1.

| | |
|---|---|
| Record Nr. | UNINA9910770274903321 |
| Autore | Pan Zhixin |
| Titolo | Explainable AI for Cybersecurity |
| Pubbl/distr/stampa | Cham : , : Springer International Publishing AG, , 2023<br>©2023 |
| ISBN | 3-031-46479-6 |
| Edizione | [1st ed.] |
| Descrizione fisica | 1 online resource (249 pages) |
| Altri autori (Persone) | MishraPrabhat |
| Lingua di pubblicazione | Inglese |
| Formato | Materiale a stampa |
| Livello bibliografico | Monografia |
| Nota di contenuto | Intro -- Preface -- Acknowledgements -- Contents -- Acronyms -- Part I Introduction -- Cybersecurity Landscape for Computer Systems -- 1 Introduction -- 2 Cybersecurity Vulnerabilities -- 2.1 Hardware Vulnerabilities -- 2.1.1 Malicious Implants (Hardware Trojans) -- 2.1.2 Supply Chain Vulnerability -- 2.1.3 Reverse Engineering -- 2.1.4 Side-Channel Leakage -- 2.2 Software Vulnerabilities -- 2.2.1 Malware Attacks -- 2.2.2 Ransomware Attacks -- 2.2.3 Spectre and Meltdown Attacks -- 2.3 Malicious Attacks on Machine Learning Models -- 2.3.1 Adversarial Attacks -- 2.3.2 AI Trojan Attacks -- 3 Detection of Security Vulnerabilities -- 3.1 Detection of Malicious Hardware Attacks -- 3.1.1 Simulation-Based Validation Using Machine Learning -- 3.1.2 Side-Channel Analysis Using Machine Learning -- 3.1.3 Heuristic Analysis Using Machine Learning -- 3.2 Detection of Malicious Software Attacks -- 3.2.1 Detection of Malware Attacks -- 3.2.2 Detection of Ransomware Attacks -- 3.2.3 Detection of Spectre and Meltdown Attacks -- 4 Summary -- References -- Explainable Artificial Intelligence -- 1 Introduction -- 2 Machine Learning Models -- 2.1 Support Vector Machine -- 2.2 Multi-Layer Perceptron -- 2.3 Decision Tree -- 2.4 Random Forest -- 2.5 Linear Regression -- 2.6 Deep Neural Network -- 2.7 Convolution Neural Network -- 2.8 Recurrent Neural Network -- 2.9 Long Short-Term Memory -- 2.10 Reinforcement Learning -- 2.11 Boosting -- 2.12 Naive Bayes -- 2.13 Zero-Shot Learning -- 3 Explainable Artificial Intelligence -- 3.1 Local Interpretability -- 3.2 Knowledge Extraction -- 3.3 Saliency Maps -- |

3.4 Integrated Gradients -- 3.5 Shapley Value Analysis -- 3.6 Layer-Wise Relevance Propagation -- 4 Summary -- References -- Part II Detection of Software Vulnerabilities -- Malware Detection Using Explainable AI -- 1 Introduction -- 2 Background and Related Work.

2.1 Malware Detection Challenges -- 2.2 Why Explainable AI for Malware Detection? -- 3 Malware Detection Using Explainable Machine Learning -- 3.1 Model Training -- 3.2 Perturbation and Outlier Elimination -- 3.3 Linear Regression -- 3.4 Outcome Interpretation -- 4 Experiments -- 4.1 Experimental Platform -- 4.2 Malware and Benign Benchmarks -- 4.3 Data Acquisition -- 4.4 RNN Classifier -- 4.5 Evaluation: Accuracy -- 4.6 Evaluation: Outcome Interpretation -- 5 Summary -- References -- Spectre and Meltdown Detection Using Explainable AI -- 1 Introduction -- 1.1 Threat Model -- 1.2 Motivation -- 2 Background and Related Work -- 2.1 Spectre and Meltdown Attacks -- 2.2 Explainable Machine Learning -- 3 Detection of Spectre and Meltdown Attacks -- 3.1 Data Collection -- 3.2 Model Training -- 3.2.1 LSTM-Based Model Training -- 3.2.2 Ensemble Boosting -- 3.3 Result Interpretation -- 3.3.1 Explainability Using Model Distillation -- 3.3.2 Explainability Using Shapely Values -- 3.4 Data Augmentation -- 4 Experiments -- 4.1 Experimental Setup -- 4.2 Comparison with Existing Spectre Detection Methods -- 4.3 Comparison with Existing Meltdown Detection Methods -- 4.4 Comparison with Existing Mitigation Techniques -- 4.5 Stability Analysis -- 4.6 Explainability Analysis -- 4.7 Efficiency Analysis -- 5 Summary -- References -- Part III Detection of Hardware Vulnerabilities -- Hardware Trojan Detection Using Reinforcement Learning -- 1 Introduction -- 2 Background and Related Work -- 2.1 Logic Testing for Hardware Trojan Detection -- 2.2 Reinforcement Learning -- 3 Test Generation Using Reinforcement Learning -- 3.1 Identification of Rare Nodes -- 3.2 Testability Analysis -- 3.3 Utilization of Reinforcement Learning -- 4 Experiments -- 4.1 Experimental Setup -- 4.2 Results on Trigger Coverage -- 4.3 Results on Test Generation Time -- 5 Summary -- References.

Hardware Trojan Detection Using Side-Channel Analysis -- 1 Introduction -- 2 Background and Motivation -- 2.1 Background: Reinforcement Learning -- 2.2 Motivation: Delay-Based Side-Channel Analysis -- 3 Reinforcement Learning-Based Path Delay Analysis -- 3.1 Overview -- 3.2 Generation of Initial Vectors -- 3.3 Generation of Succeeding Vectors -- 4 Experiments -- 4.1 Experimental Setup -- 4.2 Evaluation Results -- 5 Summary -- References -- Hardware Trojan Detection Using Shapley Ensemble Boosting -- 1 Introduction -- 2 Background and Related Work -- 2.1 Related Work for Hardware Trojan Detection -- 2.2 Ensemble Boosting -- 2.3 Shapley Values -- 3 Shapley Ensemble Boosting for Hardware Trojan Detection -- 3.1 Data Sampling -- 3.2 Model Training -- 3.3 Shapley Analysis -- 3.4 Weight Adjustment -- 3.5 Ensemble Prediction -- 4 Experiments -- 4.1 Experimental Setup -- 4.2 HT Detection Performance -- 4.3 Explainability Analysis -- 4.4 Efficiency Analysis -- 4.5 Robustness Analysis -- 5 Summary -- References -- Part IV Mitigation of AI Vulnerabilities -- Mitigation of Adversarial Machine Learning -- 1 Introduction -- 2 Background and Preliminaries -- 2.1 Attacks on Neural Networks -- 2.2 Spectral Normalization -- 3 Spectral Normalization to Defend Against Adversarial Attacks -- 3.1 Layer Separation -- 3.2 Fourier Transform -- 3.3 Activation Functions -- 3.4 Complexity Analysis -- 4 Experiments -- 4.1 Experimental Setup -- 4.2 Case Study: MNIST Benchmark -- 4.3 Case Study: ImageNet Benchmark -- 5 Summary -- References -- AI Trojan Attacks and Countermeasures -- 1 Introduction -- 2 Background and Related Work -- 3 Backdoor Attack with AI Trojans -- 3.1 Feature Extraction -- 3.2

Normal Training -- 3.3 Backdoor Training -- 3.4 Trojan Injection -- 4 Defenses Against AI Trojans -- 4.1 Pruning -- 4.2 Bayesian Neural Networks -- 4.3 Neural Cleanse.

4.4 Artificial Brain Stimulation (ABS) -- 4.5 STRIP -- 5 Experiments -- 5.1 Experimental Setup -- 5.2 Comparison of Attack Performance -- 5.3 Overhead Analysis -- 5.4 Robustness Against STRIP-Based Defense -- 6 Summary -- References -- Part V Acceleration of Explainable AI -- Hardware Acceleration of Explainable AI -- 1 Introduction -- 1.1 Graphics Processing Unit -- 1.2 Field Programmable Gate Array -- 2 FPGA-Based Acceleration of Heatmap Visualization -- 3 FPGA-Based Acceleration of Saliency Map -- 4 GPU-Based Acceleration of Shapley Value Analysis -- 4.1 Summary -- References -- Explainable AI Acceleration Using Tensor Processing Units -- 1 Introduction -- 2 Tensor Processing Units -- 3 Hardware Acceleration of Explainable AI -- 3.1 Task Transformation -- 3.2 Data Decomposition in Fourier Transform -- 4 Experiments -- 4.1 Experimental Setup -- 4.2 Comparison of Accuracy and Classification Time -- 4.3 Comparison of Energy Efficiency -- 5 Summary -- References -- Part VI Conclusion -- The Future of AI-Enabled Cybersecurity -- 1 Introduction -- 2 Summary -- 2.1 Introduction to Cybersecurity and Explainable AI -- 2.1.1 Detection of Software Vulnerabilities -- 2.2 Detection of Hardware Vulnerabilities -- 2.3 Mitigation of AI Vulnerabilities -- 2.4 Acceleration of Explainable AI -- 3 Future Directions -- 3.1 Automatic Implementation of Secure Systems -- 3.2 Detection of Malicious Implants -- 3.3 Detection of Ransomware Attacks -- 3.4 Automatic Data Augmentation -- References -- Index.