Table of Contents

List of Figures xii

List of Tables xvi

List of Abbreviations/Symbols xvii

Chapter 1: Introduction 1

1.1 Cyber Threats: The Need to Tame the Surge! 1
1.1.1 Web Attack Vectors and Web Security 4
1.1.2 Web Attack Vectors on Mobile Platform 5
1.2 Artificial Intelligence (AI) 5
1.3 Motivation & Research Gaps 7
1.4 Main Contributions 10
1.5 Thesis Organization 11

Part I: Collection, Pre-processing, Visualization and Analysis of Data for Developing Machine Learning based Web Security Solutions 13

Chapter 2: Data Capturing and Pre-processing 14

2.1 Capturing Data for Machine Learning (ML) 14
2.2 MalCrawler 15
2.2.1 Maximizing Malicious Page Seek Rate 18
2.2.2 Crawls Deep to Detect Hidden Pages 19
2.2.3 Uses Anti-cloaking Measures 19
2.2.4 Emulates Different Browsers 19
2.2.5 Emulates a User 20
2.2.6 Crawls AJAX Pages while Avoiding Entanglement 20
2.2.7 MalCrawler: Background and Literature Survey 20
2.2.8 MalCrawler: Design and Methodology 22
2.2.9 MalCrawler: Software Design 27
2.2.10 MalCrawler: Result and Analysis 29
2.3 Pre-processing of the Webpages Dataset 31
2.4 Pre-processing of the Hybrid Android Apps Dataset 32
2.5 Conclusion 32

Chapter 3: Preliminary Analysis and Visualization of Datasets 34

3.1 Webpages Dataset 34
3.1.1 Webpages Dataset: Specifications 35
3.1.2 Webpages Dataset: Importance and Relevance 35
3.1.3 Webpages Dataset: Data Description 36
3.1.4 Webpages Dataset: Self Organising Map (SOM) based Analysis of the Dataset and its Attributes 46
3.1.5 Webpages Dataset: Experimental Design and Code 49
3.2 Hybrid Apps Dataset 49

3.2.1 Hybrid Apps Dataset: Specifications 50
3.2.2 Hybrid Apps Dataset: Importance and Relevance 51
3.2.3 Hybrid Apps Dataset: Data Description 51
3.2.4 Hybrid Apps Dataset: Experimental Design and Code 56

Part II: Machine Learning Based Analysis and Solutions for Webpage Classification......... 57

Chapter 4: Identifying Malicious Webpages Using Conventional Machine Learning 58

4.1 Introduction 58
4.2 Related Work 60
4.3 Attributes for Classification 60
4.4 Attribute Extraction and Pre-processing........ 62
4.5 Experimental Setup and Evaluation Metrics........ 66
4.6 Analysis of Results........ 71
4.6.1 Classification Accuracy of Attributes 71
4.6.2 Computational Resources Used 72
4.6.3 The Top Attributes 72
4.7 Conclusion........ 73

Chapter 5: Detection of Malicious Webpages Using Deep Learning: Structured Data 75

5.1 Background 75
5.2 Related Work 76
5.3 Deep Learning Framework for Detection of Malicious Webpages........ 77
5.3.1 Understanding Deep Learning 77
5.3.2 Structured vs. Unstructured Data 78
5.3.3 Dataset and Features........ 78
5.3.4 Deep Learning Model........ 79
5.3.5 Handling Class Imbalance 85
5.3.6 Implementation 87
5.4 Results and Analysis........ 87
5.4.1 Training and Validation Results........ 87
5.4.2 Evaluation on Test Dataset 89
5.4.3 Execution Time and Computational Resource........ 90
5.4.4 Analysis 91
5.5 Conclusions 91

Chapter 6: Detection of Malicious Webpages Using Deep Learning: Unstructured Data........ 92

6.1 Background 92
6.1.1 Inspiration 92
6.1.2 Webpages Analysis: Structured vs Unstructured Data 92
6.1.3 Overview 93
6.2 Related Work 94
6.3 Hybrid Deep Learning Model........ 95
6.3.1 Hybrid Deep Learning Model 95
6.3.2 Stage-I: Unsupervised LSTM Autoencoder........ 96
6.3.3 Stage-II: Supervised Training........ 98
6.3.4 Experimental Setup........ 104
6.4 Results and Analysis........ 105
6.4.1 Stage-I (Pre-training of Autoencoder)........ 105
6.4.2 Stage-II (Classification of Webpages)........ 106
6.4.3 Utilization of Computational Resources 108
6.4.4 Analysis and Inference 109
6.5 Future Work........ 109
6.6 Deployability of Solutions Proposed........ 110

6.7 Conclusions 110

Part III: Machine Learning Based Solutions for Security of Mobile Platforms........ 112

Chapter 7: Understanding and Mitigating Threats from Hybrid Apps Using Machine Learning 113

7.1 Background 113
7.2 Related Work and Research Gap 115
7.3 Understanding Hybrid Apps & WebView Component........ 116
7.3.1 Hybrid Apps and Android WebView 116
7.3.2 Understanding WebView Class 117
7.3.3 Development of Hybrid Apps Using WebView Class 117
7.3.4 Vulnerabilities: Enabling and Using JavaScript Interface 118
7.3.5 'WebView Tool' Android App........ 119
7.4 ML Based Solutions for Hybrid Apps Security........ 119
7.4.1 Methodology........ 119
7.5 Results and Analysis........ 122
7.5.1 Training and Validation 122
7.5.2 Analyzing Active Samples from Google Play Store........ 123
7.5.3 Ranking of Attributes 123
7.5.4 Interpretation of Results and Inferences 124
7.6 Recommendations for Improving Security Of Android Hybrid Apps........ 125
7.6.1 Recommendations for Android App Developers 125
7.6.2 Recommendations for Android OS Platform 126
7.7 App For Monitoring WebView Vulnerabilities........ 127
7.8 Scope For Future Research 128
7.9 Conclusions 129

Chapter 8: Android Web Security Solution Framework Using Cross-device Federated Learning........ 130

8.1 Background 130
8.2 Related Work and Research Gap 133
8.3 Federated Learning Framework for Android Web Security........ 134
8.3.1 Federated Learning (FL) 135
8.3.2 Design and Implementation........ 135
8.3.3 Adversarial Robustness of FL Model........ 141
8.3.4 Hierarchical Federated Learning (HFL) Model 142
8.4 Experimental Setup and Results 144
8.4.1 Dataset........ 144
8.4.2 Experimental Setup........ 145
8.4.3 Results and Analysis........ 145
8.5 Deployability of Solutions Proposed........ 150
8.6 Conclusion and Future Work 150

Chapter 9: Conclusions and Future Work 152

9.1 Conclusions 152
9.2 Future Directions 154

Appendix A: Webpages Dataset........ 156

Appendix B: Hybrid Apps Dataset 160

Appendix C: Details of Software Libraries and Tools Used 167

References 172

Publications... 194

Biographies 195