Cybersecurity Solutions for Industrial Control Systems and Key Equipment

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Fan Zhang

December 2019

© by Fan Zhang, 2019
All Rights Reserved.

To life and all the people who supported me, making me a better me.

Acknowledgments

I would like to thank my committee members for taking their valuable time to review this work: Dr. Jamie Coble, Dr. J. Wesley Hines, Dr. Richard Wood, and Dr. Mark Dean. I would like to especially thank Dr. Coble for her dedication and guidance on research throughout my graduate studies at the University of Tennessee, Knoxville (UTK). She is a great research advisor, a fantastic mentor, and an amazing role model. I would like to thank Dr. Hines for his support and guidance throughout my research, and his advice for my career development. Their support, recognition, and encouragement allowed me to enjoy my three and a half-year PhD life and constantly encouraged me forward.

I would like to thank the International Atomic Energy Agency (IAEA) for allowing me to be deeply involved in cybersecurity research through a coordinated research project (CRP) J02008: Enhancing Computer Security Incident Response at Nuclear Facilities. I would like to thank all the great researchers from the 26 organizations who I cooperated with through this project, especially, Mr. Mike Rowland, Mr. Mitchell Hewes, Mr. Trent Nelson, Mr. Christopher Spirito, Mr. Paul Smith, Mr. David Allison, Mr. Jianghai Li, Mr. Rodney Busquim, Mr. Ricardo Marques, Mr. Cheol-Kwon Lee, Mr. Jaegu Song, and Mr. Marcin Dudek.

I would like to thank Dr. Richard Wood, Dr. Mark Dean from UTK, and Dr. Kenny Gross from Oracle for their advice to my academic life. I would like to thank Dr. Sacit Cetiner, Dr. Pradeep Ramuhalli, and Dr. Nageswara Rao from Oak Ridge National Laboratory for their advice during my summer internship. I would like to thank Mr. Trentin Payne and Ms. Ashly Pearson for the help of the testbed running and data collection.

Thank you to all the UTNE supportive staff who made my studies in the nuclear engineering department smooth and pleasant. Their friendly attitude and efficiency made

graduate life much easier. Thank you to the nuclear engineering department for providing me with a solid platform for study and self-improvement in diverse ways, as well as the University of Tennessee, Knoxville, for providing me a pleasant campus to study and live, and recognizing my professional promise.

I would like to thank the American Nuclear Society (ANS) and the International Society of Automation (ISA) which have honored me with scholarships and provided me with a variety of opportunities for the exchange of research ideas and professional networking.

I would like to acknowledge the Lloyds Register Foundation and the International Joint Research Center for the Safety of Nuclear Energy for partial funding of this research. I would also like to acknowledge the Oracle Corporation for partial funding of this research.

Of course, a big thanks to my loving family and friends for all of their encouragement and support in my life. They make my life more colorful and enjoyable.

Part of the material in this dissertation were already published or have been prepared for submission in journals, conference papers, and a book chapter, including journal papers in Nuclear Technology[135] and IEEE Transactions on Industrial Informatics[136]; conference papers in the ANS Winter Conference 2019[137], the 11th NPIC&HMIT Conference[133], and the International Congress on Advances in Nuclear Power Plants[134]; and in the book "Nuclear Power Plant Design and Analysis Codes"[132] . The dissertation author was the primary investigator and lead author of these papers. A big thanks to all the co-authors.

Abstract

The growing number of cyber-attacks against industrial control systems (ICSs) in recent years elevates the urgent research need for enhancing cybersecurity in industry. As ICS systems are deployed for high-value and safety-critical systems, security requirements are evolving to include resilience to cyber-attacks. Nuclear industry is adopting digital control systems for operational benefits, which brings more cybersecurity concerns as well since several cyber-incidents already occurred even before fully digitalization. Currently, common approaches to industrial cybersecurity focus on signature-based intrusion prevention and network segregation, which may be inadequate for a well-motivated cyber-attack. Therefore, a comprehensive cybersecurity solution platform with defense-in-depth concept is developed in this research to facilitate strengthening the cybersecurity of ICSs.

The platform consists of five major systems: a data collection and extraction system, a cyber-attack detection system, a cause analysis system, a cyber-attack response system, and a main control room display system, to provide robust approaches to monitor and assess the security status, and aid the decision making of ICSs based on data analytics from both information technology and operational technology domain. The first two systems were well investigated in this research while frameworks are given for the other three systems limited by data access and experiment scale. To reinforce the cyber-attack resilience of key equipment, a localized cybersecurity strategy is proposed to detect a potential cyber-attack within a controller and potentially infer the faulty signal.

In order to investigate the proposed platform and strategy, two cybersecurity testbeds were developed to conduct simulated cyber-attacks. The real-time ICS testbed contains a physical experiment facility and a supervisory control and data acquisition (SCADA) system. The hardware-in-the-loop tested integrates a programmable logic controller with a nuclear

system simulator. Various cyber-attacks have been conducted towards these two testbeds, including man-in-the-middle, denial of service, tampering, and false data injection attacks. Both cyber data and process data are collected under normal operation and attack scenarios to develop and evaluate the cyber-attack detection models. The results demonstrate both the centralized platform and localized strategies developed are effective and applicable to all ICSs. This research also demonstrates that data analytic approaches are promising in cybersecurity domain.

Table of Contents

1 Introduction **1**

1.1 Problem Statement 9

1.2 Original Contributions 11

1.3 Structure of the Document 13

2 Literature Survey **14**

2.1 ICS Cyber Incidents and Attacks 14

2.1.1 Cyber incidents and attacks occurring in the non-nuclear industry . . 15

2.1.2 Cyber incidents and attacks in nuclear industry 16

2.2 Overview of Cybersecurity Guidance for Nuclear Industry 18

2.2.1 IAEA guidance 19

2.2.2 IEC standard 22

2.2.3 NRC guidance 23

2.2.4 NEI guidance 25

2.3 Machine Learning Algorithms and Data-driven Models Used in Cyber-attack Detection 25

2.3.1 Cyber-attack detection using cyber data 28

2.3.2 Cyber-attack detection using process Data 32

2.3.3 Necessity of building an ICS testbed based on the state-of-the-art of the cyber-attack detection 35

2.4 Overview of Cybersecurity Risk Assessment for ICSs 36

2.4.1 Attack tree based vulnerabilities assessment 37

2.4.2 Quantitative cyber risk reduction using compromise graph 38

2.4.3 Bayesian network . . . 39
2.4.4 Model-data integrated risk assessment using Bayesian network . . . 41
2.4.5 Attacker vector analysis in cybersecurity assessment for digital I&C system in NPP . . . 42
2.4.6 Cybersecurity risk analysis and assessment system for digital I&C system in NPP . . . 43
2.4.7 Cybersecurity risk model using Bayesian networks for I&C systems . 43
2.4.8 Risk assessment model based on Bayesian networks and event tree . . 44
2.5 Localized Cybersecurity Research of PLC . . . 44

3 Methodology **47**
3.1 Cybersecurity Solution Platform . . . 47
3.1.1 Data collection and extraction system . . . 48
3.1.2 Cyber-attack detection system . . . 50
3.1.3 Cause Analysis System . . . 53
3.1.4 Cyber-attack response system and operator alert display system . . . 59
3.2 Supervised Cyber-Attack Models Based on Host-related Network and Host System Data . . . 59
3.2.1 KNN . . . 60
3.2.2 Decision tree . . . 61
3.2.3 Bagging . . . 62
3.2.4 Random forest . . . 63
3.2.5 Procedure of applying classification methods . . . 63
3.3 Unsupervised Cyber-Attack Detection Models . . . 64
3.3.1 AAKR model . . . 64
3.3.2 Support vector regression (SVR) model . . . 66
3.3.3 Ensemble regression model . . . 68
3.3.4 Principal component regression (PCR) model . . . 69
3.3.5 Auto-associative models . . . 71
3.3.6 AAKR cyber-attack detection using network flow data . . . 72

3.3.7 Auto-associative Cyber-attack Detection Models Using Process Data 73
3.3.8 Performance matrix . . . 74
3.4 Development of Localized Cybersecurity Models for Key Equipment . . . 75

4 ICS Cybersecurity Testbeds Description and Attack Scenarios Setup 78
4.1 ICS Cybersecurity Physical Testbed Description . . . 79
4.1.1 Physical process experimental facility . . . 80
4.1.2 SCADA system and cyber part . . . 83
4.2 Data Collection of ICS Testbed . . . 86
4.3 Attack Scenarios on ICS Testbed . . . 87
4.4 A Case Study: A Serial False Data Injection . . . 92

5 HIL Testbed Description and Attack Scenarios Setup 95
5.1 HIL Testbed Description . . . 95
5.1.1 Asherah simulator . . . 96
5.1.2 Prosys OPC UA simulation server . . . 97
5.1.3 Softing dataFEED OPC suite . . . 97
5.1.4 Real-world PLC for steam generator water level control . . . 99
5.2 Data Collection of HIL Testbed . . . 101
5.3 False Data Injection Attacks on HIL Testbed . . . 101

6 Results 104
6.1 Normal Operational Transients of ICS Testbed . . . 104
6.2 Results of Cybersecurity Attack Detection System . . . 108
6.2.1 Cyber-attack detection results of supervised classification models using cyber data . . . 108
6.2.2 Cyber-attack detection Results of unsupervised AAKR model using network flow data . . . 110
6.2.3 Cyber-attack detection results of unsupervised AAKR model using process data . . . 112

6.2.4 Cyber-attack detection results of the serial false data injection using the cybersecurity solution platform 115
6.2.5 M3 models comparison in the serial false data injection 120
6.3 Results of Cause Analysis System 122
6.4 Results of Localized Cybersecurity Models for Key Components in HIL Testbed 125

7 Conclusion and Future Work **130**

Bibliography **134**

Appendices **150**

A Detailed Data Acquisition Connection of cDAQ9188 151
B Published Works 154

Vita **156**

List of Tables

1.1 Cyber-incidents occurred in nuclear industry 5
1.2 Major characteristic difference between IT and ICS cybersecurity consideration 7

2.1 Confusion matrix 27
2.2 IDSs using machine learning algorithms based on public data sets 32

3.1 Selected Argus network flow data features description 72

4.1 List of materials for ICS testbed 82
4.2 False data injection attack steps 93

5.1 Three attack scenarios conducted towards SG in the HIL testbed 103

6.1 Performance comparison of KNN, decision tree, bagging and random Forest . 109
6.2 Computing cost comparison of KNN, Decision Tree, Bagging and Random Forest 110
6.3 Different data set description and detection results 111
6.4 Attack steps of a serial false data injection attack towards ICS testbed . . . 116
6.5 RMSE of test data (^{o}C) 120
6.6 Three attack scenarios conducted towards SG in the HIL testbed 125

List of Figures

1.1 Timeline of major ICS targeted cyber-attacks adapted from [136]. 3
1.2 Cyber threat framework [91] 4
1.3 Simplified cybersecurity level architecture in NRC RG 5.71 [120] 10

2.1 The application process and interaction of three SC45A series standards [5] . 24

3.1 Architecture and connection of ICS testbed and proposed cybersecurity solution platform [135] 49
3.2 The transient stage after a cyber-attack [69]. 51
3.3 Cyber-attack detection system with defense-in-depth concept. 52
3.4 Logic of the event discriminator. 54
3.5 The architecture of the risk assessment module. 57
3.6 Decision tree classification structure for the network data. 61
3.7 Procedure of applying the classification models. 64
3.8 The parameters for SVR models [105]. 67
3.9 The procedure of cyber-attack detection model using network flow data. . . . 73
3.10 The procedure of cyber-attack detection model using process data. 73
3.11 Logic block of the proposed localized cybersecurity strategy. 76

4.1 A typical PWR sytem [85] 79
4.2 Photo of two-loop nuclear system experimental facility. 81
4.3 Two-loop nuclear system experiment layout. 82
4.4 Testbed SCADA system and other cyber components. 83
4.5 LabVIEW model of the physical two-loop experimental facility. 84

4.6 Testbed SCADA system topology with components model number and I/O requirements. . . . 85
4.7 Main menu of Windows performance monitor. . . . 87
4.8 Basic commands for starting Argus. . . . 87
4.9 Attack patch indication for scenario C and E. . . . 90
4.10 Small LOCA due to data tampering and false data injection. . . . 92
4.11 Malicious code adds 2^{o}C to the temperature measurement. . . . 94
4.12 Two different MITM attack scenarios. . . . 94

5.1 Hardware-in-the-loop (HIL) testbed framework. . . . 96
5.2 OPC communication module in Asherah simulator. . . . 98
5.3 Prosys OPC UA server status interface. . . . 98
5.4 Softing dataFEED OPC Suite data exchange using interface. . . . 99
5.5 Data transmission sketch between Asherah and PLC. . . . 100
5.6 Attacker tampers the values in PLC. . . . 102

6.1 Process data under normal operational transients with heater power step changes from 100% to 50% and then to 100%. . . . 106
6.2 Process data under a heater power step change from 77% to 100%. . . . 107
6.3 Raw data, raw residuals, filtered residuals and fault hypothesis of data set under the MITM attack scenario . . . 112
6.4 Raw data, raw residuals, filtered residuals and fault hypothesis of data set under the DoS NI scenario . . . 113
6.5 Raw data, raw residuals, filtered residuals and fault hypothesis of testing normal data set . . . 113
6.6 The heat map of correlation coefficients. . . . 114
6.7 The detection results (fault hypothesis) of process data in data tampering and false data injection scenarios using AAKR model. . . . 115
6.8 Raw data, residuals, filtered residuals and detection result (fault hypothesis) of data under the serial false data injection attack. . . . 117
6.9 Attack detection based on process data using AAKR model . . . 118

6.10 Attack detection based on process data using AASVR model under the serial false data injection attack. 118
6.11 Attack detection based on process data using ensemble regression under the serial false data injection attack. 119
6.12 Attack detection based on process data using AAPCR model under the serial false data injection attack. 119
6.13 Prediction of the test data under normal operation using AAKR, AASVR, ensemble regression, and AAPCR model. 121
6.14 Prediction of data under the false data injection using AAKR, AASVR, ensemble regression, and AAPCR model. 122
6.15 Bayesian network for the ICS testbed. 124
6.16 Fault detection results of scenario I: water level is tampered to 15.9 constantly. 126
6.17 Fault detection results of scenario II: 0.9m was added to water level input. . 127
6.18 Fault detection results of scenario III: water level set point is tampered to 14m. 127
6.19 Inference results under scenario I: attacker alters the water level measurement input to PLC to 15.9m constantly 129

A.1 The connection of DAQ 9188 with other 5 modules. 151
A.2 The connection of NI9217. 151
A.3 The connection of NI9216. 152
A.4 The connection of NI9265. 152
A.5 The connection of NI9203. 153

List of Abbreviations

AAPCR	Auto-associative PCR
ADFA-LD	Australian Defence Force Academy Linux Dataset
ANN	Artificial Neural Network
Argus	Audit Record Generation and Utilization System
ARP	Address Resolution Protocol
BP	Bistable Processor
CADS	Cyber-attack Detection System
CDA	Critical Digital Asset
CFR	Code of Federal Regulations
COTS	Commercial-off-the-shelf
CP	Coincidence Processor
CPSs	Cyber-physical Systems
CPT	Conditional Probability Table
CRP	Coordinated Research Project
CSRI	CyberSecurity Risk Index
CSRM	Computer Security Risk Management

CUSUM	Cumulative Sum
CVE	Common Vulnerabilities and Exposures
DAG	Directed Acyclic Graph
DARPA	Defense Advanced Research Projects Agency
DDoS	Distributed Denial of Service
DoS	Denial of Service
DPI	Deep Packet Inspection
FAR	False Alarm Rate
FN	False Negative
FP	False Positive
FPBN	Fuzzy Probability Bayesian Network
HIL	Hardware-in-the-loop
HMI	Human Machine Interface
HX	Heat Exchanger
I&C	Instrumentation and Control
IAEA	International Atomic Energy Agency
ICA	Independent Component Analysis
ICS	Industrial Control System
IDS	Intrusion Detection System
IEC	International Electrotechnical Commission
INL	Idaho National Laboratory

IoC	Indication of Compromise
IoT	Internet of Things
IPS	Intrusion Prevention System
IT	Information Technology
ITP	Interface and Test Processor
KNN	K Nearest Neighbor
KPCA	Kernel Principal Component Analysis
LOCA	Loss of Coolant Accident
LSTM-RNN	Long Short Term Memory Recurrent Neural Network
M1	Module 1
M2	Module 2
M3	Module 3
MITM	Man-inthe-middle Attack
MOV	Motor Operated Valve
MSE	Mean Squared Error
MTP	Maintenance and Test Processor
NaN	Not-a-Number
NEI	Nuclear Energy Institute
NI	National Instruments
NPP	Nuclear Power Plant
NPV	Negative Predictive Value

NRC	Nuclear Regulatory Commission
NSS	Nuclear Security Series
OCTAVE	Operationally Critical Threat Asset and Vulnerability Evaluation
OPC	Open Platform Communications
OT	Operational Technology
PART	PArtial Decision Tree
PCA	Principal Component Analysis
PCR	Principal Component Regression
PCs	Principal Components
PI	Proportional-Integral
PID	Proportional Integral Derivative
PLC	Programmable Logical Controller
PPS	Plant Protection System
PWR	Pressurized Water Reactor
RB	Reactor Building
RBF	Radial Basis Function
RCS	Reactor Cooling System
RG	Regulatory Guide
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RPS	Reactor Protection System

RTDs	Resistance Temperature Detectors
SBLOCA	Small Break Loss of Coolant Accident
SC45A	Subcommittee 45A
SCADA	Supervisory Control and Data Acquisition
SG	Steam Generator
SPDS	Safety Parameter Display System
SVDD	Support Vector Data Description
SVM	Support Vector Machine
SVR	Support Vector Regression
SWaT	Secure Water Treatment Testbed
TN	True Negative
TP	True Positive
UA	Unified Architecture
UTK	University of Tennessee, Knoxville
VV	Verification and Validation
VFD	Variable Frequency Drive
WPM	Windows Performance Monitor

Chapter 1

Introduction

Industrial control systems (ICSs) are widely used in various industrial sectors and critical infrastructures. According to Presidential Decision Directive 63 of 1998, critical infrastructures are cyber-physical systems (CPSs) essential to the minimum operations of government and economic systems, including but not limited to those involving energy, transportation, water, emergency services, telecommunications, and finance [34]. Supervisory Control and Data Acquisition (SCADA) systems are one type of ICS, which are computer-based control systems composed of a set of network devices and network protocols to monitor and control physical processes. SCADA systems are highly used in critical infrastructures all over the world, such as electrical power grids, oil and natural gas pipelines, and transportation systems. Different industries may use different terms for these systems, for example, the term "instrumentation and control (I&C) system" is used for the control system in nuclear power plants (NPPs). According to the United States Nuclear Regulatory Commission (NRC), I&C system serves as the "nervous system" of the NPPs in the following aspects: provides operators with critical information of plant operation; enables operators to control the plant systems; protects the reactor core automatically during potential accidents [86]. The research conducted in this dissertation is general for all CPSs, including ICS system, SCADA system, and I&C system. These three terms may be used interchangeably across

the dissertation. A nuclear I&C system is utilized as an example in this dissertation to demonstrate the potential applications of the research. [1]

Applying digital techniques to ICSs improves the efficiency of processes (such as control precision) and connects geographically distributed subsystems. The NRC has listed several advantages of the digital I&C system: it can improve control and increase the operational efficiency, it can improve the diagnostics of automatic plant protection systems, and it is expected to improve operational safety [86]. However, with all the benefits that digitalization brings, it also introduces new security challenges that are not part of the consideration of old analog systems. One issue receiving increasing attention is that of cybersecurity, as the number of cyber-attacks towards ICSs is increasing and attacks are becoming more sophisticated. These threats include both insider threats and outsider threats. Insider threats are defined as intentional or unintentional misuse or attacks from current or former employees, contractors, vendors or other business partners who have been given access to assets such as the plant networks, systems or data. Outsider threats are an individual or a group of individuals external to the organization who do not have authorized access to its assets but seek to gain access and may cause intentional or unintentional damage once they gain access. The outsiders could range from an individual hacker with idle curiosity to advanced persistent attackers and sophisticated nation states.

Several impactful cyber-incidents and attacks will be detailed in Chapter 2; however, to illustrate how these attacks are increasing in both complexity and frequency, a timeline of ICS-focused cyber-attacks is shown in Figure 1.1. This timeline is not exhaustive and is based on public reports of identified cyber-attacks which, in some cases, are suspected to have been initiated in years previous to when they were revealed [48][13][44]. As such, it is likely that the cyber-attacks initiated in 2018 and 2019 have not been formally identified or publicly released, meaning only a few of them are shown in this timeline. A few choice threats from this timeline are the 2010 Stuxnet virus, the 2015 BlackEnergy malware, and the 2017 HatMan malware. The Stuxnet virus attacked Iranian nuclear enrichment centrifuges, damaging about 980 in total [67] and demonstrating that ICSs deployed in high-security

[1] Part of this chapter is a reformat and reprint of partial materials as they appear in [135] and [136]. The dissertation author was the primary investigator and lead author of these papers.

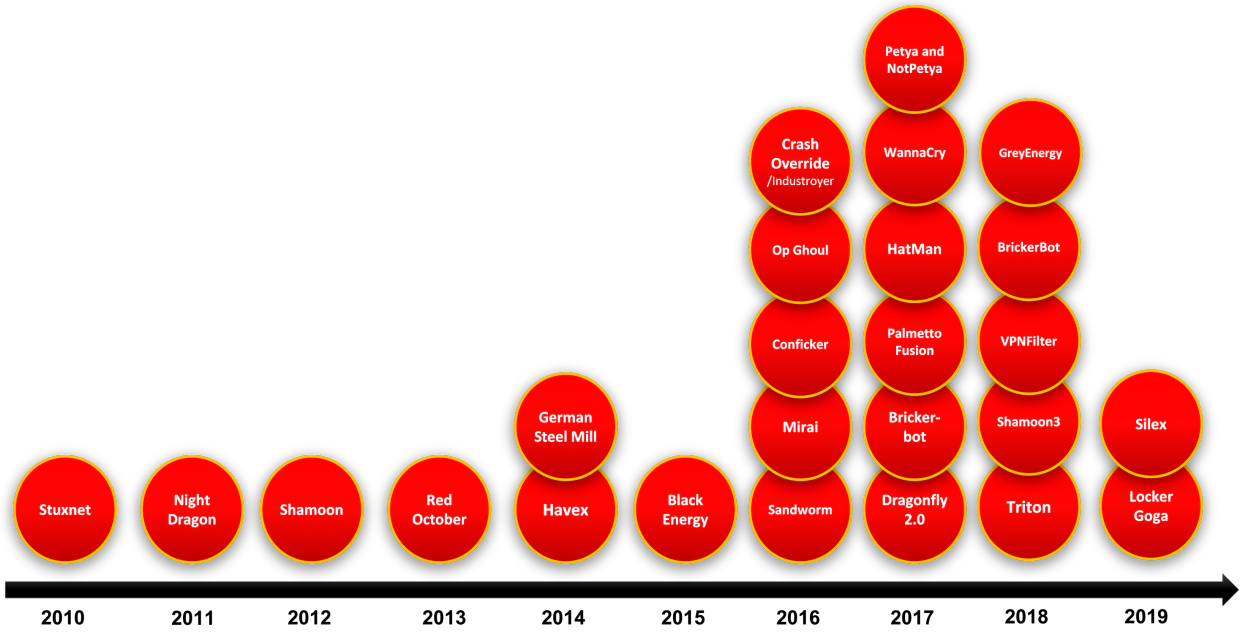


Figure 1.1: Timeline of major ICS targeted cyber-attacks adapted from [136].

facilities may still have security vulnerabilities. The BlackEnergy malware was utilized by a team of attackers to attack the Ukrainian power grid, which caused a large-scale power outage[42]. The HatMan malware was used to infect Triconex controllers, modifying in-memory firmware to add additional programming, and is believed to be behind the attack on critical infrastructure through Schneider Electric's Safety Instrumented Systems [43]. A report published by Kaspersky Lab in March 2018 revealed that about 40% of all ICSs protected by Kaspersky Lab solutions in energy organizations were attacked by malware at least once from June to December 2017 [57]. ICS-targeted cyber-attacks are increasing with time as more systems proliferate and vulnerabilities are found and more high-value assets become the targets. A survey published by SANS, a large source for information security training, points out that cyber-attacks conducted by foreign nation-states or state-sponsored parties has increase to 27.6% in 2019 [103]. The cyber-incidents occurred in nuclear industry is summarized here in Table 1.1 and will be detailed in Chapter 2. Therefore, enhancing the cybersecurity of ICSs are necessary to ensure the safe operation of industry including critical infrastructures.

Figure 1.2 shows a cyber threat framework developed by the US Government to present consistent characterization and categorization of cyber threat events, which provides a

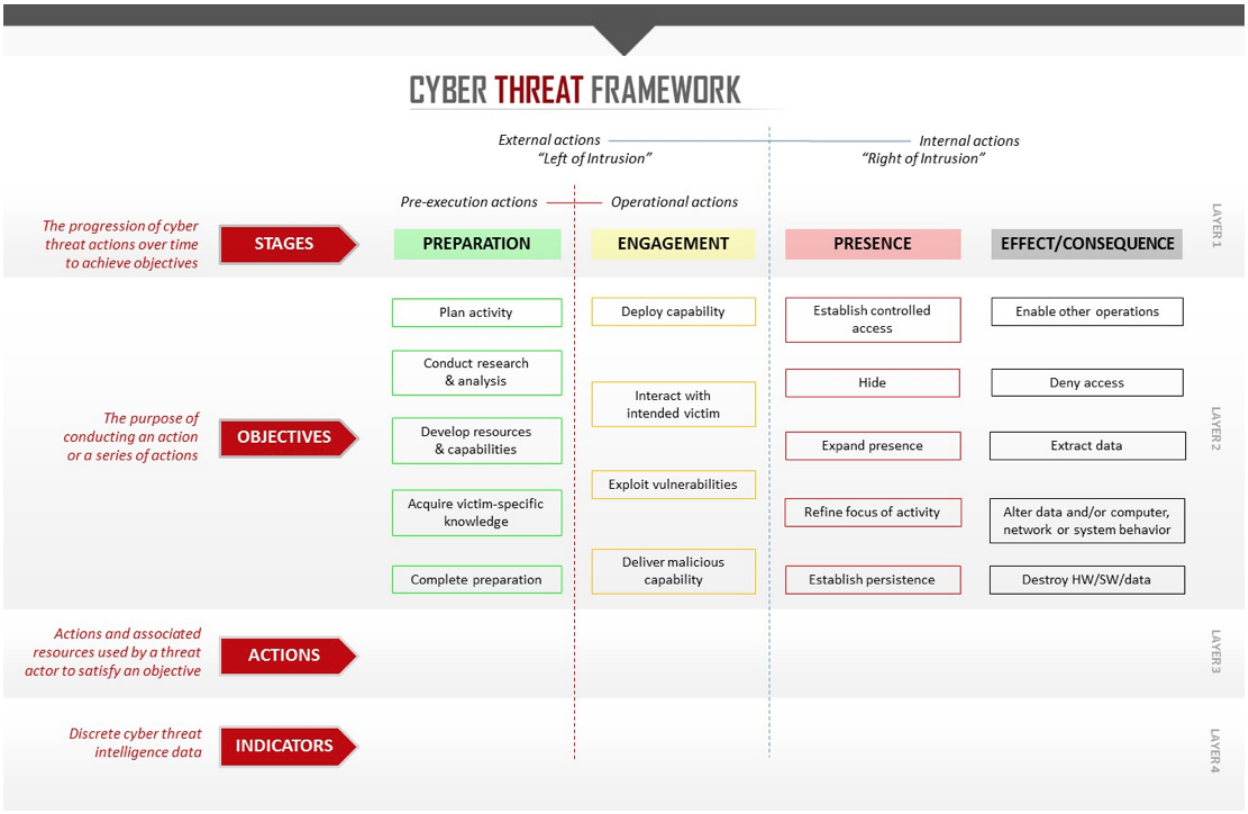


Figure 1.2: Cyber threat framework [91]

Table 1.1: Cyber-incidents occurred in nuclear industry

Year	Facility	Main cause	Consequences in brief
2003	Davis-Besse NPP [59]	Slammer worm	Safety parameter display system was inaccessible
2006	Browns Ferry NPP [63]	Ethernet broadcast storm	Loss of both main coolant
2008	Hatch NPP [123]	Software update	Automatic reactor trip
2010	Iranian nuclear enrichment facility [67]	Stuxnet malware in falsh drive	984 centrifuges were damaged
2014	Monju NPP [106]	Malware introduced by software update	The plant had been accessed over 30 times
2016	Gundermmingen NPP [124]	Malware	Two malware named Conficker and W.32Ramnit were discovered on several computers

common language for describing cyber threat activity [91]. The adversary life cycle contains four stages, starting from adversary preparation and planning, to adversary engagement with temporary non-intrusive disruptions, followed by establishing and expanding the presence on target, and finally ending with the causing of consequences for the target. Different defensive techniques should be deployed for each stage to protect the system. In the preparation stage, the attacker(s) conduct research and analysis to acquire target-specific knowledge and make plans. Information protection and intrusion prevention systems (IPSs), such as firewalls and data diodes, should be strengthened to prevent the adversary from obtaining critical information to make a feasible attack plan. Since IPSs are well-developed and easy to deploy, they are not within the scope of this research. In the engagement stage, the attacker(s) access the target to exploit vulnerabilities and deliver malicious capabilities. In the presence stage, the adversary may establish controlled access, expand the presence with lateral movement, and hide the malicious activities to establish persistence. Intrusion detection systems (IDSs) or cyber-attack detection systems should be deployed to detect the adversary in these early stages, although in the Engagement stage the IPS ideally will also

play a role in stopping initial intrusions before the IDS is needed. The Effect/Consequence stage can have multiple forms, ranging from loss/theft of data to access lockout of qualified personnel and even damage to hardware. Mitigation measurements should be deployed to minimize the consequence in the last stage.

Although the cybersecurity research in the Information Technology (IT) domain is mature, adopting these cybersecurity approaches directly is not feasible for ICSs due to the large differences between ICSs and IT systems. First of all, IT systems prioritize security objectives in the order of confidentiality, integrity, and availability, while ICSs prioritize availability first, followed by integrity and confidentiality [33]. The difference in objectives priority results in different strategies for cybersecurity in IT and ICSs. For example, to ensure the availability in ICSs, port scanning tools and encryption techniques, which are commonly used in IT domain, may not be appropriate. This is because most ICSs have low bandwidth which IT techniques tend to occupy, so that they may impact system availability and could lead to unintended component denial of service (DoS), such as the broadcast storm incident occurred in Browns Ferry NPP (will be detailed later). Other major differences between IT and Operational Technology (OT) are summarized in Table 1.2; interested readers can find details with examples in reference [128][83]. With these differences between ICSs and IT systems, dedicated research is necessary to meet the unique requirements of ICS cybersecurity. This dissertation facilitates this issue from both overall and local cybersecurity strategy development.

Since ICSs are crucial for safety-critical and valuable systems, ICS security requests both resilience to cyber-attacks and situational awareness of cyber intrusions. As shown in Table 1.2, ICSs can provide two types of data: cyber data and process data. Cyber data in this dissertation refers to all the data related to IT aspects, including the host system data and all kinds of network traffic data; process data is sensor data from field devices and operational commands, which are physical parameters related to the operational status of the system. Most cyber-attacks aim to cause severe consequences by changing the physical equipment or system operation status, so it is valuable to integrate the industrial process data into the consideration of cybersecurity research. Process data could be effective to determine if the ICS load is reasonable for the actual process as well as if process-controlled variable values

Table 1.2: Major characteristic difference between IT and ICS cybersecurity consideration

Characteristic	IT	ICS
Primary priority	Confidentiality	Availability [33]
Communication time	Delays tolerated	Time Critical
Bandwidth	High	Limited
Patching and software update	Frequent and in time	Rare and slow or impossible
Communication protocols	TCP/IP, UDP	DNP, Modbus, Fieldbus, PROFINET, S7, OPC UA
Security standards	ISO-17799, NIST SP800-53	ISA99, NIST SP800 and specific standards for different industries
Cyber forensics	Available	Not deployed or limited
Penetration testing	Yes	No for field devices, maybe for human machine interface (HMI)
System life cycle	3-5 years	Usually more than 15 years
Interoperability	Not critical	Critical
Physical components	Commercially available off-the-shelf (COTS) mostly	Small portion of COTS, more custom built or industry-qualified [128]
Functional partitioning	Administrative partitions	Levels (such as Purdue reference model)
Data types	Cyber data	Cyber data and process data
Compromise consequence	Business impacts: finance and reputation	Business impacts, equipment damage, environmental destroy, personnel safety

and change rates are appropriate. Currently, most ICS cybersecurity research or products only focus on monitoring network data, which may not be effective for some intrusions like data tampering in cases where the IDS does not catch the initial sniffing and spoofing or insider attacks. Online monitoring using data analytics have been researched as an effective method for the early detection of equipment degradation and equipment failure in industry [28][18][80][17]. The fact that cyber-attacks could cause operational change suggests that applying data analytics to process data from sensors used the industry could be a promising indicator of early-stage ICS cyber-attacks [134]. Therefore, process data can be utilized in addition to cyber data to provide multi-layer detection.

To enhance the cybersecurity of ICSs, a overall comprehensive cybersecurity solution platform is proposed and developed in this research. It mainly consists of a data collection and extraction system, a cyber-attack detection system, a cause analysis system, a cyber-attack response, and operator alert display system. It combines cyber data and process data to provide better cyber-attack detection coverage, cyber risk assessment, and cyber-attack analysis. This platform is meant to work with current cybersecurity deployment, including IPSs, COTS IDSs, and physical access monitoring camera systems, to provide a defense-in-depth architecture. In addition, a localized strategy is also proposed to improve the resilience to the cyber-attack of the key equipment through implementing a data-driven model into the physical controller.

Data analysis is the key foundation of this research. To find suitable data sets, a literature survey is performed beforehand. Chen et al. have built a testbed that integrates a power system simulator and LabVIEW to study smart grid vulnerabilities [15]. Mallouhi et al. have built a testbed for analyzing the security of SCADA systems based on a power grid simulator named Power World, and a DoS attack is analyzed [74]. Since these testbeds use a simulator instead of physical experimental facilities, they are not sufficient for research in integrating process data with network data. Idaho National Laboratory (INL) built a U.S. National SCADA Testbed containing a full-scale electric power grid dedicated to ICS cybersecurity research in order to identify, assess, and mitigate current SCADA system vulnerabilities, which has already helped identify several SCADA vulnerabilities in industry [81]. However, there is no appropriate public data set available for researchers to analyze or study intrusion detection models [134]. Therefore, two testbeds were designed and built at the University of Tennessee, Knoxville to generate different types of data needed for this research: an ICS tesbed and a hardware-in-the-loop (HIL) testbed.

The ICS testbed contains a two-loop nuclear system mock-up experimental facility and a SCADA system, to generate different types of data under a variety of scenarios. The HIL testbed integrates a pressurized water reactor (PWR) simulator and a programmable logical controller (PLC) to generate data and evaluate the developed cybersecurity models under false data injections. Different cyber-attack scenarios were designed and carried out on these two testbeds to demonstrate possible vulnerabilities in ICSs and provide different types of

data. Data was collected under normal operational transients and cyber-attack scenarios. Supervised and unsupervised data-driven models were developed for cyber-attack detection. In the proposed cybersecurity solution platform, the data collection and extraction system and the cyber-attack detection system are well developed and evaluated based on the testbed while frameworks of other three systems are designed due to the limitation of the research scale and information access. The localized strategy is developed and evaluated as a proof-of-concept research to demonstrate its promising potential and effectiveness. In addition, these two testbeds provide the best practice for future cybersecurity testbed construction.

1.1 Problem Statement

Critical infrastructures are vulnerable to cyber-attacks, as proven by various cyber-incidents and ICS-targeted cyber-attacks. Taking NPPs as examples, as they age and the majority of the industrial world moves toward digital systems, analog technologies become difficult to procure and maintain. Utilities are increasingly looking toward digital I&C systems to replace these aging analog I&C systems. Most new reactors, including both near-term Gen III+ and future advanced reactors, are being designed with significantly or fully digital I&C systems. These digital systems bring challenging cybersecurity concerns as cyber-attacks are increasing in both numbers and scales. Therefore, enhancing the cybersecurity of ICSs is an urgent issue with the progress of the digitalization.

To address cybersecurity concerns, the U.S. NRC requires a site to employ defense-in-depth strategies to protect critical digital assets (CDAs) and recommends a defensive architecture with security levels as shown in Figure 1.3 [120]. Level 4, containing CDAs that are important to safety and security functions, requires the strictest security and must be protected from lower levels. Other security levels are assigned by the power plants according to RG 5.71 [120]. The current NPP cybersecurity deployment mainly focuses on intrusion prevention through firewalls, network segregation measures such as "air gap" and data diodes that only allow one way data flow from Level 4 to Level 3 and from Level 3 to Level 2, and physical walk-down inspection of key systems. This may be effective against a subset of

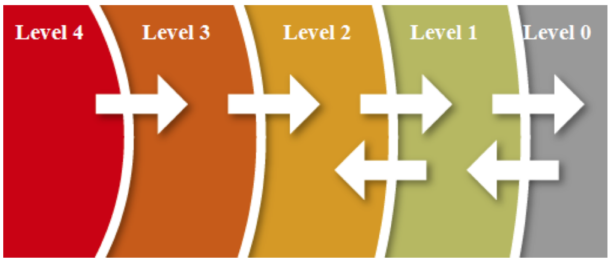


Figure 1.3: Simplified cybersecurity level architecture in NRC RG 5.71 [120]

known attacks, but given time and resources a well-motivated attacker could infiltrate an I&C system and covertly deploy novel cyber-attacks.

Another issue existing across the I&C system cybersecurity domain comes from a lack of information exchange and integration between physical processes and cyber-activities. The cyber-experts in the IT team who monitor cyber-infrastructure and the facility engineers and operators in the OT team who monitor plant processes are largely independent. Lack of cross-disciplinary training and team integration may result in insufficient understanding of the whole scenario when a cyber event happens. Robust cybersecurity approaches for I&C systems require information integration from both IT and OT systems to provide better situational awareness [135].

Due to the potentially very serious consequences of an accident, nuclear power is different from other industries. Safety objectives include controlling reactivity, maintaining core cooling, and protecting the integrity of the pressure barrier. A cyber-attack that compromises any one of these three objectives could lead to severe consequences. Even in the absence of core damage or a radiological release, cyber-attacks on NPPs could lead to economic consequences and loss of public trust, either of which would challenge continued operation. Greater situational awareness of both the plant processes and the digital I&C system can provide better assurance of security. Robust approaches to monitor and assess the security status of digital I&C systems are necessary to enhance cybersecurity, which in turn ensures the safety and security of nuclear facilities [135]. Therefore, digital I&C systems in NPPs require a cybersecurity platform which:

1. Deploys a defense-in-depth strategy;

2. Integrates process data to provide better situational awareness; and

3. Has the ability to detect unknown cyber-attacks.

Therefore, a cybersecurity solution platform is proposed to fulfill these requirements and a real-time ICS testbed was built to facilitate this development. This platform is for providing better situation awareness and cyber defense in an overall system scale, which usually uses data from across the systems or a data historian. This overall defense may be insufficient due to two reasons. A wider attack surface is created when whole I&C systems or data historians are involved, since the possibility of data compromise is introduced by long distance data transfer or multi-procedure transfer. The other reason is that centralized detection may be subject to availability of the real data. For example, in the Davis-Besse NPP incident, the display system was inaccessible to operators for a significant amount of time; during this time, if an attacker had altered some control variables, operations could have become critically unsafe. To improve the cyber-resilience of key equipment, it is valuable to develop anomaly detection and even responses to cyber-attack techniques based on the local data within the controller or the equipment before the data is transferred to the data acquisition system. This strategy minimizes the attack surface and makes equipment more resilient to some type of cyber-attacks, such as replay attacks for human machine interfaces (HMI) and maximizes the availability of the data access.

1.2 Original Contributions

ICS targeted cyber-attacks are growing rapidly but the current industry cybersecurity defense is not sufficient for sophisticated cyber-attacks and insiders due to two major reasons: there are situations in which network segregation fails, and some malicious activities leave no detectable trace or mask their cyber trace for current COTS IDSs. Therefore, enhancing the cybersecurity of ICSs to improve the resilience to cyber-attacks is necessary and urgent. This work makes several contributions to the field of cybersecurity for ICSs, with a special focus on NPP digital I&C systems. These contributions can be summarized as follows:

1. This research proposes a cybersecurity solution platform architecture integrating different types of data source which combines both IT and OT teams and performs multiple

functions, including cyber-attack detection and analysis for industry applications to provide better situational awareness and facilitate decision-making.

2. There are limited existing ICS testbeds built for cybersecurity research, and the majority of such testbeds are operated at the national lab level, which limits cybersecurity research potential for smaller educational and research institutions. A real-time ICS testbed was designed and developed in this work for such a purpose, and this dissertation serves to document the design and implementation of the ICS testbed. A HIL testbed with a nuclear reactor simulator and a PLC was developed as a proof-of-concept and best practice testbed for this research and for the guidance of future HIL testbeds as well. The ultimate goal is to accelerate the development of similar systems at other institutions to support cybersecurity and other ICS-related research.

3. There is limited research in the application of unsupervised learning algorithms for cyber-attack detection in industrial applications, which may be due to high computational cost. This work investigates computationally viable supervised and unsupervised learning algorithms to detect cyber-attacks with and without cyber traces, such as false data injection attacks launched by insiders. Therefore, this work uses machine learning to defend the cyber domain.

4. A novel decentralized cybersecurity strategy is proposed to enhance critical assets' resilience to cyber-attacks, including cyber-attack detection locally and parameter inference. This strategy deploys an empirical model in the local controller to provide faster localized cyber-attack detection with minimum attack surface. This work also shows the potential that the compromised sensed signal could be substituted by a virtual sensor inferred from other variables to reduce the cyber-attack consequence once the anomaly is detected.

1.3 Structure of the Document

The dissertation is organized as follows. Chapter 2 provides a comprehensive literature review that is relevant to this research, including ICS cyber incidents and attacks with a focus on the nuclear industry, an overview of cybersecurity standards for the nuclear industry, machine learning algorithms and data-driven models used in cybersecurity with both network data and process data, a summary of the cybersecurity risk assessment methods in ICSs, and improving the cybersecurity of PLCs. Chapter 3 describes the methodology, presenting the proposed cybersecurity solution platform architecture and the localized cybersecurity models for key equipment. Chapter 4 gives a detailed description of the cybersecurity solution platform related experiment setup. This includes a description of the developed ICS testbed, data collection methods for cyber data and process data, the designed cyber-attack scenarios and a serial false data injection. Chapter 5 presents the detailed setup related to localized cybersecurity strategy research. This includes the HIL testbed setup description, data collection, and the designed false data injection attacks. Chapter 6 presents the results, including ICS and HIL testbed operation and the effectiveness of the developed cybersecurity solution platform and localized cybersecurity strategies in enhancing the cybersecurity overall and locally. The last chapter summarizes the conclusions of this research, as well as proposes areas of future work.

Chapter 2

Literature Survey

This chapter presents a literature survey of the background of this cybersecurity research. First, ICS cyber incidents and attacks are identified in both the nuclear industry and non-nuclear industries to show the urgent need for enhancing cybersecurity in industry, including the nuclear sector. Then an overview of cybersecurity guidance for the nuclear industry is given, which points out the cybersecurity strategy and methods in high level. Based on these two reviews, data analytic methods are chosen to perform cyber-attack prevention, detection, and assessment. Then, a survey of the machine learning algorithms and data-driven models used in cyber-attack detection is given. A brief overview of the cybersecurity risk assessment for ICS systems and PLC cybersecurity is given in the end.[1]

2.1 ICS Cyber Incidents and Attacks

A significant number of cyber incidents or attacks have occurred in industry since the year 2000. Many reports on these attacks can be found online; therefore, only several typical ones from non-nuclear industry are listed in this chapter. Compared to other industries, the nuclear industry has relatively high standards for safety and facility security. Thus, all

[1]Part of this chapter is a reformat and reprint of partial materials as they appear in [135], [136], [133], and [132]. The dissertation author was the primary investigator and lead author of these publications.

public cyber incidents or attacks that have occurred in the nuclear industry are detailed in this chapter.

2.1.1 Cyber incidents and attacks occurring in the non-nuclear industry

The first confirmed cyber-attack targeting ICSs was the Australia Maroochy attack in 2000 [109]. The SCADA system of the Maroochy waste water treatment facility manages nearly 900 kilometers of sewer lines. A disgruntled employee who was refused a position in the city council tampered with the SCADA system to pump over 800,000 liters of sewage into parks and residential areas. This first cyber-attack shows how vulnerable the SCADA system was and how easily it could be attacked by insiders.

INL carried out a proof-of-concept experiment called the Aurora Generator Test in 2007 to demonstrate the possibility of component damage caused by a cyber-attack on the power grid. In this experiment, a remote computer rapidly opened and closed a diesel generator's circuit breakers out of synchronization with the grid in order to maximize stress, which destroyed the diesel generator [21].

In 2015, BlackEnergy malware was used to attack the Ukrainian power grid and caused a large-scale power outage for six hours [42]. This proves the possibility of large-scale cyber-attacks towards power grid through a simultaneous attack on three separate companies. The attackers first infiltrated the enterprise networks of the three power providers through malicious Microsoft Office documents which exploited vulnerabilities in the Office suite of software to install the BlackEnergy malware on target systems. The malware then established communication with the attackers, who began to use the malware to collect security credentials and perform network reconnaissance until the attackers were able to gain persistent access and masquerade as authorized users. Then the attackers identified and moved into the control network where the SCADA workstations were located in, and also began to learn the different power distribution management systems that were being used by the different companies on these workstations. This was all done over an estimated six-month period before the cyber-attack was carried out. When the actual

attack was enacted, the workstations were used to open the power breakers for at least 27 substations. Immediately afterwards, malicious firmware was uploaded to the serial-to-ethernet devices which communicated commands from the SCADA network to substation control systems; in this way the attackers ensured that even once operators had regained control of the workstations, restoration of power would not be possible. At the same time, the uninterruptible power supply in at least one company building had been modified beforehand by the attackers, so that when the cyber-attack was initiated, the power supplied to company buildings and/or data centers was also impacted. Finally, a DoS attack on at least one company's call centers was initiated after this attack, to both increase confusion and delay the company's ability to respond.

There are lots of the cyber-attacks that occurred in the industry and the number is increasing as more vulnerabilities are exploited. Most impactful cyber-attacks are summarized in the timeline shown in Figure 1.1. Interested readers can find details on these attacks in relevant cybersecurity reports [103] [13] [48].

2.1.2 Cyber incidents and attacks in nuclear industry

In 2003, the Davis-Besse NPP was infected with the Slammer worm, which resulted in the safety parameter display system (SPDS) being inaccessible to operators for four hours and fifty minutes [59]. The plant had been offline since February 2002, so that there was no physical consequence. In this attack,the Slammer worm traveled from a consultant's network to the corporate network and then to the process control network. It managed to do this even though the NPP had a firewall to protect the corporate network from the Internet, which would have prevented the Slammer infection. The consultant created a connection behind the firewall to his office network which opened the door for the Slammer by bypassing the firewall. Since there was no firewall between the corporate network and the process control network at that time, the Slammer infected the SPDS without obstacles. This cyber incident shows the fact that even though there are firewalls in a NPP, the NPP may still establish a connection to the Internet without knowledge of the appropriate personnel.

In 2006, the Browns Ferry NPP Unit 3 was shut down manually after variable frequency drives (VFDs) and programmable logic controllers (PLCs) initiated a DoS [116]. The VFDs

drove the circulation pumps and the PLC served as the condensate demineralizer controller. Both of these devices had embedded microprocessors that allowed communication of data over Ethernet. The Ethernet broadcast is a communication mechanism in which every device connected to the local area network receives all the broadcast packets and then determines which packet is meant for itself and ignores those packets which are not by examining every packet. In this incident, the VFDs and PLCs were flooded so that the recirculation pumps were not able to cool the reactor. Even though this cyber incident was not caused by a cyber-attack, it shows a possible attack scenario for NPPs. Together with a Slammer worm like that in the Davis-Besse NPP incident, a cyber-attack in a similar scenario could lead to a severe accident. The spread of Slammer within the control network could lead to a DoS attack on critical components such as VFDs which disable the recirculation pumps to cool the reactor core, while disabling operators' access to SPDS to prevent awareness.

In 2008, Hatch NPP Unit 2 experienced an automatic reactor trip due to seven condensate demineralizers' outlet valves closed at almost the same time which resulted in a temporary loss of feedwater with a root cause of a computer update in the corporate network [123]. An engineer was testing a software on the plant's chemistry data acquisition system server (in the corporate network) which was utilized to collect data from process control network. However, the synchronization program in the update automatically synchronized data on both process control network and corporate network, which was not realized by the engineer. When the engineer rebooted the computer, the value zero was synchronized and sent to the programmable logic controller (PLC) which was utilized to control the demineralizers. Therefore, the PLC switched to manual control with input of zero flow demand which resulted in a temporary loss of feedwater. Then the safety system received a low water level signal and automatically scrammed the reactor. Although this was not an intended cyber event, it could have been done maliciously by attackers.

In 2010, Stuxnet attacked the Iranian nuclear enrichment centrifuges and destroyed 984 of them [67]. The malware was brought into the facility on a USB drive and exploited several zero-day vulnerabilities to inject its malicious code into Siemens PLCs to speed up and slow down the centrifuges, causing them to wear out. Meanwhile, the malware faked sensor outputs to mask the attack from operators by sending normal behavior data. Stuxnet

proves that a sophisticated cyber-attack: a. does not require an Internet connection, so that the so-called "air-gap" in an NPP is not sufficient to prevent all cyber-attacks; and b. the cyber-attacks could disrupt or overwrite the monitoring system by false data injection to mask the abnormal behavior caused by an ongoing cyber-attack.

On January 2nd, 2014, one computer in the control room at Monju NPP was compromised and was shown to have been accessed over 30 times in the previous five days due to a malware attack [106]. The malware was believed to be introduced by a free application update by an employee on one of the computers in the corporate network. Some data on the compromised machine had been stolen and was sent to a command & control server located in another country. In April 2016, two malware named Conficker and W.32Ramnit, were discovered on several computers at Gundermmingen NPP in Germany [124].

These nuclear industry cyber incidents suggest that firewalls and network segregation are not sufficient for preventing cyber-attacks, especially for insiders and systems that support removable media. Indeed, the number and capability of cyber-attacks that target ICSs has increased rapidly, indicating that more cyber threats should be considered for NPPs [134]. Therefore, there is an urgent need to enhance cybersecurity for NPPs.

2.2 Overview of Cybersecurity Guidance for Nuclear Industry

Industry uses cybersecurity standards and guidance to define and deploy cybersecurity activities. There are some general industrial standards series like ISA 99 [49] and NIST SP800 [84]. This section overviews the standards and guidance for cybersecurity in the nuclear industry. They are provided by the International Atomic Energy Agency (IAEA) and International Electrotechnical Commission (IEC) internationally, and the U.S. NRC and Nuclear Energy Institute (NEI) within U.S. This section first presents a detailed overview of guidance published by IAEA since part of this dissertation work is based on an IAEA coordinated research project (CRP). Since most provided guidance has similar

basic instructions for deployment in different levels of detail, other standards and guidance are summarized at a high level.

2.2.1 IAEA guidance

IAEA is an international organization that promotes the peaceful use of nuclear energy. In regards to cybersecurity, the IAEA assists member states in improving cybersecurity capabilities at state organizations and licensees through guidance development, training courses, information exchange, and CRPs.

The cybersecurity guidance of the IAEA is published in the IAEA Nuclear Security Series (NSS). NSS provides international consensus guidance on nuclear security to support member states. There are four sets of publications in NSS: Nuclear Security Fundamentals, Nuclear Security Recommendations, Implementing Guides, and Technical Guidance. Nuclear Security Fundamentals gives objectives and principles and are the basis for Nuclear Security Recommendations, which are the applications of Fundamentals and provide the general approaches, concepts and strategies. Implementing Guides state how the recommendations should be applied, and Technical Guidance is the detailed guidance which explains specific implementation and technical subjects. The documents related to cybersecurity are NSS No.20 (Objective and Essential Elements of a State's Nuclear Security Regime) in Nuclear Security Fundamentals [37], NSS No.13 (Nuclear Security Recommendations on Physical Protection of Nuclear Material and Nuclear Facilities) in Nuclear Security Recommendations [36], NSS No.23-G (Security of Nuclear Information) in Implementing Guides [39], NSS No.17 (Computer Security at Nuclear Facilities) [35], and NSS No.33T (Computer Security of Instrumentation and Control Systems at Nuclear Facilities) [41] in Technical Guidance. IAEA also released three draft guidance, including draft technical guidance NST036 (Computer Security of Instrumentation and Control Systems at Nuclear Facilities) and draft implementing guidance NST045 (Computer Security for Nuclear Security) [40]. Major concepts and requirements from these NSSs are summarized here.

NSS No.20 is the primary publication in the IAEA NSS as a Nuclear Security Fundamental, which applies to nuclear material as well as other radioactive material and their associated facilities and activities. It defines that "the objective of a State's nuclear

security regime is to protect persons, property, society, and the environment from harmful consequences of a nuclear security event." It stresses that the cybersecurity is part of the nuclear security and defines twelve security regimes, including identification and assessment of nuclear security threats, use of risk informed approaches, and detection of nuclear security events.

NSS No.13, as the recommendation document, gives general strategies to fulfill the requirement in NSS No.20. For example, risk management, graded approach, and defence-in-depth are three strategies could be applied as risk based protection systems and measures. In terms of cybersecurity, NSS No. 13 suggests that "computer based systems that are used for physical protection, nuclear safety and nuclear material accountancy and control should be protected from compromise, which includes but is not limited to cyber-attack, manipulation and falsification"; Configuration management, access control, and record of access should be deployed for the computer systems and software to ensure the security.

"Information" is defined as knowledge regardless of its existence and expression form in NSS No.23-G. Information security is the security strategies implemented to ensure the confidentiality, integrity and availability of information. Sensitive digital assets are defined as digital assets that store, control, process, or transmit sensitive information, or that otherwise serve a significant function for the nuclear power plant. No.23-G defines four levels of security, which are unclassified, restricted, confidential, and secret from low security to high. The consequences of each level being compromised are increasing accordingly. If unclassified information is compromised, the facility is not required to deploy additional security measures. Compromise of restricted information would likely affect diplomatic relations adversely, cause distress, and make operation difficult. Compromise of confidential information would likely cause damage to areas including diplomatic relations, operational effectiveness and security; when this happens, the facility should shut down or otherwise substantially disrupt significant national operations. If secret information is compromised, it would likely cause serious damage, threaten to life directly, and raise international tension. NSS No.23-G also gives security categorization examples to guide facilities how to classify different items to the corresponding levels.

NSS No.17 proposed a graded approach to computer security with details about requirements, deploy strategies and an example of how to apply the graded approach. One implementation is categorizing system into zones and assigning different security levels. Zones are a logical and physical concept for grouping systems with similar protection demands together to make administration easy and effective. Each zone has a security level assigned. Security levels define the security protection degree required, that is, each level requires different sets of protection measures. A security level can have multiple zones assigned to it. An example is given where the generic level measures are applied to all the zones and security levels range from level 5 (level of least protection needed) to level 1 (level of most protection needed). It also points out that the risk assessment process should be fed back into and influence the graded approach. Generic levels contain basic security measures such as security gateways, IPSs, IDSs, and anomaly detection systems in place. Level 1 performs accident prevention functions, where all network data flows from systems in lower security levels are forbidden to enter level 1. Physical access to level 1 systems is strictly controlled and remote maintenance access is forbidden.

NSS No.33T focuses on the protection of I&C system cybersecurity with detailed implementation suggestions. It recommends a risk informed approach, in which a facility computer security risk management (CSRM) process should be implemented to identify the system I&C vulnerabilities to cyber-attack and the consequence of a successful compromise. The results are utilized to assign different security measures to different security levels. cybersecurity policies should be applied throughout the I&C life cycle activities, including both common elements of all life cycle phases and specific life cycle activities. Common elements include management systems, cybersecurity reviews and audits, configuration management, verification and validation, security assessments, documentation, design basis, access control, protection of the confidentiality of information, security monitoring, overall defensive computer security architecture, and defense-in-depth against compromise. Specific life cycle activities contain pre-developed item selection, design and implementation of I&C systems, integration of I&C systems, system validation, overall installation and commissioning, operations and maintenance, and modification. Detailed requirements and suggestions are given in each activity; for example, configuration management requires

that the configuration documents for I&C systems computer security measures should be protected from unauthorized access or compromise and should be classified as sensitive information. Only authorized personnel can gain access to this information on a need-to-know basis.

NST036 pointed out that cybersecurity considerations should be included in every phase of the digital I&C system life cycle [38]. It provides competent authorities and facilities with guidance on cybersecurity measures for digital I&C systems, including graded approach, security levels, and security zones. It also covers the relationship between cybersecurity and safety.

NST045 provides policy makers, competent authorities, operators and nuclear security professionals with guidance on developing, implementing and integrating cybersecurity into all aspects of nuclear security. This includes identification of sensitive digital assets, assessment of threats, vulnerabilities and impact arising, graded approach deployment with securitylevels and zones, risk assessment for defining the cybersecurity requirement,and security qualification of devices and services [40]. It defines the roles and responsibilities of different entities, including vendors, competent authorities, the State, and other relevant entities.

The suite of IAEA guidance provides overall requirements and recommendations in almost all the aspects of cybersecurity in nuclear facilities, including intrusion prevention by physical control, CDAs identification, threats and vulnerability assessment, risk assessment, graded approach deployment with security segregation, and defensive architecture deployment; intrusion detection by implementing cyber-attack detection methods; and intrusion response by identifying response organizations, making response plan and international assistance and cooperation. Therefore, the cybersecurity solutions developed in this dissertation aim at adopting some recommendations and fulfill some requirements in this suite, such as defence-in-depth architecture, cyber-attack detection, analysis, response, and risk assessment.

2.2.2 IEC standard

IEC is an international organization that promotes international co-operation on standards and is a conformity assessment body for all fields of electrotechnology. Developed

by Subcommittee 45A (SC45A) of the IEC, the IEC 62645 standard titled "Nuclear power plants - Instrumentation and control systems - Requirements for security programmes for computer-based systems" published in August 2014 addresses I&C systems in nuclear facilities and is the first IEC document targeting cybersecurity [45]. IEC 62859, titled "Nuclear power plants - Instrumentation and control systems - Requirements for coordinating safety and cybersecurity," published in the end of 2016, is the second document addressing safety and cybersecurity [46]. It aims to provide an actionable framework to coordinate safety and cybersecurity for the design, implementation and operation of I&C systems, which integrates cybersecurity into safety architecture and system. Comparing between the IEC and IAEA suite, the content contained in IEC 62645 and IEC 65859 are also included in IAEA guidance; the new IEC 63096 focusing specifically on cybersecurity controls selection and application from the included security controls catalogue, in order to prevent, detect and react to cyber-attacks targeting I&C systems, is under preparation and targeting to be issued in 2019 [5]; cybersecurity controls are not fully covered by IAEA standards currently. The application process and interaction of these three SC45A series standards are summarized in Figure 2.1 provided by reference [5].

2.2.3 NRC guidance

The U.S. NRC is an independent agency of the United States government that develops policies and federal regulations for commercial nuclear power reactors and other uses of nuclear material through licensing, inspection, and enforcement of its requirements, protecting public and environment health and safety. Therefore, the NRC actively provides regulations to protect licensed NPPs against cyber-attacks. In March 2009, the NRC released Title 10 of the Code of Federal Regulations (10 CFR) 73.54, "Protection of Digital Computer and Communication Systems and Networks," which requires NPPs to ensure that digital computer and communication systems and networks related to safety, security, and emergency preparedness (SSEP) functions are protected against cyber-attacks, up to and including design basis threats [119]. Regulatory Guide 5.71 (RG 5.71) [120] released in 2010 provides guidance on how to comply with 10 CFR 73.54. It provides a cybersecurity plan template and set of security controls. In November, 2015, NRC released 10 CFR 73.77,

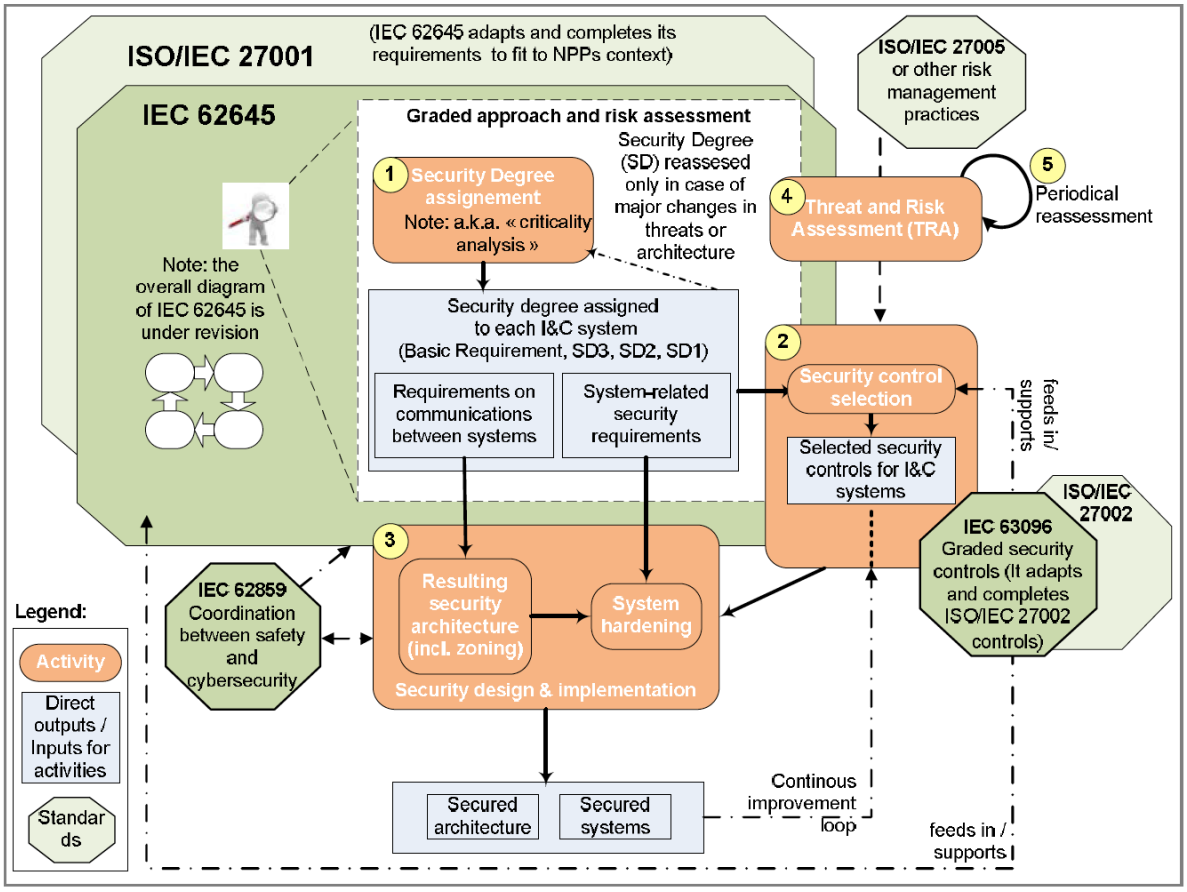


Figure 2.1: The application process and interaction of three SC45A series standards [5]

"cybersecurity Event Notifications" [121]. It requires NPPs to record and report cybersecurity events happening at NPPs, including notifying the NRC within one hour after discovery of a cyber-attack that impacts SSEP functions and related support systems and equipment.

2.2.4 NEI guidance

The nuclear industry works with the NRC to enhance cybersecurity of NPP. The NPPs in the U.S. follow NEI guidance to comply their cybersecurity plan with NRC cybersecurity requirements. NEI guidance include NEI 08-09, "cybersecurity Plan for Nuclear Power Reactors" [87], NEI 10-09, "Addressing cybersecurity Controls for Nuclear Power Reactors"[88], NEI 13-10, "cybersecurity Control Assessments" [89], and NEI 15-09, "cybersecurity Event Notifications" [90]. NEI 08-09, similar to part of RG 5.71, gives a cybersecurity plan template, including a defensive strategy with a defensive architecture and a catalog of security controls based on NIST SP 800-82 "Guide to Industrial Control System Security," and SP 800-53 "Recommended Security Controls for Federal Information Systems".

2.3 Machine Learning Algorithms and Data-driven Models Used in Cyber-attack Detection

IDSs detect unauthorized access to the system. There are three types of IDS: signature-based, anomaly-based, and hybrid. Signature-based IDSs are developed to detect known attacks using their documented behavior. This class of IDS is very effective for known attacks with low false alarm rates, but is not able to detect zero-day attacks since the IDS is not yet aware of this behavior. Anomaly-based IDSs, on the other hand, model the normal behavior using data mining techniques or machine learning algorithms and report deviations from normal behavior as an anomaly or potential attack. They are customized to the normal behavior of each system to detect attacks, including unknown attacks, making it difficult for attackers to learn the IDSs capabilities, further complicating attackers' ability to launch undetectable attacks. The very nature of this makes anomaly-based IDSs result in a high number of false positives [8]. The hybrid IDS is a combination of signature-based and

anomaly-based detection; this approach combines the accuracy of signature-based approaches for known attacks with the generalization capabilities of anomaly-based systems. Data-driven, hybrid IDSs are promising approaches to enhance ICS cybersecurity and defenders' situational awareness.

IDSs can also be categorized according to network-based or host-based approaches. A network-based IDS monitors network traffic, while a host-based IDS monitors a specific host's process activities.

Before introducing any detailed machine learning algorithms, the general procedure of building a data-driven model is given here:

1. Data pre-processing: This may include the removal of features with a large percentage of not-a-number (NaN) values and near-constant values; these features typically do not convey useful information.

2. Features selection: This usually contains three steps:

   (a) Remove some data features based on the understanding of the context of a certain data set.

   (b) Divide the data into training and testing data, sometimes divide data into training, testing and validation data.

   (c) Identify a subset of useful features based on training data, for example, reducing the dimension of the data set based on correlation analysis.

3. Model training: Train a model using the training data.

4. Model testing: Feed the test data into the trained model and get the predicted values.

5. Obtain residuals: Compare the predicted values with the real values of the test data and get the model performance.

6. Model selection: Select the best model for the application based on model performance and the application needs such as accuracy, interpretability, and computing cost. Root mean square error (RMSE) which is detailed in 3.25 is commonly utilized to assess the accuracy.

Table 2.1: Confusion matrix

	Actual class is an attack	Actual class is not an attack.
Predicted class is an attack	TP, correct detection	FP, incorrect detection
Predicted class is not an attack	FN, missed detection	TN, correct normal

7. Model performance characteristic: Characterize the model performance for validation data.

There are four types of detection results: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) as shown in Table 2.1 .

Different researchers uses different model performance assessment metrics, which are explained here first before go into the detailed research.

1. Accuracy = $(TP+TN)/(TP+TN+FP+FN)$. This assesses the classification rates, which measures the proportions of the correct results. The drawback of this assessment is significant when the classes are unbalanced. For example, in a cyber-attack detection data set, 98% of the data points are not attack and only 2% of the data points are attack. When all the data points are classified to be not an attack, this assessment still shows the model has 98% accuracy. Therefore, this assessment is not suitable for unbalanced datasets.

2. Precision= $TP/(TP+FP)$. This assesses the correct alarm rate, which is the ratio of data points correctly classified as an attack to all the data points classified as attacks.

3. Correct detection rate =$TP/(TP+FN)$. This assesses the detection rate, which is the ratio of data points correctly classified as an attack to all the data points that represent as attacks.

4. Negative predictive value (NPV) = $TN/(TN+FN)$. This assesses the ratio of data points that are correctly classified as not an attack to all the data points that represent not attacks.

5. True negative rate =$TN/(TN+FP)$. This assesses the ratio of data points that are correctly classified as not an attack to all the data points that represent not attacks.

6. False alarm rate (FAR)= $FP/(TN + FP)$. This assesses the ratio of data points that are incorrectly classified as an attack to all the data points classified as attacks.

2.3.1 Cyber-attack detection using cyber data

This section briefly explains several machine learning algorithms for intrusion detection using cyber data, including host system and network traffic data, followed by one or two example research.

Artificial Neural Networks

Artificial Neural Networks (ANNs) are composed of interconnected artificial neurons which give different weight to different inputs. The output of each layer will be the input of the next layer and the layers between the input and output layer of the model are hidden layers. ANNs were first proposed in 1958, and were further developed during the 1970s before becoming popular again in recent years with increasing commercial adoption.

In 1998, Cannady proposed a signature-based IDS using ANNs as a multi-category classifier [11]. The data set used was generated by a RealSecure network monitor, in which 3000 out of 10000 events were simulated cyber-attacks by the Internet Scanner and Satan programs. Nine features were selected after analysis: protocol identifier, source and destination port and address, ICMP type and code, raw data and its length. An ANN model was trained using the data label of normal and different attack category. The training and testing RMSE was 0.058 and 0.070, respectively. This supervised classification model using network data was an early milestone when the majority of the IDSs were rule-based.

Hodo, Bellekens, et al. trained a supervised multi-level ANN model using internet packet traces to thwart Distributed Denial of Service (DDoS) attacks against a simulated Internet of Things (IoT) Network [32]. The data set used in this experiment was traffic captured via a network tap based on a simulated 5-node (4 nodes as clients and 1 node as a server) IoT network avoiding modification of the live traffic. The attacker launched a DDoS attack against the server node by sending over 10 million packets. The model gives a result of

99.4% accuracy in classification and shows the effectiveness in detecting various DDoS/DoS attacks.

Chawla, Lee, et al. proposed an IDS based on Recurrent Neural Networks (RNNs) using sequences of system calls in host data [14]. The data set used to examine the effectiveness of the model is the Australian Defence Force Academy Linux Dataset(ADFA-LD), which consists of 833 normal training sequences, 746 attack sequences, and 4372 validation sequences. The results show 100% true detection rate and a FAR of 60%.

Bayesian network

A Bayesian network is a type of probabilistic graphical model that represents the variables and their conditional dependence using Bayesian inference [51]. Variables are presented as nodes in the network to show their relationships, which requires expert knowledge.

Kruegel and Mutz et al. proposed an event classification based on Bayesian networks to model the inter-model dependencies and aggregate additional data such as model confidence [64]. With the ability of adding additional information, this model has fewer false alarms. The MIT Lincoln Labs 1999 Defense Advanced Research Projects Agency (DARPA) data set [78] was used to evaluate the model. The results show that the approach achieved true positive rate of 100% and 0.2% false positive rate while the traditional threshold method yields about 0.4% false positive rate with the same true positive rate.

Decision trees

A decision tree is a tree-like structure which represents classifiers and branches. The internal node represents a test for an attribute; the branch represents the test result; the leaf node represents a classification decision [98]. The decision tree method will be detailed in Chapter 3.

Kruegel et al. replaced the signature detection engine of a well-known open-source tool Snort [12], with decision trees to provide faster detection [65]. Snort uses a single line description of each rule so that it detects misuse by comparing the behavior of its rules. The decision tree chose the most discriminating variable in the rule base to improve the speed. The data set used to evaluate the proposed approach in a data set with 10-day TCP dump

files in one 1999 DARPA data set. The results showed an average 40% performance gain over Snort.

K Nearest Neighbor (KNN)

KNN classification is a simple technique that classifies an object by a majority vote of its neighbors. The object is assigned to the most common class among its k nearest neighbors. The detail of KNN method will be given in Chapter 3.

Liao and Vemuri applied a KNN-based model to classify program behavior based on the frequencies of system calls for intrusion detection [71]. Twenty-four attacks within the two-week 1998 DARPA BSM audit data were utilized to examine the effectiveness of the model. The results show a 100% detection rate for the sixteen known attacks and 75% detection rate for eight novel attacks (six detected out of eight).

Ensemble learning

For a given problem, KNN and decision trees search the hypothesis space to determine a hypothesis that makes good predictions. However, identifying a good hypothesis may be nontrivial. In contrast, an ensemble method is able to combine the predictions from multiple machine learning algorithms together to determine a better hypothesis than the best one alone. Bagging and random forest are two ensemble methods that combine several decision trees which have been applied to intrusion detection in some research [8]. A brief introduction of Bagging and random forest method is given here and will be detailed in Chapter 3.

Bagging

Bagging reduces the variance of a decision tree by averaging votes when predicting a class. It creates m subsets with n samples per subset from the original data set. These n individual samples are generated from the original data set by uniformly sampling with replacement [6].

Gaikwad and Thool investigated the Bagging method with a partial decision tree (PART) as a base classifier for intrusion detection [24]. The pre-processing used genetic algorithm to choose 15 features out of 41 features from NSL-KDD data set to build the model. Compared

to Naïve Bayes, PART, C4.5, Bagging with Naïve Bayes, Bagging with C4.5, the model based on bagging and PART obtained highest accuracy of 99.7166% with true positive rate of 0.784 and false positive rate of 0.172 through cross validation.

Random forest

Random forest consists of many decision trees based on randomly picked subsets. Random forest yields predictions from all the subtrees, which have less correlation than bagging by constraining only one predictor out of a subset to be used in the split of a tree [7]. The prediction result is determined by majority voting or weighted voting. A single decision tree splits by selecting single variable while a random forest splits by selecting multiple variables at each split point. A major advantage of random forest is that if the number of trees increases, the variance of the model decreases with the same bias. However, compared to the decision tree, the disadvantage of the random forest is that it is hard to interpret.

Resende and Drummond published a literature survey of the research conducted using random forest-based IDS since 2000 [101]. In 2008, Zhang et al. [138] proposed a random forest based network IDS for misuse detection. The model was evaluated by the KDD 99 data set and the result show that the model outperforms most reported unsupervised anomaly detection approach when evaluating by the same data set, which reaches 65% detection rate with 1% false positive rate.

The IDS research using machine learnig algorithms mentioned above based on three relevant public data sets, DARPA, KDD99 [58] and NSL-KDD, are summarized in Table 2.2. These three public data sets were widely utilized in network-based IDSs development. DARPA is the base data with raw TCP/IP Dump files with 6.2 Gb training data and 3.67Gb testing data; KDD99 is the set that extracted and processed 41 features from DARPA for machine learning research with 4898431 Kb training data and 125973 Kb testing data; and NSL-KDD is a size reduced data set with duplicates removed and has 311029 Kb training data and 22544 Kb testing data. More details of these three data sets can be found in [75].

Table 2.2: IDSs using machine learning algorithms based on public data sets

Research	Algorithm	Data set	Results
Kruegel and Mutz et al. [64]	Bayesian network	DARPA	100% true positive and 0.2% false positive
Kruegel and Toth [65]	Decision tress	DARPA	40% performance gain over snort
Liao and Vemuri [71]	KNN	DARPA	100% detection rate
Gaikwad and Thool [24]	Bagging	NSL-KDD	78.4% true positive and 17.2% false positive
Zhang et al. [138]	Random forest	KDD99	65% detection rate with 1% flase positive rate.

2.3.2 Cyber-attack detection using process Data

As stated in the Introduction, using process data to detect cyber-attacks is a feasible and promising approach for cyber-attack detection in ICSs. However, there is limited research in the open literature regarding cybersecurity using process data.

Goh and Adepu et al. proposed an unsupervised learning approach based on Long Short Term Memory Recurrent Neural Network (LSTM-RNN) to detect cyber-attack based on process data [27]. Recurrent neural network were widely used in temporal sequence prediction. However, it has limitations when training on long-term temporal sequences due to the vanishing gradient problem. LSTM-RNN overcomes this shortcoming by adding memory to the network. In order to detect the small deviation of predicted values from the actual sensor data, the cumulative sum (CUSUM) method is then applied to the residuals obtained by LSTM-RNN. The upper and lower control limits are defined by the validation data. The data set used to investigate the effectiveness of this method was collected from a large scale Secure Water Treatment Testbed (SWaT), which is built by Singapore University of Technology and Design [26]. This testbed that consists of six stages to process raw water disputes for cybersecurity research. The SWaT data set was generated from the testbed under seven days of normal operation and four days with cyber-attack scenarios.

The data related to the first stage process was used in [27], which contains 10 cyber-attacks scenarios. These cyber-attacks were simulated by altering the sensor data input to the sub-control systems, including altering single points and multi points together. The

results show that 9 out of 10 attacks can be detected by the proposed approach, and can also identify a false sensor. There were three false positive alarms which were assumed to be results from previous attacks by the authors; however, these could also be a drawback of the approach. In addition, the approach is not applied to all of the process stages due to large amount of data. Therefore, an algorithm with lower computational cost is needed for faster cyber-attack detection.

Gawand, Bhattacharjee, and Roy developed a monitoring approach for detecting cyber-attacks based on process data using least square approximation and convex hull approach [25]. A simple four-tank system control model was developed using state equations to serve as a virtual testbed to verify and validate proposed model. This system contains the two pumps' speed as inputs and two water tank levels as output. Two attacks were implemented numerically in this system, and the detection results showed that that these two models were effective.

Kiss, Genge, and Haller proposed a clustering-based approach using Gaussian mixture model [60]. The Tennessee-Eastman chemical process data was used and three different cyber-attacks, including replay, Austin time junction, and DoS attack were simulated by modifying the data to examine the effectiveness of the model. The results showed that the model could detect the simulated attack effectively.

Nader and Honeine investigated two one-class classification algorithms, support vector data description (SVDD) and kernel principal component analysis (KPCA), for cyber-attack detection [82]. Real data from a SCADA gas pipeline tested and the water treatment plant was used as the training and testing data, in which the abnormal transient was regarded as the cyber-attack. The results showed that the KPCA has better detection performance than SVDD for those data.

Researchers in the nuclear engineering field have also proposed several approaches to detect cyber-attacks. Li and Huang emphasized the potential of cyber-attack detection based on process data [69]. In [69], the authors proposed using dynamic PCA to model the correlations of different variables under normal operational condition. A chi-squared detector was applied to the residuals of the predictive values and the actual values to detect deviation from normal operation. However, this method was only proposed without testing with data

or simulated attacks in this paper, which may be due to the difficulty of generating realistic data or getting access to existing data.

Eggers proposed two intrusion detection models based on principal component analysis (PCA) and independent component analysis (ICA) with static and moving window to detect simultaneous cyber and physical attacks in NPPs [22]. Eggers acquired real-time normal process data from a data historian of an NPP and modified it to simulate a hypothetical simultaneous attack scenario. In this simulated attack, the attacker created a small break loss of coolant accident (SBLOCA) physically with a simultaneous false data injection attack in which the attacker sent normal data to fool the safety system in order to prevent the reactor shut down.

After the data analysis, 29 signals collected under seven days of 100% power operation were selected to test attack detection, including the reactor coolant system (RCS) temperature, pressure, flow rate, pressurizer level, reactor building (RB) pressure, RB sump level, makeup tank water level, makeup flow, letdown flow, and radiation monitors [22]. Six hypothetical false data injection attack scenarios were simulated. In each scenario, the author modified a 15 minute interval of data of each signal to generate the hypothetical attack. The results show that the PCA and ICA with a moving window approach can detect the attacks successfully.

There is also some research in state prediction-based cyber-attack detection using process model. State-based IDSs define secure states and critical states for a process and detect a cyber-attack by comparing the current state with the critical state database. Svendsen and Wolthusen built a model to predict future states to detect physical anomalies by using physical process models together with techniques from feedback control theory [114]. Such methods require a detailed analysis and a detailed model of the process, which makes them hard to implement in real application for several reasons:

1. It may be hard to obtain a detailed model, or the model could be complex which requires high computational cost and system engineers.

2. An industrial process is a dynamic and sophisticated process. Therefore, the detailed model of the process usually varies with time. It is difficult to obtain all the transient

models. When a new operation state is introduced, the overall model needs to be updated accordingly.

3. This type of model lacks generality since it is based on a specific process model. Therefore, the model has low portability.

4. It may fail to detect zero-day attacks, when previously unknown critical states are involved in the attack.

Most of these cyber-attack detection papers are not evaluated by a realistic ICS-related cyber-attack, and none of these cyber-attack detection papers have combined cyber-data and process data together to provide better cyber-attack coverage.

2.3.3 Necessity of building an ICS testbed based on the state-of-the-art of the cyber-attack detection

Machine learning algorithms and data-driven models for cyber-attack detection using network data have been studied based on several famous public data sets, including DARPA 1998, DARPA 1999, and KDD99 [8]. Even though these data sets provide public access and an easy way to compare the effectiveness and efficiency of different models, these data sets have serious limitations [115]. In addition, not all the features included in these data sets can be obtained from industrial networks, and some techniques such as deep packet inspection (DPI) are not ideal for ICS applications as they may slow down the control network and lead to DoS. Therefore, the construction of ICS testbed with non-intrusive and easy-to-implement data collection methods is a worthwhile project to generate data as well as establish best practice for data collection systems.

The limited available research on using process data for detection may be due to the lack of data and difficulty of building a testbed. Therefore, it is valuable to build an ICS testbed with both physical and cyber components to study the behavior of an ICS by generating and analyzing adequate data. In addition, most research mainly focuses on either cyber-data or process data only, which may not be sufficient for a cyber-physical system like an ICS. A testbed which can provide the host system, network data, and process data is necessary

to study all of the characteristics an ICS can provide and can maximize the potential in broadening the cyber-attack detection scope.

In addition, to study and evaluate the performance and response of physical equipment or a subsystem, a hardware-in-the-loop testbed is a promising approach. There is currently no such HIL testbed for cybersecurity research in nuclear engineering domain described in the open literature. Therefore, a PLC integrated HIL was built to serve as a best practice and provide data for studying localized cybersecurity strategies for key components as well.

2.4 Overview of Cybersecurity Risk Assessment for ICSs

Although there is mature research for cybersecurity risk assessment in the IT world, due to the differences between IT and ICSs explained in Chapter 1, these methods are not directly applicable to ICSs. When it comes to risk assessment, a difference that needs to be given more attention is that ICSs are cyber-physical systems which involve physical processes, so a chain of consequences has to be taken into consideration. For example, in an NPP, an attacker may reduce the main coolant pump speed by a false injection attack on a controller which controls the pump speed. This reduces the flow rate of the coolant going to the reactor core and the steam generator, which in turn causes inadequate cooling of the reactor. This may cause an emergency shutdown of the reactor, or even a core melt accident. Thus, the physical process reliability and safety should be taken into consideration when assessing the risk of ICSs. These differences indicate that risk assessment in ICSs needs dedicated research. This section reviews cybersecurity risk assessment methods in ICSs first and then presents the limited research that has been done on cybersecurity risk assessment methods in the nuclear domain. Nuclear industry could refer to the risk assessment methods used in other industries while taking the cybersecurity guidance and references mentioned in Section 2.2 into consideration.

Risk assessment has three main steps: (1) identify what assets can malfunction; (2) assess the likelihood of the assets to malfunction; (3) evaluate the consequences if the assets malfunction [56]. Considerable research has been undertaken for cybersecurity risk assessment of ICSs. Poolsappasit et al. used Bayesian networks and attack models to

predict the risk of the potential future attacks [94]. Fanti et al. proposed a risk assessment method based on colored timed Petri net for highway networks [23]. A comprehensive model-based risk assessment method called Operationally Critical Threat Asset and Vulnerability Evaluation (OCTAVE) is proposed, which can assess the risk for critical assets and set an optimal security policy [1]. Attack tree [10] [53], petri net [29] [16], Markov chain [92], and Bayesian network [107] [108] [139] [93] are widely used for assessing cybersecurity risk. The interesting research which inspires the risk assessment module in the proposed cybersecurity solution platform is detailed here.

2.4.1 Attack tree based vulnerabilities assessment

An attack tree provides a structured means to assess the vulnerabilities that could lead to damage [104]. It promotes that defenders consider all reasonable attack paths so that it facilitates optimal countermeasures deployment. In addition, it provides analysis of multiple layers of the system as well as common attack path identification.

Byres et al. identifies five key factors for risk: (1) system architecture and conditions; (2) existing countermeasures; (3) difficulty of attack; (4) the probability of detection; and (5) attack cost [10]. The first two factors are evaluated by working with experts in the specific industry. The attack difficulty is identified as most important factor among the remaining three factors. Four levels indicated by “trivial, moderate, difficult, and unlikely" are used in [10]. This analysis helps to understand the path and also assesses indicator values for each top node, which is the attacker’s goal, by calculating using logical “AND" and “OR" gates. Therefore, all the attacker goals have to be identified in this method.

Reference [10] identified eleven goals for a modbus-based SCADA system and evaluates the severity of impact by giving different severity levels. The paper then gives analysis of tree example attacks for these goals and evaluates the difficulty, the impact severity and probability of detection, as well as underlying critical vulnerabilities, which is useful for choosing optimal countermeasure to prevent an avenue of attack. It also pointed out five underlying security issues for Modbus-based SCADA systems: (1) lack of confidentiality due to clear text transmission; (2) lack of integrity check; (3) no authentication; (4) simplistic framing; (5)lack of session structure.

2.4.2 Quantitative cyber risk reduction using compromise graph

In 2005, McQueen et al. proposed a novel compromise graph approach and defined time-to-compromise as a function of known vulnerabilities and attacker skill level [79]. Ten steps are involved in this methodology:

1. Specify system configuration: identify primary targets and perimeter devices.
2. Identify the portion of interest to defenders;
3. Identify and prioritize the primary target's security requirement;
4. Identify the vulnerabilities;
5. Using compromise type to categorize vulnerabilities;
6. Assess the time-to-compromise;
7. Establish compromise graphs and attack paths;
8. Find dominant attack paths;
9. Perform steps 3-8 for baseline and enhanced system; and
10. Estimate risk reduction.

A primary target is defined as any device that is able to trigger a physical event directly; a perimeter device is defined as any device that is part of the control system but can be directly reached without routing, forwarding, switching, or inspection. When assessing the risk of the cybersecurity, the probability of event occurrence are usually decomposed to the probability that the attacker launches the attack, the probability the attack is undetected, and the probability that this undetected attack is successfully conducted [73]. However, the authors pointed out that this decomposing position has assumption that if the attack is detected, there wouldn't be any consequence. This assumption is not quite reasonable for cybersecurity because even though the attack is detected, it may already have some consequence before mitigation measures could be taken. Therefore, the authors proposed a new decomposition formula which contains five conditional probabilities: the probability that

the system is within attacker's interest, the probability of the attacker conducting the cyber-attack given the system being within their interest, the probability of a perimeter breach given the system was under attack, the probability of attack success given a perimeter breach, and the probability of damage given a successful attack. This formula provides a guidance to reduce the risk by reducing each conditional probability. However, these probabilities are hard to estimate. The authors give a case study for a SCADA system by estimating the relative change in the probability of a perimeter breach and the probability of successful attack.

Identification of the system vulnerabilities involves referring to the public vulnerability libraries such as MITRE's Common Vulnerabilities and Exposures (CVE) and experts' knowledge of the specific system. Then these vulnerabilities are categorized into five edge types: reconnaissance, breach, penetration, escalation, and damage. Then the time to compromise is evaluated which depends on the attackers' skills and vulnerabilities. A formula is given to estimate the time to compromise [10]. Then compromise graphs are developed in which nodes are used to indicate the potential attack stages. The attack paths with minimum time-to-compromise required are chosen as dominant attack paths.

2.4.3 Bayesian network

Literature shows that Bayesian networks are widely used in cybersecurity risk assessment because a Bayesian network conveys lots of information [52][107][139]. Therefore, this subsection introduces Bayesian networks briefly as the basis of the research detailed later. A Bayesian network is a directed acyclic graph (DAG) which is composed with nodes as variables and directed arcs which represents the conditional dependencies [52]. The probability distribution of the nodes are usually represented by a conditional probability table (CPT). Usually in graphic method with large number of nodes, the joint probability is difficult to calculate due to the scale; Bayesian networks solve this problem by splitting the joint distribution to simple nodes multiplication. The joint probability distribution is calculated using the following equation:

$$\Pr[V] = \Pr[X_1, X_2, ..., X_n] = \prod_{X_i \in \mathcal{V}}^{n} \Pr[X_i | \pi(X_i)] \tag{2.1}$$

where X_i are set of variables and $\pi(X_i)$ is a the parent node of X_i.

Usually when Bayesian networks are used in risk assessment, different researchers define different kinds of nodes and then build the visual structure of Bayesian network by enumerating the possible attack scenarios considering the system vulnerabilities. Two risk assessment research using Bayesian network are detailed here.

Zhang et al. proposed a fuzzy probability Bayesian network (FPBN) approach to overcome the challenge of acquiring sufficient cyber-attack knowledge from historical data [139]. That paper has several original contributions and provided significant inspiration for the risk assessment model in Chapter 3.

The first step of the cybersecurity risk assessment in [139] is attack evidence collection. The attack evidences are collected by both IDSs based on network data and the anomaly evidences based on process data. Then the evidences are fed into the FPBN inference engine, together with the probability provided by experts from different domains. Linguistic probabilities provided by the expert team are convert to a fuzzy number, which is based on the membership function. Then a Bayesian network is built by analyzing the system; a multilevel Bayesian network is used to present attacks, system functions, hazardous incidents, and system assets. Two iteration equations are given to calculate the information parent nodes pass to child nodes and child nodes pass to parent nodes at the $(t+1)$th iteration. When the iteration completes, the risk is calculated as the product of the probability and asset losses, which could be equipment damage, staff safety, and environment pollution, and their corresponding probabilities.

Using process data as part of the attack evidence is a novel contribution in [139]. However, this could lead to bias when the an evidence caused by the the system fault is counted into cyber-attack evidence. Therefore, reference [139] proposed a noise evidence filter to filter out the anomalies caused by system faults to avoid incorrect risk assessment. A confidence index is proposed to describe the confidence degree of whether a process evidence is caused by cyber-attack or not. It is defined as the multiplication of all the false negative rates of

the IDS for specific attacks. That is, if the joint probabilities of the IDS missing a specific attack is smaller than a predefined threshold, then the evidence is labeled as a system fault. This means that if the possibility of the IDSs failing to catch a specific attack is very low, the anomaly in process data should be caused by system fault instead of the cyber-attack. A simulated experiment is conducted based on a stirred-tank reactor to demonstrate the proposed research and a multilevel Bayesian network is built. A simulated evidence event is utilized to demonstrate the effectiveness of the proposed model and a dynamic risk curve is given for the event in a timeline with a colored degree of confidence.

Reference [139] led to a novel suggestion to integrate the process data as well as the consideration of overestimating the risk due to system faults. However, the noise filter implemented in [139], which only depends on the IDSs' false negative rate, is not robust to distinguish cyber events and operational events.

2.4.4 Model-data integrated risk assessment using Bayesian network

Peng et al. proposed a model-data integrated risk assessment method that applies Bayesian network to perform quantitative assessment and optimizes the Bayesian network dynamically using data-driven parameter learning with online real-time data [93].

Three kinds of nodes are proposed: vulnerability nodes are used to indicate the vulnerabilities that the attacker may exploit, function nodes are used to indicate the function of the process which may be targeted by attackers, and accident nodes are used to indicate the possible accidents under attack. The risk is calculated as a product of the probability of the accident corresponding to the leaf node and the severity of the accident.

Tennessee Eastman, a process model [3], is used in this paper to build a simulated ICS in Matlab to assess the proposed approach. After analyzing the vulnerabilities of the simulated ICS based on CVE, a Bayesian network is constructed and the dynamic risk is then calculated and given in a risk-time curve.

2.4.5 Attacker vector analysis in cybersecurity assessment for digital I&C system in NPP

Song et al. [111] proposed a method to perform the cybersecurity assessment in accordance with the RG 5.71, which provides suggestions to the system designers in developing procedures for I&C systems. Five steps are suggested in the paper to perform the analysis of attack factors:

1. Analyze I&C system architecture at a conceptual level. An I&C system consists of safety systems and non-safety systems. The data transmission from non-safety systems to safety systems is not allowed. Although this provides protection to CDAs from being compromised by outside attackers, external digital devices or medias could be connected to these CDAs during NPP maintenance and tests. Therefore, maintenance should be taken into the cybersecurity assessment scope.

2. Model the targeted system. This procedure simplifies the functions, roles, and data communication to analyze. For example, redundancy should be simplified as a single train.

3. Assign security level according to RG 5.71. In this procedure, different security levels are assigned to difference device according the requirements and recommendations in the guidance.

4. Analyze the attack vectors. Four major attack entry points, including network, user, external device, and external media, are identified in this procedure. Attack vectors are initiated from the entry points.

5. Investigate known vulnerabilities and penetration test. CVE is helpful for known vulnerabilities. Penetration tests are conducted towards a testbed to confirm the analysis results and verify the consequences.

Reference [111] performs risk analysis for plant protection system (PPS), including penetration testing on a testbed. It identifies the three most critical technical security controls that should be deployed to the PPS: access control, monitoring and logging, and encryption. This approach could also be extended to non-safety systems since their attack vectors are similar.

2.4.6 Cybersecurity risk analysis and assessment system for digital I&C system in NPP

Lee et al. proposed a cybersecurity risk analysis and assessment system to analyze the security requirements with consideration of the life cycle of an I&C system in an NPP [68]. Four steps are identified to be integrated into the life cycle:

1. Identify CDAs. This includes target system selection, security modeling of the target system, and determining the security levels of CDAs.

2. Analyze the essential requirements of CDAs. In this procedure, common cybersecurity requirements should be considered by referring all types of the references issued by NRC, NEI, NIST, IAEA etc.

3. Analyze the vulnerability. Assess the vulnerabilities according to known vulnerabilities database, such as CVE.

4. Reporting. This includes risk assessment results, security requirements results, and recommended practices used in IT.

2.4.7 Cybersecurity risk model using Bayesian networks for I&C systems

Jinsoo Shin et al. [108] proposed an activity-quality model. Activity-quality model analyzes if the security requirements issued by different regularities are complied with. The numerical values “5, 4, 3, 2, 1" and weights “0.9, 0.7, 0.5, 0.3, and 0.1" are assigned to “very well, well, so so, bad, very bad" to describe the vulnerability. A checklist is derived by reviewing the guidance and a 34-item checklist is illustrated in the paper for Reactor Protection System (RPS). The checklist should be evaluated in each phase of the cybersecurity life cycle, including ‘establish program’, ‘integrate’, ‘continuous monitoring’, ‘review’, ‘change control’ and ‘record’.

Architecture analysis model is used to analyze cyber threats and mitigation measures. Numerical values “0.9, 0.7, 0.5, 0.3, and 0.1" are assigned to “Very Low Attack Occurrence Probability", “Low Attack Occurrence Probability 0.7", “Medium Attack Occurrence Probability 0.5", “High Attack Occurrence Probability 0.3", and “Very High Attack

Occurrence Probability 0.1". A Cyber Security Risk Index (CSRI) is defined as the probability of cyber-attack occurrence and a formula is given to calculate the CSRI value of each node. Through back propagation of the Bayesian network, when assuming that a cyber-attack occurred, the key elements corresponding to the attack can be identified.

2.4.8 Risk assessment model based on Bayesian networks and event tree

Shin et al. proposed a cybersecurity risk model with the combination of a Bayesian network model and an event tree model [107]. The paper provides a case study based on an RPS. A single channel in RPS consists of a bistable processor (BP), a coincidence processor (CP), a maintenance and test processor (MTP), and an interface and test processor (ITP). Five possible sequential countermeasures are identified as: (1) firewall; (2) online monitoring; (3) advanced online monitoring; (4) intrusion prevention system; and (5) IDS. Four threats are identified as virus, Trojan, worm, and DoS. A Bayesian network is constructed considering the 34-item checklist for RPS in [108]. A Venn diagram is used to represent the coverage of the mitigation measures, and then the probability that each mitigation measure works can be calculated. Then the weight of detection by a certain mitigation measure (since a threat can be detected by different mitigation measures) and failure of the mitigation measure has to be estimated according to experience and experiment. An event tree is used to represent the multiple-layer mitigation measures, which provides intuitive and quantitative information.

2.5 Localized Cybersecurity Research of PLC

Localized cybersecurity research enhances the cybersecurity of the controller device for key components. One such device, the PLC, has a microprocessor and can control actuators according to a set of pre-programmed rules. PLCs are widely utilized in industry since they can run for a very long time without major maintenance. However, PLCs are vulnerable to cyber-attacks due to two reasons: one is that most PLCs run with clear text protocols, so attackers can see the payload easily, and the other is that PLCs typically have very

low memory which constrains strong authentication and encryption techniques. These two characteristics are more common in the older version of the PLCs; for example, the Siemens S7-1200 is a widely used PLC in past 10 years that only has four megabytes of integrated memory, seventy-five kilobytes of which is available for the user to implement a customized program.

Several research studies have shown that PLCs are vulnerable to cyber-attacks. Sandaruwan et al. showed that an attacker can capture a single authenticate packet and reverse engineer it to find out the password for a PLC [102]. In [112], a PLC worm is demonstrated which can spread internally, does not require any standard PCs, and is not noticed by antivirus products. The paper also provides some potential detection methods since the initial infection causes PLCs to stop for ten seconds and also caused some abnormal network traffic behavior. Another method that an attacker could use to inject malicious code is demonstrated in [61] which would not cause the PLC stop, thus being even harder to detect. In [61], a prototypical port scanner and a socket secure proxy is developed which runs in a PLC to make it a backdoor of the ICS to scan the network and report back to the attackers.

To better protect PLC from cyber-attacks, companies which manufacture PLCs have added more security features to their protocols, firmware, and hardware over the years, such as the encrypted S7Comm-Plus protocol and integrity checks. However, a study presented in Blackhat 2019 demonstrated even the latest version S7 protocol is vulnerable [4]. Researchers first demonstrated starting and stopping of the latest S7-1500 PLCs remotely, then demonstrated download control logic on a remote PLC, and finally showed a stealth program injection attack, which allows attackers to modify the PLC control logic while displaying the original source code to the engineering station. In this situation, the PLC is running with different logic from the control logic displayed to operators [4].

Several studies related to enhancing the cybersecurity of PLCs have been conducted. Snapshotter, for example, was proposed to detect and prevent attackers from running malicious code on the PLCs [54]. The researchers used a forward secure logging mechanism to maintain data integrity and used a new encryption key for each log. The method was implemented on the OpenPLC framework as a proof of concept. However, this method requires an additional agent, which increases the attacker surface. Allison developed a C++

program called Goosewolf that can be implemented in the PLC Linux section and performs intrusion detection [2]. It does not interfere with the original PLC program. Currently, to the best of this author's knowledge, this is the only cybersecurity study to the date that has actually put a program into a PLC. This study shows the potential of implementing program which does not interrupt the original PLC function and does not increase the attack surface. Therefore, the localized cybersecurity strategy proposed in this dissertation is promisingly applicable, although only proof-of-concept research is conducted within the scope of this dissertation due to equipment limitations in the available testbeds.

Chapter 3

Methodology

To aid cybersecurity situational awareness and decision-making with the consideration from both the IT and OT sides, a novel solution platform framework is proposed for ICS, including I&C systems in NPP. This framework is introduced first in this chapter, followed by the supervised and unsupervised cyber-attack detection models. In the end, a localized cybersecurity strategy for key components is proposed to provide additional security for these low level devices. It is worthwhile to point out that cybersecurity not only takes intended cyber-atacks into account, but also unintentional cyber incidents that may result from human factors. Here, the term cyber event is used to include these two.[1]

3.1 Cybersecurity Solution Platform

Figure 3.1 shows the set up of the ICS tesbed on the left, the architecture of the cybersecurity solution platform on the right, and their connection with the data collection and extraction system in the middle. The cybersecurity platform consists of five major systems: a data collection and extraction system, a cyber-attack detection system, a cause analysis system, a cyber-attack response system, and a main control room display system. The data collection and extraction system pulls data from I&C systems and extracts the

[1]Part of this chapter is a reformat and reprint of partial materials as they appear in [135], [136], [137], and [133]. The dissertation author was the primary investigator and lead author of these papers.

useful data features and then feeds them to the cyber-attack detection system. The cyber-attack detection system consists of supervised models and unsupervised models to detect the cyber-attack with wide range and high accuracy. Supervised models detect and classify known cyber-attacks with historic patterns, which is a dictionary-based detection method. Unsupervised models reconstruct the signals first with the model built with fault-free data, then obtain residuals between reconstructed signals and real measurement fed by the data collection and extraction system, and report any anomaly deviated from normal operation. These models are able to detect cyber-attacks not included in the pre-defined cyber-attack database. The supervised models provide high accuracy since the model has seen similar attack pattern before and unsupervised models broaden the attack detection range to unknown attacks. The cause analysis system analyzes the detection results of the cyber-attack detection system to locate the anomaly, distinguish the cyber-attack from normal component degradation and operational events, and update the risk dynamically as well. The cyber-attack response system provides response suggestions to operators to mitigate the cyber-attack consequences. The main control room display system provides operators with necessary information to help with decision-making. The data collection and extraction system and cyber-attack detection system have been fully developed and detailed in this dissertation, along with the framework and the proof-of-concept research of the cause analysis system. The development of a cyber-attack response system and main control room display requires expert experience from different domains, such as plant cybersecurity experts and human factor experts, which fall out of the scope of this dissertation.

3.1.1 Data collection and extraction system

The proposed cybersecurity solution platform uses four types of data collected from the ICS, including host system data, network traffic data, network flow data, and process data, to perform analysis and defend the ICS. In traditional IT cybersecurity, there are host-based and network-based IDSs. A host-based IDS monitors a specific host's process activities while a network-based IDS monitors network traffic data [125]. These two data sources are commonly used since they carry most of the relevant information in IT infrastructure. These sources provide lots of information but with a huge number of features, which is time- and

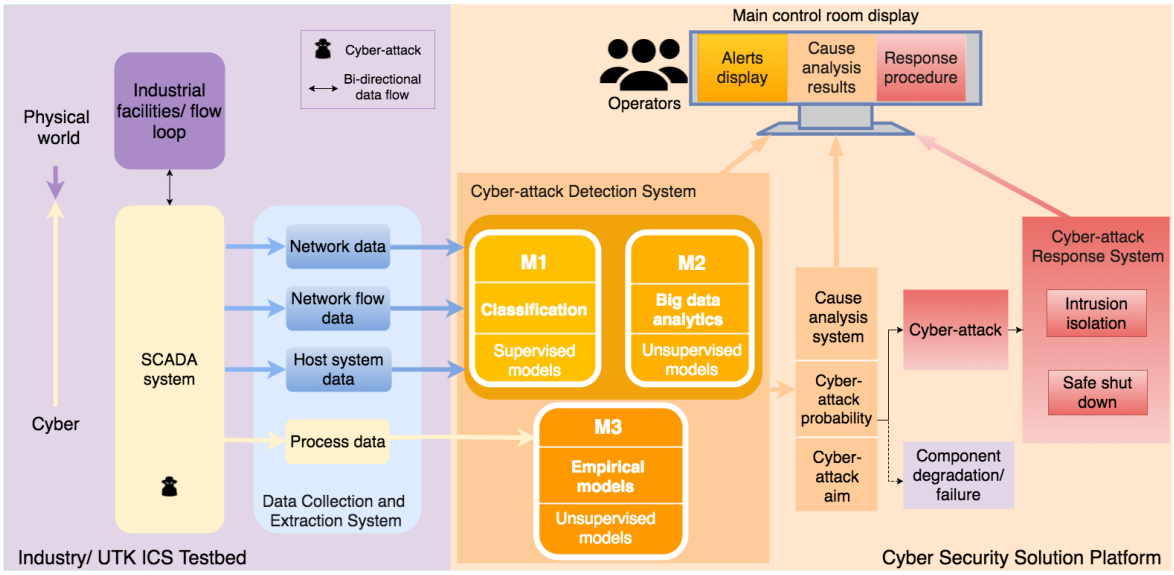


Figure 3.1: Architecture and connection of ICS testbed and proposed cybersecurity solution platform [135]

resource-consuming to parse and down-select to only the useful ones. In order to provide faster detection, data extraction time is also crucial. Network flow data summarizes the statistics of a continuous communication of two entities with lumped information, such as start time, end time, load, and transfer rate, which reduces the great quantity of data to a well-organized set of relevant characteristics. Therefore, network flow data is another data source selected for the platform to provide relatively small data space. Audit Record Generation and Utilization System (Argus) is an open source software that decodes the packets and summarizes them into network flow data [97]. The details of Argus are given in Chapter 4.These three sources, host system data, network traffic data, and network flow data, are the cyber data collected.

Since most, if not all, ICS-targeted cyber-attacks aim to cause disruption or damage in the physical system, ICSs provide another data source, called process data, to indicate anomalies. Process data includes the sensor data from the field, such as temperature, flow rate, and pressure, and the control commands that control systems send to field actuators, such as valve position. Process data are integrated to the platform as the fourth data source due to several valuable advantages:

1. It serves as an independent data source for cyber-attack detection, which is valuable when cyber data fails to detect cyber-attacks that may not leave a cyber-trace, such as false data injection launched by insiders.

2. Even when the intrusions are detected through cyber data and engineers take action to evict the intruder or eliminate the virus, the impact could already happen in the physical facility without operator awareness. For example, an intruder may have successfully altered a set point of a controller before being expelled. Such a case is demonstrated in a simulated attack scenario in Chapter 4. Under this scenario, process data is able to indicate the physical impact after the elimination of the cyber-attack action, which prevents further physical consequences.

3. The early detection of the physical anomaly gains defenders precious time to take countermeasures. Figure 3.2 shows operational performance after a cyber-attack [69], which indicates that the operational performance drop in the beginning of the attack is much less significant than later. This shows that early detection based on all available data before system performance dramatically drops is crucial to mitigate consequences.

4. The alarming signal in process data has the potential to indicate the anomaly location. For example, if the flow rate measurement shows an anomaly, the pump or the valve which corresponds to the flow rate should be checked first for effected operation.

In summary, the data collection and extraction system pulls out four types of data and extracts the useful features: the host system data, network data, network flow data, and process data, and feeds them into in the cyber-attack detection system to provide better detection coverage.

3.1.2 Cyber-attack detection system

The cyber-attack detection system contains three modules, which were developed by applying machine learning algorithms. Figure 3.3 demonstrates the defense-in-depth concept of this system. Taking a nuclear power plant as example, the first defense layer is the traditional intrusion prevention layer, including firewalls, data diodes, and gateways, which

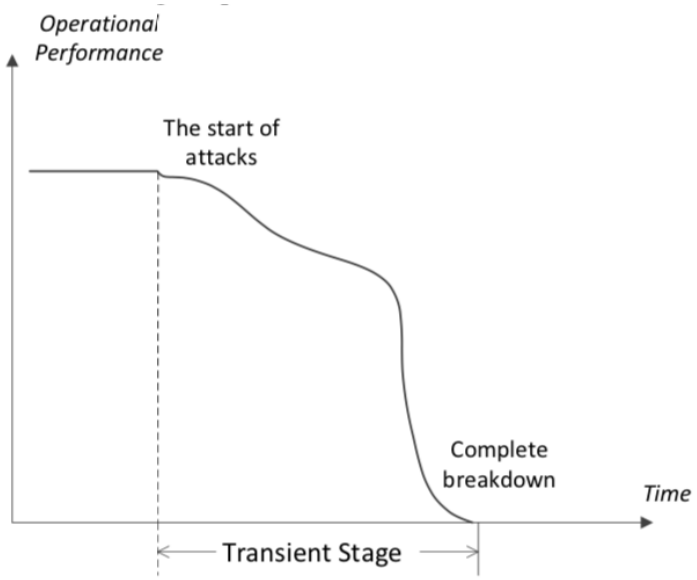


Figure 3.2: The transient stage after a cyber-attack [69].

are already widely applied in the industry. However, there are situations that the attackers could bypass these defense lines. The second defense layer consists of data-driven models for cyber-attack detection based on network traffic and system data, including the classification model indicated by M1 and the big data analytics models indicated by M2. Module 1 (M1) is built on supervised classification models using selected network and host system data to detect known cyber-attacks with historic patterns, which is quite similar to current dictionary-based detection systems. The goal of this module is to provide a fast detection of cyber-attacks in the cyber-attack database; this approach is proven to be successful for known attacks but is not able to detect cyber-attacks with new signatures and unknown attacks. Module 2 (M2) applies unsupervised modeling and data analytics to the network and host system data to detect new exploits that are not included in the attack database in M1.

For advanced cyber-attacks that may not be detectable by monitoring network and host system data, such as tampering and false data injection attacks by an insider, monitoring process data and system operation may provide useful information to detect the effects of cyber-attacks. Therefore, Module 3 (M3) is proposed, which monitor process data and system operation and detect small deviations from normal operation using unsupervised

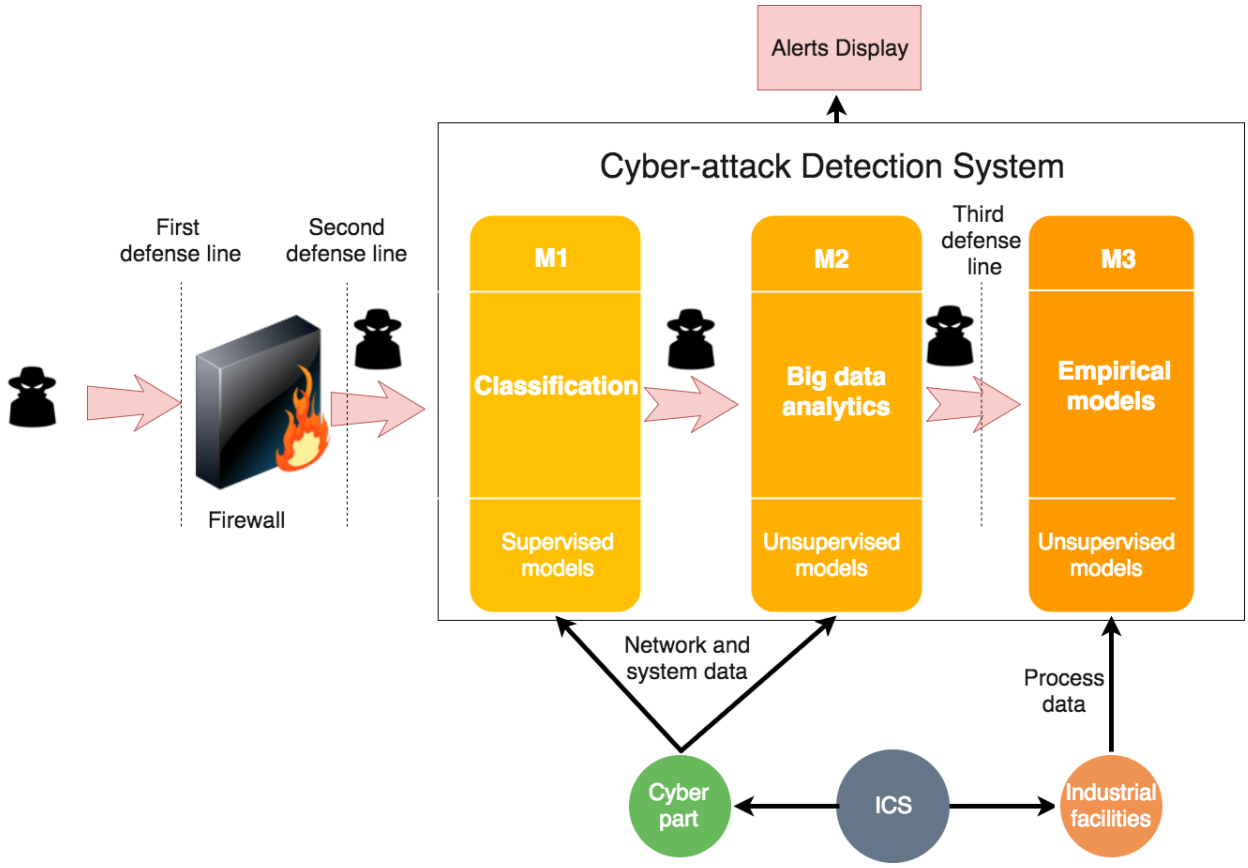


Figure 3.3: Cyber-attack detection system with defense-in-depth concept.

models. With the contextual information, M3 is able to detect commands and measurement tampering attacks and false data injection.

Therefore, in this system, M1 and M2 provide early detection of attackers when the attacks cause abnormal activities and indications in the cyber data, such as some man-in-the-middle attack (MITM) and DoS attacks. If the secondary layer fails to detect malicious activities, the last defense line monitors process data and uses unsupervised models in M3 to detect deviation from normal operation, potentially due to a cyber-attack. This multi-layer detection system improves the robustness of overall attack detection and is sensitive to both known and zero-day exploits. The detailed model algorithms will be given in Sections 3.2 and 3.3.

3.1.3 Cause Analysis System

The cause analysis system incorporates several subsystems: an event discriminator to separate cyber events from operational events, a dynamic risk assessment module to update the risk and potential consequences, and an attack isolation system to determine the location and security level of compromised components.

Event discriminator

Although M3 provides advantageous information for attack detection, the process data deviations detected by M3 can result from operational events, such as component degradation or failure. Since the risk assessment and response are different for cyber event and operational events, distinguishing between them plays a crucial role in cybersecurity decision-making. An event discriminator is proposed to analyze the anomalous indicators to evaluate the probability that the event is a result of cyber-attack.

Figure 3.4 shows the inference logic of the event discriminator. Once an anomaly is detected by M3, the discriminator checks if there is abnormal behavior in the cyber data first. Since abnormal behavior in the cyber data may not occur concurrently with the process anomaly during an attack, a time interval T_{max} is proposed to indicate the maximum time lag between an anomaly in the cyber data and a resultant effect in the physical process,

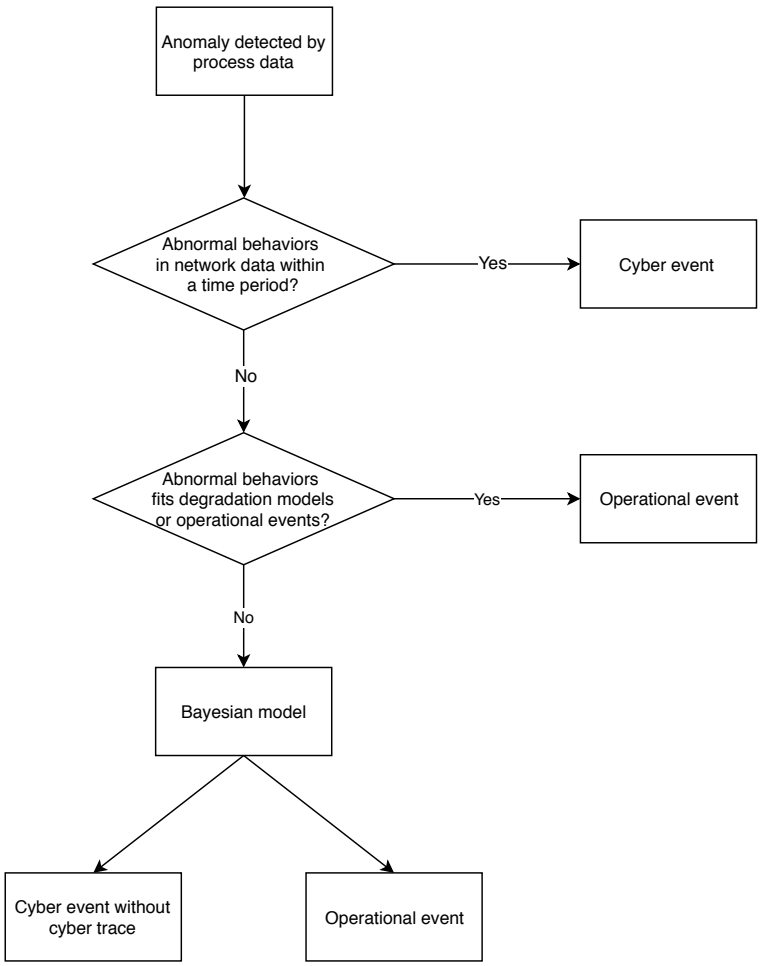


Figure 3.4: Logic of the event discriminator.

which could be specified by IT and OT experts. If there is an abnormal behavior in cyber data within T_{max}, the deviation in the process data is categorized as a cyber event. If there is no evidence seen in the cyber data, the anomaly in process data is examined by the equipment and system degradation database. If the process data and/or the residual between expected values and measurement follows a degradation trend in the database, the event is categorized as a normal equipment or process degradation. The degradation database is built based on historic data across the industry and is updated constantly with NPP operational information. The assumption is that if the data follows a degradation trend in the database then it is an operational event, because cyber-attacks are expected to cause faster damage than normal degradation.

The most difficult situation is when the anomaly is neither associated with abnormal behavior in cyber data nor follows a degradation model. Under this situation, a model based on Bayes' rule is utilized to give a decision. Letting A denote the abnormal behavior in process data, S denote that the event is an operational event, and C denote that the event is a cyber attack, Eq. 3.1 can be utilized to calculate the ratio α, which is the likelihood of an event being an operational event to a cyber event given that abnormal behavior in process data is detected:

$$\alpha = \frac{P(S|A)}{P(C|A)} = \frac{\frac{P(A|S)\cdot P(S)}{P(A)}}{\frac{P(A|C)\cdot P(C)}{P(A)}} = \frac{P(A|S)P(S)}{P(A|C)P(C)} \tag{3.1}$$

where $P(S|A)$ and $P(C|A)$ are the posterior probabilities of the system having an operational event or cyber event given that the an anomaly is detected in the process data, respectively; $P(S)$ and $P(C)$ are the prior probabilities that the system is having an operational event and a cyber event, respectively. $P(A|S)$ is the probability that an anomaly occurs given the system is having an operational event, which could be obtained by simulation with a simulator with heavier weights on the operational events documented in the Licensee Event Reports from NRC database [122]. $P(A|C)$ is inferred by experts using a Bayesian network which will be given later in the dynamic risk assessment section. If α is bigger than 1, then the event is diagnosed as an operational event while if α is less than 1, the event is diagnosed as a cyber event.

After the discriminator examines the potential attack detected by M3, the results of all cyber event detection modules are fed into the dynamic risk assessment system, the attack isolation system, and the attack response system.

Dynamic risk assessment

Even though there are security levels implemented in the NPP as required by NRC and some other IPSs deployment, vulnerability and attackers could be introduced to the higher security levels through maintenance. Cybersecurity guidance requires the nuclear facility perform vulnerability and risk assessment as discussed in Section 2.2. This means that a risk re-assessment should be performed after maintenance or a change of system architecture, components, and devices, which may be costly in terms of economics, labor, and time. In

contrast, a dynamic risk assessment tool could update the risk automatically and provide important information for decision-making. Risk assessment gives a numeric risk value that supports risk-informed decision-making. Considerable research has been undertaken for cybersecurity risk assessment of ICS as detailed in Chapter 2. Most cybersecurity risk assessment starts with vulnerability analysis for every digital device in the system. Common vulnerability libraries such as CVE are used for finding the entry points and the vulnerabilities. To assess the risk accurately, ideally IT experts should find all possible attack paths for a system. However, it is impossible to enumerate all the possibilities. Unknown cyber-attacks, such as zero-day attacks as well as unintentional cyber incidents will lead to underestimation of the risk if they are excluded. Therefore, unsupervised cyber-attack detection based on process data should be integrated to evaluate the cybersecurity risk to count the unknown cyber-attacks as well.

Usually in the risk assessment domain, the risk is calculated by the product probability and consequence. Unlike a physical device which can refer to a failure probability from product history, it is not realistic to find the probability of a cyber-attack occurring. A Bayesian network solves this problem by using conditional probability instead of the absolute probability. The conditional probability is the probability that component B will be compromised given malicious activity A is detected. The Bayesian network also provides a visual representation for complex dependencies and both quantitative and qualitative information, which helps indicate and analyze the causal relationship in the cyber event. Bayesian networks have already been adopted in risk assessment; Shin et al. studied using the Bayesian network to assess the cybersecurity risk for the reactor protection system [107]. Considering cybersecurity regulatory guidance, this study demonstrates the effectiveness of the Bayesian network for cyber-risk assessment. However, it only considers known attacks and is a static model which cannot update the risk according to normal plant configuration changes or attack movements. Zhang et al. proposed a FPBN approach to perform dynamic risk assessment [139], which addresses the issue of getting a reasonable CPT by using fuzzy probability. However, the research in [139] does not consider unknown cyber-attacks.

To address this need, a dynamic risk assessment framework is proposed for the cybersecurity solution platform as shown in Figure 3.5, which can update the risk

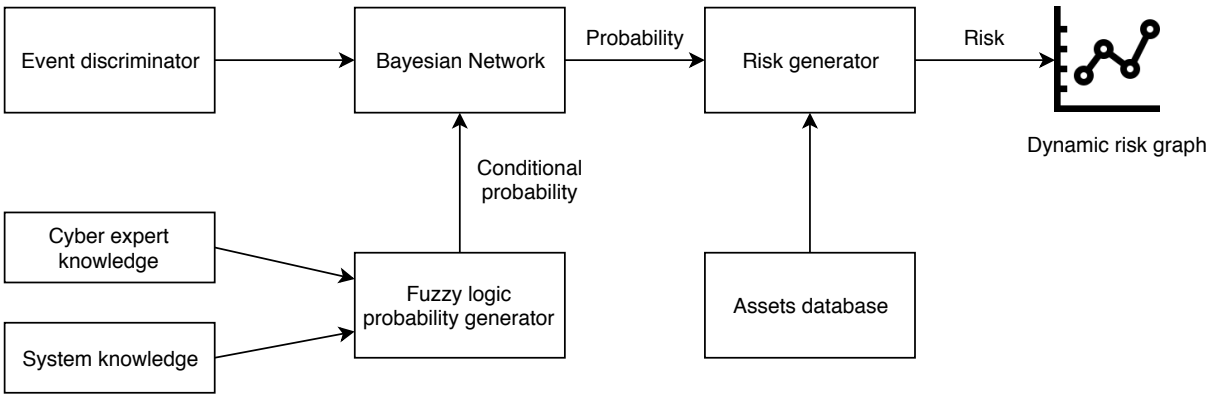


Figure 3.5: The architecture of the risk assessment module.

according to network architecture change, system change, new countermeasure deployment, maintenance, as well as the attack movements. It applies a multi-level Bayesian network, integrates the results from the event discriminator, and uses fuzzy probability extracted from expert knowledge. As shown in Figure 3.5, the event discriminator in the previous section provides the cyber event to the Bayesian Network, which excludes the system faults and covers unknown cyber-attacks as well. A CPT should be generated based on the knowledge from both cybersecurity and system experts beforehand. The knowledge could be obtained by querying experts on the probability of other components being compromised or failing to perform normal operations if the cyber-attack detection system indicates that an end point A is compromised. Only linguistic description of experts are required, such as if the failure of B based on A being compromised is impossible, unlikely, could be, very likely, and definitely. Then these fuzzy languages go through the fuzzy logic module, where they are transferred to crisp numbers and the weights can be given to different experts, to produce a conditional probability.

This expert knowledge can only be given according to the detailed network architecture and system configuration in a facility. For example, in an NPP, an indication of compromise (IoC) in Level 1 (in figure 1.3) indicates zero probability of CDAs failure which are in Level 4 since there is a uni-direction data gateway installed between them. Once an IoC is detected within Level 4, the probability of key component failure is given by the possible lateral movement path depending on the network architecture and system configuration. Because

this system relies on domain specific expert knowledge, only a framework is given in this dissertation.

The assets database contains four types of assets: data, equipment, human, and environment. The risk generator assesses the risk from these four aspects. For the data, including different types of data, configuration and design document, and maintenance document, it is the product of the probability that the data will be stolen and the value of the data in terms of how significant the consequence would be if the data is maliciously used. For equipment, it is the product of the probability of equipment and system degradation or damage and the financial cost to fix the equipment and system (including equipment purchase cost and labor cost). For human assets, it is the product of the probability of human injury and the injury level, where the injury level is recoverable, unrecoverable, and fatal. For the environment, it is the product of the probability of environmental destruction and the cost to fix the damage. Different weights are assigned to these four types of risk and a total risk is calculated. Since some risk is impossible to evaluate with money, a number without a unit is used here. Four types of risk results and the total score will be displayed together to the operators.

Attack isolation system

In addition to the event discriminator and dynamic risk analysis, the cause analysis system identifies the attack location, which is important for decision-making in response. The risk given by the dynamic risk assessment module is the input of the attack isolation system as well. The different risk score for each aspect and the total score can aid the decision-making. Based on this information, operators can analyze if a cyber-attack can be restricted to certain areas and mitigated by intrusion isolation without reactor shut down since unnecessary reactor shut down will cause significant financial costs. If the intrusion isolation is not enough to restrict the attack, safe shut down will be triggered to ensure the safety of NPP. For example, if the cause analysis system identifies a malicious access in the Level 1 network and has no indication of compromise of the control network, the risk will be isolated only to documents potentially being stolen, and the response will be that the IT team expels the malicious IP. If the attacker has already compromised Level 4 and

causes process data deviation in key parameters, such as the steam generator water level, then the risk will be equipment and system damage with high score, and human risk and environmental destruction with low score, and an emergency shut down will be necessary to prevent severe accident.

3.1.4 Cyber-attack response system and operator alert display system

In the event of a cyber-attack, a cyber-attack response system provides guidance to engineers and operators to minimize the consequences of the attack. Because the IT team and NPP operators are usually in separate departments, the signals in the alarm system in the main control room should be sent to IT team as well to promote stronger coordination between IT and OT teams in case of a cyber event. However, this research requires a collaboration between the cybersecurity, system, and human factors experts, which is not within the scope of this dissertation.

3.2 Supervised Cyber-Attack Models Based on Host-related Network and Host System Data

In an industrial system, such as an NPP, there are several important hosts, such as engineering workstation with SCADA master, which sends commands to the field based on sensor data acquired. Therefore, these hosts are primary cyber-attack targets for attackers to deploy further malicious activities such as changing operational parameters. In the developed ICS testbed, detailed in Chapter 4, host-related Network data and host system data is collected by Windows Performance Monitor [62]. All the features available in Windows Performance Monitor were collected to determine which features may be most useful for data-driven detection models. On average, there are 47,000 features collected; the exact number of features depends on the counters configured in the Windows Performance Monitor. The data pre-processing procedure includes the removal of features with a large percentage of NaN values and near-constant values; these features typically do not convey useful information.

Features related to software that were not directly relevant to the cyber operation and attacks in this research were also removed; these features do not statistically vary as a result of the postulated attack scenarios. 142 features related to memory, computer processes, and network behavior that contribute useful information for attack detection were selected for the final dataset. The dataset used for detection is collected under three cyber-attacks: MITM, DoS attack to the engineering workstation, and DoS attack to the National Instruments (NI) cDAQ9188. The observations under the attack period were labeled as 1, indicating they are under attack scenarios whereas the observations under normal condition were labeled as 0, indicating there is no cyber-attack. The data set was divided into training data and test data for training the classifier and for testing its performance, respectively [136].

KNN [131], decision tree [117], bagging [127], and random forest [72] are four classification methods investigated in this research. These four approaches were selected based on suggestions from a survey of data mining methods for cybersecurity intrusion detection [9].

3.2.1 KNN

KNN classification assigns a data point to the most common class among its K nearest neighbor points. Although KNN is a simple method, surprisingly it can often generate classifiers close to the optimal Bayes classifier [50].

In a classification problem, we would like to know the conditional distribution of Y (class) given X (input observation). Given a test data point x_0, the KNN classifier finds its K closest neighbor points first, denoted by N_0, then estimates the conditional probability of this observation being class j (denoted by Pr(Y=j|X=x_0)) by calculating the fraction that its K neighbors N_0 belongs to class j as show in Eq. (3.2):

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j). \tag{3.2}$$

In the end, the KNN algorithm classifies x_0 to the class with largest probability (Pr). The classification results largely depend on the choice of K. Therefore, cross-validation is conducted to choose the most optimal K in the specific application. Usually, the K which produces the smallest test error is chosen.

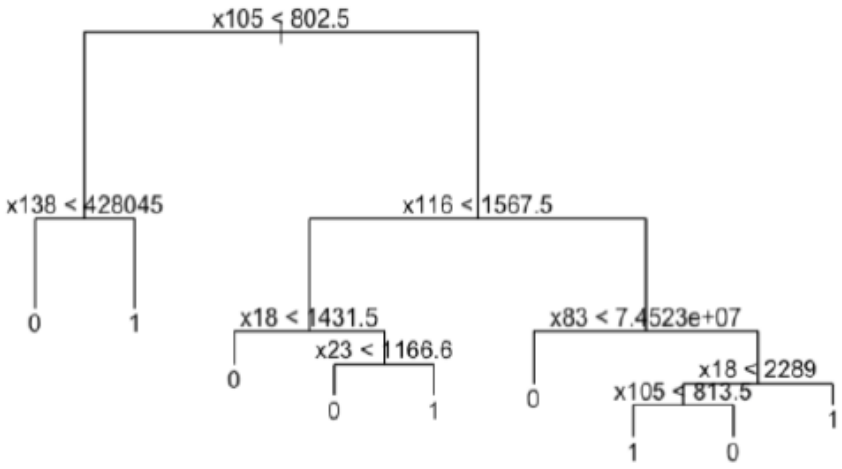


Figure 3.6: Decision tree classification structure for the network data.

3.2.2 Decision tree

Decision tree based methods are usually upside down tree-like structures which represent classifiers and branches. Figure 3.6 shows an example, which is the decision tree classification results of the network data detailed in Chapter 4.

The internal node, such as the 105^{th} variable x105, represents a test for an attribute; the branch, left and right, represents the test result of false and true, respectively; the leaf node, such as "0" and "1" in the bottom, represents a classification decision after considering all attributes. In this application, class "0" is the class that observations are not under a cyber-attack while "1" indicates the class that observations are under a cyber-attack.

Any predictor could divide the space to two regions and then another predictor could divide the subspace into sub-regions continuously, which could yield numerous classification methods. However, ideally we would like to always find the one which divides the space into minimized regions. Unfortunately, this is infeasible computationally; a decision tree handles this problem by recursive binary splitting, which begins at the top of the tree as one region. Then it splits further down by make the best split at each step, rather than looking forward to pick a split which will lead to a more optimal tree in future steps. This is done by finding a predictor X_j and a threshold s which minimizes the classification error rate which is defined as the fraction of the mis-classified training observations. That is for any X_j and s, the pair

of half-planes is define as:

$$R_1(j,s) = \{X|X_j < s\} \; and \; R_2(j,s) = \{X|X_j \geqslant s\}, \tag{3.3}$$

and then the algorithm searches the values j and s to minimize sum of the mis-classified fraction in both regions. To prevent overfitting to the training data, cross-validation is performed to select the optimal number of internal nodes, also known as tree size, which yields the lowest cross-validation error.

One main characteristic of a decision tree is that it splits at each variable by maximizing the information gain, which gives it advantages and disadvantages. The main advantage is that a decision tree has high classification accuracy and intuitive interpretation, which is also useful for variable ranking and selection. The major disadvantage is that when it comes to a data set with levels such as a categorical variable, to maximize the information gain, the decision tree prefers variables with more levels.

3.2.3 Bagging

Bagging reduces the variance of a decision tree to reduce the overfitting issue [50]. Given n independent observations with variance σ^2, the variance of the mean of these n observations is σ^2/n. Therefore, generating prediction models and results based on training sets from the population and then averaging them can reduce the variance and hence increase the prediction accuracy. However, this is not always feasible due to data size limitation. Bootstrap is a sampling method where a certain number of the samples are sampled from the whole observations with replacement. Bagging is a method that generates B different bootstrapped training subsets and builds decision trees for each of them and then takes a majority vote of the predictions to obtain the final decision [127]. That is, the final decision of the class is the major class among those B predictions. However, Bagging may yield highly correlated B subsets. Assuming there are several strong predictors in the data set, in the bagging method, most of the trees will select the strong predictors as the top nodes, which yield similar bagged trees. Therefore, the predictions from each of these trees are very

likely to be highly correlated. By taking average in bagging, the variance will not be reduced significantly due to this high correlation.

3.2.4 Random forest

Random forest consists of many decision trees based on randomly picked subsets and has less correlation than bagging [72]. In the bagging method, a number of decision trees are built based on bootstrapped training subsets. In contrast, when splitting a tree, random forest randomly chooses m predictors out of total p predictors to be the split predictor candidates. Then only one predictor out of m candidates can be used in the split in a tree. A new sample of m is selected at each split, which avoids considering a majority of the available predictors. The prediction result is determined by majority voting or weighted voting. Usually, $m=\sqrt{p}$. This yields much less correlated predictions, which is an advantage over bagging. Bagging and random forest usually outperform single decision tree in terms of prediction accuracy. However, they are difficult to interpret due to the combination of trees and the voting process. Therefore, decision tree is used to perform the feature selection as well.

3.2.5 Procedure of applying classification methods

Figure 3.7 shows the procedure of building models using these four different classification methods. First, the data were shuffled to eliminate the auto-correlation and then the data set was randomly divided into 50% training data and 50% test data. The four classification models are trained using training data. Then each model's hyper-parameters were optimized to obtain the lowest misclassification rate in the optimizer: the number of the neighbors K for KNN; the number of internal node for decision tree; the bootstrap subsets number B for bagging, and the m for random forest. After that, each model was evaluated using the test data by comparing the predicted class and actual class to obtain the accuracy of the model. The accuracy assessment considers true positive (correct detection), false positive (incorrect detection), true negative (correct normal), and false negative (missed detection) rates.

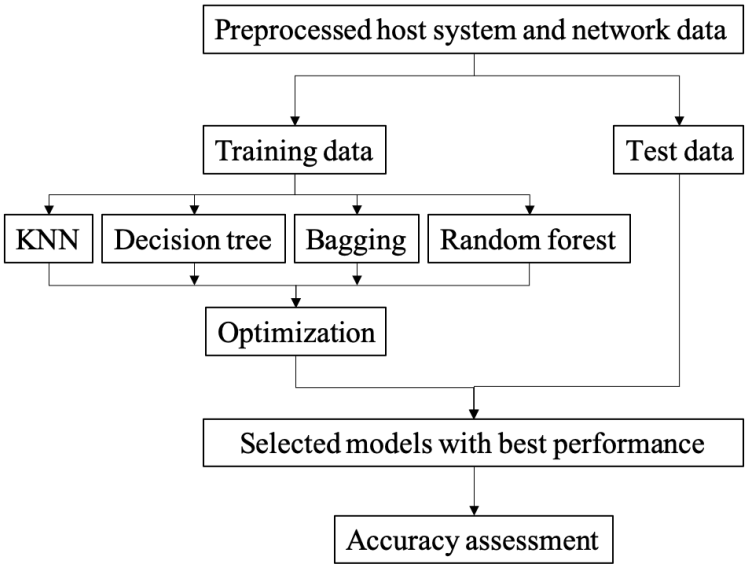


Figure 3.7: Procedure of applying the classification models.

3.3 Unsupervised Cyber-Attack Detection Models

To build a reasonable unsupervised model, a normal data set was divided into training, test, and validation sets for building a model, optimizing the model parameters, and characterizing model performance, respectively.

3.3.1 AAKR model

AAKR is a non-parametric, memory-based modeling algorithm that calculates the similarity of an input data point to a known memory matrix [30]. The memory matrix is usually selected from the training data to represent the training space when the training data set is large. The memory matrix X_m is shown in Eq. (3.4):

$$X_m = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \tag{3.4}$$

where p is the total number of features and n is the total number of observations. A query vector, defined as a vector $Q(1,p)$ as shown in Eq. (3.5), is the a new data point which needs to be predicted.

$$Q = \begin{bmatrix} q_1 & q_2 & \dots & q_p \end{bmatrix}, \tag{3.5}$$

where q and x are the corresponding elements in a query vector and a memory vector. The similarity between the query vector and the memory vector is used to predict an error-corrected estimate of the query vector. In this study, the Euclidean distance is utilized to calculate the distance between the query vector and each memory vector X_i. Let d_i be the Euclidean distance between a query vector Q and a memory vector X_i as shown in Eq. (3.6):

$$d_i = \sqrt{(x_{i1} - q_1)^2 + (x_{i2} - q_2)^2 + \cdots + (x_{ip} - q_p)^2}. \tag{3.6}$$

This distance calculation is performed for each memory vector from the memory matrix to obtain the distance matrix D as:

$$D = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_i \\ \vdots \\ d_n \end{bmatrix} \tag{3.7}$$

After the distance matrix D is obtained, weights w_i are calculated by converting distance through a Gaussian kernel function with bandwidth h as shown in Eq. (3.8):

$$w_i = \exp\Big(- \frac{d_i^2}{h^2} \Big). \tag{3.8}$$

The predicted value $\hat{x}_i$ of a query vector is calculated as a weighted average of the memory vectors, as shown in Eq. (3.9):

$$\hat{x}_i = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}. \tag{3.9}$$

An independent test data set is utilized in optimization to evaluate the hyper parameter: the kernel bandwidth h and the number of the memory vectors n. A brute-force search of the reasonable bandwidth values and the number of the memory vectors is conducted to find the best model which yields the lowest mean squared error (MSE) on the test data, as shown in Eq. (3.10):

$$MSE_p = \frac{1}{n_t} \sum_{i=1}^{n_t} [\hat{x}_{i,p} - x_{i,p}]^2, \tag{3.10}$$

where $\hat{x}_{i,p}$ is the i^{th} observation's predicted (or expected) value of the p^{th} feature, $x_{i,p}$ is the i^{th} observation measurement of the p^{th} feature, and n_t is the total number of observations of the test data set. AAKR depends on the relationship among the inputs variables to make a prediction, thus, well correlated inputs are required.

3.3.2 Support vector regression (SVR) model

Support vector regression (SVR) is based on support vector machine (SVM) theory. Linear regression finds the parameter estimations to minimize the error, which may be largely impacted from one or several outliers (or influential observations in some references). The outlier could drag the regression towards itself which yields a less general model. In contrast, SVMs for regression gives a less sensitive model.

The SVR of any variable can be expressed as shown in Eq. (3.11) [19]:

$$\hat{y}_i = f_i(x) = \phi(x_i) w_i + b_i = \sum_{j=1}^{l} y_j \alpha_j \phi(x_i) \cdot \phi(x_j) + b_i, \tag{3.11}$$

where the vector w is the weight, b is the bias, i is the index of the i^{th} observation, $\phi(x)$ is the support vector, j is the index of the support vector, l is the number of support vectors, y_i and $\hat{y}_i$ are the regression target and the predicted value of the regression, respectively, and

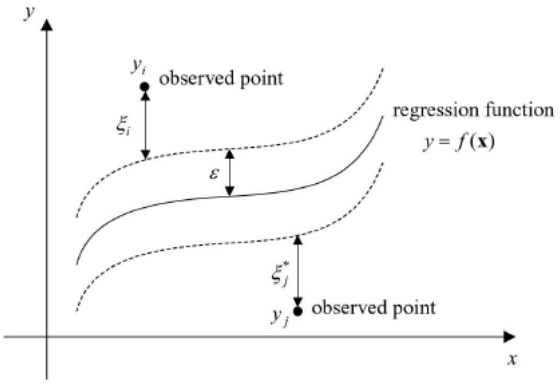


Figure 3.8: The parameters for SVR models [105].

α_i is the coefficient for the weight. The objective function of SVR is shown in Eq. (3.12) [20]:

$$U \sum_{i=1}^{n} L[y_i - \hat{y}_i] + ||w||^2, \tag{3.12}$$

where the first part and second part of the equation measure error and generality, respectively. U is a user-defined parameter to adjust the objective function. Large U makes the objective function put more emphasis on the error while small U puts more emphasis on the norm of the weights which yields a more general regression [20]. L is an ϵ-insensitive loss, which is defined as:

$$L = \begin{cases} 0, & |y_i - \hat{y}_i| < \epsilon \\ |y_i - \hat{y}_i| - \epsilon, & \text{otherwise.} \end{cases}, \tag{3.13}$$

where ϵ is a user-defined insensitive margin. Figure 3.8 shows the parameters for SVR, where ξ_i and ξ_j^* are the difference between observed points and the values on ϵ band. If the observed point is inside the 2ϵ band, ξ_i and ξ_j^* are zero which makes α_i in Eq. (3.11) zero. If the observed point is outside the ϵ bands, then ξ_i and ξ_j^* are nonzero and α_i is nonzero. Therefore, the observed points within the ϵ band have no impact on the regression equation $f_i(x)$. This means only a subset of the training data are utilized for prediction, which are called the support vectors since they support the regression function.

Therefore, minimizing Eq. (3.12) is equivalent to minimizing the following equation:

$$U(\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\xi_j^*) + \frac{1}{2}(w^t w) \tag{3.14}$$

for all $i = 1, 2, \ldots, n$. The constraints are shown as follows:

$$\begin{cases} y_i - (w^t\phi(x) + b) \leq \epsilon + \xi_i & \epsilon, \xi_i, \xi_i^* \geq 0 \\ w^t\phi(x) + b - y_i \leq \epsilon + \xi_i^* & \epsilon, \xi_i, \xi_i^* \geq 0 \end{cases}. \tag{3.15}$$

In Eq. (3.11), $y_i\phi(x)$ is a dot product of a new observation and support vector, which could be written as a more general equation as:

$$f(q) = b_i + \sum_{i=1}^{n}\alpha_i K(x_i, q), \tag{3.16}$$

where x_i is the element of inputs in the model, q is the new query observation, and $K(x_i, u)$ is called the kernel function. There are different types of kernel functions that can be used to generalize the regression with nonlinear relationships. Radial basis function (RBF) [66] as shown in Eq. (3.17) is a Gaussian kernel with a scaling parameter σ which is used in this paper.

$$K(x_i, q) = exp(-\sigma||x_i - q||^2) \tag{3.17}$$

3.3.3 Ensemble regression model

Ensemble regression models are the combination of several tree-based regression. The basic of the decision tree is described in Section 3.2.2. Therefore, only the regression part is described in this section.

Different from tree-based classification problem which minimizes the classification error rate, the tree-based regression minimizes the summed error between the predicted values and targets. The split with the largest reduction in error is chosen at each step. For j^{th} predictor X_j and a certain value s, the pair of half-planes is define as:

$$R_1(j,s) = \{X|X_j > s\} \text{ and } R_2(j,s) = \{X|X_j \leqslant s\}, \tag{3.18}$$

where $R_1(j,s)$ and $R_2(j,s)$ are the two different regions separated by half-planes. The algorithm searches the X_j and s to minimize the following objective function:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R2})^2, \tag{3.19}$$

where $\hat{y}_{R1}$ and $\hat{y}_{R2}$ are the mean response for the training observations in $R_1(j,s)$ and $R_2(j,s)$, respectively. Then the process continues in each separated regions until it reaches a stopping criterion. One of the criteria is that the total regression error is smaller than a certain value ϵ, shown as follows:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})^2 \leqslant \epsilon. \tag{3.20}$$

Then a linear model is created within each region using the split variables [66]. A top-down, greedy approach is performed in decision tree that the best split is made at each step rather than looking into future splits, which is subject to over-fitting the training model [50]. Similar to the ensemble tree classification methods described before (bagging and random forest), ensemble regression is also a combination of several tree-regression to yield better results.

3.3.4 Principal component regression (PCR) model

Principal component analysis (PCA) linearly transforms the input space into an orthogonal space. Through this transformation, a set of possibly correlated input variables are mapped into a set of linearly uncorrelated principal components (PCs), which reduces the dimensions of input space by remove linearity without much loss of information [129]. PCR is a regression where the new input spaces based on PCs are utilized as inputs of the linear regression instead of original input space [55]. The steps of PCR are given here:

1. Find a new variable t_1 which is a linear combination of the original input space X (with n variables) which maximum variance as shown in Eq. (3.21)

$$t_1 = Xp_1 = \sum_{j=1}^{n} x_j p_{j1} = x_1 p_{11} + x_2 p_{21} + ... + x_n p_{n1}, \tag{3.21}$$

where t_1 is called the scores of the first PC and p_1 is called the first loading vector.

2. Find another variable t_2 which is still a linear combination of the original input space X that has maximum variance and is orthogonal to t_1 as shown in Eq. (3.22)

$$t_2 = Xp_2 = \sum_{j=1}^{n} x_j p_{j2} = x_1 p_{12} + x_2 p_{22} + ... + x_n p_{n2}, \tag{3.22}$$

3. Continue finding t_k which has maximum variance of the information left and is orthogonal to t_{k-1}. Eventually m t_k can be found as shown in Eq. (3.23). The original input space X is converted to k uncorrelated variables and arranged in the order of maximum to minimum variance. This can be presented by matrix form as shown in Eq. (3.23)

$$\begin{bmatrix} t_{11} & t_{12} & \dots & t_{1k} \\ t_{21} & t_{22} & \dots & t_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \dots & t_{mk} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} * \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1k} \\ p_{21} & p_{22} & \dots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \dots & p_{nk} \end{bmatrix}, \tag{3.23}$$

4. Mathematically, PCs are calculated by finding the eigenvalues and eigenvectors of the covariance matrix of X. The eigenvalues are proportional to the amount of the variance and the corresponding eigenvectors are the PCs. Usually the useful information is contained by the first several PCs and other PCs are mostly noise. Therefore, we could choose the first several PCs that explain majority of the variance (for example, 99%) to reduce the dimension of the input.

5. Use the reduced space based on PCs as new input to perform the linear regression for each of the original input variable.

In order to give equal weights to all the variables, standardization of the input space is usually performed when the variables are not in the same unites or the variance of the variables are very different. The advantage of PCR is low computing cost for new test observation since the input space is reduced.

3.3.5 Auto-associative models

In a practical application of cyber-attack detection using process data, such as an NPP, there are a significant number of sensors have to be monitored together. These sensors are the input of the cyber-attack detection models. Generally, for a regression problem, feature or predictor selection is important because it could impact the accuracy of the regression results but also has a high computing cost. If feature selection is performed for each sensor monitored in an NPP, there is lots of feature selection needed, which could require tremendous engineering work and may be unrealistic. Therefore, all the regression models introduced above are applied in an auto-associative manner to reduce this computing and engineering cost. Auto-associative models are the models that the predicted variables are the input variables. As shown in equation Eq. (3.24), the input space X is fed into each regression model to predict each variable X_1, X_2, ..., X_n of the input space.

$$X = (X_1, X_2, \ldots, X_n) \longrightarrow \begin{cases} \text{Regression}_1 \rightarrow & \hat{X}_1 \\ \text{Regression}_2 \rightarrow & \hat{X}_2 \\ \quad \vdots & \quad \vdots \\ \text{Regression}_n \rightarrow & \hat{X}_n \end{cases}, \tag{3.24}$$

where X is the input space with n variables X_1, X_2, ..., X_n, $Regression_i$ presents the regression models, including SVR, ensemble regression, and PCR, and $\hat{X}_1$, $\hat{X}_2$, ..., $\hat{X}_n$ are the predicted values for X_1, X_2, ..., X_n.

This auto-associative manner allows the feature selection and the pre-regression process only need be conducted once. For example, for auto-associative PCR (AAPCR), producing

Table 3.1: Selected Argus network flow data features description

Feature name	Feature description
DstPkts	Destination to source packet count
Rate	Packets per second for both source and destination
Dstrate	Packets per second for destination

PCs based on input space with selected features only needs to be done once, and then the PCs are utilized for linear regression for each variable, resulting in n regression models.

3.3.6 AAKR cyber-attack detection using network flow data

AAKR was utilized to detect the cyber-attack based on network flow data. The collection of the network flow data will be detailed in Chapter 4. The feature selection for the cyber-attack detection model is described here first. Since the solution platform should be utilized together with commercial firewalls in real applications, the features related to IP addresses and port are not within the selected features. If the intended cyber-attack is related with noticeable abnormal IP addresses and ports, the mature firewalls and COTS IDSs could prevent or detect it. Three well-correlated features in Argus network flow data, which are called DstPkts, Rate, and Dstrate, were selected as the inputs. The correlation coefficients between DstPkts and Rate, DstPkts and Dstrate, and Rate and Dstrate are 0.896, 0.849 and 0.774, respectively. Table 3.1 summarizes the description of these three features.

Figure 3.9 shows the procedures of cyber-attack detection using AAKR based on network flow data. Selected network flow data is fed in to the trained AAKR model to produce the predicted values of themselves. Then the residuals are obtained by calculating the differences between the predicted values and the actual values. After trying different filters, a five-order one-dimensional median filter is applied to the residuals to reduce noise. "DstPkts" and "Rate" yield large residuals when under the attack scenarios. Therefore, these two features are used to generate the fault hypothesis, which are also the attack alarms. Threshold selection was engineered; that is, different thresholds were applied to the residuals of these two features based on the scale and then the threshold which gave the best alarm results was selected. Cross validation was conducted using different sets of data to make sure that the thresholds were generalized for MITM and DoS attacks.

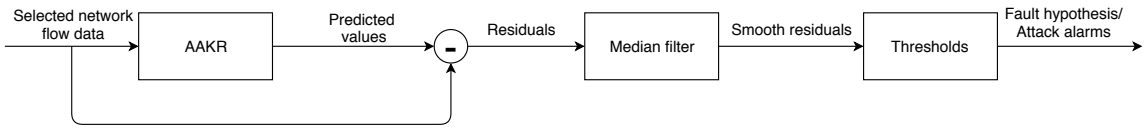


Figure 3.9: The procedure of cyber-attack detection model using network flow data.

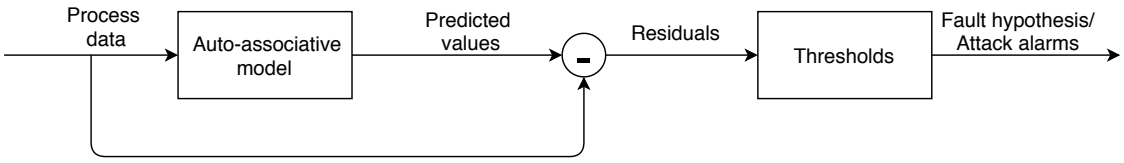


Figure 3.10: The procedure of cyber-attack detection model using process data.

3.3.7 Auto-associative Cyber-attack Detection Models Using Process Data

The collection of process data in the ICS testbed will be detailed in Chapter 4. As stated above, the auto-associative models only require one feature selection before training, which reduces computing cost and human resources. Therefore, cyber-attack detection models are used in an auto-associative manner. The feature selection of the process data will be detailed in Chapter 6. As will be demonstrated in chapter 4, the cyber-attack may cause deviations in several process variables, therefore applying a robust model to faulty input to yield correct prediction is necessary. Therefore, several auto-associative models, including AAKR, AASVR, auto-associative ensemble regression, and AAPCR were utilized to detect the cyber-attack based on process data and their performance will be compared and discussed.

Figure 3.10 shows the procedure of cyber-attack detection using process data based on different auto-associative models. Process data is fed into the model to produce the predicted values and generate the residuals. Different from the sparse network flow data discussed previously, the residuals in process data do not need a filter. Therefore, engineered thresholds are applied directly to the monitoring system residuals to generate the attack alarms.

3.3.8 Performance matrix

All the cyber-attack detection models should produce accurate, repeatable, and robust results so that they can be utilized in the real-world application. This section talks about the performance matrix for the unsupervised detection models which assesses the accuracy, sensitivity, first detection time, and false alarms of the models.

Accuracy measures the ability of the model to accurately predict the values under normal condition. Here the accuracy is characterized by the RMSE between the predicted values and real values of the fault free test data. The RMSE of j^{th} variable x_j could be calculated by Eq. (3.25)

$$RMSE_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}[\hat{x}_{i,j} - x_{i,j}]^2}, \tag{3.25}$$

where $\hat{x}_i$ is the i_{th} observation's predicted value of the j_{th} variable, $x_{i,j}$ is the i_{th} observation measurement of the j_{th} variable, and m is the total number of observations of the test data set.

Since all the unsupervised detection models developed in this research are in the auto-associative manner, they are potentially subject to high sensitivity. Sensitivity is generally defined as a measure of the change in the prediction of the j^{th} variable x_j produced by the change of the model input space. Sensitivity assessment includes auto-sensitivity and cross-sensitivity. Auto-sensitivity measures how well a model is able to make correct predictions of the faulty variable when it is included in the input of the model [31]. Basically, it measures how much the prediction of faulty j^{th} variable x_j will be impacted by itself. Cross-sensitivity measures how well a model is able to make correct predictions of a certain variable when one or more other faulty variables are included in the input space. That is, it measures how much the prediction of j^{th} variable x_j will be impacted by faulty k^{th} variable x_k or even more faulty variables [31]. Since usually a cyber-attack can impact one or more sensor measurements, at least one faulty variable is included in the input under the cyber-attack scenario, and only robust models with low sensitivity can produce correct expected values. Usually, to assess the sensitivity, an artificial fault could be added in the fault-free data set and its impact can show the sensitivity. In this research, since several sets of fault data (data

under attack scenarios) are available, sensitivity is assessed directly by the prediction of the model under the cyber-attack scenarios produced from the ICS test bed.

First detection time measures how fast a model can detect the attack after the attack occurred. In this research, the first detection time is assessed by how many attack observations are not alarmed before the first true alarm the model gives.

False alarm rate is important in the cyber-attack detection since a high false alarm rate could mislead the operators and cause panic. In this research, the false alarm rate is assessed by the fraction of false alarms in total normal observation under test.

3.4 Development of Localized Cybersecurity Models for Key Equipment

The unsupervised cyber-attack detection models using process data focuses on detecting the abnormal signals based on analyzing the relationship between the various signals measured by different types of sensors. Most of these models are usually built for the whole ICS or I&C system using data across the systems or a data historian. Thus, these models may be unsuccessful in two ways: a) a wide attack surface is created when the whole I&C system or a data historian are involved since more possibility of data compromise is introduced by long distance data transfer or multi-procedure data transfer; and b) centralized detection may be subject to availability to the real data. For example, in the Davis-Besse incident, the display system was inaccessible to operators for a long time; during this time, if the attacker had altered some control variables, operations could have become critically unsafe or unreliable. This suggests that key equipment should have localized resilience and have some capability to respond to the cyber-attack in order to maintain safety and operability in the scenario where operators lose access to data.

Therefore, a localized cybersecurity strategy is proposed to enhance the security of key equipment. This strategy deploys data-driven models into the control hardware of the key equipment to detect the cyber-attack and respond to the cyber-attack. This localized strategy has a smaller attack surface since data is transferred locally before the

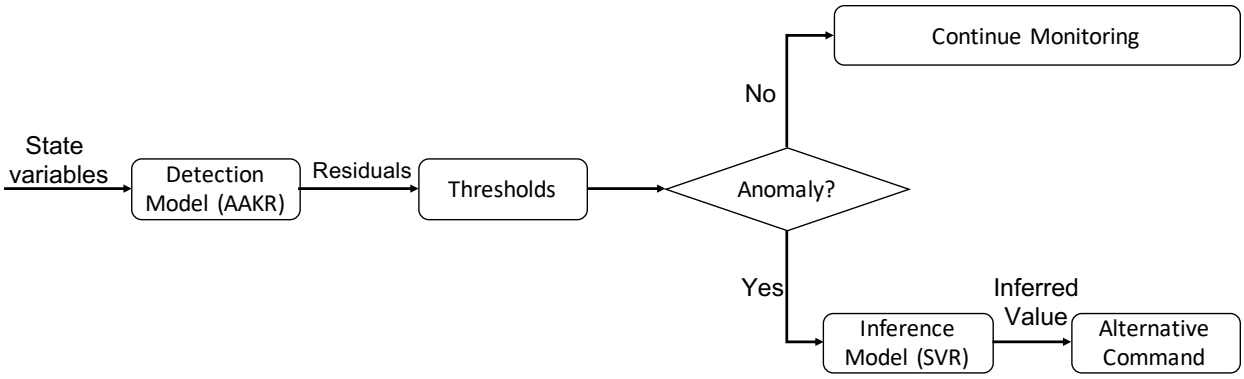


Figure 3.11: Logic block of the proposed localized cybersecurity strategy.

data acquisition stage of I&C system. A low computational cost model was developed to be implemented into controllers in the future, such as in PLCs to detect a potentially altered signal. Potentially, an inference model could be applied to correct the compromised control variable to add cyber-attack resistance to key equipment.

Figure 3.11 shows the logic block of the proposed localized cybersecurity strategy. AAKR is utilized to reconstruct the state variables and generate the residuals by obtaining the difference between the predicted value and real value of state variables at each data point (in each time step). Then pre-determined thresholds are applied to detect anomalies. When the residual exceeds the thresholds, a deviation is confirmed at the corresponding time step. When there are several deviations within a period, such as when 3 out of 5 continuous data points have deviations, then an anomaly is detected. If an anomaly is detected, which indicates false operation states of the component, an inference model is applied to predict the real state variable; this prediction can be utilized as a virtual sensor. Then an alternative command will be generated based on the virtual sensor. If no anomaly is detected, the localized cybersecurity model will continue monitoring.

There are two types of sensors in terms of functionality: indication sensors and control sensors. Indication sensors are sensors that only indicate the reading of the measurement. Control sensors are sensors that not only indicate the measurement but also are the input of controllers. That is, the reading of control sensors impact the command issued by controllers. Control sensors are more impactful in a process than indication sensors and could lead to

serious consequence if compromised. Thus, control sensors should be inferred to provide robust measurement under cyber-attack situations.

Chapter 4

ICS Cybersecurity Testbeds Description and Attack Scenarios Setup

Both the ICS testbed in this chapter and the HIL testbed in Chapter 5 are based on considerations in the nuclear industry since this research was conducted in the nuclear engineering department. A brief introduction of a typical PWR design is given here to provide some background.[1]

Figure 4.1 is a layout of a PWR given by the NRC [85]. A PWR is a type of nuclear reactor that includes a primary loop and secondary loop. The primary loop mainly consists of the reactor core where the fission reaction takes place and generates heat, the main coolant pump to force coolant water to circulate through the reactor core, the steam generator (SG), the SG primary side, and the pressurizer to maintain pressure in the closed loop. The Reactor Cooling System (RCS) in the primary loop, indicated by the red color in Figure 4.1, is the system that takes the heat from reactor core and transfers it to the secondary side through the SG without leak of radioactive materials. The secondary loop is indicated by the blue color in Figure 4.1, and mainly consists of the SG secondary side, turbines, condenser, and feedwater pump. The feedwater pump forces cold coolant from the condenser into the SG to be heated to steam by drawing heat from primary side. The steam produced in the SG goes

[1]Part of this chapter is a reformat and reprint of partial materials as they appear in [135], [136], [133], and [134]. The dissertation author was the primary investigator and lead author of these papers.

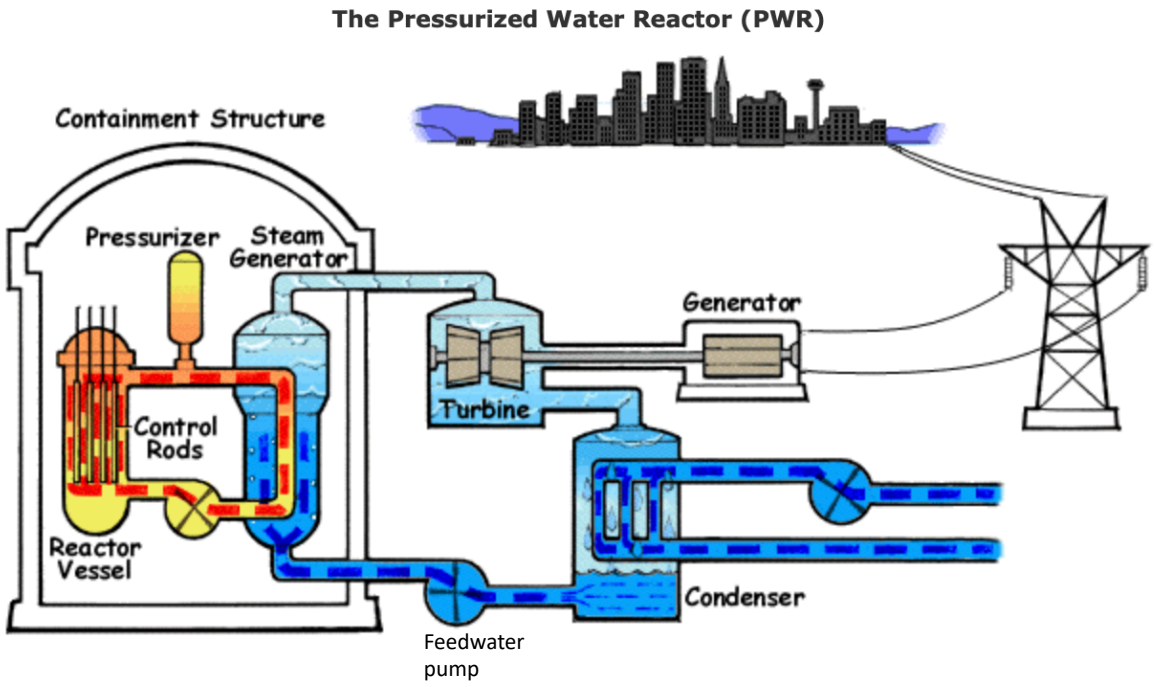


Figure 4.1: A typical PWR sytem [85]

to different levels of turbines to generate electricity. Then the exhausted steam is condensed into water in the condenser and pumped back to the SG by the feedwater pump.

This section presents the ICS testbed and the related experiments in detail. This includes a description of three major parts, the designed cyber-attacks, and the data collection.

4.1 ICS Cybersecurity Physical Testbed Description

For studying the ICS behavior and to generate the data needed for this research, an ICS testbed was built at UTK. This ICS testbed contains a two-loop system as the physical field experimental facility, as well as a LabVIEW-based SCADA system and a LAN based cyber part, which provides the capability to study industrial process behavior, carry out the postulated cyber-attacks and collect different types of data.

4.1.1 Physical process experimental facility

A physical facility was integrated into the testbed as the physical industrial process system, which receives commands from and sends data to the SCADA master. Figure 4.2 is a photo of the physical facility. As shown in the sketch in Figure 4.3, it emulates the thermal-hydraulic part of a typical two-loop nuclear system. The primary loop contains an adjustable 9 kW heater, a variable speed pump, a Coriolis flow meter, an upper water tank, a lower water tank, heat exchanger (HX) primary side, and 4 temperature sensors. The heater represents the reactor core, whose power is controlled by LabVIEW with open loop control and is able to vary between 0 and 100% of nominal power. The pump represents the main coolant pump and is controlled by a variable frequency drive (VFD), which is controlled by a LabVIEW proportional-integral (PI) controller. The upper and lower water tanks represent the reactor vessel upper and lower plenums. The HX represents the SG which transfers the heat from the primary loop to the secondary loop. Four resistance temperature detectors (RTDs) are installed in the core inlet and outlet and HX primary side inlet and outlet to monitor primary loop temperature. A Coriolis flow meter is located at the core inlet to measure the primary flow rate [134].

The secondary loop contains the HX secondary side, a motor operated valve (MOV), a bypass valve, a magnetic flow meter and two RTDs. The coolant in the secondary side is the forced water flow with inlet temperature of about 14.7^{o}C. The bypass valve provides the capability to simulate the HX degradation, which is not utilized in this research. Two RTDs are installed at the inlet and outlet of the HX secondary side; the magnetic flow meter is installed at the heat exchanger outlet, respectively. Two PI controllers are used for controlling flow rates in both the primary (by controlling VFD frequency) and secondary (by controlling MOV position) sides to maintain temperatures at the inlet and outlet of the heater at desired set points of 26^{o}C and 30^{o}C, respectively. This control logic mimics a reactor design with constant core inlet and outlet temperatures to reduce degradation of reactor fuel and core internals caused by changing thermal gradients across the core [134]. Table 4.1 lists the detail of major components installed in the ICS testbed, including function and model number, to provide documentation and reference of future testbeds construction.

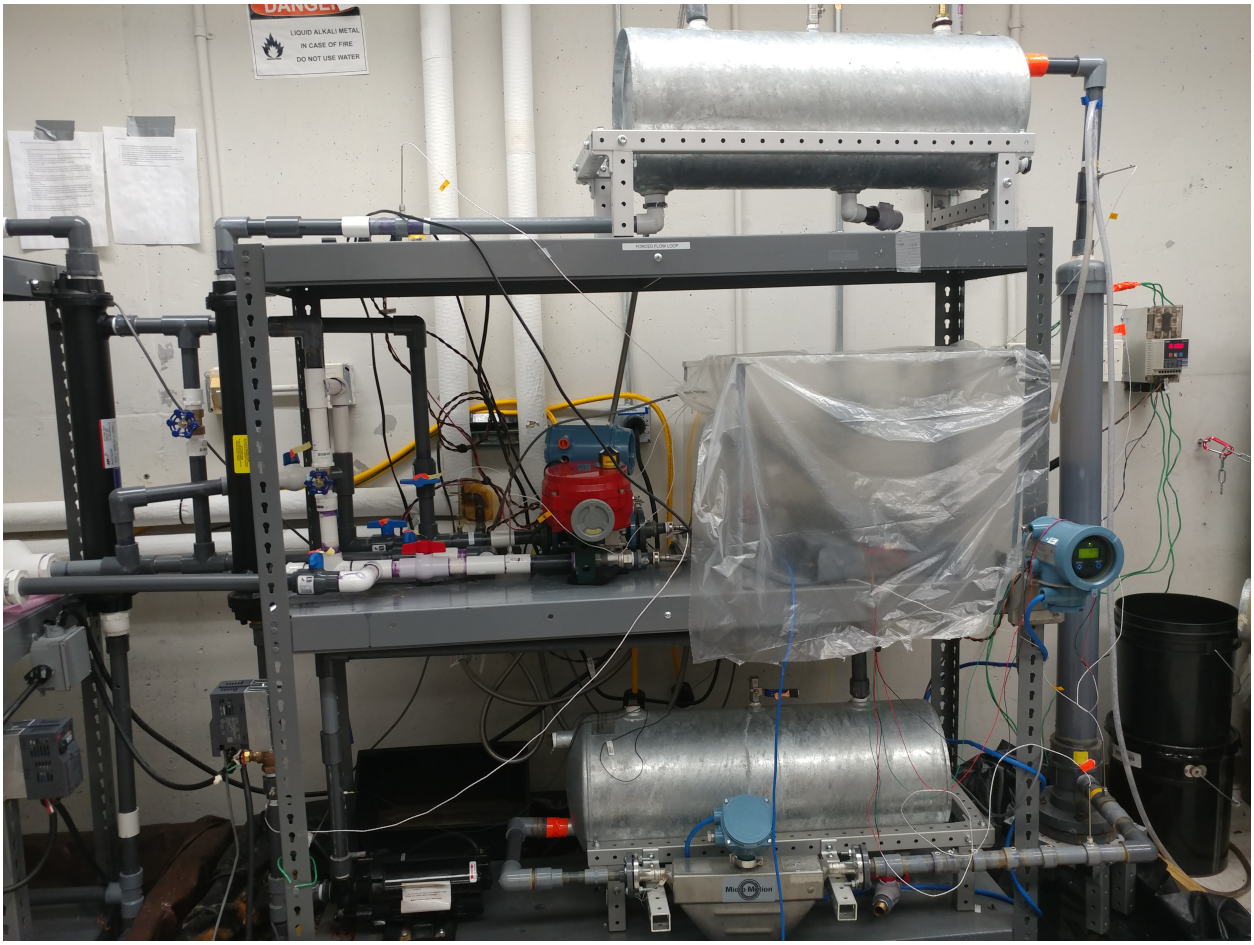


Figure 4.2: Photo of two-loop nuclear system experimental facility.

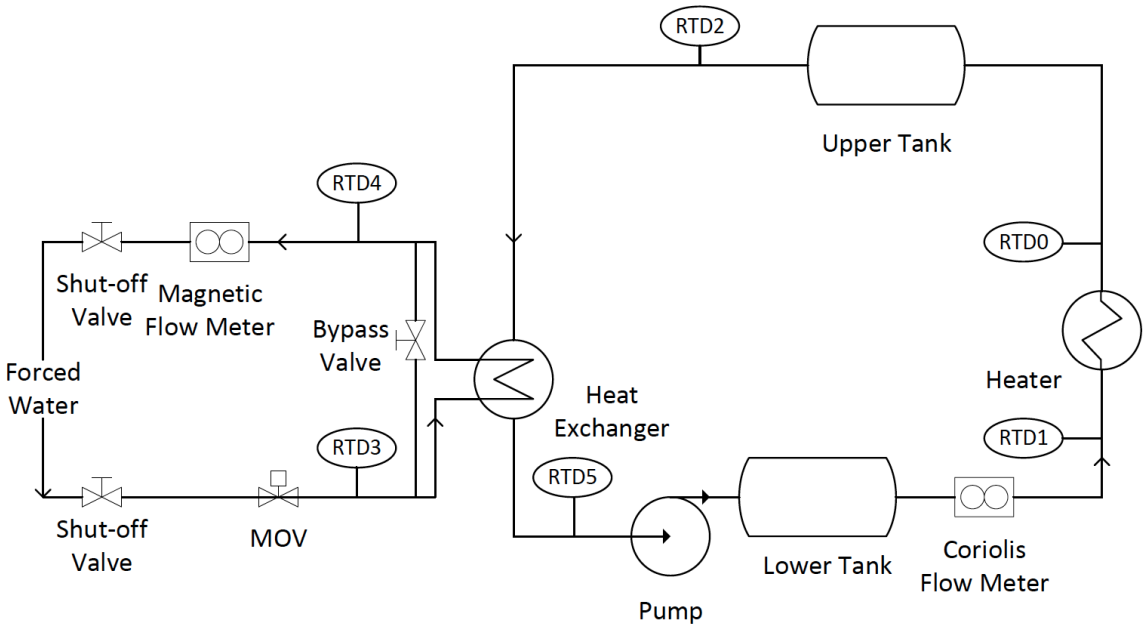


Figure 4.3: Two-loop nuclear system experiment layout.

Table 4.1: List of materials for ICS testbed

Equipment	Function	Model number
Heater	Controlled by LabVIEW to provide different heat power percentage.	SE Thermal EPS3- 40(H)
3-phase Pump	Serves as coolant pump in the primary loop.	DAYTON2ZWP2
VFD	Controlled and monitored in LabVIEW to send output signals (pump frequency) to control the pump speed.	MITSUBISHI FRD720-025-NA
MOV	Controlled by LabVIEW (On/Off) to change the flow rate of secondary side.	Bray AC Valve 70-Seris
RTD	Measure temperature in different location.	Omega PRTF-10-2-100-1/8-6-E
Coriolis flow meter	Measure the primary loop flow rate.	Emerson Micro Motion F050S
Magnetic flow meter	measure the secondary loop flow rate.	Emerson Rosemount 8732E

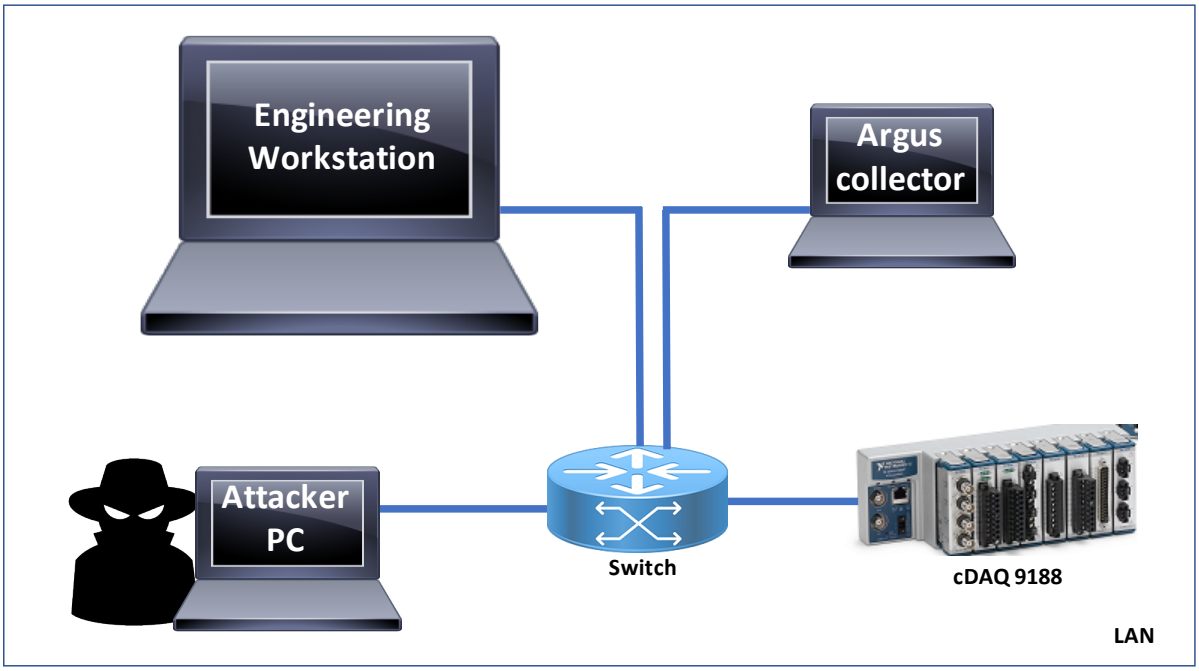


Figure 4.4: Testbed SCADA system and other cyber components.

4.1.2 SCADA system and cyber part

Figure 4.4 shows the cyber part of the testbed, which contains the SCADA system, Argus collector and an attacker computer. A switch with port mirroring function connects all the components into a LAN. The SCADA system is the system which sends control commands to the physical loop and records the sensor data from the physical loop as well. The SCADA system consists of NI data acquisition assistant device cDAQ9188, switch, and engineering workstation. The engineering workstation was configured as the SCADA master and a NI cDAQ9188 chassis with various data collection and control signal output modules were utilized as the SCADA slave [133]. The cyber part was built to achieve two functions: building a LAN for performing cyber-attacks towards the testbed, which prevents the cyber-attacks in this research from impacting the devices or network outside this LAN; and collecting different types of data for the proposed research, including host-related network data, host system data, network flow data, and process data.

The cDAQ9188 is the interface between the physical and cyber aspects of the system. It is a compact DAQ Ethernet chassis which controls the timing, synchronization, and data transfer between C Series I/O modules and an external host. LabVIEW is a system design

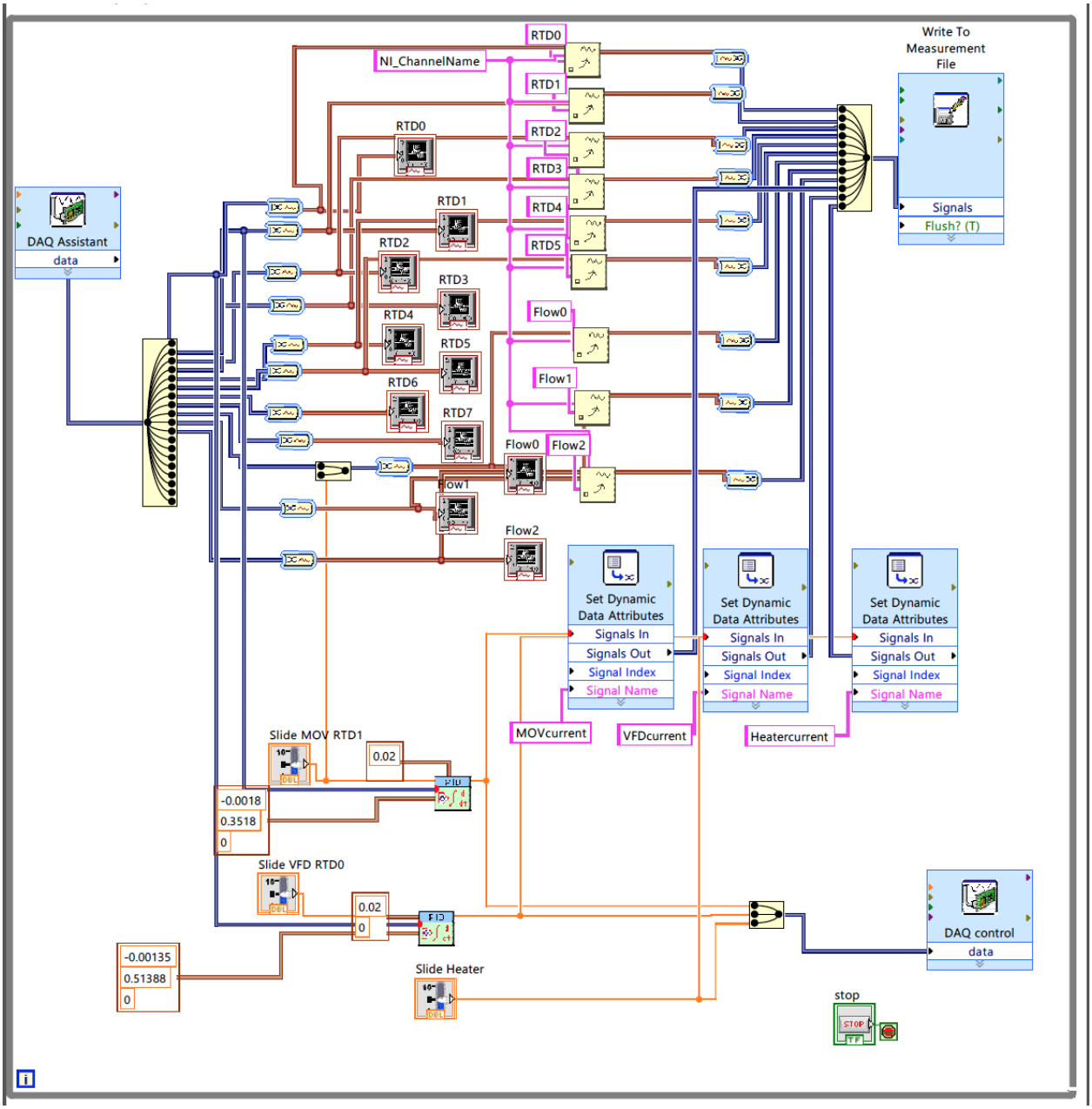


Figure 4.5: LabVIEW model of the physical two-loop experimental facility.

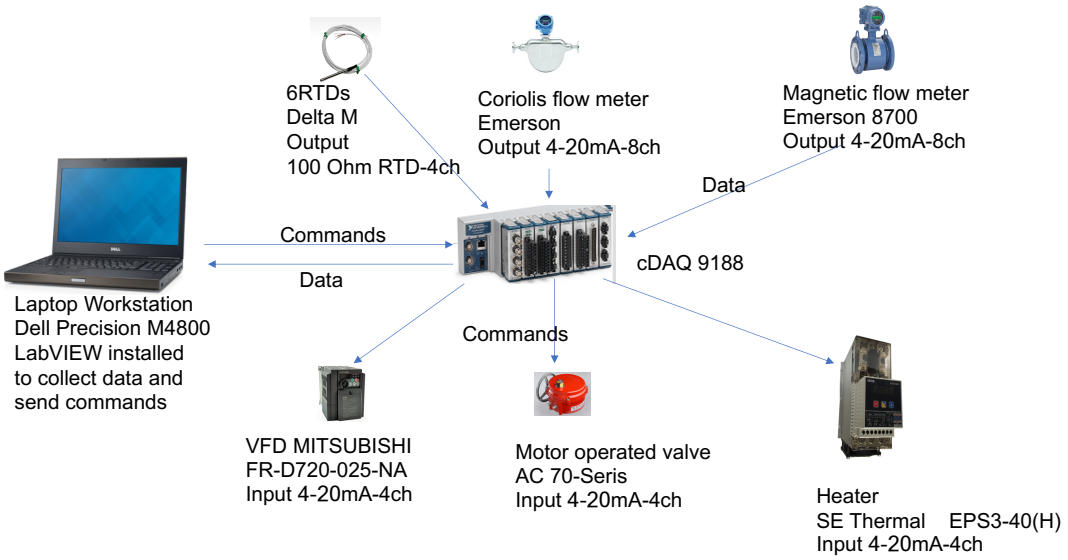


Figure 4.6: Testbed SCADA system topology with components model number and I/O requirements.

platform for a graphical language from NI, which is commonly used for data acquisition and instrument control [118]. LabVIEW was installed on the engineering workstation to make it a SCADA master. With the help of cDAQ9188, a model for the two-loop physical system was programmed using LabVIEW as shown in Figure 4.5 to record sensor data and send control commands to actuators. The "DAQ assistant" module is for data acquisition, the "DAQ control" module is to control actuators, and the "Write To Measurement File" module on the upper right corner achieves the data collection. The "Slide Heater", "Slide VFD RTD0", and "Slide MOV RTD1" are the open controller and PI controllers which control the heater power, VFD frequency, and MOV position, respectively.

Figure 4.6 shows the SCADA system topology and gives the components model number and I/O requirements as well. Temperature and flow rate values are recorded by LabVIEW. The VFD frequency, heater power, and the MOV position are controlled by LabVIEW commands. Data is written into a file by LabVIEW. The Kali Linux operating system is used on the attacker PC to launch different attacks within the LAN network used in this research. Kali Linux is a special purpose Linux distribution aimed at advanced penetration

testing and security auditing that includes state-of-the-art software tools that assist security engineers in performing security audits [95]. Argus collector summarizes the network flow data collected by the Argus flow generator.

Different component are mapped to different level according to Purdue ICS model. Level 0 contains the sensors and actuators in the physical experimental loop, including all the RTDs, flow meters, heater, VFD, and MOV. Level 1 consists of cDAQ9188, while Level 2 contains the engineering workstation which is the SCADA master and the Argus collector which collects network flow data. The attacker PC is out of the level scope and is assumed to have gained access to the switch as shown in Figure 4.4 and spread out in the LAN through the switch.

4.2 Data Collection of ICS Testbed

The testbed has the capability of providing network traffic data and host system data of the SCADA master, network flow data of the entire network, and process data of the flow loop. The network traffic data and host system data of the SCADA master was collected using Windows Performance Monitor installed in the engineering workstation [62]. The network flow data was summarized in Argus collector [97]. The process data was collected by LabVIEW installed in the engineering workstation as described in the section above.

Windows Performance Monitor (WPM) was utilized to collect host system data and host-related network data in this testbed since the engineering workstation runs on the Windows operating system. Figure 4.7 shows the main menu of WPM. WPM can monitor different aspects of the computer like ram or storage, and provides a user-defined collector to collect interested aspects. In this research, all the aspects available were collected to find the useful ones.

Argus is a piece of open source software designed to summarize network flow data from packets [97]. Argus captures and summarizes the dynamics and semantics of each flow for numerous packets, which reduces the quantity of data by a significant amount. This also provides an efficient way for users to store, process and analyze large amounts of network data, since detailed statistics for each bidirectional flow are generated, such as load, duration,

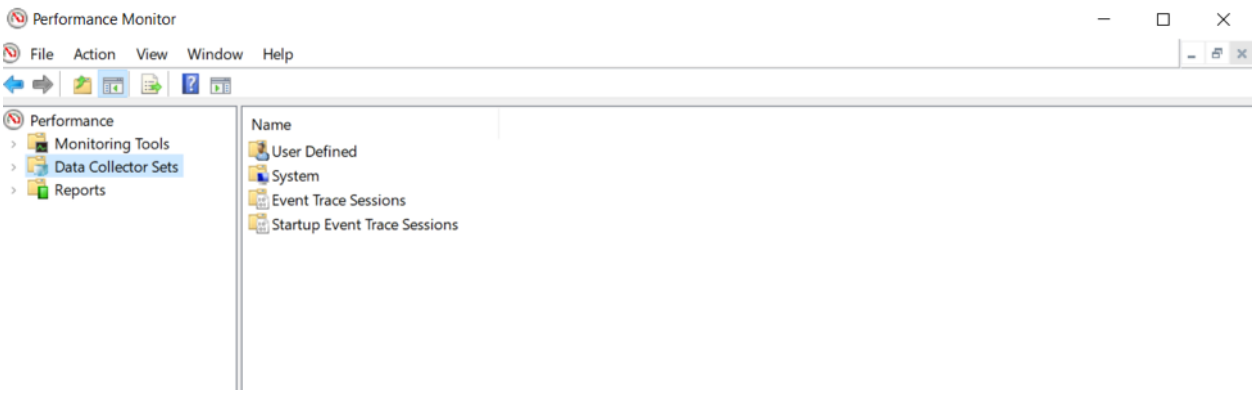


Figure 4.7: Main menu of Windows performance monitor.

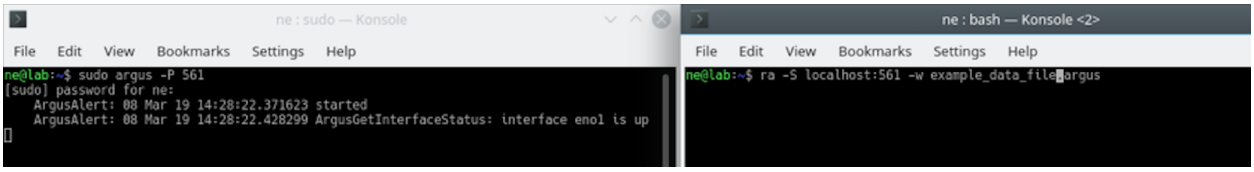


Figure 4.8: Basic commands for starting Argus.

and rate. The switch in the testbed can mirror the entire LAN traffic to the Argus collector. Argus is Linux based and the basic commands for starting Argus are shown in Figure 4.8.

Process data was collected by LabVIEW with the frequency of 0.25 Hz as shown in Figure 4.5. Normal data and abnormal data are essential to build data-driven models. Normal behavior data was collected under normal testbed operation without any cyber-attack in order to build a training model that would learn system characteristics during normal operation. Abnormal data was collected under proposed cyber-attack scenarios to: a) build supervised classification models by providing attack patterns in M1; and b) evaluate the effectiveness of built unsupervised models in M2 and M3.

4.3 Attack Scenarios on ICS Testbed

Based on the ICS testbed described in the section before, several cyber-attack scenarios were designed and conducted, including packet sniffing using MITM, DoS, data exfiltration, data corruption, false data injection, tampering, and simultaneous cyber-attacks which lead to small loss of coolant accident (LOCA). Some of these are for demonstrating the possibility of different cyber-attacks while others were used for generating data under the attack

scenarios for building and evaluating the cyber-attack detection models and the cybersecurity solution platform. All the scenarios discussed here assumed that the attacker gained access to the LAN, which was simulated in the testbed by connecting the attacker PC to the LAN through a switch as shown in Figure 4.4.

Cyber-attack scenario A is packet sniffing using a MITM attack, which leverages an Address Resolution Protocol (ARP) poisoning attack. The ARP resolves the IP addresses to the MAC addresses. The ARP poisoning attack uses the lack of authentication in ARP advertisements to manipulate the IP/MAC address mapping tables of the victims to reroute packets through the attacker [126]. To launch this attack, a network discovery was conducted first by Nmap in Kali and then this MITM attack was launched using Ettercap in Kali. In this scenario, the attacker PC listened to the communications between two parties (engineering workstation and cDAQ9188) without their knowledge and learned information for preparation of further attacks.

Cyber-attack scenario B is a DoS attack against the cDAQ9188 using a spoofed IP address. The DoS attack flooded the cDAQ9188 with rapid superfluous requests to exhaust its capacity and render it unavailable, which prevented legitimate requests from getting through [130]. This attack was launched using an Hping3 flood in Kali. This attack successfully interrupted the connection between the engineering workstation and cDAQ9188.

The number of phishing attacks have increased in recent years. Two real-world phishing attacks learned from a cybersecurity company in the EDGE2018 summit which caused financial and reputation loss will be introduced here. For privacy considerations, the names of companies with compromised staff and the investigating team are omitted. In the first phishing attack, an account executive was going to set up email forwarding from Office 365 mailbox to her Hotmail account to receive company emails when she prepared to leave for vacation. However, she found out the forward set up already existed to an unrecognized Gmail account. She reported this to the company and the company hired an investigation team for root cause analysis. The investigation team found that a suspicious email received by the account executive five months earlier turned out to be a credential harvesting phish. She had clicked the credential harvesting link in the email which was found out by the investigate team through hard disk. The investigating team also determined that more than

4500 emails containing hundreds of social security numbers, routing and account numbers, and bank balances had been forwarded to the malicious Gmail account.

In the other phishing attack, a high-level executive discovered that the email address that he was corresponding with changed to an almost identical domain name, which prompted him to hire an investigation team. The investigation team scraped the hard disk and determined that he had clicked the credential harvesting link in a phishing email with a similar domain name. Through this phishing attack, the attacker had logged into his email and set up email forwarding to a Gmail account. The attacker then used information obtained from other forwarded emails to craft what appeared to be a legitimate financial request and response. Through this, the attacker was able to wire a significant amount of money from the company. These two real-world attacks demonstrate that phishing attacks towards employees are real and achievable. It is therefore not infeasible that attackers in the future may use phishing attacks on engineers and other employees in industry to obtain information on system structure, steal credentials to gain access to the corporate network, and even gain access to control networks.

Based on these two phishing attacks, cyber-attack scenario C was designed. It is a data exfiltration where the attacker steals and transmits critical information from the system. The attacker uses the phishing methods described above to steal information from the engineering workstation, including the LabVIEW model and data collected from the cDAQ9188 via a server-client protocol established between the engineer workstation and an external malicious server. Figure 4.9 shows the file direction of different attack scenarios. A malicious client script is installed in the engineering workstation to simulate the attacker installing malicious client scripts on a compromised machine in the system using phishing and social engineering attacks. A malicious server is deployed off the premises by the attacker which listens to any attempts to connect by a client. The malicious script in the engineering workstation then connect to the malicious server via the existing network infrastructure. Critical process data is then transmitted through the network periodically and silently once the data is available on the engineering workstation.

Cyber-attack scenario D is data corruption. The attacker installs a malicious program in the engineering workstation, which corrupts the LabVIEW model or the data collected from

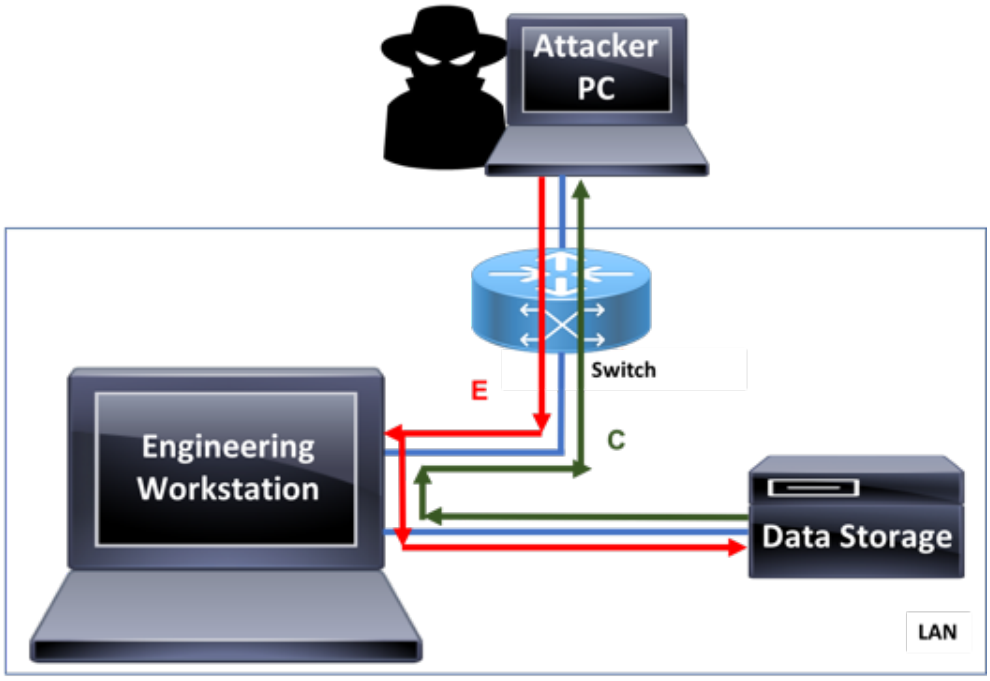


Figure 4.9: Attack patch indication for scenario C and E.

the cDAQ9188 by adding random generated numbers. This allows the attacker to disrupt industrial operations.

Cyber-attack scenario E is malicious data injection. The attacker injects false data into the engineering workstation as indicated by E in Figure 4.9. The false data could also be injected to the data storage to cover the attacker's tracks during a false data injection attack to hinder forensic investigation.

Cyber-attack scenario F is a false data injection. It could be launched by an insider (such as a disgruntled employee) or an external attacker. A disgruntled employee who has access to the system could conduct this attack without leaving cyber-traces in the network data. An external attacker can use a spear phishing technique to deceive someone into thinking they are an insider. In this scenario, the insider is referred to as the "carrier", who has access to the intranet of the power plant to utilize as a vessel to bypass the "air gap" between the outside world and the intranet. A detailed scenario is given here: the "carrier" is a subcontractor who is responsible for updating software of the intranet computers. The attacker identified the "carrier" by social engineering attack and then launched a spear phishing attack to the "carrier" by sending a specially crafted email, which contained a fake link of a news

article. The "carrier" then inadvertently downloaded a virus by clicking this link to read some interesting piece of "news". This virus is programmed to obfuscate its code to bypass antivirus software in the "carrier's" laptop. Once executed, it changes the system host's configuration to point to a malicious update website that is designed to mimic the authentic website of the software vendor. It also updates the certificate authorities to accept a faux HTTPS certificate. When the "carrier" visits the software website to download the updates, he is redirected to the attacker's mock website with the malicious program, which is then downloaded to his laptop. Since the host configuration and certificate authorities are altered, the laptop does not give any warnings about this specious download.

When the carrier goes to the power plant to do his software update job, he executes the malicious software update program. Even though the laptop is being run as a standard user, it has an exploit to leverage a known and unpatched vulnerability in the operating system the power plant uses in order to escalate its privileges. Using these higher privileges, it set the Ethernet interface of the laptop to promiscuous mode, which enables the interface to obtain all network packets regardless of the intended recipients. It then performs a host scan to identify all the hosts that are in the network including their interface MAC addresses. Using a vendor MAC address look-up table embedded in the malicious software update, it identified the cDAQ9188 and its IP address. It then launches a MITM attack to redirect the traffic between all other available hosts and the cDAQ9188 through ARP poisoning.

Then, the external attacker listens to the connection between the engineering workstation and cDAQ9188 long enough to identify which command it needs to alter. The attacker then modifies the commands to achieve its malicious goals. In this scenario F, the false command injection first brings the heater power to 0 kW, which simulates a forced reactor scram. Then the attacker brings the heater power to 6.3 kW and alters the core outlet temperature set point from 30^{o}C to 28.4^{o}C, which is aimed at accelerating the degradation or damage of critical assets by making the system run at an operational status outside of its design [136].

Cyber-attack scenario G is an extended version of scenario F that gives a demonstration of simultaneous cyber-attacks which leads to a small LOCA. In this attack, the attacker stops the pump and closes the MOV by false data injection. This results in zero flow rate in both the primary and secondary loop while the heater power remains at 70%. Since the

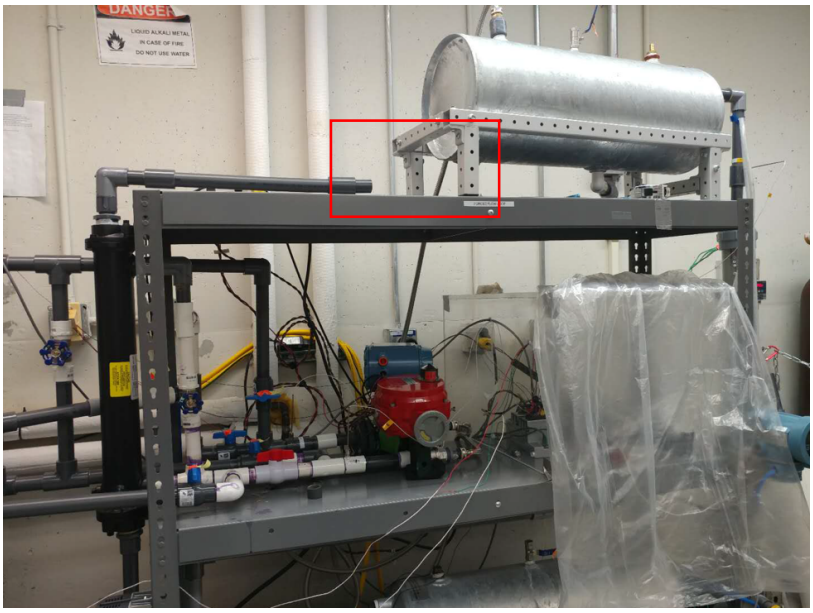


Figure 4.10: Small LOCA due to data tampering and false data injection.

heat is not dissipated, the joints of the pipes expand due to overheat, which leads to LOCA. Figure 4.10 shows that the leaking part of the joint was cut off to repair the testbed after this simulated attack, indicating the severity of this possible attack.

Among the cyber-attack scenarios described above, scenarios A, B, and F were used to generate data for studying data-driven models while other scenarios are demonstrations of how cyber-attacks could be launched in ICSs.

4.4 A Case Study: A Serial False Data Injection

This section presents a serial false data injection attack conducted toward the ICS testbed to test the effectiveness of the proposed cybersecurity platform. The attack used the same method to obtain access to the control network as described in scenario F. To make the attack scenario clear, Table 4.2 summarizes all the attack steps with both time step and observation index (denoted by Obs Index) in LabVIEW records. First, the testbed was running under normal operation without any cyber-attack. However, the attacker had already gained access to the LAN by bypassing the IPSs before this time but did not conduct

Table 4.2: False data injection attack steps

Start time (Obs Index)	End time (Obs Index)	Attack description
600s (150)	630s (158)	Network discovery
840s (210)	1020s (255)	MITM by Ettercap
1020s (255)	1020s (255)	Malicious code injection
1200s (300)	2400s (600)	LabVIEW model run with the malicious code

any malicious activity and did not trigger any alarms. Then, at 600s, the attacker launched reconnaissance within the LAN to learn the network architecture, including the devices' IP addresses and manufacturers. This step provided the attacker with the information necessary to develop an attack plan for the target. For example, after the attacker figured out that the engineering workstation is a Dell workstation, the attacker can research the vulnerabilities this engineering workstation has with online information. Then the attacker can use these vulnerabilities to further attack the system. After network discovery and attack development, the attacker launched a MITM attack towards the SCADA master and slave by ARP poisoning through Ettercap at 840s. This malicious action enabled the attacker to sit between the SCADA master and slave without their awareness. At 1020s, the attacker injected a malicious code into the SCADA master engineering workstation and left the system. This malicious code found the core outlet temperature acquisition channel and added a value of 2^oC to the measurement before the signal goes into the PI controller as shown in Figure 4.11.

This attack scenario is shown as the blue arrow with MITM in Figure 4.12, where the attacker injected the malicious code and then left the system. There is another type of the attack scenario as shown in the purple arrow with MITM in Figure 4.12, where the attacker stays in the system while performing the false data injection attack, such as changing a control command, altering set point, or tampering a measurement. The difference between the two scenarios is whether or not the adversary (which could be indicated by malicious IP address) is still present in the system. The one with blue arrow shows that the adversary is no longer in the system when the malicious code is executing, which is the focus of this research to demonstrate another valuable point of process data and the necessity and the effectiveness of the proposed defense-in-depth detection architecture.

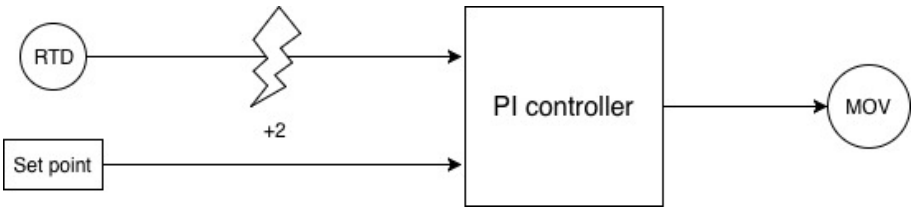


Figure 4.11: Malicious code adds 2^{o}C to the temperature measurement.

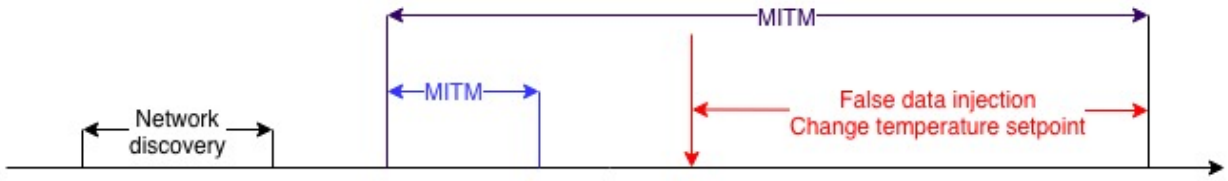


Figure 4.12: Two different MITM attack scenarios.

This scenario assumes that the dictionary-based detection module M1 did not catch the network discovery due to the short duration of these attacks relative to the collection frequency or the attack signatures are not in the attack database. This is a reasonable assumption for most advanced cyber-attacks which bypass commercial IPSs and IDSs. With this in mind, network flow data and process data are collected under this scenario to evaluate the effectiveness of proposed M2 and M3 models using other data sources.

Chapter 5

HIL Testbed Description and Attack Scenarios Setup

This section presents the HIL testbed and the related experiments in detail. This HIL testbed majorly serves as HIL testbed best practice and generate data for localized cybersecurity strategy development.[1]

5.1 HIL Testbed Description

Resilience of key equipment is important in an industrial facility, such as NPP. In order to investigate the impact of false data injection on key equipment and develop better cybersecurity strategies, a HIL testbed containing a NPP simulator and PLC was developed. HIL simulation consists of a simulated system and a real world device or system under test. It provides a real-time simulation to evaluate the device or subsystem under test in-situ. It provides an easy way to inject perturbations, such as malfunctions, to test the fault tolerance of the component or system. In this way, the behavior of real world device or subsystem can be observed together with the whole system behavior. HIL simulation has been utilized in power grid [100][76], vehicle design [140][77], traffic system [70][141], and aerospace [113][47] applications for a long time. In the nuclear industry, it has been utilized for verification

[1]Part of this chapter is a reformat and reprint of partial materials as it appears in [137]. The dissertation author was the primary investigator and lead author of the paper.

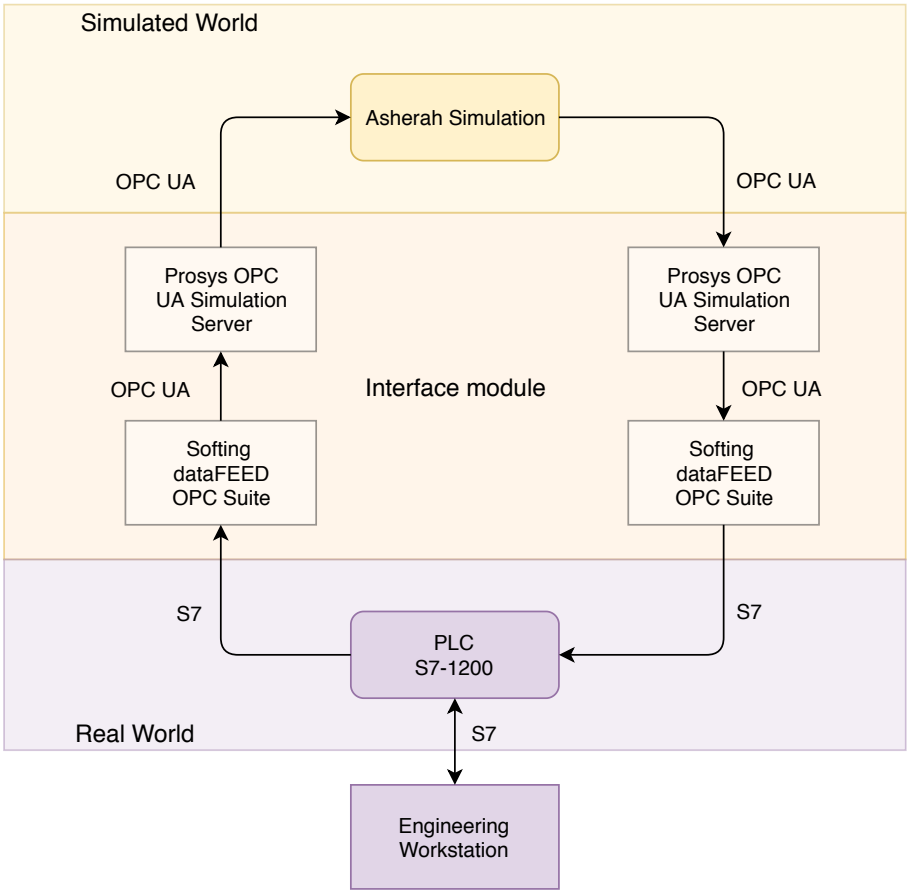


Figure 5.1: Hardware-in-the-loop (HIL) testbed framework.

and validation (V&V) of I&C systems in NPPs [99]. Therefore, a HIL simulation testbed is promising for testing cybersecurity strategies locally by injecting cyber-attacks.

Figure 5.1 shows the framework of the HIL testbed, which contains a NPP simulator named Asherah in the simulated world, a Prosys OPC UA Simulation Server and a Softing dataFEED OPC Suite for an interface module, and a Siemens S7-1200 PLC in real world. An engineering workstation is connected to the PLC to achieve system configuration and data acquisition for the PLC. All the parts are located in the same LAN and are assigned their own IP addresses.

5.1.1 Asherah simulator

Asherah is a two-loop PWR simulator developed with MATLAB SIMULINK by the University of Sao Paulo. Asherah simulates the PWR design with all key components,

such as the reactor core, coolant pump, pressurizer, and SG. It is able to simulate normal operation with power change from 80% to 110% nominal power (800 MWe). It provides an Open Platform Communications (OPC) read/write module as shown in Figure 5.2 to provide the I/O for communication with Prosys OPC UA server through OPC UA protocol. Variables on the left side of the yellow box show the potential variables that can be read by the OPC Unified Architecture (UA) server and the variables on the right side show the potential variables that can be written to Asherah. Considerable work has been done in this research to modify the existing Asherah model to achieve the I/O connection according to the needs of this research.

5.1.2 Prosys OPC UA simulation server

The Prosys OPC UA simulation server is a data server that can connect different data sources and enables data exchange with OPC UA protocol as shown in Figure 5.3. It is a multi-platform stand-alone OPC UA server with easy-to-configure endpoints [96]. OPC UA is a standardized communication protocol for Industrial Internet security and is widely adopted for Industry 4.0 solutions, which provides inter-operability in process automation and beyond.

5.1.3 Softing dataFEED OPC suite

Softing dataFEED OPC Suite is software designed to connect OPC UA and OPC classic communication and has several key advantages [110]. It has the read and write access to the data in controllers from leading manufacturers without needing to modify control programs, and the configuration is easy. It also has an integrated OPC UA Server, which provides communication to the PLCs of most manufacturers. Therefore, it is run between the Prosys server (via OPC UA) and Siemens PLC (via S7) to achieve data exchange between the PLC and Asherah. The Softing dataFEED utilized in this testbed is in demo mode with limited functionality, which satisfies the needs of this testbed. Figure 5.4 shows the data exchange interface of the Softing dataFEED OPC Suite, which displays part of the data exchanged

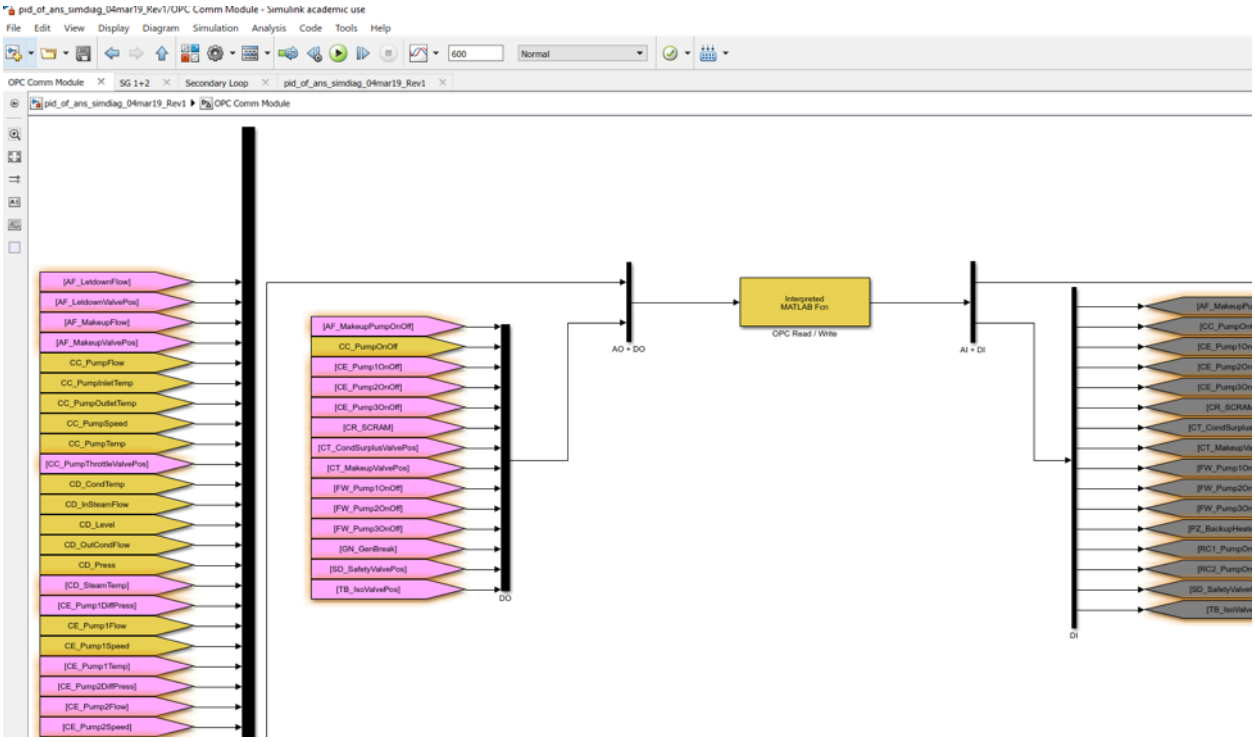


Figure 5.2: OPC communication module in Asherah simulator.



Figure 5.3: Prosys OPC UA server status interface.

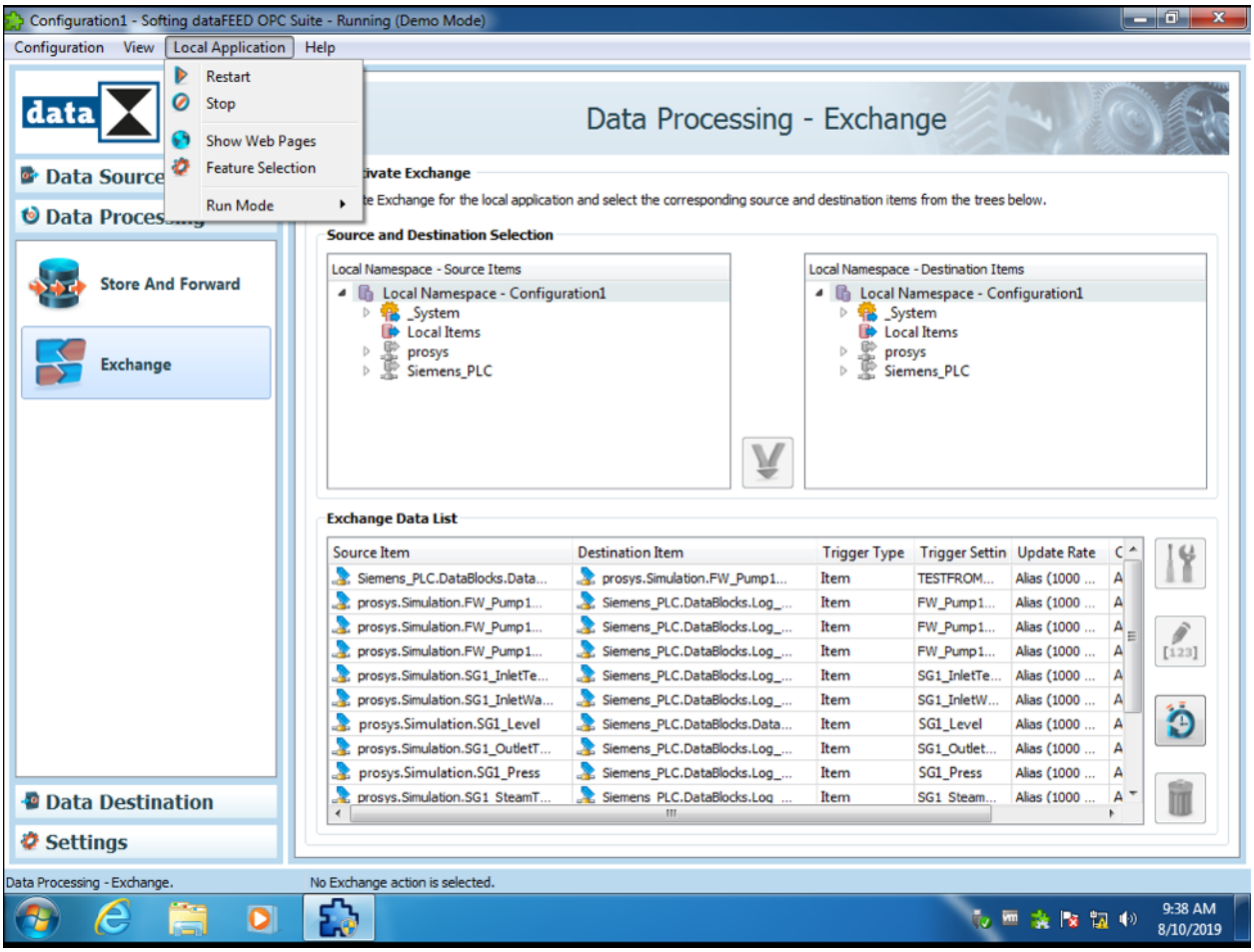


Figure 5.4: Softing dataFEED OPC Suite data exchange using interface.

between the PLC and Prosys in the exchange data list. The data exchange can be started or stopped from the local application as shown in Figure 5.4.

5.1.4 Real-world PLC for steam generator water level control

As key equipment in an NPP, SG plays the important role of generating steam to produce electricity. There are two level separators located in the top of the SG which separate the steam and water to provide satisfying steam quality. The control of the SG water level is significant in the safe operation of the NPP. It is crucial to maintain the SG water level within a specific set range. In a PWR, a higher water level than the set range may cause water to overflow from the separator while a lower water level than the set range may expose and potentially damage the heat transfer tubes by thermal stress caused by unevenly heated

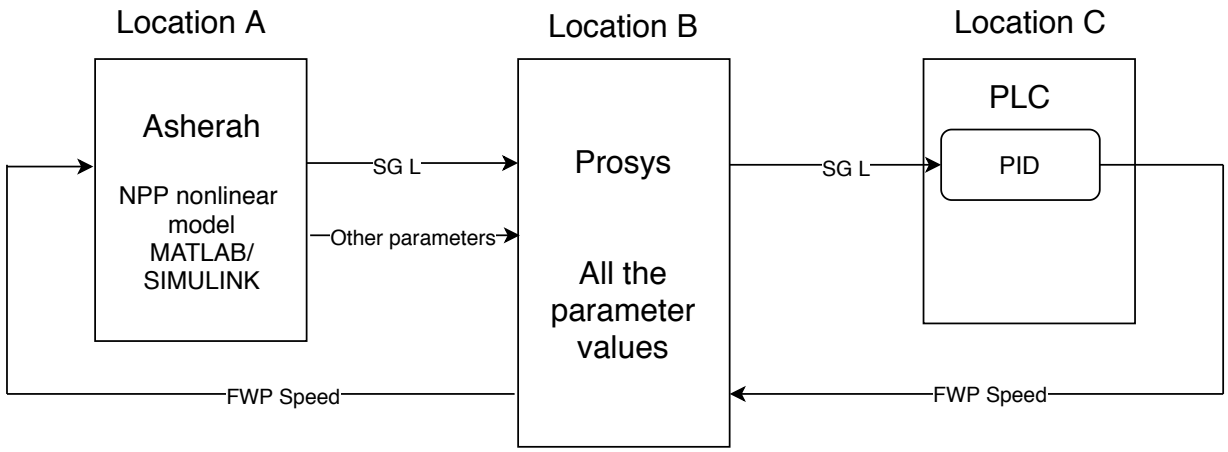


Figure 5.5: Data transmission sketch between Asherah and PLC.

tubes. Therefore, SG control is selected as the key function to be studied and tested with the HIL testbed.

The Siemens S7-1200 PLC is utilized here as the real-world device and is programmed to control the SG water level in the Asherah simulator. Since the shrink and swell phenomena is not modelled in Asherah, an one-element controller is adopted to control the SG water level at a reference level. As shown in Figure 5.5, the PLC takes the SG water level measurement (denoted by SG L) from Prosys (which is updated by Asherah output) as its input to a programmed proportional integral derivative (PID) controller and outputs the feedwater pump speed command (denoted by FWP Speed) to the Prosys server. Under the normal operation, the feedwater pump speed stays around 50% of the maximum pump speed. Then Asherah takes the feedwater pump speed command from Prosys as one of its inputs and calculates all the state variables in the whole simulated PWR system, including SG water level. Then Asherah sends variables, which are configured in the OPC communication module (shown in Figure 5.2), to the Prosys server which are accessible for the PLC. The Prosys server updates these variables with the frequency of 1 Hz. In Asherah, the reference water level of the SG is 15m. Based on this simulator design, 13.9m is regarded as the low-low level and 16m is the high-high level, which is calculated proportional to the AP1000 SG design.

5.2 Data Collection of HIL Testbed

This HIL testbed provides the ability to capture the behavior of the physical parameters, such as reactor power temperature, pressure, flow rate, pump speed, and water level in the Asherah model. As shown in Figure 5.5, this HIL testbed is capable of providing data collection at three locations: Asherah SIMULINK model (location A), Prosys server (location B), and PLC (location C). This feature allows the state variable monitoring as well as data manipulation in Prosys or PLC to simulate different false data injection scenarios.

To build local cybersecurity strategy models, state variables related to the SG are collected. Five variables, feedwater pump speed command, reactor power, SG inlet flow rate, SG water level, and SG pressure, are fed into the PLC to be utilized as inputs of the AAKR model. Other state variables of Asherah can be collected from MATLAB and Prosys server to aid the data analysis.

5.3 False Data Injection Attacks on HIL Testbed

In order to study the proposed localized cybersecurity kit, three simulated false data injection scenarios were conducted towards the SG in this HIL testbed. The attack scenario is based on an imaginary NPP called Asherah and an adversary team with highly-skilled anti-nuclear hackers that is unsatisfied with the current state of the government. Therefore, this adversary team is constantly seeking an opportunity to attack Asherah and disrupt operation to: a) cause public panic and b) cause financial loss to the current government. They have identified several potential insiders through information gained from social networks. The Asherah NPP announced that a scheduled offline maintenance is occuring in 15 days. The adversary team thought this was a good chance and selected a nuclear mechanical technician whose name is Sily and works in a subcontractor company, based on information found on his professional account page. Sily's social media posts also indicated that Sily does not like the current government. One post detailed: "Old manager asked me to do some heavy maintenance work back in Asherah NPP! It has been a while without going out but I guess this month's bill is easy to pay! Great!" The reasons that Sily was selected by

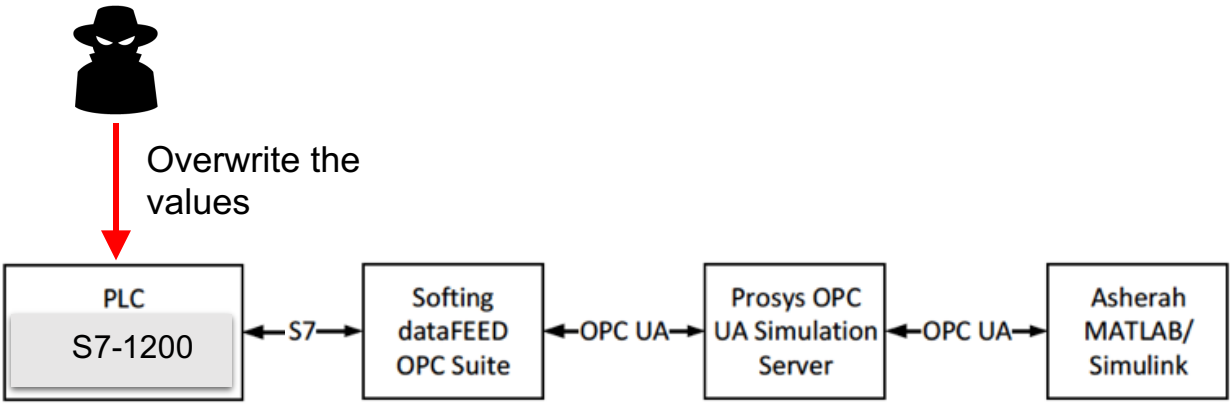


Figure 5.6: Attacker tampers the values in PLC.

the adversary team were: a) He is going to do maintenance work in the Asherah NPP; b) He is short of money; and c) He does not like the current government. The adversary team conducted social engineering attacks on Sily and promised Sily \$30k after Sily put a LAN Turtle into a maintenance list kiosk. On the day that Sily went to Asherah, Sily found a "good" opportunity to plug the LAN Turtle into a kiosk. This LAN Turtle enabled the adversary team access the Asherah NPP business network from the Internet. After a period of reconnaissance, the adversary team injected a malware program to a host on the business network. This malware exploited firewall vulnerabilities and traveled to the engineering workstation, which is connected to a PLC. At this point, the malware executed automatically, being programmed to overwrite the values sent to the PLC as shown in Figure 5.6.

Three attack scenarios were conducted through three malware programs that alter the SG water level measurement so that the PLC received the compromised water level measurement. These three attack scenarios are summarized in Table 5.1. In scenario I, PI controller input in the PLC was compromised with constant values. This was conducted by overwriting SG level measurement to 15.9m constantly in PLC. This made the PLC controller issue commands based on a constant fake value no matter how the process changed. In scenario II, the malicious code added 0.9m to the real SG water level measurement at the input of the PLC, which alters the input of the PI controller to X+0.9m (X being the real SG water level measurement). In this situation, the PLC reduces the pump speed until the actual SG water level reaches 14.1m. Meanwhile, the malicious program masked the malicious

Table 5.1: Three attack scenarios conducted towards SG in the HIL testbed

Scenario No.	Attack description	PLC input
scenario I	Overwrite the SG level measurement to 15.9m	15.9m
scenario II	Inject a malicious code to add a 0.9m to the SG water level measurement	X+0.9m
scenario III	Simultaneous attack: alter the water level set point to 14m; mask the SG water level set point with 15m	14m, eventually

action by faking the normal reading of 15m to the operator, which is similar to the Stuxnet attack where the attacker changed the spin speed of the centrifuges and masked it from the operators by presenting the normal speed in HMI displays. In scenario III, the malicious program altered the water level set point in the PLC to 14m and masked the SG water level set point with 15m. Therefore, the PLC reduced the pump speed and went back to around 50% once the water level reduced to 14m.

Chapter 6

Results

This chapter presents results from the ICS testbed and HIL testbed to demonstrate the effectiveness of the cybersecurity solution platform and localized cybersecurity strategy. Data was collected under normal operational transients and under cyber-attack scenarios to build data-driven models and perform evaluation. For the cybersecurity solution platform, the results of normal ICS testbed operation are given to show that the built testbed functions well and is able to provide the data needed for cybersecurity research. Then the cyber-attack detection results on this ICS are presented, including the supervised classification models for host-related network and system data, the unsupervised models for network flow, and the unsupervised models for process data. Finally, the results for the serial false data injection scenario are presented to show the effectiveness of the cybersecurity solution platform with wide cyber-attack detection coverage and functionality using the defense-in-depth concept. For the HIL testbed, the results of the localized detection models and an inference model are presented as proof-of-concept research.[1]

6.1 Normal Operational Transients of ICS Testbed

Several normal operational transients are tested to ensure the SCADA system's ability to maintain a steady state and meet control expectation under normal operation. The

[1]Part of this chapter is a reformat and reprint of partial materials as they appear in [135], [136], [133], [134], and [137]. The dissertation author was the primary investigator and lead author of these papers.

capability of operating with a steady state is essential for building data-driven models which learn patterns under normal operation, so that abnormal activities can be detected by the constructed unsupervised models.

Figure 6.1 shows the change of all the process data under normal operational transients with step changes of the heater power from 100% nominal power to 50% and back to 100% with 10% power increments. These process variables include heater power, temperatures of RTDs, flow rates of the primary and secondary loop indicated by the current that the SCADA master received from the Coriolis flow meter and magnetic flow meter, respectively, and the position of MOV and frequency of the VFD indicated by the current received. The top left subplot shows the heater power level profile specified. There are some overshoot due to manual control in the beginning of every power change. The top right subplot shows the temperature readings of RTDs located in different places, as indicated in Figure 4.3. RTD1 and RTD0 are core inlet and outlet temperature, which are supposed to be maintained at set points of 26^{o}C and 30^{o}C, respectively. As shown in this subplot, they are maintained at their set points most of the time with some noise, except for fluctuations at 50% of the heater power. This shows that under low power, this ICS testbed is not able to maintain a steady state. This is because the design of this physical facility considers nominal power. RTD2 is installed between the upper delay tank outlet and the HX primary side inlet; therefore, as shown in this subplot, it follows the temperature of core outlet (RTD0) with a short time delay. RTD5 is located between the HX primary side outlet and the lower delay tank inlet; therefore, as shown in this subplot, it follows the temperature of core inlet (RTD1) with a short time delay. The water in the HX secondary side is forced water from the building system to provide an unlimited heat sink. RTD3 measures the HX secondary side inlet temperature; therefore, the RTD3 temperature is not controllable by the physical system and fluctuates according to the other activities in the building. RTD4 measures the HX secondary outlet temperature, which depends on the temperature measured by RTD3 as well as the heat removed from the primary side coolant to maintain heater inlet and outlet temperatures at their set points. The bottom right subplot shows the currents, which are the control commands, sent to the MOV and VFD determined by the PI controllers. The bottom left subplot shows the primary and secondary flow rate meter currents; as shown

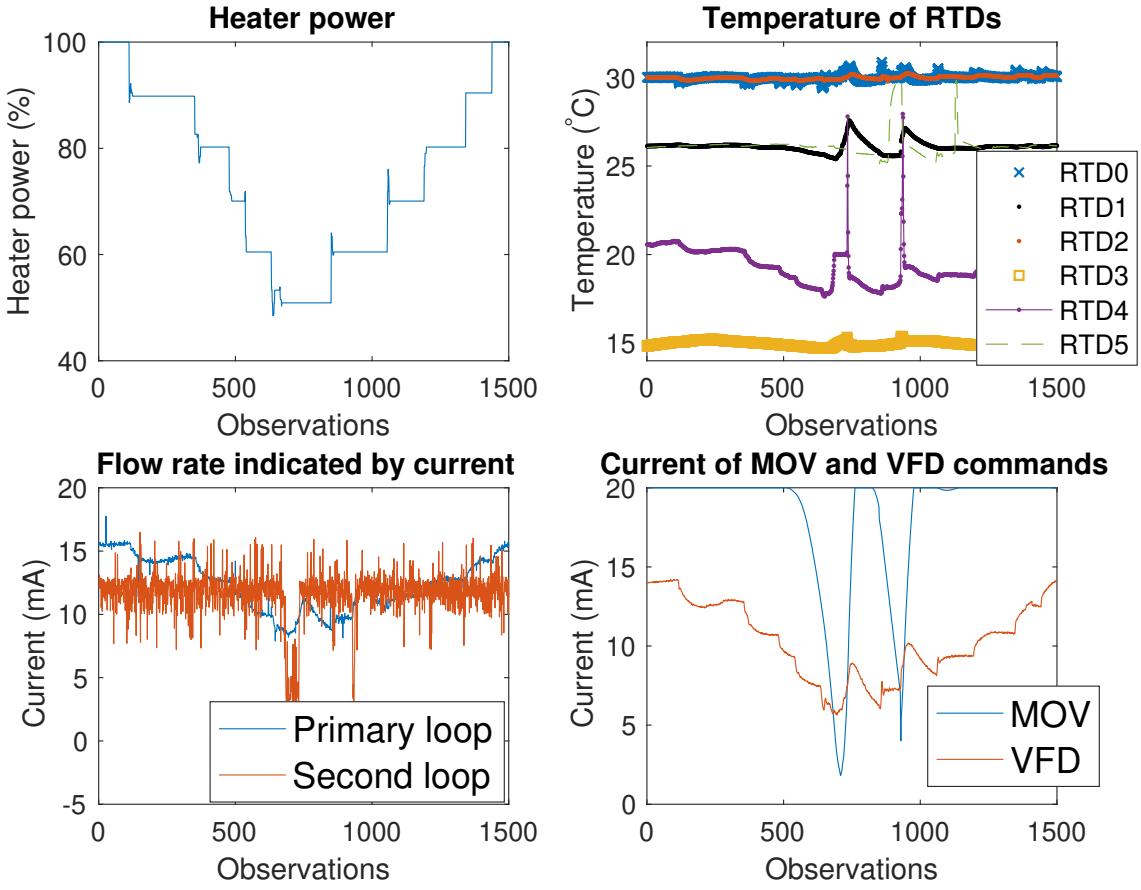


Figure 6.1: Process data under normal operational transients with heater power step changes from 100% to 50% and then to 100%.

in this subplot, the magnetic flow meter on the secondary side has significantly more noise than the coriolis meter on the primary side due to the measurement modality used.

To illustrate that the testbed can maintain steady state for a long time when the heater power is greater than 60% of nominal power, another power change experiment is presented here. Figure 6.2 shows process data under a heater power step change from 77% to 100% of the nominal power; this step change occurred around observation 220 (top left subplot). The MOV and VFD command signal subplot (bottom right) shows that the MOV and VFD took the expected actions to maintain the heater inlet and outlet temperature at the set points. The 20mA (0.02A) MOV current means the MOV is fully open. The VFD command current varies from 4mA to 20 mA to change its frequency which in turn changes the speed of the pump. Correspondingly, the flow rates in both the primary and secondary sides changes

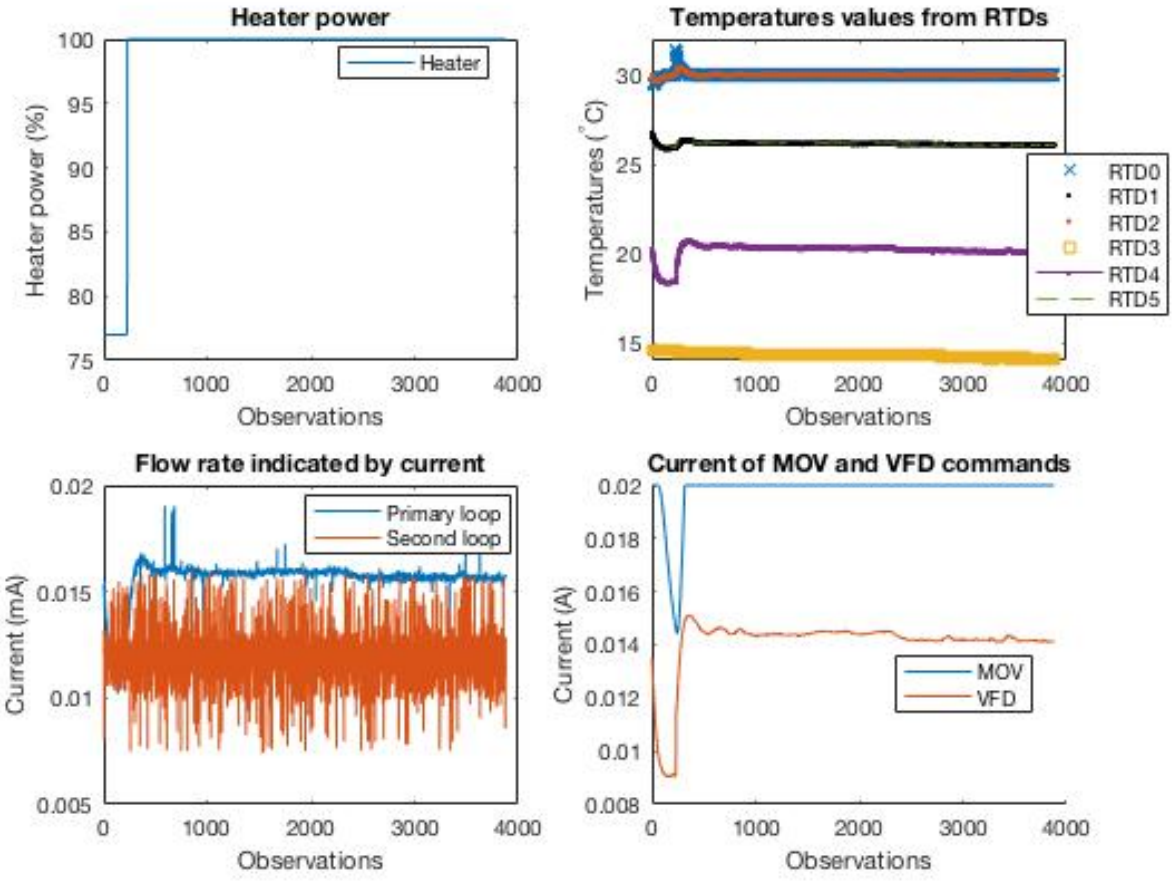


Figure 6.2: Process data under a heater power step change from 77% to 100%.

as shown in the bottom left subplot. Finally, the temperature data shown in the top right subplot demonstrates that all the RTDs were able to reach their steady temperatures. RTD1 was maintained at the set point of 26^{o}C and RTD0 was maintained at the set point of 30^{o}C. RTD2 followed the temperature of the core outlet temperature RTD0. Once the upper tank water buffer volume was heated up, the temperature value of RTD2 matched RTD0. Similar to RTD2, the RTD5 temperature also matched RTD1 when steady-state was reached. RTD3 indicates the temperature of forced water provided by building system, which varied slightly as shown in the yellow lines. RTD4 indicates the HX secondary outlet temperature, which maintains almost 20.1^{o}C with slight changes following RTD3. Therefore, a good steady state can be maintained.

As shown in Figure 6.1 and 6.2, the temperatures can be mainly maintained at their set points (RTD0 and RTD1) and steady states (RTD2 and RTD5), which shows that the

functionality of the SCADA system meets expectations and is able to provide the capability to conduct cybersecurity research.

6.2 Results of Cybersecurity Attack Detection System

6.2.1 Cyber-attack detection results of supervised classification models using cyber data

This section analyzes the host-related network data and host system data under a MITM attack and DoS attacks carried out on the ICS testbed using the four supervised classification models: KNN, decision tree, bagging, and random forest. In the MITM attack, the attacker sat between and re-routed the packets between engineering workstation and cDAQ9188 using ARP poisoning. Two DoS attacks were carried out: one was a Dos attack on the cDAQ9188 where the attacker flooded the cDAQ9188 to make it unavailable to the engineering workstation. The Dos attack on the engineering workstation was one in which the attacker flooded the engineering workstation to make it unavailable. The attack scenario is described here: the attacker launched the MITM attack first and stopped it, where the attacker gained some information about the system. Then after a period of normal operation without any cyber-attack, the attacker Dos attacked the cDAQ9188 and then stopped, and finally after another period of normal operation without any cyber-attack, the attacker DoS attacked the engineering workstation and then stopped.

The data collection system collects both normal data and cyber-attack data sets. The normal data set was collected under normal operation without any cyber-attack while the cyber-attack data set was collected under the attack scenario described above. The collected cyber-attack data set contains 1438 observations, of which 179 observations are MITM attacks, 53 observations are DoS attacks to the engineering workstation, and 105 observations are DoS attacks to the cDAQ9188. The remaining 76.36% of observations are normal operation intervals between specific cyber-attacks. Then the data extraction system extracted the useful features and fed the data to the four classification models. The performances of the four models are summarized in Table 6.1, including True positive, false

Table 6.1: Performance comparison of KNN, decision tree, bagging and random Forest

Classification Methods	True Positive	False Negative	False Positive	True Negative
KNN	98.84%	1.16%	0.54%	99.46%
Decision tree	94.80%	5.20%	1.25%	98.75%
Bagging	98.27%	1.73%	0.0%	100.0%
Random forest	97.69%	2.31%	0.0%	100.0%

negative, false positive, and true negative rate. KNN gave the best true positive rate of 98.84% while the decision tree gave the lowest true positive rate of 94.8%. The bagging and random forest yielded true positive rates higher than 97% and zero false positive rates. This low false positive rates make these two methods very suitable for applications requiring false alarms be as low as possible. False negative rate results show that the KNN yielded the lowest missed alarm rate, which means it catches cyber-attacks efficiently. The decision tree yielded the poorest performance with lowest true positive rate, highest false negative rate, and highest false positive rate. It is expected that bagging and random forest will outperform the decision tree in terms of accuracy since these two methods are techniques developed by combining single tree to improve classification performance. However, as mentioned in chapter 2, decision tree has the advantage of easy interpretation over bagging and random forest. It is also useful in performing feature selection since it presents a clear tree structure with a single variable used to separate the data in each branch as shown in Figure 3.6.

In addition to accuracy performance, the computing cost is another important characteristic in cyber-attack detection since a timely detection is crucial. Table 6.2 shows the computing costs (shown in milliseconds) of these four models, which was measured by user time for R code to determine whether a signal observation was from attack scenarios or not. R version 3.4.3 was utilized to do this classification on a MacBook Air with a 1.7 GHz dual-core CPU and 8 GB of RAM. All other active programs were terminated prior to the execution of the classification. The results show that among the four methods, the decision tree has the lowest computing cost, followed by bagging, random forest, and then KNN. The classification speed of bagging, random forest and KNN was also much faster than the

Table 6.2: Computing cost comparison of KNN, Decision Tree, Bagging and Random Forest

Classification method	User time (ms)
KNN	0.1558
Decision tree	0.0473
Bagging	0.0612
Random forest	0.0779

sample rate of the network data in this application (1Hz), which is a fast sample rate already. Therefore, computing costs of these three classification models are acceptable.

6.2.2 Cyber-attack detection Results of unsupervised AAKR model using network flow data

This section analyzes the Argus network flow data using an unsupervised AAKR model. The data under two types of cyber-attacks, a MITM attack and a DoS attack, were fed into the model to evaluate the effectiveness of the model. Another normal data set which was collected under normal operation was fed into this AAKR model as well to demonstrate that the model gives no false alarms when the system is not under the cyber-attack. Therefore, three data sets are utilized: a data set collected under the MITM attack (denoted by MITM) scenario, a data set collected under the scenario where the attacker DoS attacked the NI cDAQ9188 (denoted by DoS NI), and a normal data set for testing(denoted by Normal2). To demonstrate that the model can distinguish between normal operation and cyber-attacks, in both the MITM and DoS NI data set, the cyber-attacks were launched during the middle of the whole time period, that is the system operates normally without cyber-attack in the beginning and the end of the collection interval. Table 6.3 shows the total number of observations, the index of observations under attacks and the index of observations given by the AAKR cyber-attack detection model. The data collection and extraction system collected Argus data and extracted selected features and then fed the data into the unsupervised AAKR model.

Table 6.3: Different data set description and detection results

Attack scenarios	Total number of observations	Index of observations under attack	Index given by attack detection results
MITM	218	51-145	51-147
DoS NI	192	54-110	56-111
Normal2	139	None	None

Figures 6.3, 6.4, and 6.5 show the cyber-attack detection results of the MITM, DoS NI and Normal2 data sets given by the AAKR model, respectively. In each figure, the red rectangular background indicates the attack scope; that is, the observations in this scope are the network flow under attack. The upper left subplot shows the raw data measurements of the three selected features; it suggests that a model is necessary for cyber-attack detection, because the change of data caused by attacks are not detectable by simple thresholds since the data in the red scope does not appear different from the data outside the red background. The lower left subplot shows the raw residuals, which is the difference between the predicted values and observed values of all three selected features. The upper right subplot shows the filtered residuals, which are the results of a five-order one-dimensional median filter applied to the raw residuals. Then an engineered threshold is selected and applied to the filtered residual to generate the fault hypothesis of cyber-attack detection results. The fault hypothesis results are shown in the lower right subplot, in which "1" indicates the observation is under a cyber-attack while "0" indicates the observation is from normal operation. The blue box in the lower right subplot indicates the detected scope given by the AAKR model. The fault hypothesis in Figures 6.3, 6.4, and 6.2 shows no false alarms. However, as shown in Figure 6.3 and 6.4, the alarms for the attacks are not continuous, especially in the DoS NI scenario. There were many missed alarms during the attack, because during the attack, the values of selected features are sometimes within normal behavior scope. In this case, to reduce missed alarms, one may try to investigate additional features and may find some anomalous behavior in these missed alarm observations. However, the situation may also be that there is no abnormal behavior shown in the network flow data during these missed alarms. Therefore, the detection scope is adopted here to indicate the attack interval instead of the single alarm. Therefore, in Figure 6.3 and 6.4, the blue box overlaps with the red box.

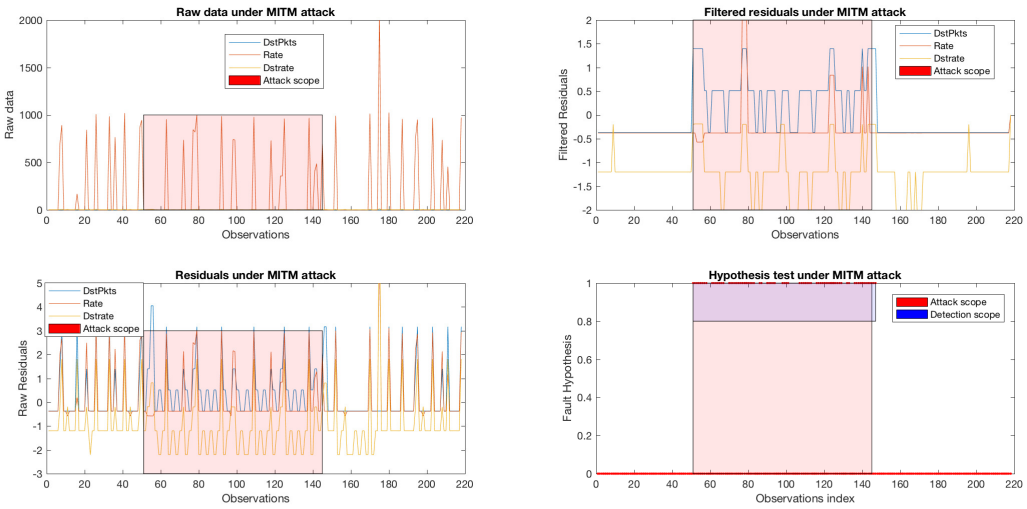


Figure 6.3: Raw data, raw residuals, filtered residuals and fault hypothesis of data set under the MITM attack scenario

As shown in the last column in Table 6.3, under the MITM attack scenario, the detected scope is 51 to 147 while the scope of the observations under attack is 51 to 145. Under the DoS attack NI cDAQ9188 scenario, the detected scope is 56 to 111 while the scope of observations under attacks is 54 to 110. For the normal testing data, the model gives zero alarms. Therefore, the proposed AAKR model yields satisfying results.

6.2.3 Cyber-attack detection results of unsupervised AAKR model using process data

This section presents the cyber-attack detection results of the unsupervised AAKR model using process data collected under attack scenario F: simultaneous tampering and false data injection attack. Nine highly correlated process sensors (V1-V9) are extracted by the data extraction system as inputs for the AAKR model: temperatures measured by RTD0, RTD1, RTD2, RTD4, and RTD5, primary flow rate (Coriolis flow meter), control signal given for MOV current, control signal given for VFD frequency, and heater power level. Figure 6.6 is a heat map plot indicating the correlation coefficients between these signals with the order from V1 to V9. The HX secondary side inlet temperature (RTD3) is not sufficiently correlated to

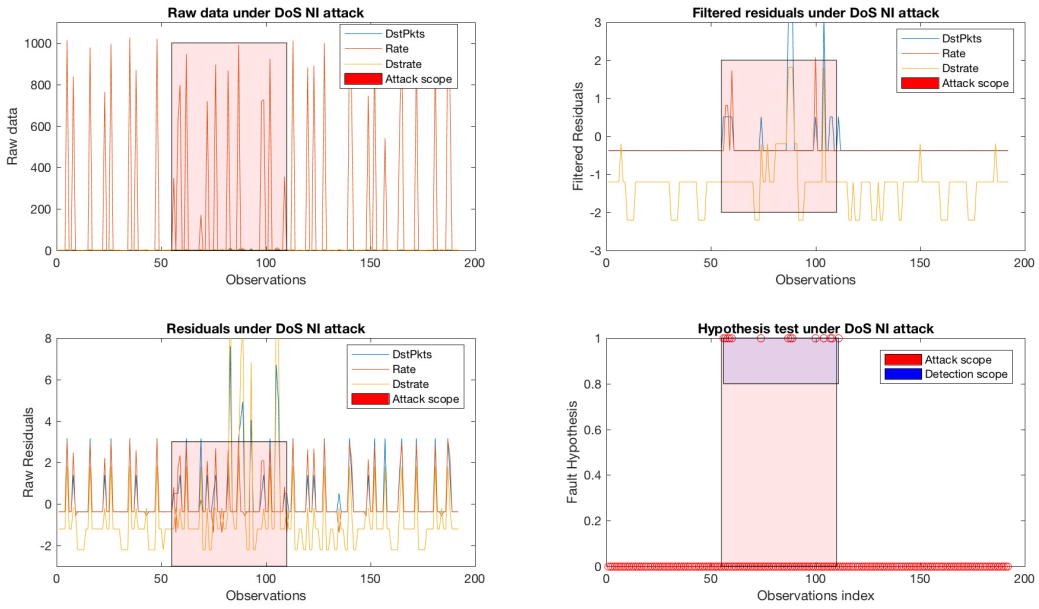


Figure 6.4: Raw data, raw residuals, filtered residuals and fault hypothesis of data set under the DoS NI scenario

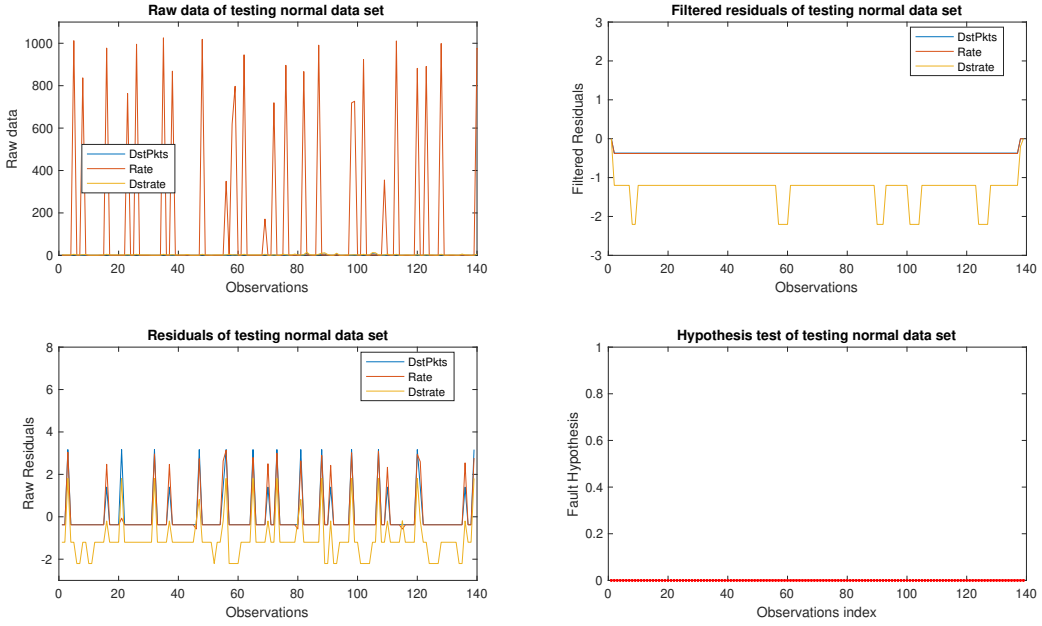


Figure 6.5: Raw data, raw residuals, filtered residuals and fault hypothesis of testing normal data set

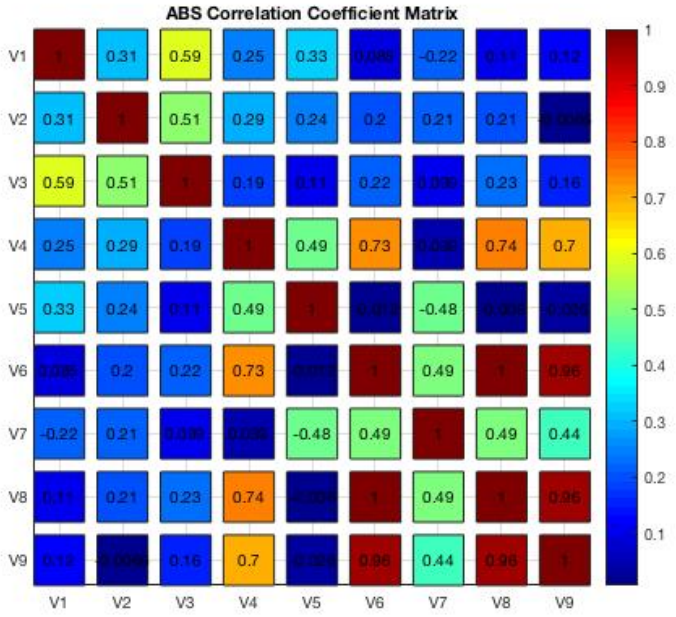


Figure 6.6: The heat map of correlation coefficients.

the other signals, because it only depends on the building supply water temperature. The data set used in this section contains 3000 observations with sample frequency of 0.25 Hz.

The selected signals were reconstructed using AAKR model. The residuals between the reconstructed values and real measurements were obtained. Figure 6.7 shows the temperatures of RTD 0-5 and the thresholds of the residuals in the top subplot. The bottom subplot shows the fault hypothesis of the observations, where the fault hypothesis is indicated in red points. "0" indicates normal operation while "1" indicates abnormal activity and the red box in background indicates the attack scope. If the residual of any of the 5 RTDs' observation exceeds the threshold, it is determined to be an abnormal observation. The first 250 observations are under normal operation, then at observation 251, the attacker built the connection to the cDAQ9188 and changed the heater power to zero. At observation 600, the attacker tampered with the core outlet set point and injected the command to change power to 70% of nominal power, which is within normal operational parameters. The process monitoring results suggest that the AAKR model is effective in detecting the process change caused by cyber-attacks.

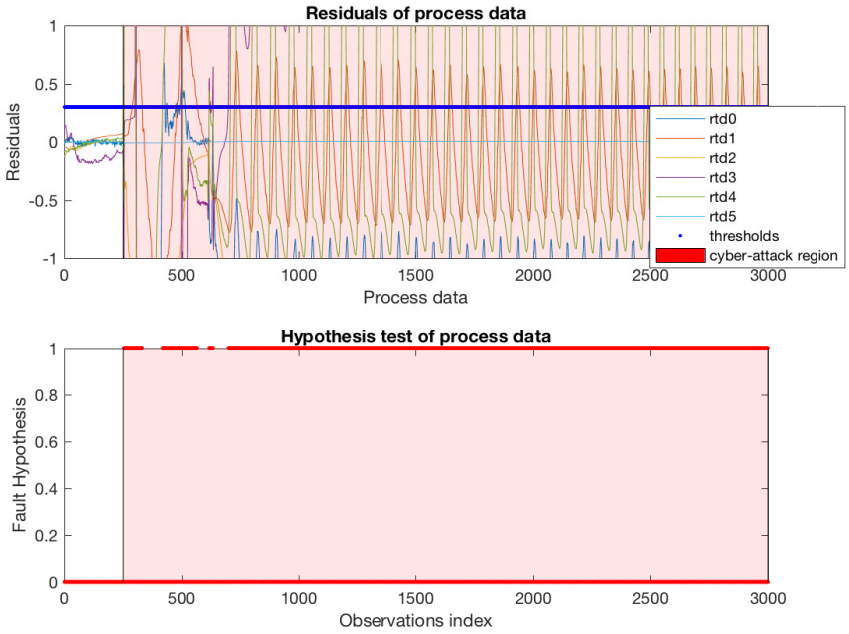


Figure 6.7: The detection results (fault hypothesis) of process data in data tampering and false data injection scenarios using AAKR model.

6.2.4 Cyber-attack detection results of the serial false data injection using the cybersecurity solution platform

The detection results of data collected under the serial false data injection described in Section 4.4 are presented here to evaluate the effectiveness of the cybersecurity solution platform, including the data collection and extraction system, the multi-layer cyber-attack detection system, and the cause analysis system. The attack steps of this scenario are relisted here in Table 6.4 to make this chapter easy to read.

All four types of the data: host system data, host-related network data, network flow data, and process data were collected from the ICS testbed by the data collection system and the selected feature were extracted by the data extraction system and fed into the cyber-attack detection system. It is assumed that M1 did not catch the network discovery and MITM attacks. This could happen due to two potential reasons: one is that the duration of these attacks is short compare to the host system and host-related network data sample rate, and therefore only few data points are collected. Or, the attack signatures were not included in the attack database.

Table 6.4: Attack steps of a serial false data injection attack towards ICS testbed

Start time (Observation Index)	End time (Observation Index)	Attack description
600s (150)	630s (158)	Network discovery
840s (210)	1020s (255)	MITM
1020s (255)	1020s (255)	Malicious code injection
1200s (300)	2400s (600)	LabVIEW model runs with the malicious code

Since M1, M2 and M3 operate simultaneously, M2 and M3 perform cyber-attack detection regardless of the detection results of M1. Figure 6.8 shows the cyber-attack detection results of M2 based on Argus network flow data. In each subplot, the abscissa indicates the sequence number index of Argus network flow data and the red background indicates the attack scope. The upper left subplot shows the raw data measurements of selected features, which indicates that the data changes caused by attacks are not detectable by simple thresholds. The lower left subplot shows the raw residuals between the observed and predicted value. The upper right subplot shows the filtered residuals. The lower right subplot shows the fault hypothesis, where "1" indicates there is evidence that the data point is a result of an attack and "0" indicates it is a normal observation. After the attack was detected as shown in the lower right subplot, the start and the end time of the sequences were traced back. The results show that the detection window covers the entire duration of the MITM attack (840-1020s). Therefore, M2 is effective to detect this attack without prior knowledge about the attack signature.

Recall Figure 4.12, which shows the scenario simulated in this dissertation is indicated by the blue arrow, that is the attacker injected a malicious code and then left the network but the malicious code remained. This malicious code altered the controller input measurement by adding 2^{o}C to the measured value. This lead to a compromised input to the PI controller although the attacker was already not in the system. Therefore, even though the MITM attack was detected by the M2 module, the operator may believe the system is safe since the attacker is no longer in the system. Without the M3 unsupervised cyber-attack detection model using process-data, the operators may not be immediately aware of the false data injection attack that is altering the operation of the system.

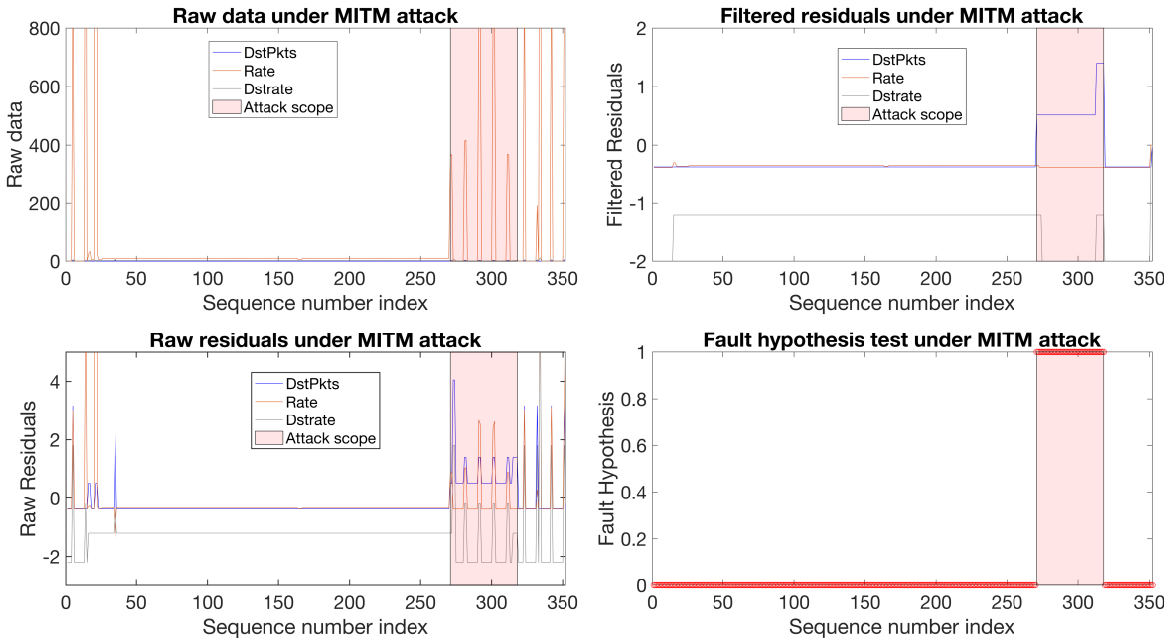


Figure 6.8: Raw data, residuals, filtered residuals and detection result (fault hypothesis) of data under the serial false data injection attack.

The cyber-attack detection results of the AAKR model, AASVR model, ensemble regression model, and AAPCR in M3 are presented in Figures 6.9, 6.10, 6.11, and 6.12, respectively. The top subplot shows the residuals of five sets of process data: core outlet temperature (RTD0), core inlet temperature (RTD1), HX primary side outlet temperature (RTD2), HX secondary side outlet temperature (RTD4), and HX primary side inlet temperature (RTD5). The blue lines are the thresholds for all the RTDs. The bottom subplot shows the fault hypothesis based on the residuals and the thresholds. When a residual is lower than threshold, the fault hypothesis is "0" , which indicates the observation is likely from normal operation. When a residual is outside of the threshold, the fault hypothesis is "1", which indicates the observation is abnormal. The red box indicates the time interval that the malicious code was affecting system operation.

As indicated in Table 4.2, at the 300th observation in the process data, the malicious code started adding 2^{o}C to the measurement; As shown in Figure 6.10, the AASVR model detected the anomaly immediately. The AAKR model detected the anomaly after 18 observations as shown in 6.9. The sample frequency of the LabVIEW system is 0.25 Hz, so the detection

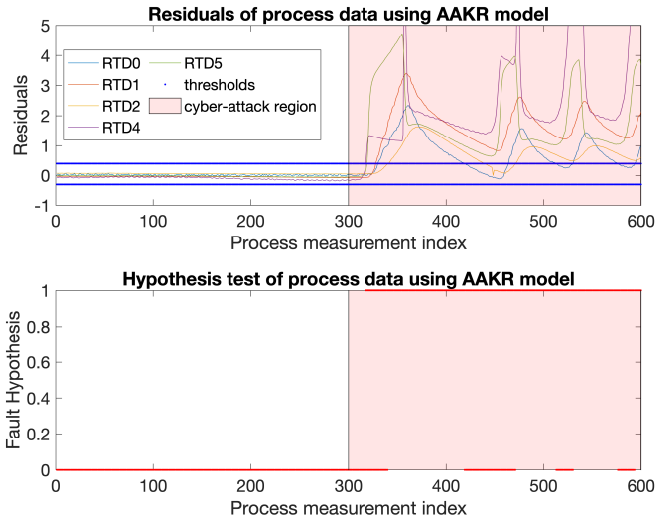


Figure 6.9: Attack detection based on process data using AAKR model

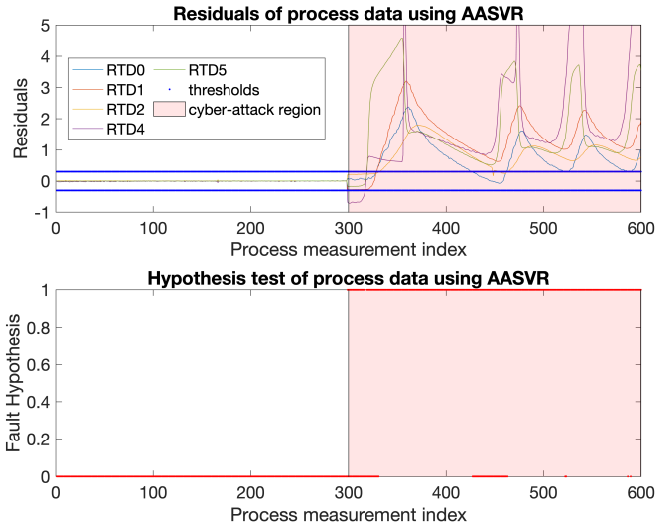


Figure 6.10: Attack detection based on process data using AASVR model under the serial false data injection attack.

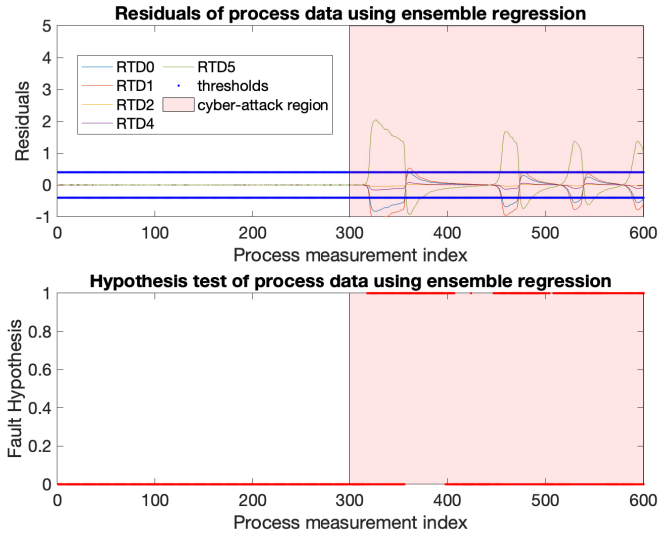


Figure 6.11: Attack detection based on process data using ensemble regression under the serial false data injection attack.

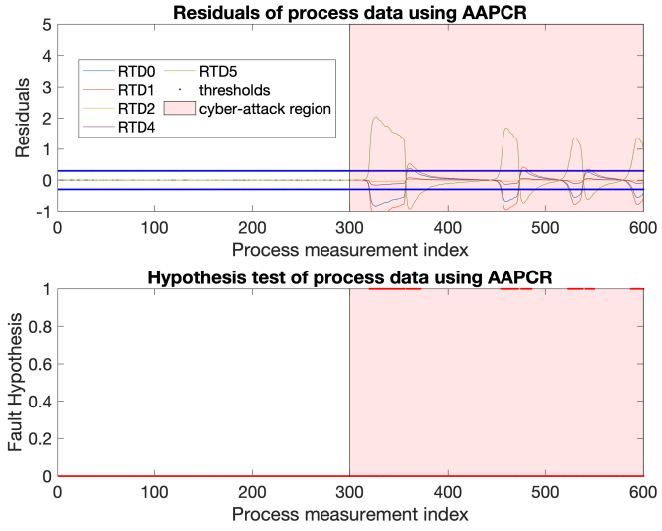


Figure 6.12: Attack detection based on process data using AAPCR model under the serial false data injection attack.

Table 6.5: RMSE of test data (^{o}C)

Model	RTD0	RTD1	RTD2	RTD4	RTD5
AAKR	0.0877	0.2351	0.2294	0.3618	0.1703
AASVR	0.0463	0.1544	0.1786	0.0973	0.1112
Ensemble regression	0.0108	0.0152	0.0486	0.0242	0.0176
AAPCR	0.0034	0.0048	0.0002	0.0007	0.0085

delay is 72 seconds. Ensemble regression model and AAPCR model have a similar first detection time. Therefore, AASVR has the best performance in terms of the early detection, which gives the cause analysis system, cyber-attack response system and operator precious time to produce the analysis results and respond to mitigate the consequences of the cyber-attack. It is worthwhile to mention that as introduced in Chapter 3, Argus summarizes the network flow information as it occurs; these data do not have a consistent sample rate, but are timestamped to allow time resolution of anomaly and attack detection. Conversely, LabVIEW collects data at a fixed sample rate (in this case 0.25 Hz). Therefore, the observation indices on the abscissa in Figures 6.8, 6.9 and 6.10 do not directly corresponds to each other, but the system can relate them by tracing back the time an observation occurred.

6.2.5 M3 models comparison in the serial false data injection

Table 6.5 shows the RMSE of the fault-free test data for RTD0, RTD1, RTD2, RTD4, and RTD5, which shows that the AAPCR gives the best results with lowest RMSE. The predicted value and real value of the test data under normal operation using AAKR, AASVR, ensemble regression, and AAPCR are shown in Figure 6.13. Both the RMSE and the prediction plots of test data show that the AASVR and AAPCR models yield good predictions based on normal operation data.

A model with smaller sensitivity has better robustness. The false data injection impacted most of the sensors, which means it changed most of the variables in the model to faulty variables. Obviously, a robust model is needed in cyber-attack detection, so that the prediction under the cyber-attack would not be impacted too much by the cyber-attack. Otherwise, the prediction may be dragged towards the fault values and the detection will

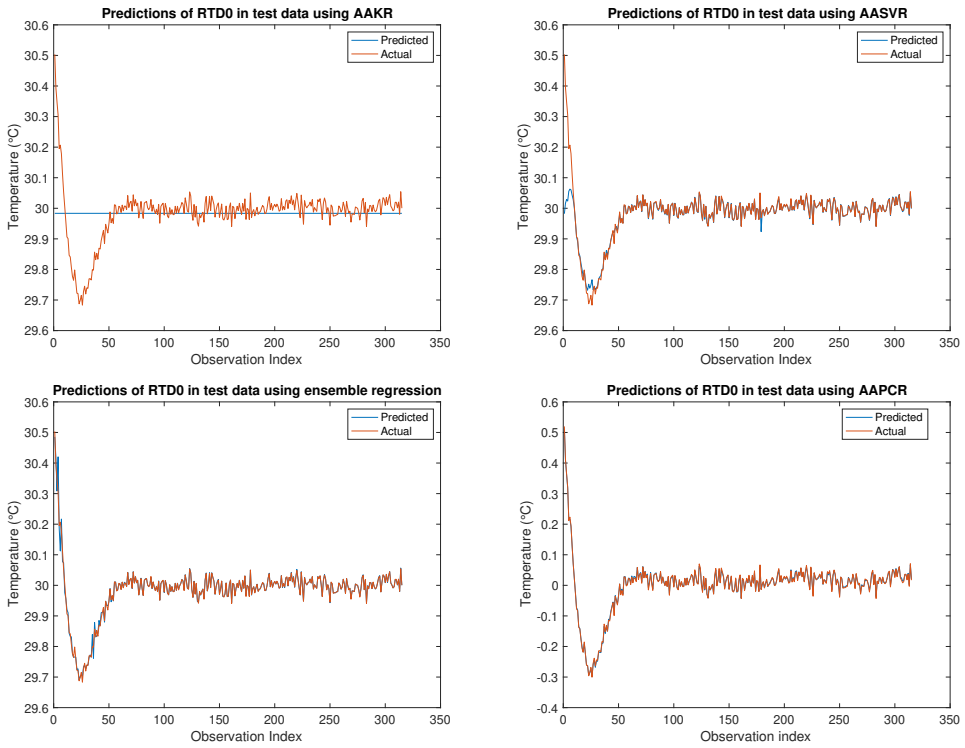


Figure 6.13: Prediction of the test data under normal operation using AAKR, AASVR, ensemble regression, and AAPCR model.

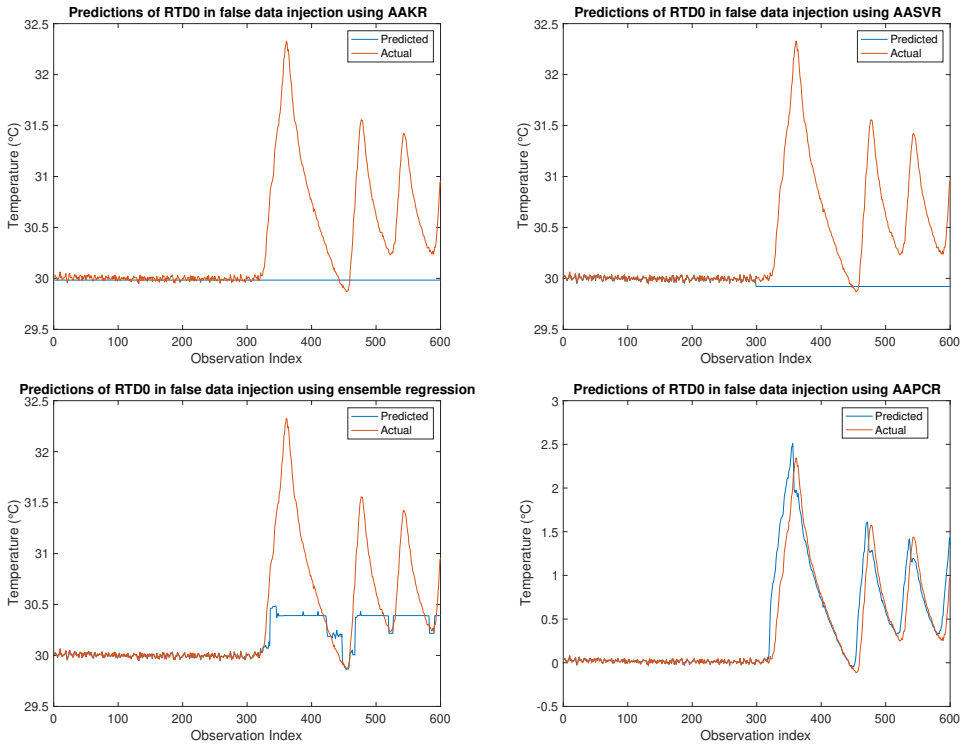


Figure 6.14: Prediction of data under the false data injection using AAKR, AASVR, ensemble regression, and AAPCR model.

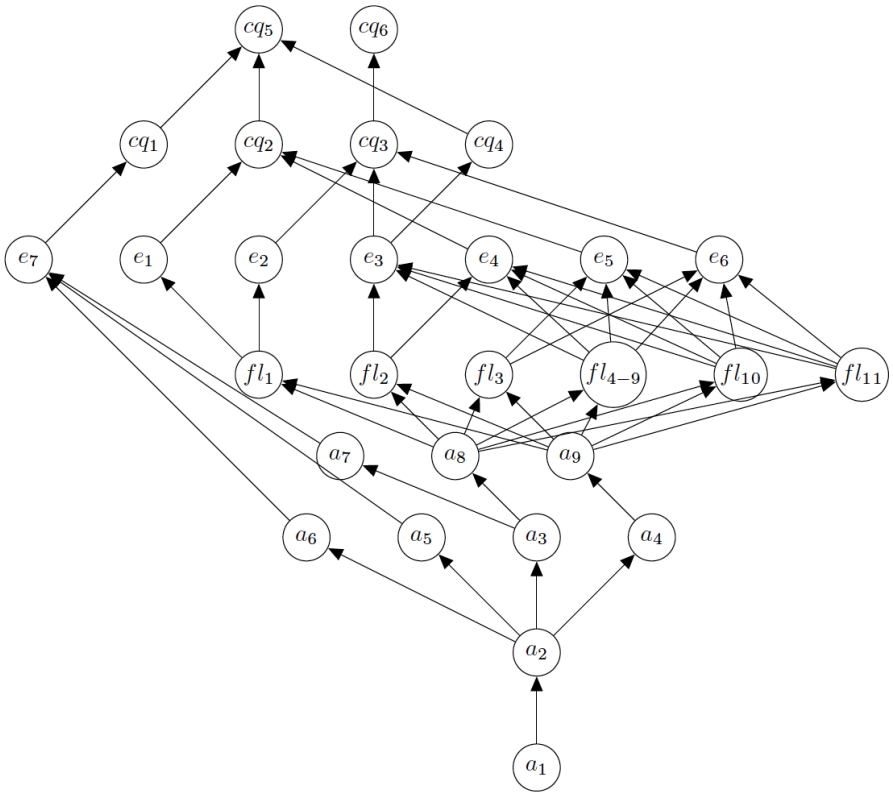
fail. Since it is impossible to know which variable the attacker is going to attack in advance, it is impossible to exclude faulty sensors in the detection models. Figure 6.14 shows the prediction of the AAKR, AASVR, ensemble regression, and AAPCR under the false data injection scenario in subplot, respectively, which shows that the AASVR is robust to the faulty input while ensemble regression and AAPCR are not. The prediction of the ensemble regression and AAPCR are dragged by the faulty sensors with different levels and would not be able to detect the cyber-attack effectively. Therefore, AASVR is the best robust model in M3 under this scenario considering time to detection, accuracy, and sensitivity.

6.3 Results of Cause Analysis System

The results of the cyber-attack detection system were fed into the cause analysis system. In this scenario, the results of the cyber-attack detection system M1 that yielded no alarm, M2 that detected a MITM attack, and M3 that detected an anomaly in all RTDs were fed

into subsystems in the cause analysis system at the same time: the event discriminator and the dynamic risk assessment module. Since M2 detected MITM abnormal behavior in the cyber data prior to M3 detecting an anomaly in all the RTDs, and the time interval is less than T_{max}, according to the predefined logic, the event discriminator categorized this event as a cyber event. In addition, the residuals of RTDs given by M3 did not follow a degradation trend in the database and all five RTDs gave anomalies which did not match degradation or equipment fault behavior. Therefore, the discriminator gave the decision that the event was a cyber event with 100% confidence. If no cyber abnormal was detected within T_{max} interval, the Bayesian model is utilized to infer if the event is more likely to be a cyber event or an operational event.

The dynamic risk assessment module received this decision as well and fed all the information to the Bayesian Network. Risk can be generated with expert knowledge and assets data base. Figure 6.15 shows the dynamic Bayesian network for the ICS testbed. There are four types of nodes: a nodes are cyber-attack nodes, which are the purely IT cyber-attack steps that an attacker launched; fl nodes are function loss nodes, which show what actions an attacker performed to change the physical process and indicate cyber-physical attacks; e nodes show the direct effects of the cyber-physical attacks on the OT side; cq nodes show the physical consequences based on the effects. This Bayesian network is able to update the probability dynamically with the status of nodes changing. For example, If a_2 was detected, the probability that each node would happen is calculated based on the CPT first, the probability of the node a_4 happen is 0.4. Then when the detection system detected a_4, the probability of a_4 becomes 1. Another advantage of the Bayesian network is that it gives a direct visual explanation of the casual relationship which is useful for cyber event forensics.



Legend of nodes:

a_1 Network scanning

a_2 Vulnerability scanning

a_3 MITM between engineering workstation and cDAQ9188

a_4 Use authentication vulnerability on engineering workstation

a_5 DoS attack engineering workstation

a_6 DoS attack cDAQ9188

a_7 Delete packets

a_8 Tamper packets

a_9 Obtain administrator authority of engineering workstation

fl_1 Heater power control

fl_2 VFD frequency control

fl_3 MOV position control

fl_4 Tamper RTD0

fl_5 Tamper RTD1

fl_6 Tamper RTD2

fl_7 Tamper RTD3

fl_8 Tamper RTD4

fl_9 Tamper RTD5

fl_{4-9} Tamper any RTD

fl_{10} Tamper Coriolis flow meter

fl_{11} Magnetic flow meter

e_1 Heater power increase

e_2 Heater power reduce

e_3 VFD frequency increase

e_4 VFD frequency reduce

e_5 MOV opening reduce

e_6 MOV opening increase

e_7 Devices remain the last status

cq_1 Operator lose control of the flow loop

cq_2 Flow loop overheat

cq_3 Flow loop run on low power

cq_4 Pump damage

cq_5 Pipe break

cq_6 Insufficient load

Figure 6.15: Bayesian network for the ICS testbed.

Table 6.6: Three attack scenarios conducted towards SG in the HIL testbed

Scenario No.	Attack description	PLC input	PLC action and final SG water level
scenario I	Overwrite the SG level measurement to 15.9m	15.9m	Reduces pump speed constantly until SG dry out with 0m
scenario II	Inject a malicious code to add a 0.9m to the SG water level measurement	X+0.9m	Reduces pump speed until the SG level reaches 14.1m
scenario III	Simultaneous attack: alter the water level set point to 14m; mask the SG water level with 15m	14m eventually	Reduces pump speed until the SG level reaches 14m

6.4 Results of Localized Cybersecurity Models for Key Components in HIL Testbed

This section first presents the detection results of the localized AAKR model under the three false data injection attack scenarios on the HIL testbed described to evaluate the effectiveness of the proposed localized cybersecurity model. Then the preliminary result of a SVR inference model under scenario I is presented and discussed.

Table 6.6 shows the attack description of the three attack scenarios, as well as the PLC input, action and the final SG water level. Under the attack scenario I, the input of the PI controller is overwritten to 15.9m constantly by the attacker, causing the PLC to to think that the water level is higher than the set point of 15m and compensating by keeping the pump speed low. Therefore, the real SG water level keeps reducing until SG dry out with zero water level and the simulator crashes. Under scenario II, the PLC reduces the pump speed until the SG water level measurement is reduced to 14.1m due to the PI controller input experiencing a 0.9m increase from the malicious code. Under scenario III, the PLC reduces the pump speed until the SG water level measurement is reduce to 14m, which is the set point altered by attacker.

Five state variables collected in the PLC are fed into the AAKR model, including feedwater pump speed command, reactor power, SG inlet flow rate, SG water level, and SG pressure. To simulate advanced attacks and demonstrate the robustness of this localized

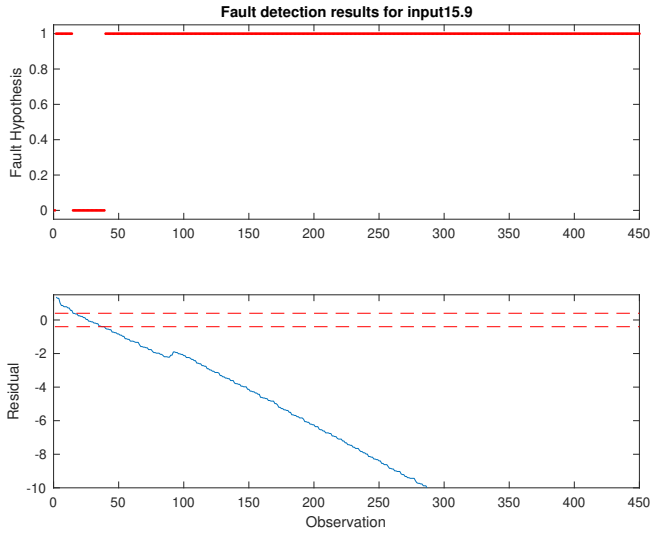


Figure 6.16: Fault detection results of scenario I: water level is tampered to 15.9 constantly.

model, the attacker altered the SG level values to 15m (normal value) in the display to the operator to mask the attack from operators in all three scenarios. Figures 6.16, 6.17, and 6.18 show the fault detection results of the scenario I, II, and III, respectively. In each figure, the bottom subplot shows the residual of a signal with blue lined and the thresholds with the red lines; the upper subplot shows the fault hypothesis based on the residual and the thresholds relationship. When the residual exceeds the threshold, the fault hypothesis is "1" and indicates a fault state while "0" means the residual is within the threshold and indicates a normal state. Different threshold values are assigned to three different signals: SG inlet flow rate, SG level, and SG pressure. The final alarms are the combination of alarms generated from all three of these signals. Figure 6.16, 6.17, and 6.18 only show the signal with the most significant residual to give a clear display: SG pressure is used in Figure 6.16, SG inlet flow rate is used in Figure 6.17, and SG water level is used in Figure 6.18. Figures 6.16, 6.17, and 6.18 show that this localized strategy can detect the cyber-attack effectively.

Once the anomaly is detected, the measurements of control variables used for controller input are not trusted any more since they are likely compromised. In this research, the SG water level is the control variable and is determined as compromised. Then the inference model is utilized to predict the real measurement and then fed into the controller as input instead to generate the correct command. Theoretically, SG water level may be inferred

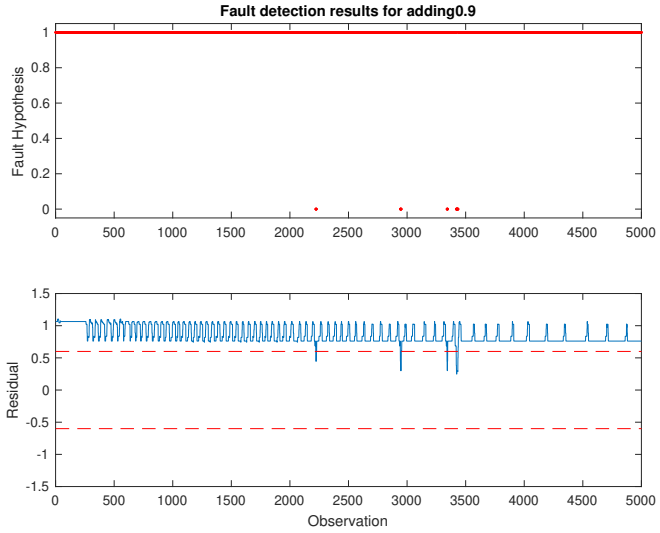


Figure 6.17: Fault detection results of scenario II: 0.9m was added to water level input.

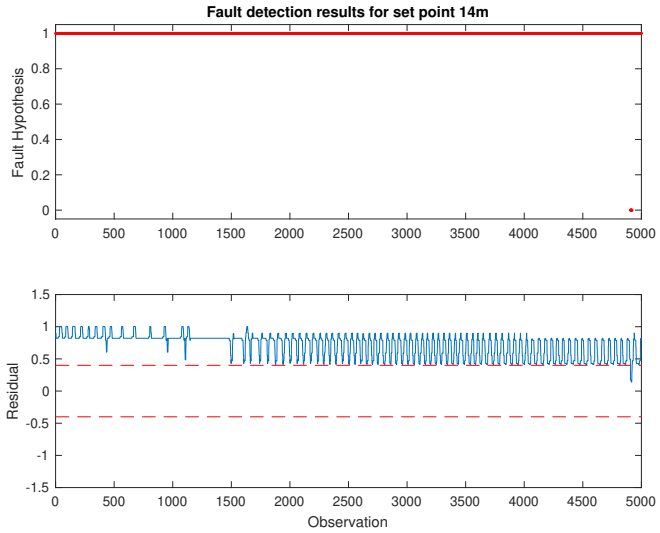


Figure 6.18: Fault detection results of scenario III: water level set point is tampered to 14m.

from SG-related variables by a pre-trained regression model, such as SG inlet flow, SG pressure, and SG related temperature values. Thus, a SVR model is built that uses the PLC speed command, reactor power and SG pressure to predict the real SG water level. This variable selection is based on variable availability in the Asherah SIMULINK model and the relationship between variables.

In scenario I, the adversary team altered the water level measurement input to the PLC to 15.9m constantly but masked the values shown to the operator with a normal 15m display. Figure 6.19 shows the compromised water level, real water level extracted from the Asherah model, and inference results. The green line shows a constant value of 15.9m which is a compromised measurement that the PLC received. Since the PLC got a positive difference between water level measurement and the reference level, it maintained a low feedwater pump speed to bring down the water level measurement. However, the actual water level kept reducing as shown in the red line, until the SG dried out in this testbed because the PLC constantly received 15.9m as input. Under this advanced attack scenario, once the fault was detected in the beginning of the attack as shown in Figure 6.16, the control sensor SG water level could no longer be trusted. Then the inference value from the inference model could be utilized as a virtual sensor. The blue line in 6.19 shows the predicted value by the inference model. This gives the correct prediction during the first 100 observations when the SG water level drops from 15m to about 13m. After that, the inference model gives an incorrect prediction. This is believed to be due to the limitations of the Asherah simulator, which yields insufficient or low-fidelity data in the current HIL testbed. Not all SG-related variables which are measured in real NPPs are available or well modeled in the Asherah model, such as SG steam related variables. Additionally, SG pressure does not change as water level changes under part of the simulated attacks, which is not indicative of physical operation. This lead to insufficient or incorrect data for algorithms to learn. Finally, the "new" data generated under a cyber-attack may fall outside the range of training data. For the trained SVR (and several other regression models), the training and test MSE under normal operation are reasonably small which indicates the model is able to handle the regression within the training data range. However, when the models were applied to the "new" data under attacks, the models could not infer the real water level correctly.

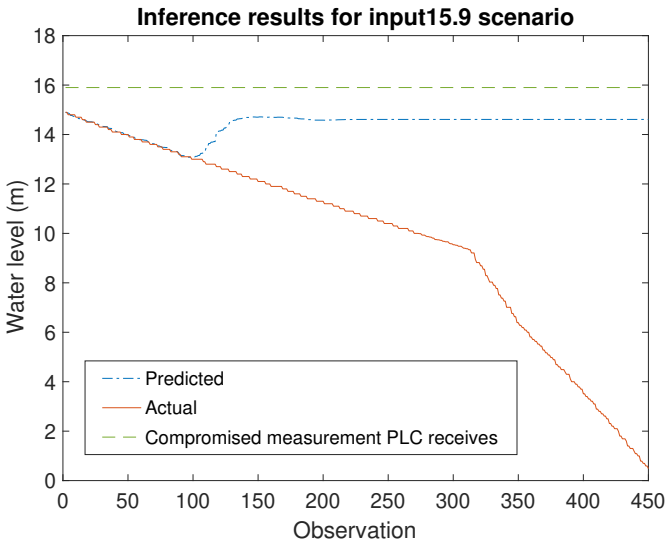


Figure 6.19: Inference results under scenario I: attacker alters the water level measurement input to PLC to 15.9m constantly

Since any model is not meant to do extrapolation (unless the process is quite stable, which is not the case under any anomaly situation), the inference is not correct. The first two potential reasons can be fixed by conducting research on a high fidelity HIL testbed, which is planned in the near future based on this best practice HIL tesbed. The last reason may be more difficult for the inference model application, because in order to build a comprehensive training data set, all possible operations need to be simulated to produce data for algorithms to learn on, which may be impossible. Especially, different from equipment degradation or fault which may still yield sensor measurements with a relationship based on underlying physical processes, under cyber-attacks, the sensors may be compromised and altered to an unlimited range of values. Another limitation of the inference model is that if the attackers alter several sensors simultaneously, the inference will likely be incorrect.

Chapter 7

Conclusion and Future Work

ICS-targeted cyber-attack are growing fast with more complexity. It is important to enhance the cybersecurity of ICSs to ensure safe and economic operations in industry. This paper proposed two major research focuses to facilitate the challenging cybersecurity problems especially brought by the increasing use pf digital ICSs: a robust cybersecurity solution platform to improve the resilience to cyber-attacks overall, and a localized cybersecurity strategy to enhance robustness to cyber-attacks locally.

The cybersecurity solution platform contains five sub-systems to perform cybersecurity functions with a defense-in-depth concept. The data collection and extraction system collects different types of cyber data as well as process data and parses the useful features for subsequent analysis. The extracted features are fed into a multi-layer cyber-attack detection system, which contains three detection modules to perform supervised and unsupervised data detection for both known and unknown cyber-attacks. The results of the cyber-attack detection system are fed into a cause analysis system with a event discriminator, a dynamic risk assessment module and an attack identification module. The event discriminator distinguish between cyber events and operational events since the response to these two could be quite different. The dynamic risk assessment module utilizes a Bayesian network and fuzzy probability to provide dynamic risk assessment of documents, equipment, human assets, and the environment. The attack identification module infers the compromised points location. The results of the risk and identification module are the input of the cyber-attack

response system and the operator alert system to facilitate decision making. The cyber-attack response system and the operator alert system are not developed in this dissertation due to the research scope. This platform combines the cyber data which is currently under monitoring by the IT team in the NPPs and the process data which is monitored by power plant operators. The advantages of the process data integration includes providing independent data sources for attack detection, early stage indication of compromise before severe physical damage, and provides information for attack target identification by pointing out the fault signals. This approach promotes better cooperation between OT and IT team.

The localized cybersecurity strategy could be implemented into the controller itself locally, which significantly reduces the attack surface. Unsupervised data-driven models are applied to detect the fault state caused by false data injection attacks. Once the attack or anomaly is detected, the control sensor measurement is not trusted anymore and is inferred by a pre-trained supervised inference model with other state variables. The inference is utilized as a virtual sensor instead of the control sensor as the input of the controller to generate the correct command. This method could aid in the resilient operation of key equipment under some operational events like sensor failure as well.

To generate data to develop the models and evaluate the effectiveness of the developed platform, an ICS testbed was built. The ICS testbed contains a physical experimental facility to simulate a two-loop nuclear reactor thermal hydraulic system, a SCADA system to perform the data acquisition and control, and some cyber equipment and a LAN to provide a realistic mock-up of an ICS and data collection capability. Different types of cyber-attacks including MITM, DoS, and false data injection were conducted towards the testbed to demonstrate the possibility of cyber-attack and generate data for cybersecurity research. Four types of data were collected by data collection and extraction system under both normal operation and cyber-attack scenarios: host-system data, network data, network flow data, and process data. Supervised and unsupervised cyber-attack detection models were built using these four types of the data. A serial false data injection attack was conducted and utilized to demonstrate the effectiveness of the proposed cybersecurity solution platform. The results show that the data collection and extraction system, cyber-attack detection system, and cause analysis system are effective to enhance the cybersecurity of ICSs.

To develop and evaluate the proposed localized cybersecurity strategy, a HIL testbed consisting of an NPP simulator and a PLC was built. The PLC is programmed to control the SG water level at the desired set point. It takes the SG water level measurement as input and outputs commands to adjust the feedwater pump speed. Three false data injection attacks were launched towards the HIL testbed, in which the attacker altered the SG water level measurement or set point to mislead the PLC to issue a false feedwater pump speed command, which in turn lead the NPP to an unplanned operation state. Meanwhile, the attacker masked the compromised SG water level measurement from operators by overwriting the SG water level using normal measurements. An unsupervised AAKR model was constructed using five SG related state variables as inputs to detect the abnormal operation state of the SG. Once the anomaly was detected, the SG water level measurement was inferred from other state variables and was utilized as the input of the PLC to issue correct command. The results show that this strategy improves the resilience of the key equipment SG to the cyber-attack.

This dissertation serves as proof-of-concept research for both the cybersecurity solution platform and localized strategy and provides best practices of building the cybersecurity testbeds. The development of each subsystem or module was limited by the resource available. In the cybersecurity solution platform, the data collection and extraction system and cyber-attack detection system are well developed and evaluated. However, these two systems are facility specific, which means new systems have to be constructed based on the specific application in the future. The framework of the cause analysis system is given in this research. However, due to lack of ground truth of the risk assessment, it is hard to evaluate the proposed dynamic risk assessment model, which is an issue currently in all risk assessment research. The cyber-attack response system and operator alert system is not studied in this research since experts in facility operation and human factors have to be involved. Therefore, a HIL testbed with a high fidelity NPP simulator (utilized for human factor research), PLCs, and simulated network architecture with different security levels will be constructed and cooperation with experts from various domain will be used to perform further research on the cybersecurity solution platform and localized cybersecurity strategy.

For future developments on the proposed cybersecurity solution platform and localized cybersecurity strategy, a HIL testbed with high-fidelity commercial NPP simulator, S7-1500 PLCs, and plant data acquisition and control workstation will be constructed. Realistic traffic between the PWR simulator and PLCs will be built to mimic traffic expected in a typical NPP I&C system. This HIL testbed will provide the capability to conduct the following future work:

1. Define and design cyber-attacks of interest to generate cyber data and process data.

2. Test the effectiveness of the cyber-attack detection models developed in this dissertation and continue investigating and developing new algorithms for cyber-attack detection.

3. Further develop and evaluate the event discriminator to attribute detected anomalies to either cyber or physical causes to provide actionable information to engineers and operators.

4. Further develop the dynamic risk assessment module with experts in both IT and OT sides and evaluate it with new HIL testbed.

5. Develop the cyber-attack response system and operator alert system with cybersecurity and human factors experts.

6. Further develop and evaluate the anomaly detection models and inference models for a localized cybersecurity strategy based on high-fidelity data.

7. Implement the cyber-attack detection and inference models into the S7-1500 PLC without impacting the original PLC function and identify what are the limitations and benefits.

Bibliography

[1] Alberts, C., Dorofee, A., Stevens, J., and Woody, C. (2003). Introduction to the octave approach. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST. 37

[2] Allison, D. (2019). Utilising the Multi-Functional Platform of the Siemens S7-1518 MFP Programmable Logic Controller for Security Applications. Master's thesis, Queen's University Belfast. 46

[3] Bathelt, A., Ricker, N. L., and Jelali, M. (2015). Revision of the tennessee eastman process model. *IFAC-PapersOnLine*, 48(8):309–314. 41

[4] Biham, E., Bitan, S., Carmel, A., Dankner, A., Malin, U., and Wool, A. (2019). Rogue7: Rogue engineering-station attacks on s7 simatic plcs. In *Black Hat USA 2019*. Blackhat. 45

[5] Bochtler, J., Quinn, E., and Bajramovic, E. (2017). Development of a new iec standard on cybersecurity controls for i&c in nuclear power plants–iec 63096. *NPIC & HMIT*. xiii, 23, 24

[6] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140. 30

[7] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. 31

[8] Buczak, A. L. and Guven, E. (2016a). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2):1153–1176. 25, 30, 35

[9] Buczak, A. L. and Guven, E. (2016b). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2):1153–1176. 60

[10] Byres, E. J., Franz, M., and Miller, D. (2004). The use of attack trees in assessing vulnerabilities in scada systems. In *Proceedings of the international infrastructure survivability workshop*, pages 3–10. Citeseer. 37, 39

[11] Cannady, J. (1998). Artificial neural networks for misuse detection. In *National information systems security conference*, volume 26. Baltimore. 28

[12] Caswell, B. and Beale, J. (2004). *Snort 2.1 intrusion detection*. Elsevier. 29

[13] Center for strategic & international studies (2019). Significant cyber incidents. https://www.csis.org/programs/technology-policy-program/significant-cyber-incidents. 2, 16

[14] Chawla, A., Lee, B., Fallon, S., and Jacob, P. (2018). Host based intrusion detection system with combined cnn/rnn model. In *Proceedings of Second International Workshop on AI in Security*. 29

[15] Chen, B., Pattanaik, N., Goulart, A., Butler-Purry, K. L., and Kundur, D. (2015). Implementing attacks for modbus/tcp protocol in a real-time cyber physical system test bed. In *Communications Quality and Reliability (CQR), 2015 IEEE International Workshop Technical Committee on*, pages 1–6. IEEE. 8

[16] Chen, T. M., Sanchez-Aarnoutse, J. C., and Buford, J. (2011). Petri net modeling of cyber-physical attacks on smart grid. *IEEE Transactions on Smart Grid*, 2(4):741–749. 37

[17] Coble, J., Ramuhalli, P., Meyer, R., and Hashemian, H. (2013). Online sensor calibration assessment in nuclear power systems. *IEEE Instrumentation & Measurement Magazine*, 16(3):32–37. 7

[18] Coble, J. B., Meyer, R. M., Ramuhalli, P., Bond, L. J., Hashemian, H., Shumaker, B., and Cummins, D. (2012). A review of sensor calibration monitoring for calibration interval extension in nuclear power plants. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States). 7

[19] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297. 66

[20] Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161. 67

[21] East, S., Butts, J., Papa, M., and Shenoi, S. (2009). A taxonomy of attacks on the dnp3 protocol. In *International Conference on Critical Infrastructure Protection*, pages 67–81. Springer. 15

[22] EGGERS, S. L. (2018). *Adapting Anomaly Detection Techniques For Online Intrusiondetection In Nuclear Facilities*. PhD thesis, University of Florida. 34

[23] Fanti, M. P., Iacobellis, G., and Ukovich, W. (2015). A risk assessment framework for hazmat transportation in highways by colored petri nets. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):485–495. 37

[24] Gaikwad, D. and Thool, R. C. (2015). Intrusion detection system using bagging with partial decision treebase classifier. *Procedia Computer Science*, 49:92–98. 30, 32

[25] Gawand, H. L., Bhattacharjee, A., and Roy, K. (2017). Securing a cyber physical system in nuclear power plants using least square approximation and computational geometric approach. *Nuclear Engineering and Technology*, 49(3):484–494. 33

[26] Goh, J., Adepu, S., Junejo, K. N., and Mathur, A. (2016). A dataset to support research in the design of secure water treatment systems. In *International Conference on Critical Information Infrastructures Security*, pages 88–99. Springer. 32

[27] Goh, J., Adepu, S., Tan, M., and Lee, Z. S. (2017). Anomaly detection in cyber physical systems using recurrent neural networks. In *2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)*, pages 140–145. IEEE. 32

[28] Gross, K. C., Singer, R. M., and Humenik, K. E. (1993). Expert system for online surveillance of nuclear reactor coolant pumps. US Patent 5,223,207. 7

[29] Henry, M. H., Layer, R. M., Snow, K. Z., and Zaret, D. R. (2009). Evaluating the risk of cyber attacks on scada systems via petri net analysis with application to hazardous liquid

loading operations. In *2009 IEEE Conference on Technologies for Homeland Security*, pages 607–614. IEEE. 37

[30] Hines, J., Garvey, D., Seibert, R., Usynin, A., and Arndt, S. (2008). Technical review of on-line monitoring techniques for performance assessment (nureg/cr-6895) vol. 2, theoretical issues. *Published May*. 64

[31] Hines, J. W. and Seibert, R. (2006). Technical review of on-line monitoring techniques for performance assessment. volume 1. state-of-the-art. 74

[32] Hodo, E., Bellekens, X., Hamilton, A., Dubouilh, P.-L., Iorkyase, E., Tachtatzis, C., and Atkinson, R. (2016). Threat analysis of iot networks using artificial neural network intrusion detection system. In *2016 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–6. IEEE. 28

[33] Homeland Security Centre for the Protection of National Infrastructure (2011). Cyber security assessments of industrial control systems. https://www.ccn-cert.cni.es/publico/InfraestructurasCriticaspublico/CPNI-Guia-SCI.pdf. 6, 7

[34] House, W. (1998). White paper: The clinton administration's policy on critical infrastructure protection: Presidential decision directive 63. *White Paper*. 1

[35] IAEA (2011a). Computer security at nuclear facilities. https://www-pub.iaea.org/books/iaeabooks/8691/Computer-Security-at-Nuclear-Facilities. 19

[36] IAEA (2011b). Nuclear security recommendations on physical protection of nuclear material and nuclear facilities. https://www-pub.iaea.org/MTCD/Publications/PDF/Pub1481_web.pdf. 19

[37] IAEA (2013). Objective and essential elements of a state's nuclear security regime. https://www-pub.iaea.org/books/iaeabooks/10353/objective-and-essential-elements-of-a-state-s-nuclear-security-regime. 19

[38] IAEA (2014). Computer security of instrumentation and control systems at nuclear facilities-draft technical guide. http://www-ns.iaea.org/downloads/security/security-series-drafts/tech-guidance/nst036.pdf. 22

[39] IAEA (2015). Security of nuclear information. https://www.iaea.org/publications/10774/security-of-nuclear-information. 19

[40] IAEA (2016). Computer security for nuclear security-draft implementing guide. https://www-ns.iaea.org/downloads/security/security-series-drafts/implem-guides/nst045.pdf. 19, 22

[41] IAEA (2018). Computer security of instrumentation and control systems at nuclear facilities. https://www-pub.iaea.org/books/iaeabooks/11184/Computer-Security-of-Instrumentation-and-Control-Systems-at-Nuclear-Facilities. 19

[42] ICS Alert (2016). Cyber-attack against ukrainian critical infrastructure. https://www.us-cert.gov/ics/alerts/IR-ALERT-H-16-056-01. 3, 15

[43] ICS-cert (2017). Mar-17-352-01 hatman - safety system targeted malware (update a). https://ics-cert.us-cert.gov/MAR-17-352-01-HatMan-Safety-System-Targeted-Malware-Update. 3

[44] ICSCER Team and others (2016). Ics-cert year in review. https://ics-cert.uscert.gov/Year-Review-2016. 2

[45] IEC (2014). Iec 62645:2014: Nuclear power plants - instrumentation and control systems - requirements for security programmes for computer-based systems. https://webstore.iec.ch/publication/7311. 23

[46] IEC (2016). Iec 62859:2016: Nuclear power plants - instrumentation and control systems - requirements for coordinating safety and cybersecurity. https://webstore.iec.ch/publication/26131. 23

[47] Inayat, I., Amin, R. u., and Ali, M. M. (2018). Hardware-in-the-loop flight motion simulation of flexible variable sweep aircraft. *Aircraft Engineering and Aerospace Technology*, 90(9):1428–1437. 95

[48] Indegy (2019). A historical timeline of cyber attacks on critical infrastructure. https://info.indegy.com/indegy-infographic-timeline-cyber-attacks-lp. 2, 16

[49] ISA (2019). Isa99, industrial automation and control systems security. https://www.isa.org/isa99/. 18

[50] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer. 60, 62, 69

[51] Jensen, F. V. et al. (1996a). *An introduction to Bayesian networks*, volume 210. UCL press London. 29

[52] Jensen, F. V. et al. (1996b). *An introduction to Bayesian networks*, volume 210. UCL press London. 39

[53] Ji, X., Yu, H., Fan, G., and Fu, W. (2016). Attack-defense trees based cyber security analysis for cpss. In *2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 693–698. IEEE. 37

[54] Jin, C., Valizadeh, S., and van Dijk, M. (2018). Snapshotter: Lightweight intrusion detection and prevention system for industrial control systems. In *2018 IEEE Industrial Cyber-Physical Systems (ICPS)*, pages 824–829. IEEE. 45

[55] Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303. 69

[56] Kaplan, S. and Garrick, B. J. (1981). On the quantitative definition of risk. *Risk analysis*, 1(1):11–27. 36

[57] Kaspersky Lab ICS-CERT (2018). Threat landscape for industrial automation systems in h2 2017. 3

[58] KDD (1999). Kdd cup 1999 data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. 31

[59] Kesler, B. (2011). The vulnerability of nuclear facilities to cyber attack; strategic insights: Spring 2010. *Strategic Insights, Spring 2011*. 5, 16

[60] Kiss, I., Genge, B., and Haller, P. (2015). A clustering-based approach to detect cyber attacks in process control systems. In *2015 IEEE 13th international conference on industrial informatics (INDIN)*, pages 142–148. IEEE. 33

[61] Klick, J., Lau, S., Marzin, D., Malchow, J.-O., and Roth, V. (2015). Internet-facing plcs as a network backdoor. In *2015 IEEE Conference on Communications and Network Security (CNS)*, pages 524–532. IEEE. 45

[62] Knop, M., Schopf, J., and Dinda, P. (2002). Windows performance monitoring and data reduction using watchtower. In *11th IEEE Symposium on High-Performance Distributed Computing (HPDC11)*, volume 35. 59, 86

[63] Krebs, B. (2008). Cyber incident blamed for nuclear power plant shutdown. *Washington Post, June*, 5:2008. 5

[64] Kruegel, C., Mutz, D., Robertson, W., and Valeur, F. (2003). Bayesian event classification for intrusion detection. In *IN: PROCEEDINGS OF ACSAC 2003, LAS VEGAS, NV*, page 14. 29, 32

[65] Kruegel, C. and Toth, T. (2003). Using decision trees to improve signature-based intrusion detection. In *International Workshop on Recent Advances in Intrusion Detection*, pages 173–191. Springer. 29, 32

[66] Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, volume 26. Springer. 68, 69

[67] Langner, R. (2011). Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy*, 9(3):49–51. 2, 5, 17

[68] Lee, C.-K. (2013). Introduction of a cyber security risk analysis and assessment system for digital i&c systems in nuclear power plants. *IFAC Proceedings Volumes*, 46(9):2140–2144. 43

[69] Li, J. and Huang, X. (2016). Cyber attack detection of i&c systems in npps based on physical process data. In *2016 24th International Conference on Nuclear Engineering*, pages V002T07A011–V002T07A011. American Society of Mechanical Engineers. xiii, 33, 50, 51

[70] Li, P. and Mirchandani, P. B. (2016). A new hardware-in-the-loop traffic signal simulation framework to bridge traffic signal research and practice. *IEEE Transactions on Intelligent Transportation Systems*, 17(9):2430–2439. 95

[71] Liao, Y. and Vemuri, V. R. (2002). Use of k-nearest neighbor classifier for intrusion detection. *Computers & security*, 21(5):439–448. 30, 32

[72] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22. 60, 63

[73] Major, J. A. (2002). Advanced techniques for modeling terrorism risk. *The Journal of Risk Finance*, 4(1):15–24. 38

[74] Mallouhi, M., Al-Nashif, Y., Cox, D., Chadaga, T., and Hariri, S. (2011). A testbed for analyzing security of scada control systems (tasscs). In *Innovative Smart Grid Technologies (ISGT), 2011 IEEE PES*, pages 1–7. IEEE. 8

[75] Manharlal, S. V. (2018). *Enhancing Performance Of Intrusion Detection Systems Using Data Fusion Techniques*. PhD thesis, Gujarat Technological University. 31

[76] Maniatopoulos, M., Lagos, D., Kotsampopoulos, P., and Hatziargyriou, N. (2017). Combined control and power hardware in-the-loop simulation for testing smart grid control algorithms. *IET Generation, Transmission & Distribution*, 11(12):3009–3018. 95

[77] Mayyas, A. R., Kumar, S., Pisu, P., Rios, J., and Jethani, P. (2017). Model-based design validation for advanced energy management strategies for electrified hybrid power

trains using innovative vehicle hardware in the loop (vhil) approach. *Applied Energy*, 204:287–302. 95

[78] McHugh, J. (2000). Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Transactions on Information and System Security (TISSEC)*, 3(4):262–294. 29

[79] McQueen, M. A., Boyer, W. F., Flynn, M. A., and Beitel, G. A. (2006). Quantitative cyber risk reduction estimation methodology for a small scada control system. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 9, pages 226–226. IEEE. 38

[80] Meyer, R. M., Bond, L. J., and Ramuhalli, P. (2014). Online condition monitoring to enable extended operation of nuclear power plants. In *Progress of Nuclear Safety for Symbiosis and Sustainability*, pages 161–176. Springer. 7

[81] Morris, T., Vaughn, R., and Dandass, Y. S. (2011). A testbed for scada control system cybersecurity research and pedagogy. In *Proceedings of the Seventh Annual Workshop on Cyber Security and Information Intelligence Research*, page 27. ACM. 8

[82] Nader, P., Honeine, P., and Beauseroy, P. (2014). l_p-norms in one-class classification for intrusion detection in scada systems. *IEEE Transactions on Industrial Informatics*, 10(4):2308–2317. 33

[83] Neitzel, L. and Huba, B. (2014). Top ten differences between ics and it cybersecurity. *InTech*, 61(3):12–18. 6

[84] NIST (2019). Computer security resource center. https://csrc.nist.gov/publications/sp800. 18

[85] NRC (2017). The pressurized water reactor (pwr) overview. https://www.nrc.gov/reading-rm/basic-ref/students/animated-pwr.html. xiii, 78, 79

[86] NRC (2019). Nrc: Digital instrumentation and control. https://www.nrc.gov/reactors/digital.html. 1, 2

[87] Nuclear Energy Institute (2010). Nei 08-09 rev 6 compliance for nuclear facilities. https://logrhythm.com/solutions/compliance/nei-08-09-rev-6. 25

[88] Nuclear Energy Institute (2011). Addressing cyber security controls for nuclear power reactors. 25

[89] Nuclear Energy Institute (2015). Cyber security control assessments. 25

[90] Nuclear Energy Institute (2016). Cyber security event notifications. 25

[91] OFFICE of the DIRECTOR of NATIONAL INTELLIGENCE (2019). Cyber threat framework. https://www.dni.gov/index.php/cyber-threat-framework. xiii, 4, 5

[92] Orojloo, H. and Azgomi, M. A. (2014). A method for modeling and evaluation of the security of cyber-physical systems. In *2014 11th International ISC Conference on Information Security and Cryptology*, pages 131–136. IEEE. 37

[93] Peng, Y., Huang, K., Tu, W., and Zhou, C. (2018). A model-data integrated cyber security risk assessment method for industrial control systems. In *2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS)*, pages 344–349. IEEE. 37, 41

[94] Poolsappasit, N., Dewri, R., and Ray, I. (2012). Dynamic security risk management using bayesian attack graphs. *IEEE Transactions on Dependable and Secure Computing*, 9(1):61–74. 37

[95] Pritchett, W. L. and De Smet, D. (2013). *Kali Linux Cookbook*. Packt Publishing Ltd. 86

[96] Prosys OPC (2019). Prosys opc. https://www.prosysopc.com/products/opc-ua-simulation-server/. 97

[97] Qosient, L. (2004). Argus. *see: http://www. qosient. com/argus/index. htm*. 49, 86

[98] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106. 29

[99] Rankin, D. J. and Jiang, J. (2011). A hardware-in-the-loop simulation platform for the verification and validation of safety control systems. *IEEE Transactions on nuclear science*, 58(2):468–478. 96

[100] Ren, W., Steurer, M., and Baldwin, T. L. (2009). An effective method for evaluating the accuracy of power hardware-in-the-loop simulations. *IEEE Transactions on industry applications*, 45(4):1484–1490. 95

[101] Resende, P. A. A. and Drummond, A. C. (2018). A survey of random forest based methods for intrusion detection systems. *ACM Computing Surveys (CSUR)*, 51(3):48. 31

[102] Sandaruwan, G., Ranaweera, P., and Oleshchuk, V. A. (2013). Plc security and critical infrastructure protection. In *2013 IEEE 8th International Conference on Industrial and Information Systems*, pages 81–85. IEEE. 45

[103] SANS (2019). Sans 2019 state of ot/ics cybersecurity survey. https://www.sans.org/media/vendor/State-of-OT-ICS-Cybersecurity-Survey.pdf. 3, 16

[104] Schneier, B. (1999). Attack trees. *Dr. Dobb's journal*, 24(12):21–29. 37

[105] Seo, I.-Y. and Kim, S.-J. (2008). An on-line calibration monitoring technique using support vector regression and principal component analysis. In *2008 International Conference on Computational Intelligence for Modelling Control & Automation*, pages 663–669. IEEE. xiii, 67

[106] Shin, J., Heo, G., and Son, H. (2017a). Risk-informed cyber security evaluation of nuclear facilities. In *10th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPICHMIT 2017) (San Francisco, CA, USA, June 2017)*. 5, 18

[107] Shin, J., Son, H., and Heo, G. (2017b). Cyber security risk evaluation of a nuclear i&c using bn and et. *Nuclear Engineering and Technology*, 49(3):517–524. 37, 39, 44, 56

[108] Shin, J., Son, H., Heo, G., et al. (2015). Development of a cyber security risk model using bayesian networks. *Reliability Engineering & System Safety*, 134:208–217. 37, 43, 44

[109] Slay, J. and Miller, M. (2007). Lessons learned from the maroochy water breach. In *International Conference on Critical Infrastructure Protection*, pages 73–82. Springer. 15

[110] Softing (2019). datafeed opc suite. https://data-intelligence.softing.com/us/products/opc-software-platform/datafeed-opc-suite/. 97

[111] Song, J.-G., Lee, J.-W., Park, G.-Y., Kwon, K.-C., Lee, D.-Y., and Lee, C.-K. (2013). An analysis of technical security control requirements for digital i&c systems in nuclear power plants. *Nuclear Engineering and Technology*, 45(5):637–652. 42

[112] Spenneberg, R., Brüggemann, M., and Schwartke, H. (2016). Plc-blaster: A worm living solely in the plc. *Black Hat Asia*, 16. 45

[113] Sun, W., Cai, X., and Meng, Q. (2016). Testing flight software on the ground: Introducing the hardware-in-the-loop simulation method to the alpha magnetic spectrometer on the international space station. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 815:83–90. 95

[114] Svendsen, N. and Wolthusen, S. (2009). Using physical models for anomaly detection in control systems. In *International Conference on Critical Infrastructure Protection*, pages 139–149. Springer. 34

[115] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. (2009). A detailed analysis of the kdd cup 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–6. IEEE. 35

[116] Tesfaye, M. (2012). Secure reprogramming of a network connected device: Securing programmable logic controllers. 16

[117] Thangaparvathi, B., Anandhavalli, D., and Shalinie, S. M. (2011). A high speed decision tree classifier algorithm for huge dataset. In *Recent Trends in Information Technology (ICRTIT), 2011 International Conference on*, pages 695–700. IEEE. 60

[118] Travis, J. and Kring, J. (2007). *LabVIEW for everyone: graphical programming made easy and fun*. Prentice-Hall. 85

[119] US Nuclear Regulatory Commission (2009). Nrc: 10 cfr 73.54 protection of digital computer and communication systems and networks. https://www.nrc.gov/reading-rm/doc-collections/cfr/part073/part073-0054.html. 23

[120] US Nuclear Regulatory Commission (2010). Regulatory guide 5.71, cyber security programs for nuclear facilities. https://www.nrc.gov/docs/ML0903/ML090340159.pdf. xiii, 9, 10, 23

[121] US Nuclear Regulatory Commission (2015). Nrc: 10 cfr 73.77 cyber security event notifications. https://www.nrc.gov/reading-rm/doc-collections/cfr/part073/part073-0077.html. 25

[122] US Nuclear Regulatory Commission (2019). U.s. nrc licensee event reports. https://lersearch.inl.gov/LERSearchCriteria.aspx. 55

[123] Varuttamaseni, A., Bari, R. A., and Youngblood, R. (2017). Construction of a cyber attack model for nuclear power plants. Technical report, Brookhaven National Laboratory (BNL), Upton, NY (United States). 5, 17

[124] Venkatachary, S. K., Prasad, J., and Samikannu, R. (2018). Cybersecurity and cyber terrorism - in energy sector - a review. *Journal of Cyber Security Technology*, 2(3-4):111–130. 5, 18

[125] Vigna, G. and Kemmerer, R. A. (1998). Netstat: A network-based intrusion detection approach. In *Proceedings 14th Annual Computer Security Applications Conference (Cat. No. 98EX217)*, pages 25–34. IEEE. 48

[126] Wagner, R. (2001). Address resolution protocol spoofing and man-in-the-middle attacks. *The SANS Institute*. 88

[127] Wang, B. and Pineau, J. (2016). Online bagging and boosting for imbalanced data streams. *IEEE Transactions on Knowledge & Data Engineering*, (1):1–1. 60, 62

[128] Weiss, J. (2010). *Protecting industrial control systems from electronic threats*. Momentum Press. 6, 7

[129] Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52. 69

[130] Wood, A. D. and Stankovic, J. A. (2002). Denial of service in sensor networks. *computer*, 35(10):54–62. 88

[131] Xing, E. P., Jordan, M. I., Russell, S. J., and Ng, A. Y. (2003). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528. 60

[132] Zhang, F. (2020). *Nuclear Power Plant Cybersecurity, In Jun Wang, Xin Li, Chris Allison, Judy Hohorst Eds. Nuclear Power Plant Design and Analysis Codes*. Elsevier. v, 14

[133] Zhang, F., Coble, J., and Hines, J. W. (2019a). Data-driven model application for attack detection of scada system. In *11th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPICHMIT 2019), Orlando, FL, USA, February 9-14, 2019*. v, 14, 47, 78, 83, 104

[134] Zhang, F., Hines, J. W., and Coble, J. (2018a). Industrial control system testbed for cybersecurity research with industrial process data. In *International Congress on Advances in Nuclear Power Plants (ICAPP 2018), Charlotte, NC, USA, April 8-11, 2018*. v, 7, 8, 18, 78, 80, 104

[135] Zhang, F., Hines, J. W., and Coble, J. B. (2019b). A robust cybersecurity solution platform architecture for digital instrumentation and control systems in nuclear power facilities. *Nuclear Technology*, 0(0):1–12. v, xiii, 2, 10, 14, 47, 49, 78, 104

[136] Zhang, F., Kodituwakku, H. A. D. E., Hines, W., and Coble, J. B. (2019c). Multi-layer data-driven cyber-attack detection system for industrial control systems based on network, system and process data. *IEEE Transactions on Industrial Informatics*, 5(7):4362–4369. v, xiii, 2, 3, 14, 47, 60, 78, 91, 104

[137] Zhang, F., Payne, T., Wesley, H. J., and Jamie, C. (2019d). Enhancing the resilience of key equipment to false data injection attacks in npps. In *2019 ANS Winter Conference*. American Nuclear Society (ANS). v, 47, 95, 104

[138] Zhang, J., Zulkernine, M., and Haque, A. (2008). Random-forests-based network intrusion detection systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(5):649–659. 31, 32

[139] Zhang, Q., Zhou, C., Tian, Y.-C., Xiong, N., Qin, Y., and Hu, B. (2018b). A fuzzy probability bayesian network approach for dynamic cybersecurity risk assessment in industrial control systems. *IEEE Transactions on Industrial Informatics*, 14(6):2497–2506. 37, 39, 40, 41, 56

[140] Zhu, Y., Hu, H., Xu, G., and Zhao, Z. (2009). Hardware-in-the-loop simulation of pure electric vehicle control system. In *2009 International Asia Conference on Informatics in Control, Automation and Robotics*, pages 254–258. IEEE. 95

[141] Zulkefli, M. A. M., Mukherjee, P., Sun, Z., Zheng, J., Liu, H. X., and Huang, P. (2017). Hardware-in-the-loop testbed for evaluating connected vehicle applications. *Transportation Research Part C: Emerging Technologies*, 78:50–62. 95

Appendices

A Detailed Data Acquisition Connection of cDAQ9188

This section presents the detailed connection of DAQ 9188 with 5 modules as shown in figure A.1, which are utilized to enable the SCADA master to collect measurement data and control the actuators.

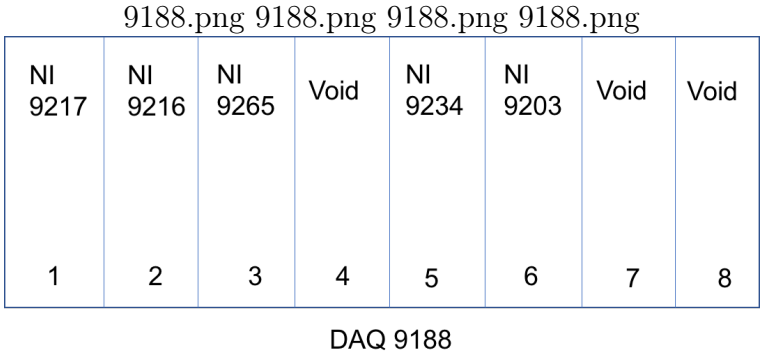


Figure A.1: The connection of DAQ 9188 with other 5 modules.

Figure A.2 and A.3 show the connection sketch of NI 9217 and NI 9216, respectively. NI 9217 acquires data from RTD 0, 1, 2 and 3 while NI 9216 acquires data from RTD 4, 5, 6 and 7 (RTD6 and 7 is installed in the potential intermediate loop).

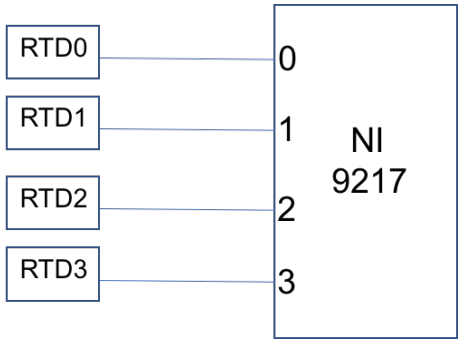


Figure A.2: The connection of NI9217.

NI 9265 is a 4-Channel C Series Current Output Module. Figure A.4 is the connection sketch of NI 9265. It is used for interfacing and controlling current-driven actuators (MOV, VFD and heater) at high rates. It requires an external power supply and includes a channel-to-earth ground double isolation barrier for safety and noise immunity.

NI-9203 is a 20 mA, 8-Channel C Series Current Input Module. It is used for high-performance control and monitoring. It features programmable input ranges and variable

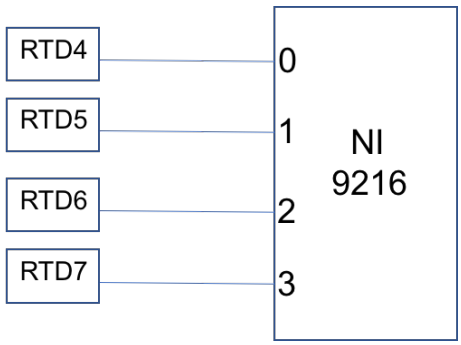


Figure A.3: The connection of NI9216.

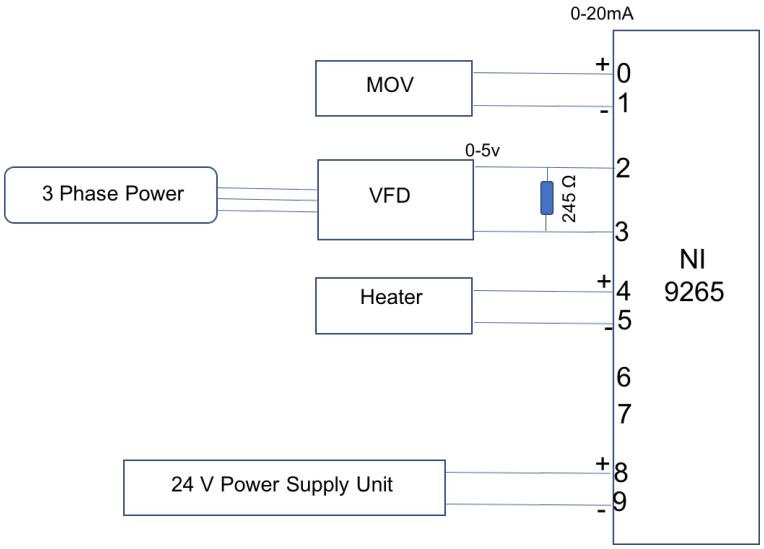


Figure A.4: The connection of NI9265.

connectivity options. Figure A.5 shows the connection sketch of NI 9203. It is used for collecting data from flow meter 0, 1 and 2 in this test bed. (flow meter 2 is for potential intermediate loop in the future) NI-9203 includes a channel-to-earth ground double-isolation barrier for safety and noise immunity.

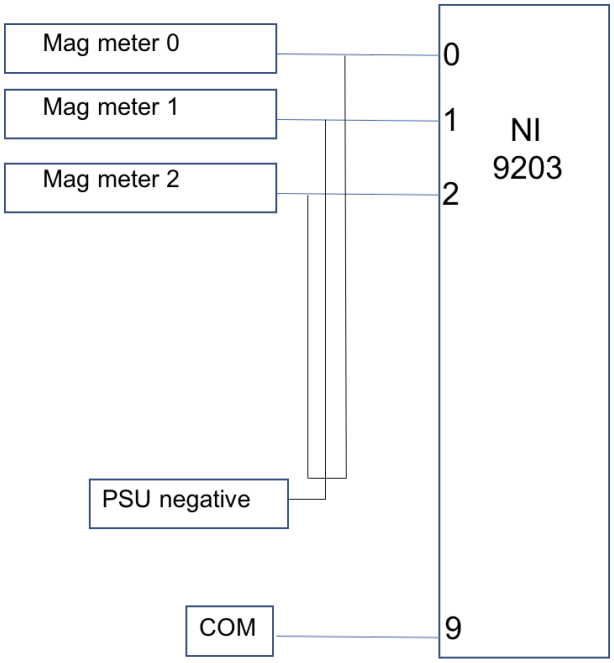


Figure A.5: The connection of NI9203.

B Published Works

1. F. Zhang. Nuclear Power Plant Cybersecurity, in Jun Wang, Xin Li, Chris Allison, Judy Hohorst Eds. Nuclear Power Plant Design and Analysis Codes. Elsevier, 2020.

2. F. Zhang, T. Payne, J. W. Hines, J. Coble, "Enhancing the Resilience of Key Equipment to False Data Injection Attacks in NPPs," In ANS Winter Meeting 2019, Washington, DC, USA, November 17 - 21, 2019.

3. F. Zhang, J.W. Hines, J. Coble, "A Robust Cybersecurity Solution Platform Architecture for for Digital Instrumentation and Control Systems in Nuclear Power Facilities," Nuclear Technology, 2019, Digital Object Identifier: 10.1080/00295450.2019.1666599.

4. F. Zhang, H. Kodituwakku,W. Hines, J. Coble, "Multi-Layer Data-Driven Cyber-Attack Detection System for Industrial Control Systems Based on Network, System and Process Data," IEEE Transactions on Industrial Informatics, vol 5, No.7, pp 4362-4369, 2019, Digital Object Identifier: 10.1109/TII.2019.2891261.

5. F. Zhang, J.W. Hines, J. Coble, "A Robust Cybersecurity Solution Platform Architecture for Digital I&C Systems in NPP," In ANS Student Conference 2019, Virginia Commonwealth University, Richmond, VA, USA, April 4 - 6, 2019.

6. F. Zhang, J.W. Hines, J. Coble, "Data-driven Model Application for Attack Detection of SCADA System," In 11th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC&HMIT 2019), Orlando, FL, USA, February 9 - 14, 2019.

7. F. Zhang, J.W. Hines, J. Coble, J, "Industrial Control System Testbed for Cybersecurity Research with Industrial Process Data," In International Congress on Advances in Nuclear Power Plants (ICAPP 2018), Charlotte, NC, USA, April 8 - 11, 2018.

8. F. Zhang, J.W. Hines, J. Coble, "Empirical Models Integration for Enhanced Cybersecurity of Industrial Control System," In ANS Student Conference 2018, The University of Florida, Gainesville, FL, USA, April 5 - 7, 2018.

9. F. Zhang, S. Boring, J.W. Hines, J. Coble, K. Gross, "Combination of Unquantization Technique and Empirical Modeling for Industrial Applications," In ANS Winter Meeting 2017,Washington, DC, USA, October 29 - November 2, 2017.

10. F. Zhang and J. Coble, "A Dynamic Model of Once-through Steam Generator (OTSG) for Prototypical Advanced Reactor," In 10th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC&HMIT 2017), San Francisco, CA, June 11-15, 2017.

Vita

Fan Zhang went to North China Electric Power University, China, for her undergraduate education to study Nuclear Engineering, which appeals to her as a clean and economical source of energy for the sustainable development of human society. She graduated with the Beijing Outstanding Student honor with her Bachelor of Science degree in July 2013. As an undergraduate, Fan participated in several engineering project competitions based on her interests, such as "Generating power in subway tunnel using flexible-blade wind turbines."

Due to her excellent performance in academic study and research, she was recommended to the Master's program for Nuclear Energy Science and Engineering in the same university with entrance exam exemption. She then conducted research related to thermal hydraulics, instrumentation and control systems and safety analysis under the supervision of Dr. Daogang Lu, and was awarded the National Scholarship for Master's Students based on research merit. She obtained her Master's degree with her thesis titled "Full scope modeling and analysis on the secondary circuit of Chinese Large-Capacity Advanced PWR Based on RELAP5 Code." in April 2016. As a Master's student, she published three peer-reviewed journal papers.

In August 2016, Fan started her doctoral program in Nuclear Engineering under the supervision of Dr. Jamie Coble at the University of Tennessee, Knoxville. She enrolled in the Intercollegiate Graduate Statistics Program (IGSP) to obtain a master degree in Statistics as well. Her research focused on the cybersecurity of industrial control systems and nuclear power plant online monitoring using data analytic methods. She took initiative to pursue research funding opportunities, and helped secure UTK involvement in an International Atomic Energy Agency (IAEA) Coordinated Research Program "Enhancing Computer security Incident Analysis and Response Planning at Nuclear Facilities." She co-authored

a DoE-awarded Nuclear Energy University Program (NEUP) proposal, "A Cyber-Attack Detection Platform for Cyber Security of Digital Instrumentation and Control Systems." She has published two journal papers and presented 7 conference papers. She has been invited to Toronto, Vienna, Beijing, Daejeon, and Fredericton by IAEA to present her research and for technical consultation as well. As a graduate student, Fan has received four scholarships from the American Nuclear Society (ANS) and two scholarships from the International Society of Automation (ISA). She was awarded the University of Tennessee Extraordinary Professional Promise Award. She won two awards from the 2019 ANS Student Conference: Top Paper Submission and Best Presentation in cybersecurity. In Summer 2019, Fan led a team that participated in a Data Challenge for the Smoky Mountains Computational Sciences Engineering Conference, which was awarded best paper. She is an active member of ANS, Alpha Nu Sigma, Women in Nuclear, ISA, and the Institute of Electrical and Electronics Engineers (IEEE).