2.8	Mobile malware dataset timeline	93
3.1	Chapter three outline	98
3.2	Research design	99
3.3	Main phases of research framework	101
3.4	Research process	103
3.5	Experimental design	106

3.6	Data pre-processing phase	108
3.7	Process of attribute selection evaluation	110
4.1	Chapter four outline	115
4.2	Example declaration of permissions in Android manifest file	118
4.3	Declaration of intent filter in manifest file	119
4.4	Permission's frequency used in malware & benign applications	121
4.5	Intent's frequency used in malware & benign application	124
4.6	Correlation between permission and intent (same family)	127
4.7	Graph of correlation between permission (same family) and intent (same family)	128
4.8	Correlation between permission and intent (different family)	128
4.9	Graph of correlation between permission and intent (different family)	129
4.10	Permission and intent in data theft	130
4.11	Permission and intent in gain privilege access	131
4.12	Permission and intent in Short Messaging System (SMS)	132
4.13	Permission and intent in network and Wi-Fi	133
4.14	Permission and intent in phone activity	134
4.15	Intersection attribute technique framework	135
4.16	Venn diagram concept	136
4.17	Example of the manifest file	138
4.18	Example of attribute vector	139
5.1	Chapter five outline	148
5.2	ROC curve for Dataset 1, D1	152
5.3	ROC curve for Dataset 2, D2	153
5.4	ROC curve for Dataset 3, D3	154

5.5	ROC curve for Dataset 1, D1	158
5.6	ROC curve for Dataset 2, D2	159
5.7	ROC curve for Dataset 3, D3	160
6.1	The objectives and contributions mapping	170

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Screenshot of Dex2Jar tool	210
B	Android malware application manifest file	211
C1	Attribute tuning source code using Python	212
C2	Removing extra attributes source code using Python	214

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Screenshot of Dex2Jar tool	210
B	Android malware application manifest file	211
C1	Attribute tuning source code using Python	212
C2	Removing extra attributes source code using Python	214

LIST OF ABBREVIATIONS

ACC	-	Accuracy
ANOVA	-	Analysis of Variance
APK	-	Android Package KiT
APP	-	Application
AVD	-	Android Virtual Device
CFS	-	Correlation-based Feature Selection
CS	-	Chi Square
FN	-	False Negative
FNR	-	False Negative Rate
FP	-	False Positive
FPR	-	False Positive Rate
GR	-	Gain Ratio
IAT	-	Intersection Attribute Technique
IG	-	Information Gain
MMD	-	Mobile Malware Detection
MMS	-	Multimedia Messaging System
NB	-	Naïve Bayes
RO	-	Research Objective
ROC	-	Receiver Operating Characteristic Curve
RP	-	Research Problem
RQ	-	Research Question
SMS	-	Short Messaging System
SVM	-	Support Vector Machine
TN	-	True Negative
TNR	-	True Negative Rate
TP	-	True Positive
TPR	-	True Positive Rate

UID	-	Unique ID
VM	-	Virtual Machine
XML	-	Extensible Markup Language

LIST OF SYMBOLS

Σ	-	Sum
$\cap$	-	Intersection
r	-	Pearson's Correlation Coefficient

LIST OF PUBLICATIONS

Indexed Journal

1. Ismail, N.S. et al. 2022. Generate Optimal Number of Features in Mobile Malware Classification using Venn Diagram Intersection. *International Journal of Computer Science and Network Security* 22(7), pp. 1–8.

2. Ismail, N. S., Saad, H., Yusof, R., and Abdollah, M. F., 2017. General android malware behaviour taxonomy. *Defence S and T Technical Bulletin*, 10(2), pp. 160–168.

Non-Indexed Journal

1. Ismail, N. S., Yusof, R., Saad, H., and Abdollah, M. F., 2019. Intersection Features for Android Botnet Classification. *International Journal of Recent Technology and Engineering*, 8(4), pp. 4422–4427.

Conference Proceedings

1. Saad, H. and Ismail, N.S. and Faizal, M.A. and Yusof, R. and Abdullah, R.S., 2016. Enhancement taxonomy and analysis on Android malware detection. *Proceedings of the 5th International Cryptology and Information Security Conference 2016, CRYPTOLOGY*, pp. 96-206.