Contents

1 Introduction **1**

1.1 The Web and Its Users 1

1.2 Web-based Attacks 2

1.2.1 Steps in Launching a Web-based Attack 4

1.3 Types of Web-based Attacks 5

1.3.1 Cross-site Scripting Attacks 6

1.3.2 HTTP Flooding Attacks 10

1.3.3 Attacks in Critical Infrastructure 13

1.4 Sophistication of Attacks 18

1.5 Motivation of Attackers 18

1.6 Defense Approaches 19

1.6.1 Machine Learning for Web-based Attack Defense 20

1.7 Motivation and Objectives 21

1.8 Contributions 25

1.9 Thesis Organization 26

2 Background **28**

2.1 Networks 28

2.2 Network Communications 29

2.3 Internet and Its Issues 30

2.4 Network Attacks 31

2.5 Web-based Attacks 32

2.5.1 Cross-site Scripting Attacks 32

2.5.2 Cyber Physical Systems 46

2.5.3 HTTP Flooding Attacks . . . 56
2.6 Raw and Feature Data . . . 58
2.6.1 A Generic Pipeline of Dataset Generation . . . 59
2.6.2 Types of Datasets . . . 63
2.7 Machine Learning Approaches . . . 65
2.7.1 Supervised Learning and Its Significance . . . 66
2.7.2 Unsupervised Learning and Its Significance . . . 69
2.7.3 Ensemble Learning Approaches . . . 70
2.8 Cost Effective Methods for Attack Analysis . . . 72
2.8.1 Feature Selection Methods . . . 72
2.9 Validation Measures . . . 75

3 Datasets Used **81**
3.1 Introduction . . . 81
3.1.1 Desired Characteristics of a Dataset . . . 83
3.2 Benchmark Datasets . . . 85
3.2.1 Cross-site Scripting Attack Repositories . . . 85
3.2.2 HTTP Flooding Attack Datasets . . . 86
3.2.3 Cyber Physical Systems Datasets . . . 88
3.2.4 Other Security Datasets . . . 90
3.3 Generated Datasets . . . 93
3.3.1 XSS Attack Dataset . . . 93
3.3.2 Proposed Dataset Generation Framework . . . 94
3.3.3 Stages and Modules . . . 94
3.3.4 Tasks and Sub-tasks . . . 96
3.3.5 XSSD: The Dataset and Its Characteristics . . . 98
3.3.6 Performance Evaluation and Validation . . . 103
3.4 Discussion . . . 104

4 Ensemble Learning Methods **106**
4.1 Introduction . . . 106
4.1.1 Motivation . . . 107

4.1.2 Contribution 107
4.2 Background 107
4.2.1 Ensemble Learning 107
4.3 Data-centric Supervised Ensemble Proposed Framework 113
4.3.1 Preprocessing Engine 114
4.3.2 Parameter Tuning Engine 115
4.3.3 Ensemble Engine 116
4.3.4 Performance Analysis Engine 117
4.4 Experimental Results 117
4.4.1 Bagging Results 118
4.4.2 Boosting Results 118
4.4.3 Bagging-Boosting Results 123
4.4.4 Stacking Results 123
4.4.5 Observations 123
4.4.6 Discussion 127

5 MICC-UD: A Mutual Information and Correlation-based Feature Selection Method 129

5.1 Introduction 129
5.1.1 Feature 130
5.1.2 Related Work 130
5.1.3 Limitations of the Existing Approaches 132
5.1.4 Motivation 133
5.1.5 Contribution 133
5.2 Problem Formulation 134
5.3 Background 134
5.3.1 Mutual Information for Feature Selection 135
5.3.2 Correlation Co-efficient for Feature Selection 135
5.4 MICC-UD: Proposed Method 136
5.4.1 Preprocessing Engine 137
5.4.2 Feature Selection Engine 137

5.4.3 Optimal Feature Subset Identification using Recursive Feature Elimination . . . 139
5.4.4 Proposed Algorithm . . . 140
5.5 Complexity Analysis . . . 141
5.6 Experimental Results . . . 142
5.6.1 Results and Analysis . . . 143
5.6.2 Comparison with Existing Works . . . 146
5.7 Discussion . . . 147

6 INFS-MICC: An Incremental Feature Selection Method 150
6.1 Introduction . . . 150
6.1.1 Incremental Data . . . 151
6.1.2 Necessity of Incremental Feature Selection . . . 151
6.1.3 Related Work . . . 151
6.1.4 Motivation . . . 154
6.1.5 Contribution . . . 154
6.2 Problem Definition . . . 155
6.3 Background . . . 156
6.3.1 Mutual Information for Feature Selection . . . 156
6.3.2 Correlation for Feature Selection . . . 157
6.4 INFS-MICC: Proposed Method . . . 157
6.4.1 Relevance and Independent Feature Subset Finding . . . 159
6.5 Experimental Results . . . 160
6.5.1 Comparison with other Feature Selection Methods . . . 165
6.5.2 Discussion . . . 167

7 FSRA: A Feature Selection and Rank Aggregation Framework for CPS Attack Classification 172
7.1 Introduction . . . 172
7.1.1 Related Work . . . 174
7.1.2 Motivation . . . 177
7.1.3 Contributions . . . 177

7.2 Problem Formulation 178
7.3 CPSAD: Proposed Attack Detection Framework 178
7.3.1 Detection Module 179
7.3.2 FSRA: Proposed Ensemble Feature Selection Method 180
7.3.3 Alarm Generation Module 183
7.3.4 Feedback Analyzer Module 184
7.3.5 Reference or Rule generation Module 184
7.4 Experimental Results 184
7.4.1 Comparison with Existing Methods 191
7.5 Discussion 194

8 Conclusion **197**
8.1 Concluding Remarks 197
8.2 Future Work 199

Appendices **202**
A Combining continuous outputs 202
B Hyper-parameter values 203
B.1 Bagging 203
B.2 Boosting 204

Glossary **237**