Contents

1 Introduction **1**

1.1 Introduction 1

1.2 Malware and Its Types 3

1.3 Stages of a Malware Attack 3

1.4 Defense Approaches 5

1.4.1 Signature-based Detection 5

1.4.2 Anomaly-based Detection 6

1.5 Motivation 7

1.6 Objectives 8

1.7 Contributions 10

1.8 Organization of Thesis 11

2 Background **13**

2.1 Networks 13

2.2 Network Components and Protocols 14

2.3 Network Architectures and Topologies 14

2.4 Network Devices 16

2.5 Operating Systems 17

2.6 Security Vulnerabilities 18

2.7 Malware Authors 19

2.8 Malware Attack Categories 19

2.8.1 Propagation Methods 19

2.8.2 Evasion Methods 21

2.8.3 Targeted Platform 22

2.8.4 Infected Objects 23

2.9 Sandboxes and Its Types 24

2.10 Honeynet System 25

2.10.1 Honeypots 25

2.10.2 Honeynet Architecture 25

2.10.3 Deployment Strategies 26

2.11 Malware Analysis 26

2.11.1 Static Analysis 26

2.11.2 Dynamic Analysis 27

2.11.3 Automated Analysis 28

2.11.4 Manual Code Reversing 28

2.11.5 Recent advances in Malware Analysis 29

2.12 Representation of Malware Data 30

2.13 Malware as a Service (MaaS) Model 31

2.14 Feature Selection 33

2.14.1 Filter Approach 33

2.14.2 Wrapper Approach 34

2.14.3 Embedded Approach 34

2.14.4 Hybrid Approach 34

2.15 Machine Learning 34

2.15.1 Supervised Learning 35

2.15.2 Unsupervised Learning 35

2.15.3 Reinforcement Learning 35

2.15.4 Ensemble Learning 36

2.16 Deep Learning 36

2.16.1 Aritifical Neural Networks 36

2.16.2 Convolutional Neural Network (CNN) 37

2.16.3 Generative Adversarial Networks 39

2.16.4 Graph Neural Networks 39

2.17 Evaluation Metrics 40

2.17.1 Accuracy 40

2.17.2 Precision 40

2.17.3 Recall 41

2.17.4 F1 Score 41

2.17.5 AUC-ROC 41

2.18 Discussion 42

3 Malware Dataset Generation and Evaluation 43

3.1 Introduction 43

3.1.1 Importance of Malware Datasets and Their Desired Characteristics 44

3.1.2 Motivation 46
3.1.3 Contributions 46
3.2 Background 47
3.2.1 Representation of Malware and Their behavior 47
3.2.2 Benchmark Datasets 49
3.3 TUMALWD: Windows Malware Dataset Creation Framework . . . 51
3.3.1 Phase 1: Data Collection and Storage 51
3.3.2 Phase 2: Data Analysis 53
3.3.3 Phase 3: Feature Engineering 53
3.3.4 TUMALWD: The Dataset and Its Characteristics 55
3.4 TUANDROMD: The Proposed Android Malware Dataset Creation Framework 58
3.4.1 Phase 1: Data Collection and Storage 58
3.4.2 Phase 2: Data Analysis 59
3.4.3 Phase 3: Feature Engineering 59
3.4.4 TUANDROMD: The proposed dataset and its characteristics 59
3.5 Performance Evaluation and Validation 61
3.6 Additional Malware Feature Datasets 62
3.6.1 Image-Based Malware Feature Dataset 62
3.6.2 Function Call Graph (FCG) Dataset 62
3.7 Discussion 63
4 An Enhanced Feature Selection Method for Imprecise Data 64
4.1 Introduction 64

4.1.1 Motivation 65

4.1.2 Contribution 66

4.2 Background 66

4.2.1 Filter Methods 67

4.2.2 Wrapper Methods 67

4.2.3 Embedded Methods 67

4.2.4 Hybrid Methods 68

4.2.5 Supervised Method 69

4.2.6 Unsupervised Method 69

4.2.7 Semi-supervised Method 70

4.3 Problem Statement 71

4.4 Proposed Method: FSR 71

4.4.1 Rough Set 71

4.4.2 Significance of FSR 79

4.5 Performance Analysis 79

4.5.1 Datasets and Preprocessing 80

4.5.2 Result Analysis and Performance Comparison 80

4.6 Discussion 87

5 A Cost-Effective Method for Ransomware Detection 88

5.1 Introduction 88

5.1.1 Motivation 89

5.1.2 Contribution 89

5.2 Background 90

5.3 Problem Statement 95

5.4 Proposed Framework 95

5.4.1 Data Collection 96

5.4.2 Data Preprocessing 97

5.4.3 Ensemble Feature Selection 97

5.4.4 Classification of Ransomware family and Its Variants 102

5.5 Results 103

5.5.1 Computation of weights for classifiers using NSGA-II 103

5.5.2 Weighted Majority Based Combination Function 104

5.5.3 Classification of Ransomware Variants 105

5.5.4 Classification of Ransomware Family 105

5.5.5 Result analysis and Performance comparison 105

5.6 Discussion 109

6 An Ensemble Approach for Effective Malware Detection 110

6.1 Introduction 110

6.1.1 Motivation 111

6.1.2 Contribution 112

6.2 Background 112

6.2.1 Feature selection 112

6.2.2 Ensemble Feature Selection 113

6.2.3 Markov Chain 114

6.3 Problem Statement 115

6.4 Proposed Method 115

6.4.1 Extraction of indicator of compromises 116

6.4.2 Selection of key indicators 116

6.4.3 Complexity Analysis 119

6.4.4 Optimal Features 119

6.5 Performance Analysis 121

6.5.1 Datasets and preprocessing 121

6.5.2 Classification Performance Analysis and Comparative Evaluation 121

6.6 Discussion 125

7 Parallel k-Nearest Neighbors for Enhanced Malware Detection 126

7.1 Introduction 126

7.1.1 Motivation 127

7.1.2 Contribution 128

7.2 Background 129

7.2.1 Introduction to K-Nearest Neighbors (K-NN) 129

7.2.2 Proximity Measures 129

7.2.3 Introduction to CUDA 130

7.3 Problem Statement 132

7.4 Proposed Work 133

7.4.1 Distance Measures 133

7.4.2 Sequential KNN algorithm 136

7.4.3 TUKNN Algorithm 136

7.4.4 Complexity Analysis 137

7.5 Implementation and results 137

7.5.1 Dataset Used 138

7.5.2 Results and Observation 139

7.6 Discussion 166

8 A Deep Learning Based Malware Detection Method 168

8.1 Introduction 168

8.1.1 Motivation 169

8.1.2 Contribution 170

8.2 Background 170

8.2.1 Neural Networks 170

8.2.2 Convolutional Neural Networks (CNNs) 172

8.3 Problem statement 174

8.4 The Proposed Method 175

8.4.1 Feature Data Generation 175

8.4.2 Data Preprocessing 178

8.4.3 Selection of CNN architectures 179

8.4.4 Transfer Learning and Fine-tuning Approaches 179

8.5 Experimental Results 181

8.5.1 Datasets 181

8.5.2 Results 182

8.6 Discussion 186

9 A Graph Neural Network Based Malware Detection Method 188

9.1 Introduction 188

9.1.1 Motivation 189

9.1.2 Contribution 190

9.2 Background 190

9.2.1 Graph 190

9.2.2 Function Call Graph 191

9.2.3 Graph Neural Networks 191

9.2.4 Node Feature Generation 194

9.2.5 Types of Tasks in Graph Learning 195

9.2.6 Loss Function 196

9.3 Problem Statement 196

9.4 Proposed Method 197

9.4.1 Function Call Graph Generation 197

9.4.2 GNN method 200

9.5 Experimental Results 203

9.5.1 Dataset 203

9.5.2 Results 204

9.6 Discussion 208

10 Conclusion and Future Direction 210

10.1 Conclusion 210

10.2 Future Directions 213