Dedicated to Maa and Deta...

Declaration

I, Upasana Sarmah, hereby declare that the thesis entitled *Detection of Web-based Attacks using Machine Learning Techniques* submitted to the Department of Computer Science and Engineering under the School of Engineering, Tezpur University, in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy is based on bona-fide work carried out by me. The results presented in this thesis have not been submitted in part or in full, to any other university or institute for award of any degree or diploma.

Upasana Sarmah

(Upasana Sarmah)

Tezpur University

Certificate

This is to certify that the thesis entitled ***"Detection of Web-based Attacks using Machine Learning Techniques"*** submitted to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a record of research work carried out by **Upasana Sarmah** under my personal supervision and guidance. All helps received by her from various sources have been duly acknowledged. No part of this thesis has been reproduced elsewhere for award of any other degree.

(D. K. Bhattacharyya)

Prof. D.K. Bhattacharyya
Deptt. Of Computer Sc. & Engg.
Tezpur University

Signature of Research Supervisor
(Dr. Dhruba Kumar Bhattacharyya)
Designation: Professor
School: Engineering
Department: Computer Science and Engineering

UCCS UNIVERSITY ·OF· COLORADO
AT COLORADO SPRINGS

University of Colorado

Certificate

This is to certify that the thesis entitled ***"Detection of Web-based Attacks using Machine Learning Techniques"*** submitted to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a record of research work carried out by **Upasana Sarmah** under my co-supervision and guidance.
All helps received by her from various sources have been duly acknowledged. No part of this thesis has been reproduced elsewhere for award of any other degree.

Jugal Kumar Kalita

Signature of Research Co-Supervisor
(Dr. Jugal Kumar Kalita)
Designation: Professor
College of Engineering and Applied Science
Department of Computer Science
University of Colorado
Colorado Springs, CO 80918, USA

Tezpur University

Certificate

This is to certify that the thesis entitled ***"Detection of Web-based Attacks using Machine Learning Techniques"*** submitted by **Upasana Sarmah** to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science has been examined by us on _____21/11/2024________ and found to be satisfactory.
The Committee recommends for award of the degree of Doctor of Philosophy.

(D. K. Bhattacharyya)

Prof. D.K. Bhattacharyya
Deptt. Of Computer Sc. & Engg.
Tezpur University

Signature of Principal Supervisor

Date: 21/11/2024

Acknowledgment

The tenure of my research has been an intense learning experience which does not merely extend over research skills, but also inculcates a number of social skills on a personal level. It gives me immense pleasure to take this opportunity to express my deep sense of gratitude to my esteemed supervisor Prof. Dhruba Kumar Bhattacharyya. I am proud to be a student of such erudite, honest and considerate individual. His discipline and constant motivation was a key driver of my endeavor to achieve the goals of my research. He has been a constant source of inspiration throughout my PhD tenure. I sincerely appreciate his problem handling tactics, life long research experience and whole heartedly thank him for the trouble he took up to arrange facilities for my research works. This thesis is a result of his guidance, constant support, invaluable suggestions and encouragement. I also convey my heartiest thanks and gratitude to my co-supervisor Prof. Jugal Kumar Kalita, University of Colorado, USA, for all his help, support and guidance in shaping my Ph.D thesis upto this extent. It is my privilege to thank the authorities of Tezpur University and the Department of Computer Science and Engineering for providing me the facilities during the pursuit. Their co-operation and support will always be revered. I deeply acknowledge Dr. Sanjib Kr. Deka and Dr. Debojit Boro, members of my doctoral research committee for their valuable suggestions, inspirations and co-operations during the entire research tenure. I am also indebted to all the faculty members of the department, with special mention to Prof. Sarat Saharia for his relentless support during his tenure as Head of the Department.

Everyday I feel blessed for the untiring moral support of my parents, Maa and Deta (who is not with me today but I know he is proud wherever he is). Maa and Deta have been and will always be my pillars of strength. This work is whole heartedly dedicated to them. I am thankful to my seniors Dr. Pooja Sharma, Dr. Nazrul Hoque, Dr. Prakash Chouhan, Dr. Hussain Ahmed Chowdhury, Dr. Monisha Devi and Dr. Pooja Dutta for giving me valuable insights whenever I was in need of it. Special mention to Dr. Ram Charan Baishya who helped me and showed me the way not only during my research tenure but also in life. I am grateful to my lab mates Khushboo, Prayakhi, Sushmita, Minhazur, and Rishang for their constant moral support. I owe a lot to my best friends Parthajit, Annushree and Suroshikha for being there for me whenever I needed that extra push. Words are less to express how much grateful I am to them. This note of acknowledgment can never be complete without a mention to a few more individuals, my constants Tanmaya and Urma, my brothers Rimpu, Baba, Angshu and Pranjal for their constant love and appreciation. I am also grateful to all the research scholars of the department, and the office staff (Golap Da, Pronoti Baidew and Bobita Baidew) for all their help and support. Last but not the least, I thank the almighty for everything.

Upasana Sarmah

(Upasana Sarmah)

List of Tables

1.1 Real World XSS Worms Over the Years 9

1.2 Real World HTTP Flooding Attacks 12

1.3 Real World Attacks on Critical Infrastructure 15

1.3 Real World Attacks on Critical Infrastructure 16

1.3 Real World Attacks on Critical Infrastructure 17

2.1 Summary of XSS Filters 44

2.2 Existing Approaches for Detection of XSS Attacks 78

2.3 Summary of Machine Learning Approaches 79

2.4 Existing Approaches for Detection of HTTP Flooding Attacks . . . 80

3.1 HTTP-Flood Traffic Generator Tools 87

3.2 Brief Description of UNSW-NB15 Dataset 88

3.3 A Brief Overview of SWaT Dataset 89

3.4 Common Payload Features for Gas Pipeline and Water Storage Dataset 90

3.5 Some Examples of Permission-based Features in TUANDROMD . . 91

3.6 Different Ransomware Families 92

3.7 Script-based Features and their Description 100

3.8 URL-based Features and their Description 102

3.9 AUC Values for all Five Classifiers 105

3.10 Summary of All the Datasets Used 105

4.1 Description of Datasets Used 118

4.2 Stacking Ensemble Results 123

4.3 Stacking Ensemble Results for Ransomware Multiclass Dataset . . . 126

5.1 Symbol Table for the Proposed Method (MICC-UD) 134
5.2 Datasets Used 142
5.3 Top 10 Ranked Features of the XSSD Dataset 146
5.4 Comparison of F1-scores for all Datasets 147
5.5 Class-specific Comparison of F1-scores for Ransomware Multiclass Dataset 149

6.1 Symbol Table for the Proposed Method (INFS-MICC) 155
6.2 Datasets Used 161
6.3 Top 10 Ranked Features of the HTTP Flooding Datasets 170

7.1 Symbol Table for the Proposed Method 178
7.2 Ranked List of Features for the Datasets 185
7.3 SWaT Dataset Aggregated List of Features 185
7.4 Aggregated Ranked List of Features for Gas Pipeline and Water Storage Dataset 186
7.5 Top 10 ranked features as given by the proposed method 196
7.6 Comparison with Existing Methods 196

1 Hyper-parameter Values for 2-class Security Datasets (Bagging) . . 203
2 Hyper-parameter values for ransomware multiclass datasets (Bagging) 204
3 Hyper-parameter values for 2-class security datasets (Adaboost, GB and XGB) 205
4 Hyper-parameter values for 2-class security datasets (LGB and HGB) 205
5 Hyper-parameter values for Ransomware multiclass dataset (Adaboost, GB and XGB) 206
6 Hyper-parameter values for Ransomware multiclass dataset (LGB and HGB) 207

List of Figures

1.1 Vulnerability Count of Different Vulnerabilities 3

1.2 Typical Steps in a Web Attack 4

1.3 XSS Vulnerability Count Over the Years 7

1.4 DDoS Attacks Over the Years 11

2.1 A Computer Network 29

2.2 Network Communications and the Internet 31

2.3 Steps in Persistent and Non-persistent XSS Attacks 33

2.4 Malicious XSS JavaScript 34

2.5 A Crafted Malicious Link 34

2.6 XSS Actors 35

2.7 XSS Types 35

2.8 Steps in an XSS Attack 36

2.9 Reflected XSS Attack Model 38

2.10 Stored XSS Attack Model 38

2.11 DOM-based XSS Attack Model 39

2.12 DOM-based XSS Attack Crafted URL 39

2.13 Malicious Link 40

2.14 Obfuscated Malicious URL 40

2.15 Shortened Malicious URL 40

2.16 Taxonomy of the Detection Approaches 41

2.17 Steps in Static Analysis 42

2.18 Steps in Dynamic Analysis 42

2.19 Layers of Cyber Physical Systems 48

2.20 Critical Infrastructure Thrust Areas 50

2.21 Layer-wise Attacks in CPSs 53

2.22 HTTP Flood Attack Scenario 57

2.23 Example of Raw Data 59

2.24 Example of Feature Data 59

2.25 Example of Feature Selection 62

2.26 Example of a Labeled Dataset 63

2.27 Example of Supervised/Unsupervised Learning Technique 67

2.28 Some Supervised Learning Algorithms 67

2.29 Supervised Learning Framework 69

2.30 Example of Supervised and Unsupervised Learning 70

2.31 Filter Method 73

2.32 Wrapper Method 74

2.33 Embedded Method 74

2.34 Hybrid Method 75

3.1 Worldwide Internet Users from 2005 to 2022 82

3.2 Top 10 Countries with Largest Digital Population 82

3.3 Statistics of Malware Attacks from 2015 to 2022 91

3.4 Dataset Generation Framework 94

3.5 Example of WARC Content 95

3.6 Task and Sub-tasks of the Modules 96

3.7 A Snapshot of XSSD (Script-based Features) 103

3.8 A Snapshot of XSSD (URL-based Features) 103

3.9 Performance Evaluation for the Script Dataset 104

3.10 Performance Evaluation for the URL Dataset 104

3.11 ROC Curve for the Script Dataset 104

3.12 ROC Curve for the URL Dataset 104

4.1 A Typical Ensemble Learning Framework 108

4.2 Bagging Example 112

4.3 Boosting Example 113

4.4 Stacking Example 114

4.5 Data-centric Supervised Ensemble Framework 114
4.6 Preprocessing Engine 115
4.7 Parameter Tuning Engine 116
4.8 Ensemble Engine 117
4.9 Bagging Results for Security Datasets 119
4.10 Bagging Results for Ransomware Multiclass Dataset 120
4.11 Boosting Results for Security Datasets 121
4.12 Boosting Results for Ransomware Multiclass Dataset 122
4.13 Bagging-Boosting results for Security Datasets 124
4.14 Bagging-Boosting results for Ransomware Multiclass Datasets . . . 125
4.15 Stacking Results for Security Datasets 127
4.16 Stacking Results for Ransomware Multiclass Datasets 127

5.1 Proposed framework for MICC-UD 136
5.2 MICC-UD Results for all Datasets 144
5.3 Class-specific Results for Ransomware Multiclass Dataset 145

6.1 Proposed Framework for INFS-MICC 158
6.2 RFE with XGBoost Classifier for CICIDS Dataset (Accuracy) . . . 161
6.3 RFE with Adaboost Classifier for CICIDS Dataset (Accuracy) . . . 162
6.4 RFE with Gradient Boosting Classifier for CICIDS Dataset (Accuracy)162
6.5 RFE with Random Forest Classifier for CICIDS Dataset (Accuracy) 162
6.6 RFE with Extra Trees Classifier for CICIDS Dataset (Accuracy) . . 162
6.7 RFE with Gradient Boosting Classifier for CICIDS Dataset (F1-score)163
6.8 RFE with Adaboost Classifier for CICIDS Dataset (F1-score) . . . 163
6.9 RFE with XGBoost Classifier for CICIDS Dataset (F1-score) 163
6.10 RFE with Random Forest Classifier for CICIDS Dataset (F1-score) 164
6.11 RFE with Extra Trees Classifier for CICIDS Dataset (F1-score) . . 164
6.12 RFE with Gradient Boosting Classifier for UNSW Dataset (Accuracy)164
6.13 RFE with Adaboost Classifier for UNSW Dataset (Accuracy) 165
6.14 RFE with XGBoost Classifier for UNSW Dataset (Accuracy) 165
6.15 RFE with Random Forest Classifier for UNSW Dataset (Accuracy) 165

6.16 RFE with Extra Trees Classifier for UNSW Dataset (Accuracy) . . 165

6.17 RFE with Gradient Boosting Classifier for UNSW Dataset (F1-score) 166

6.18 RFE with Adaboost Classifier for UNSW Dataset (F1-score) 166

6.19 RFE with XGBoost Classifier for UNSW Dataset (F1-score) 166

6.20 RFE with Random Forest Classifier for UNSW Dataset (F1-score) . 167

6.21 RFE with Extra Trees Classifier for UNSW Dataset (F1-score) . . . 167

6.22 RFE with XGBoost Classifier for HTTP Flood Dataset (Accuracy) 167

6.23 RFE with Adaboost Classifier for HTTP Flood Dataset (Accuracy) 168

6.24 RFE with Gradient Boosting Classifier for HTTP Flood Dataset (Accuracy) 168

6.25 RFE with Random Forest Classifier for HTTP Flood Dataset (Accuracy) 168

6.26 RFE with Extra Trees Classifier for HTTP Flood Dataset (Accuracy) 168

6.27 RFE with Gradient Boosting Classifier for HTTP Flood Dataset (F1-score) 169

6.28 RFE with Adaboost Classifier for HTTP Flood Dataset (F1-score) . 169

6.29 RFE with XGBoost Classifier for HTTP Flood Dataset (F1-score) . 169

6.30 RFE with Random Forest Classifier for HTTP Flood Dataset (F1-score) 170

6.31 RFE with Extra Trees Classifier for HTTP Flood Dataset (F1-score) 170

6.32 F1-score Comparison for CICIDS Dataset 170

6.33 F1-score Comparison for HTTP-Flood Dataset 171

6.34 F1-score Comparison for UNSW Dataset 171

7.1 CPSAD: An attack detection framework 179

7.2 FSRA: Feature Selection using Rank Aggregation 181

7.3 RFE with XGBoost classifier for SWaT dataset (Accuracy) 187

7.4 RFE with Random Forest classifier for SWaT dataset (Accuracy) . 187

7.5 RFE with Extra trees classifier for SWaT dataset (Accuracy) 187

7.6 RFE with Adaboost classifier for SWaT dataset (Accuracy) 188

7.7 RFE with Gradient Boosting classifier for SWaT dataset (Accuracy) 188

7.8 RFE with XGBoost classifier for SWaT dataset (F1-score) 188

7.9 RFE with Adaboost classifier for SWaT dataset (F1-score) 189

7.10 RFE with Extra trees classifier for SWaT dataset (F1-score) 189

7.11 RFE with Gradient Boosting classifier for SWaT dataset (F1-score) 189

7.12 RFE with Random Forest classifier for SWaT dataset (F1-score) . . 189

7.13 RFE with Gradient Boosting classifier for GAS pipeline dataset (Accuracy) 190

7.14 RFE with Adaboost classifier for GAS pipeline dataset (Accuracy) . 190

7.15 RFE with Extra trees classifier for GAS pipeline dataset (Accuracy) 190

7.16 RFE with XGBoost classifier for GAS pipeline dataset (Accuracy) . 191

7.17 RFE with Random Forest classifier for GAS pipeline dataset (Accuracy) 191

7.18 RFE with XGBoost classifier for GAS pipeline dataset (F1-score) . 191

7.19 RFE with Extra trees classifier for GAS pipeline dataset (F1-score) 192

7.20 RFE with Random Forest classifier for GAS pipeline dataset (F1-score) 192

7.21 RFE with Gradient Boosting classifier for GAS pipeline dataset (F1-score) 192

7.22 RFE with Adaboost classifier for GAS pipeline dataset (F1-score) . 192

7.23 RFE with XGBoost classifier for Water pipeline dataset (Accuracy) 193

7.24 RFE with Random Forest classifiers for Water pipeline dataset (Accuracy) 193

7.25 RFE with Extra trees classifier for Water pipeline dataset (Accuracy) 193

7.26 RFE with Adaboost classifier for Water pipeline dataset (Accuracy) 194

7.27 RFE with Gradient Boosting classifier for Water pipeline dataset (Accuracy) 194

7.28 RFE with XGBoost classifier for Water pipeline dataset (F1-score) . 194

7.29 RFE with Adaboost classifier for Water pipeline dataset (F1-score) 195

7.30 RFE with Extra trees classifier for Water pipeline dataset (F1-score) 195

7.31 RFE with Gradient Boosting classifier for Water pipeline dataset (F1-score) 195

7.32 RFE with Random Forest classifiers for Water pipeline dataset (F1-score) 195

LIST OF ABBREVIATIONS

OWASP	Open Worldwide Application Security Project
CVE	Common Vulnerabilities and Exposures
HTTP	HyperText Transfer Protocol
CI	Critical Infrastructure
IP	Internet Protocol
ZAP	Zed Attack Proxy
XSS	Cross-site Scripting
OSI	Open Systems Interconnection
CMIM	Conditional Mutual Information Maximization
MRMR	Maximum Relevance and Minimum Redundancy
MIFS	Mutual Information-based Feature Selection
GB	Gradient Boosting
XGB	Extreme Gradient Boosting
RF	Random Forest
EXTT	Extra Trees
Adaboost	Adaptive Boosting
RFE	Recursive Feature Elimination
URL	Uniform Resource Locator