

Industrial Datasets with ICS Testbed and Attack Detection Using Machine Learning Techniques

Sinil Mubarak¹, Mohamed Hadi Habaebi^{1,*}, Md Rafiqul Islam¹, Asaad Balla¹, Mohammad Tahir², Elfatih A. A. Elsheikh³ and F. M. Suliman³

¹IoT & Wireless Communication Protocols Lab, International Islamic University Malaysia, Selangor, 53100, Malaysia

²Sunway University, Selangor, 47500, Malaysia

³Department of Electrical Engineering, College of Engineering, King Khalid University, Abha, 61421, Saudi Arabia

*Corresponding Author: Mohamed Hadi Habaebi. Email: habaebi@iiu.edu.my

Received: 08 June 2021; Accepted: 16 July 2021

Abstract: Industrial control systems (ICS) are the backbone for the implementation of cybersecurity solutions. They are susceptible to various attacks, due to openness in connectivity, unauthorized attempts, malicious attacks, use of more commercial off the shelf (COTS) software and hardware, and implementation of Internet protocols (IP) that exposes them to the outside world. Cybersecurity solutions for Information technology (IT) secured with firewalls, intrusion detection/protection systems do nothing much for Operational technology (OT) ICS. An innovative concept of using real operational technology network traffic-based testbed, for cyber-physical system simulation and analysis, is presented. The testbed is equipped with real-time attacks using in-house penetration test tool with reconnaissance, interception, and firmware analysis scenarios. The test cases with different real-time hacking scenarios are implemented with the ICS cyber test kit, and its industrial datasets are captured which can be utilized for Deep packet inspection (DPI). The DPI provides more visibility into the contents of OT network traffic based on OT protocols. The Machine learning (ML) techniques are deployed for cyber-attack detection of datasets from the cyber kit. The performance metrics such as accuracy, precision, recall, F1 score are evaluated and cross validated for different ML algorithms for anomaly detection. The decision tree (DT) ML technique is optimized with pruning method which provides an attack detection accuracy of 96.5%. The deep learning (DL) techniques has been used recently for enhanced OT intrusion detection performances.

Keywords: SCADA; industrial control system; intrusion detection system; machine learning; anomaly detection

1 Introduction

Operational technology (OT) process is a part of critical infrastructure and often targeted by criminal organizations. OT networks were traditionally kept separate or “air-gapped” from IT networks. However, new business requirements associated with the efficiency benefits of digitalization are forcing increased



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

connectivity between IT and OT networks, thereby increasing the attack surface and hence the cyber risk. One growing trend is the use of cyberattacks to target critical infrastructure and strategic industrial sectors [1].

This trend is raising fears that attackers could trigger a breakdown in the essential systems that keep societies from functioning (electricity transmission & water distribution, oil & gas refineries, process industries) [2]. The analysis of machine learning algorithms for SCADA systems can be performed with (i) Mathematical modeling of the system, (ii) Testbeds (iii) Industrial datasets. Many universities have difficulties in building cyber range lab (OT) and industrial use case set up for ICS cybersecurity research centers on their university campus due to financial constraint.

Most of the datasets are outdated and are only associated with information technology systems. The KDD cup 1999 dataset was developed in 1998 and last updated in 2008. The 1999 DARPA dataset created to validate the effectiveness of IDS, was found to contain unintended patterns. Gao's dataset was also found to have obvious unintended patterns, unsuitable and outdated for IDS research.

The contribution of this paper is to introduce an innovative and cost-effective real-time network traffic testbed, which can mimic the industrial control system operation. The packet capture file can be utilized to develop anomaly-based Machine learning algorithms for intrusion detection in ICS systems. An intrusion detection system (IDS) integrated with machine learning (supervised and unsupervised techniques) can improve the detection rates of ICS SCADA system attacks [3]. A variety of machine learning methods use mathematical techniques to learn, profile, classify and predict unusual results. For supervised methods, a pre-labelled data feature is required, and unsupervised methods do not require pre-labeled data to be analyzed.

This paper is subdivided into the following sections: Section I deals with ICS OT details with the public datasets available for machine learning analysis. Section II highlights the innovative ICS portable kit prototype and setup that is available at IIUM lab facility and its industrial network data capture with real-time control system operation for a normal scenario, along with traditional machine learning techniques which can be deployed for attack detection. Section III describes the methodology of generating ICS OT real-time network dataset along with simulation of advanced ethical hacking attacks using cyber test kit. Section IV details the anomaly detection capability of ML algorithms with the industrial datasets. The paper then concludes with future works for deep learning ML algorithms, based on the OT traffic data for different scenarios.

2 Related Work

Many researchers use publicly available data sets to analyze machine learning strategies. Some public database shortcomings are highlighted as follows. The proposed framework in [4] is limited to Modbus/TCP-based content. Data size in [5] is very limited in online testing activities and only multi-machine learning strategies are developed. The database used in [6] is out of date and does not include current threats, whereas the data size used in [7] is small (about 1000 cases) and limited to single cyber- attack. In [8], the database is out of date, and the attacks were related to the IT domain.

The Singapore University of Technology has developed water treatment testbed with SCADA network traffic and attack scenarios [9]. Recently, real-time data sets that include standard and 35 types of cyber-attacks were introduced to train and test machine learning techniques. Comparison of the performance of the algorithm is done only with supervised machine learning techniques. The results show a high rate of false detection in algorithms [10]. The performance of ML algorithms may provide a satisfactory outcome on the public dataset. However, the payload/data frame is manipulated with datasets labels and manually randomize and parameterize the attacks, to mimic the operational/attack scenario [11].

The research gap was identified in a comprehensive study of the analysis of the effects of malicious activities using protocols pertaining to OT industrial processes in [12].

2.1 ICS Cyber Testbed Prototype

The ICS cyber portable test kit can overcome the main concern for researchers—lack of common framework and dependence on public datasets which usually do not include all types of attacks to train and test the algorithms.

In collaboration with industrial partner (Necon Automation), IIUM has developed an in-house innovative portable ICS cyber test-kit for research & training purposes [13]. The package consists of a PLC system, HMI system, Process simulation modules, Ethernet switch, Physical sensor, and Attacker system. The ICS portable kit package provides real industrial network flow data for research & training, and machine learning software development. The ICS kit prototype is depicted in Fig. 1.

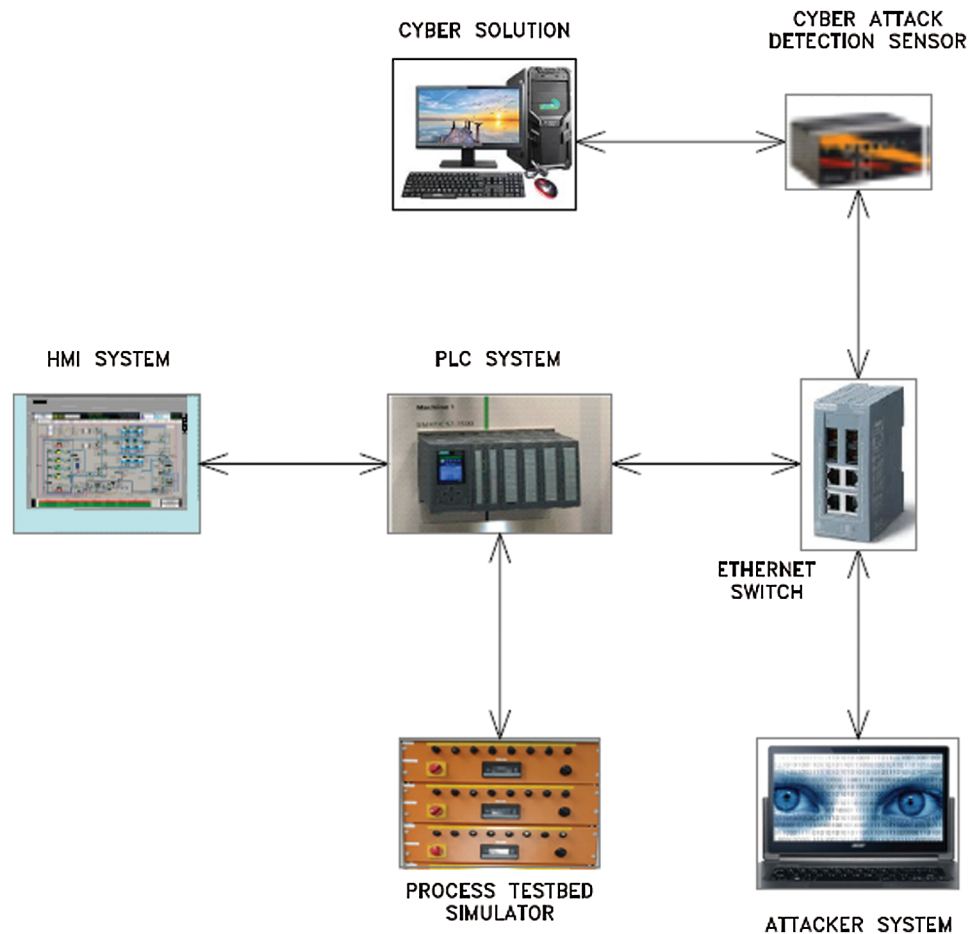


Figure 1: ICS portable testbed prototype

2.2 ICS Cyber Testkit Components

The ICS portable kit package is made up of the following components, as depicted in Fig. 2.

- PLC system (IP address-192.168.101.22): SIMATIC S7-1200 Siemens for process operation simulation.
- HMI system (IP address-192.168.101.23): SIMATIC Basic panel Siemens for a graphical interface.
- Ethernet switch (IP address: 192.168.101.111): Scalance XB213 Siemens with mirror port facility to copy the operational technology network traffic.
- Process simulation modules—Analogue, digital inputs/outputs for field process simulation.
- Physical Sensor: Data collection of network traffic with Deep Packet Inspection (DPI).
- Attacker system: Launch OT attacks, Penetration test software with Kali Linux installed on virtual platform with a license.

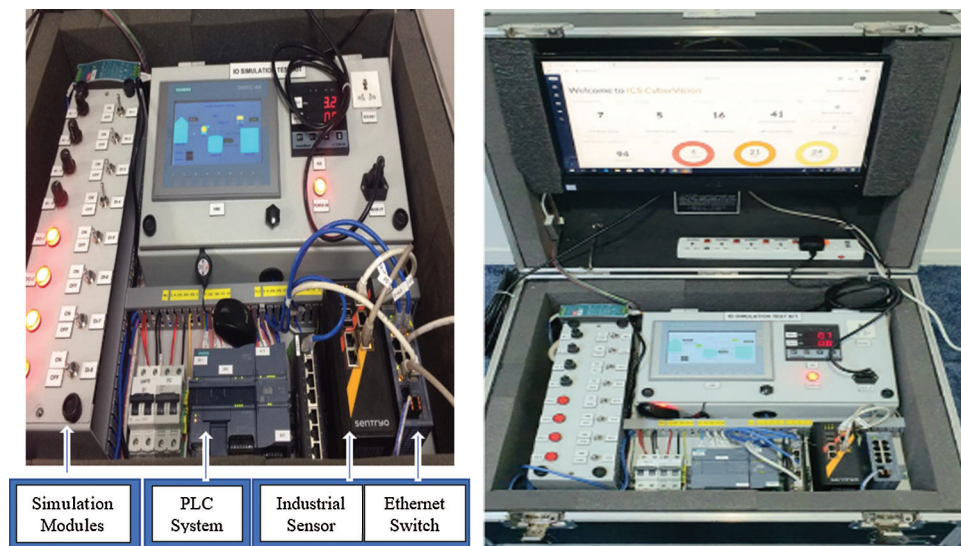


Figure 2: ICS portable testbed components

2.3 Machine Learning Techniques

Machine learning algorithms are widely implemented in the intrusion detection system (IDS) as first line of defence. Different machine learning supervised techniques can improve the detection rates of attacks for SCADA systems as mentioned in [Tab. 1](#).

Table 1: Machine learning—supervised methods

Algorithm	Technique
Logistic regression	Non-linear probability prediction for binary classification output
Naive Bayes	Conditional probability
K-nearest neighbor	Instance-based learning based on similarity
SVM	Map non-linear to linear hyper plane
Decision tree	More stable and accurate, easy interpretation for both classification & regression

3 ICS Cyber Kit Network Traffic Dataset Methodology

The ICS scenarios can be performed using the developed kit, and OT real network traffic can then be obtained for further analysis. To ensure proper work environment, a network capture (PCAP file) of at least 10 min with normal traffic and some actions on the PLC system is necessary [14]. The operation of a wastewater treatment system is simulated in HMI/PLC control system for the real-time network traffic data for normal scenario, as shown in Fig. 3.

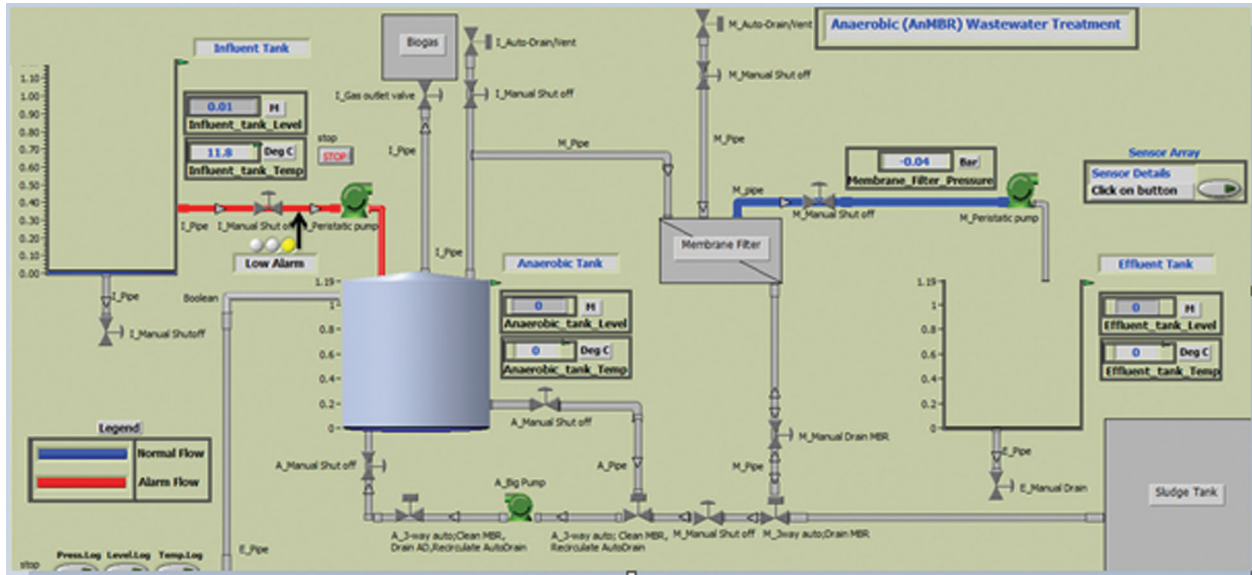


Figure 3: HMI/PLC wastewater treatment system

The below scenarios are deployed to ensure the proper collection of OT network traffic data:

- Stable mode of control system operation is maintained for at least 10 minutes.
- Set-point changes of process variables and command execution are carried out with simulation module along with forcing input/output register bits in PLC.

The network traffic can be captured with two possibilities, as follows: (1) PCs need to be connected to the OT network, and network traffic can be captured using PCAP capture software (Wireshark), but the data size can be large. (2) Physical sensor provides automatic capture and extracts meaningful information only from OT network flows and supports long capture, and 100% passive also understands OT protocols.

The PCAP file is retrieved from the physical sensor and the network traffic data with protocol packet length versus time for normal operation is identified. The TCP protocol traffic data between PLC (IP address 192.168.101.22), HMI (192.168.101.23), and Ethernet switch (192.168.101.111) is represented in below Figs. 4a and 4b [15].

The sensor performs only passive discovering on the network, as active scanning may have an unpredictable reaction and may cause a huge risk on OT networks. The DPI engine extracts the following properties.

- Inventory information of all OT network components (e.g., PLC and controllers)
- Metadata such as packet sizes, number of packets, packet rates and timing.
- Identification properties: Rack slot, MAC address, protocol ID, TCP & UDP port.

- Inventory properties: Vendor, model name, firmware/hardware version, sub-module location/slot.
- Process control information: Messages exchanged between process control devices, read and write commands.
- Basic PLC control: Program download commands, start and stop commands.

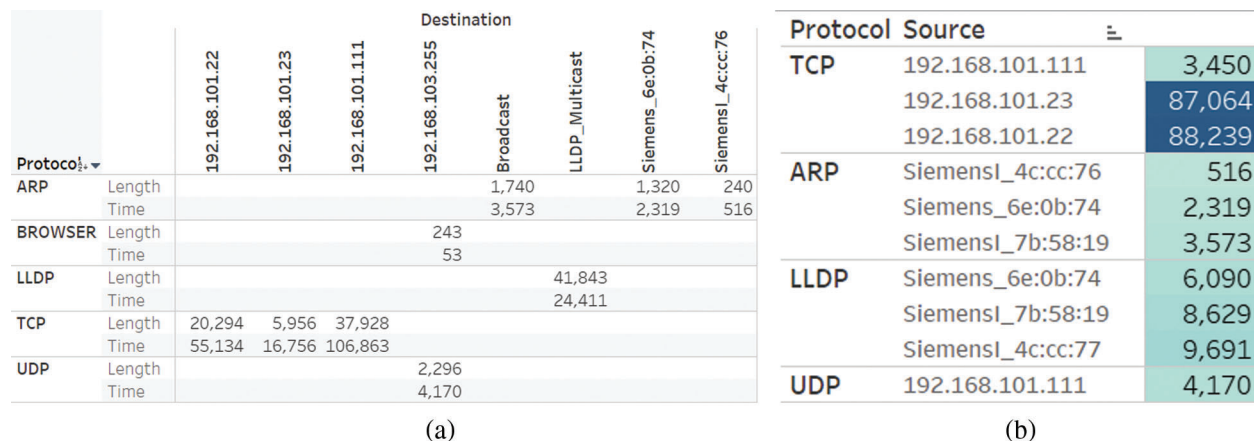


Figure 4: (a) Protocol destination vs. length, (b) Protocol source vs. time

3.1 ICS OT Real-time Data Results with Attacks Simulation

The data workflow of industrial OT traffic data packet capture with ICS OT kit is shown in Fig. 5.

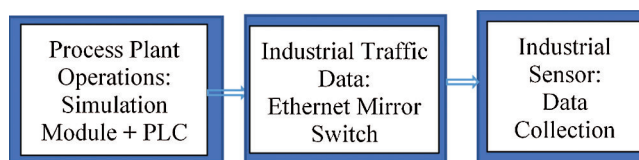


Figure 5: ICS data collection workflow

Industrial Process Plant Operations: The ICS kit is possessed with input/output modules to simulate/mimic the field devices and its process variables. The process variables correlated to transmitters, pumps, valves are connected to PLC systems and the wastewater system graphical user interface is demonstrated with HMI panel.

Mirrored Ethernet Switch: The ICS components of kit (PLC, HMI) are connected to Siemens Ethernet switch and mirror port functionality is enabled.

Industrial Sensor: OT Sensors are used in passive mode and connected to an OT Ethernet switch, which is then configured to redirect all the OT traffic to a mirror port. The port allows full capture of the traffic without disturbance for the system and the process, and the network traffic data can then be used for machine learning analysis.

3.2 Industrial Hacking Techniques

The test cases with different real-time hacking methods, as mentioned in Figs. 6a and 6b, is implemented with the ICS cyber test kit, and its industrial datasets are captured.

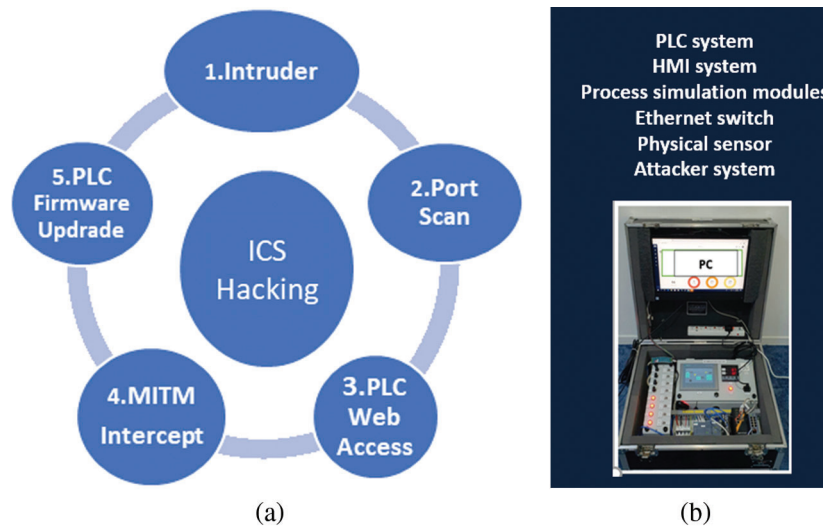


Figure 6: (a) Industrial hacking scenarios, (b) ICS cyber test kit

Reconnaissance/Port Scan Attacks: The test kit is attacked with an intruder laptop having KaliLinux penetration tool, and industrial data is collected, through sensor for machine learning analysis. Reconnaissance by scanning using internal systems through the lateral network using Nmap scripts and Metasploit tools can be demonstrated with insecure port scan activity, as shown in Fig. 7. The ICS components with PLC IP addresses (192.168.101.22), MAC addresses along with its model can be identified with the port scan hacking method.

```
root@kali:~# nmap -sP 192.168.101.1-255
Starting Nmap 7.70 ( https://nmap.org ) at 2020-02-17 00:58 EST
mass_dns: warning: Unable to determine any DNS servers. Reverse DNS is disabled
Try using --system-dns or specify valid servers with --dns-servers
Nmap scan report for 192.168.101.22
Host is up (0.0016s latency).
MAC Address: E0:DC:A0:4C:CC:76 (Siemens Industrial Automation Products Chengdu)
Nmap scan report for 192.168.101.23
Host is up (0.00053s latency).
MAC Address: E0:DC:A0:7B:58:19 (Siemens Industrial Automation Products Chengdu)
Nmap scan report for 192.168.101.25
Host is up (0.0039s latency).
MAC Address: 20:87:56:27:8A:51 (Siemens AG)
Nmap scan report for 192.168.101.252
Host is up (0.00026s latency).
MAC Address: E4:E7:49:97:46:96 (Unknown)
Nmap scan report for 192.168.101.200
Host is up.
Nmap done: 255 IP addresses (5 hosts up) scanned in 1.52 seconds
```

Figure 7: Nmap scan with Kali linux tool

PLC Web Server Access: Some ICS platforms run web applications which could be exploited for its web app vulnerabilities like a typical web application server [16]. More detailed information from Wireshark and the whole HTTP communication in clear text can be obtained, if any of the packets selected and the option to Analyze/Follow/HTTP Stream is selected. The laptop with webserver using HTTP service can access the Siemens S7-CPU for diagnostics functions, shown in Fig. 8.

MITM Attack with Ettercap Tool for ARP Sniff: Intercepting unencrypted communication between ICS systems to hijack a communication channel, for malicious activities can be demonstrated with kali Linux Ettercap option [17]. ARP poisoning MITM is executed between the switch using insecure telnet protocol service and intruder laptop using port no.23. The Siemens Ethernet switch (telnet) (IP address

192.168.101.25), and sniffer laptop (IP address 192.168.101.220) having KaliLinux software with Ettercap tool can capture the clear text password of the switch which is using unencrypted protocols. The insecure protocol such as telnet is configured on the Siemens switch as demonstrated in Figs. 9a and 9b.

No.	Time	Source	Destination	Protocol	Length	Info
547	8.752115456	192.168.101.250	192.168.101.22	HTTP	519	GET /Portal/Portal.mwsl?PriNav=Bgz HTTP/1.1
596	8.853302198	192.168.101.250	192.168.101.22	HTTP	481	GET /CSS/S7Web.css HTTP/1.1
600	8.854115576	192.168.101.250	192.168.101.22	HTTP	510	GET /Scripts/update.js HTTP/1.1
601	8.854825748	192.168.101.250	192.168.101.22	HTTP	486	GET /CSS/S7WebPrint.css HTTP/1.1
602	8.854829439	192.168.101.250	192.168.101.22	HTTP	521	GET /Scripts/jquery-1.11.2.min.js HTTP/1.1
641	8.909787316	192.168.101.22	192.168.101.250	HTTP	60	HTTP/1.1 200 OK (text/html)
654	8.963819576	192.168.101.22	192.168.101.250	HTTP	155	HTTP/1.1 304 Not Modified
660	9.001939872	192.168.101.22	192.168.101.250	HTTP	148	HTTP/1.1 304 Not Modified
666	9.026812445	192.168.101.22	192.168.101.250	HTTP	155	HTTP/1.1 304 Not Modified
671	9.048782694	192.168.101.250	192.168.101.22	HTTP	496	GET /ClientArea/Bgz.mwsl?PriNav=Bgz HTTP/1.1
673	9.048789423	192.168.101.22	192.168.101.250	HTTP	148	HTTP/1.1 304 Not Modified
674	9.050049453	192.168.101.250	192.168.101.22	HTTP	491	GET /ClientArea/BgzSecNav.mwsl HTTP/1.1
707	9.131354984	192.168.101.250	192.168.101.22	HTTP	482	GET /CSS/S7Web.css HTTP/1.1

> Internet Protocol Version 4, Src: 192.168.101.250 (192.168.101.250), Dst: 192.168.101.22 (192.168.101.22)

> Transmission Control Protocol, Src Port: 47631 (47631), Dst Port: http (80), Seq: 1, Ack: 1, Len: 465

▼ Hypertext Transfer Protocol

> GET /Portal/Portal.mwsl?PriNav=Bgz HTTP/1.1\r\n

Accept: text/html, application/xhtml+xml, */*\r\n

0020 65 16 ba 0f 00 50 ce 9e 25 eb 00 03 00 ce 50 18 e...P...%...P...

0030 ff ff 83 a6 00 00 47 45 54 20 2f 50 6f 72 74 61GET /Porta

0040 6c 2f 50 6f 72 74 61 6c 2e 6d 77 73 6c 3f 50 72 1/Portal .mwsl?Pr

0050 69 4e 61 76 3d 42 67 7a 20 48 54 54 50 2f 31 2e iNav=Bgz HTTP/1.

0060 31 0d 0a 41 63 63 65 70 74 3a 20 74 65 78 74 2f 1·Accep t: text/

0070 68 74 6d 6c 2c 20 61 70 70 6c 69 63 61 74 69 6f html, ap plicatio

0080 6e 2f 78 68 74 6d 6c 2b 78 6d 6c 2c 20 2a 2f 2a n/xhtmll+ xml, */*

0090 0d 0a 52 65 66 65 72 65 72 3a 20 68 74 74 70 3a ··Refere r: http:

00a0 2f 2f 31 39 32 2e 31 36 38 2e 31 30 31 2e 32 32 //192.16 8.101.22

00b0 2f 50 6f 72 74 61 6c 2f 50 6f 72 74 61 6c 2e 6d /Portal/ Portal.m

Figure 8: Wireshark information between PLC and laptop

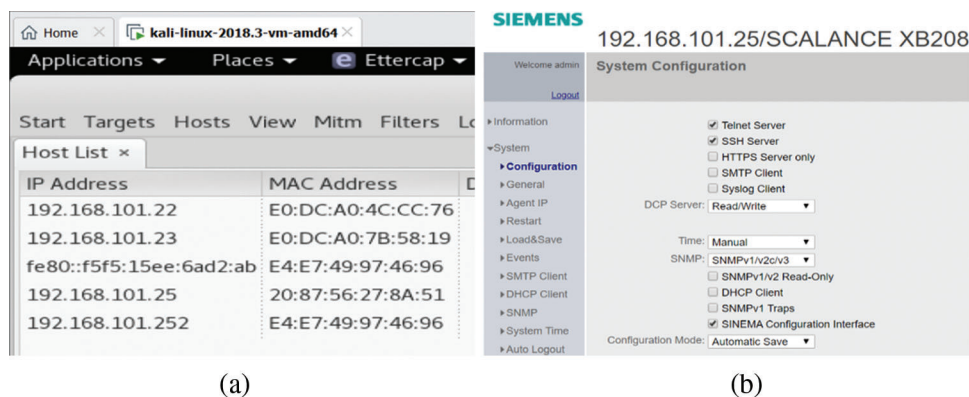
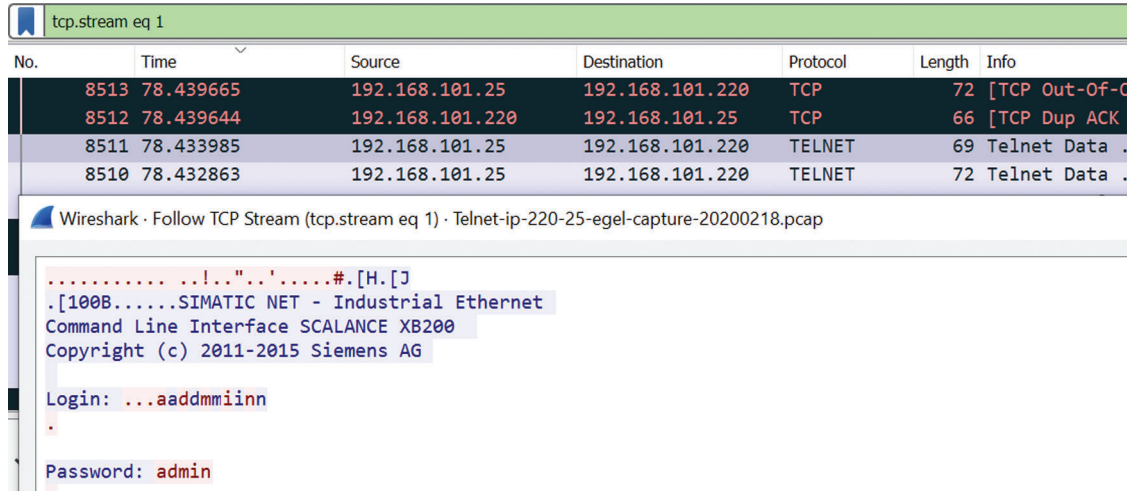


Figure 9: (a) Kali linux software with Ettercap tool, (b) Ethernet switch configured with telnet

The password of the Ethernet switch can be easily extracted as shown in Fig. 10 with Ettercap sniffing tool, and the data is captured for machine learning analysis.



No.	Time	Source	Destination	Protocol	Length	Info
8513	78.439665	192.168.101.25	192.168.101.220	TCP	72	[TCP Out-Of-Order]
8512	78.439644	192.168.101.220	192.168.101.25	TCP	66	[TCP Dup ACK]
8511	78.433985	192.168.101.25	192.168.101.220	TELNET	69	Telnet Data .
8510	78.432863	192.168.101.25	192.168.101.220	TELNET	72	Telnet Data .

Wireshark · Follow TCP Stream (tcp.stream eq 1) · Telnet-ip-220-25-egel-capture-20200218.pcap

```

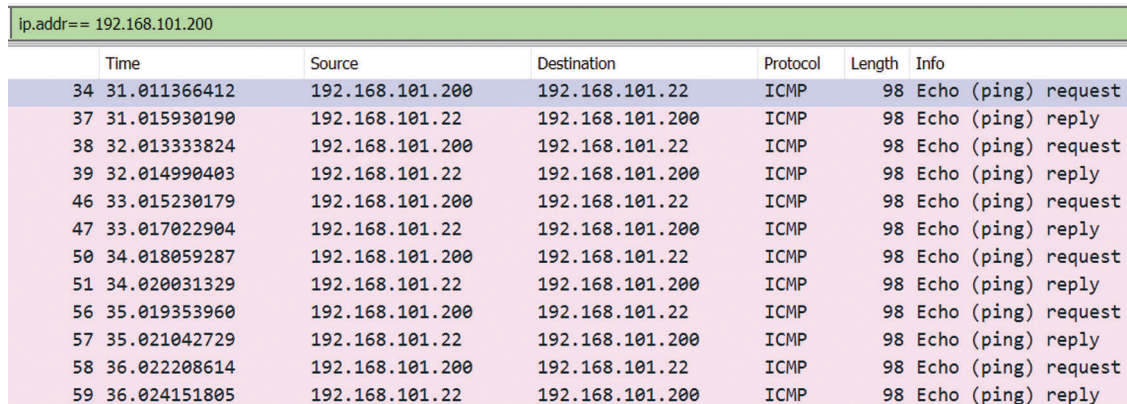
.....!..".'.#.[H.]
.[100B.....SIMATIC NET - Industrial Ethernet
Command Line Interface SCALANCE XB200
Copyright (c) 2011-2015 Siemens AG

Login: ...aaddmiinn
.
Password: admin

```

Figure 10: Wireshark information between PLC and laptop

PLC Firmware Upgrade: The dataset can be obtained for the pinging scenario with programming laptop (IP address 192.168.101.220) with Siemens PLC (IP address 192.168.101.22), as shown in Fig. 11.



No.	Time	Source	Destination	Protocol	Length	Info
34	31.011366412	192.168.101.200	192.168.101.22	ICMP	98	Echo (ping) request
37	31.015930190	192.168.101.22	192.168.101.200	ICMP	98	Echo (ping) reply
38	32.013333824	192.168.101.200	192.168.101.22	ICMP	98	Echo (ping) request
39	32.014990403	192.168.101.22	192.168.101.200	ICMP	98	Echo (ping) reply
46	33.015230179	192.168.101.200	192.168.101.22	ICMP	98	Echo (ping) request
47	33.017022904	192.168.101.22	192.168.101.200	ICMP	98	Echo (ping) reply
50	34.018059287	192.168.101.200	192.168.101.22	ICMP	98	Echo (ping) request
51	34.020031329	192.168.101.22	192.168.101.200	ICMP	98	Echo (ping) reply
56	35.019353960	192.168.101.200	192.168.101.22	ICMP	98	Echo (ping) request
57	35.021042729	192.168.101.22	192.168.101.200	ICMP	98	Echo (ping) reply
58	36.022208614	192.168.101.200	192.168.101.22	ICMP	98	Echo (ping) request
59	36.024151805	192.168.101.22	192.168.101.200	ICMP	98	Echo (ping) reply

Figure 11: Wireshark data PLC and programming tool

The unsecured ICS component allows enumeration of system parameters, leads to the disclosure of system firmware versions and their vulnerabilities. Siemens programming tool is used to browse and discover PLC configuration and PLC firmware upgradation is performed as mentioned in Fig. 12.

4 Industrial Datasets with Operational Scenarios

The following ICS datasets with various scenarios and attack vectors are obtained with the test kit:

- Datasets include normal operational scenario with its ICS components and protocols dedicated to OT ICS domain such as TCP, UDP, ARP, ICMP has 68965 instances for normal behaviour profile.
- Dataset for MITM attack scenario with Ettercap tool has 52600 instances.

- Dataset for telnet communication for PLC/HMI Ethernet switch has 13076 instances.
- Dataset for web-server access of PLC system with programming tool laptop for system diagnostic scenario has 21435 instances.
- Dataset for ping scenario of programming tool laptop with PLC System along with PLC firmware change activity is also obtained for analysis.



Figure 12: PLC firmware upgrade with programming tool

The test kit can also be attacked and tested with ethical hacking penetration tools to obtain the industrial datasets which are unique and innovative and is represented in Fig. 13.

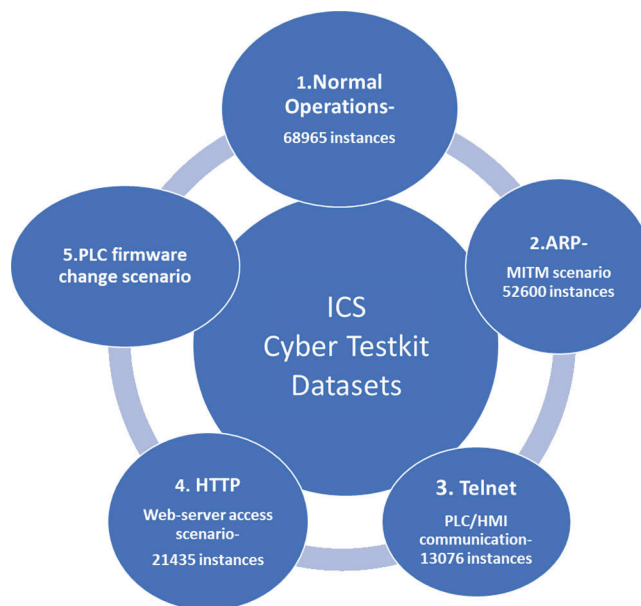


Figure 13: ICS cyber kit datasets with industrial scenarios

The different abnormal conditions of S7 communication modbus TCP flags is analyzed based on the SYN, ACK, RST flag values as shown below in Fig. 14.

```

  Flags: 0x012 (SYN, ACK)
    000. .... = Reserved: Not set
    ...0 .... = Nonce: Not set
    .... 0... = Congestion Window Reduced (CWR): Not set
    .... .0.. = ECN-Echo: Not set
    .... ..0. = Urgent: Not set
    .... ...1 = Acknowledgment: Set
    .... ....0 = Push: Not set
    .... ....0.. = Reset: Not set
  Syn: Set
iso-tsap(102) > 21289 [RST, ACK] Seq=1 Ack=1 Win=2920 Len=0 (RST=1, ACK=1)
[TCP Retransmission] iso-tsap(102) > 45218 [PSH, ACK] Seq=28 Ack=1 Win=8192 Len=88
[TCP Keep-Alive] iso-tsap(102) > 52277 [ACK] Seq=0 Ack=1 Win=2920 Len=0
[TCP Dup ACK 1970#6] iso-tsap(102) > 7462 [ACK] Seq=1 Ack=1 Win=2920 Len=0
45218 > iso-tsap(102) [ACK] Seq=1044 Ack=3223 Win=2920 Len=0 (SYN=1, ACK=1)
iso-tsap(102) → 33578 [SYN, ACK] Seq=0 Ack=1 Win=2920 Len=0 MSS=1460
13872 > iso-tsap(102) [SYN] Seq=0 Win=32767 Len=0 (SYN=1, ACK=0)

```

Figure 14: Modbus TCP protocol analysis

The datasets mentioned in Tab. 2, include the industrial protocols and services such as Transmission Control Protocol (TCP) for services, with dedicated destination port: HTTP (80), Telnet (23), Modbus (502), ISO-on-TCP- Siemens S7comm (102). These type of industrial OT datasets is very difficult to get from the industries due to its sensitivity and criticality [18].

Table 2: Industrial cyber kit datasets instances with scenarios

Datasets	Instances	OT Protocol	Scenario
Dataset#1	68965	S7,TCP	Normal
Dataset#2	52600	ARP	Attack
Dataset#3	13076	Telnet	Attack
Dataset#4	372	ICMP	Attack
Dataset#5	21435	HTTP	Attack

4.1 Anomaly Prediction of Industrial Datasets with Machine Learning Algorithms

The present drawback of public datasets to represent the real operational technology traffic can be overcome by our compact and portable ICS test kit. The industrial datasets with different scenarios are profiled with deep packet inspection based on protocol-based behavior analysis for the normal scenario of the ICS network traffic and different machine learning algorithms are deployed for anomaly detection in ICS OT domain.

The industrial datasets are pre-processed, profiled and the abnormality is analyzed with DPI. Metadata processing with dataframe feature selection is carried out before applying the algorithm. Metadata cascade matrix [30608,16] dataframe is generated by combining the info column of each OT protocols-["TCP","MRP","ARP","Telnet","Modbus","HTTP"]. The pre-processed metadata is cleaned, normalized for the easiness of algorithm analysis and split to train and test datasets to model with ML algorithms.

The following machine learning (ML) steps are deployed with the industrial datasets obtained from cyber kit to identify the best machine learning model for anomaly attack detection as shown in Fig. 15.

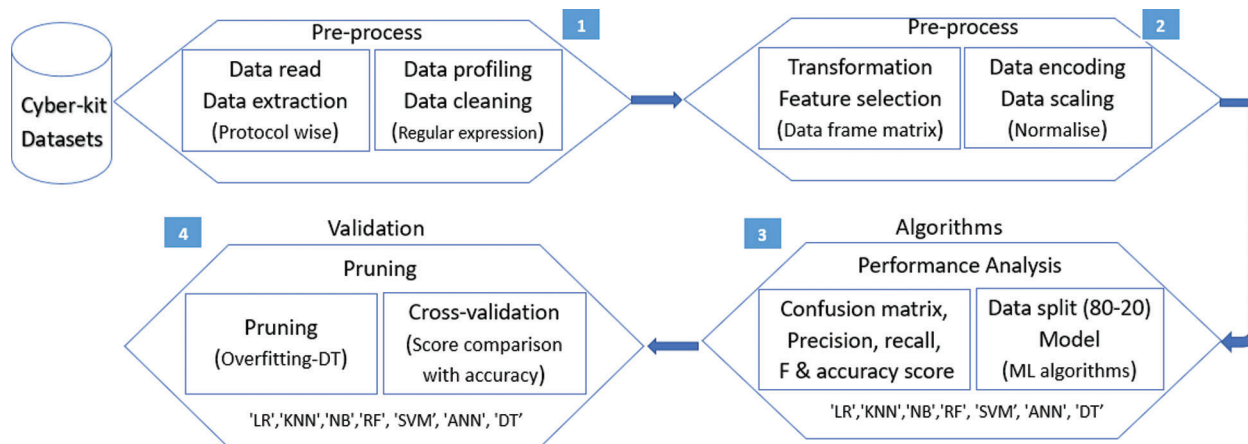


Figure 15: Machine learning steps for anomaly detection with test kit datasets [19]

The performance metrics of machine learning techniques for logistic regression, KNN, naive bayes, random forest, SVM (rbf), SVM (sigmoid), ANN, decision tree is evaluated along with its cross-validation score with varying test datasets for anomaly detection as shown in Fig. 16.

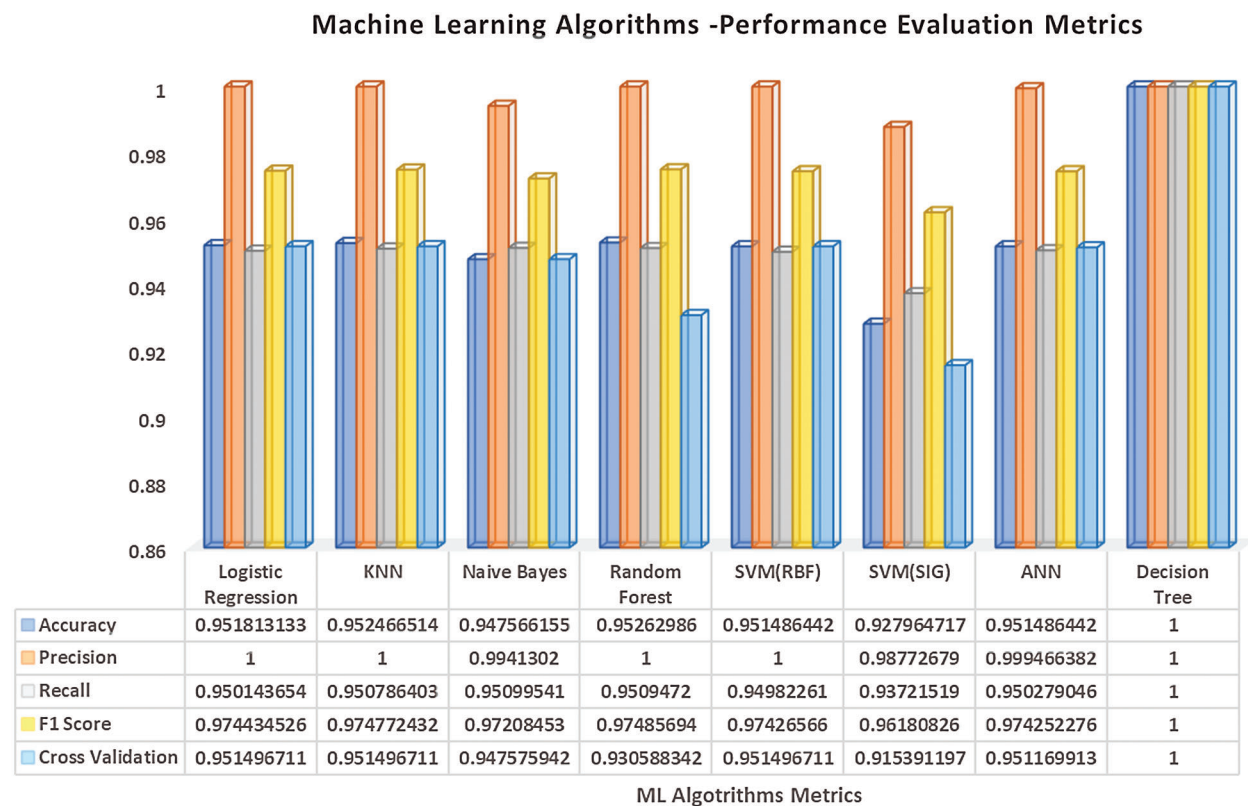


Figure 16: Machine learning algorithms performance metrics evaluation

Decision tree model has two options—entropy method for information gain where the root node is identified, and another option is the gini method for the impurity measurement. The decision tree (DT) with gini criterion method provides 100% accuracy performance due to over fitting error of test data variance, which is not acceptable.

The pruning method is performed to reduce the overfitting issue of decision tree (DT) model, by reducing the branch nodes to optimize the overfitting issue by using different cost complexity (alpha values). The corresponding training and test accuracies are identified with new decision tree (DT) models and the accuracy with cost complexity value at 4 is acceptable with 96.5% as shown in [Fig. 17](#), with less chance of overfitting.

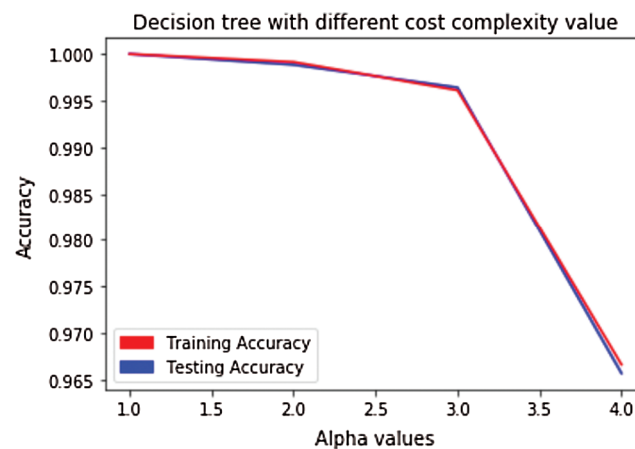


Figure 17: Pruning method of decision tree for accuracy optimization

4.2 Anomaly Data Analysis

The abnormality of industrial OT network traffic dataset with ICS cyber test kit is analysed based on the communication flags for intrusion detection. The analysis also provides information regarding asset discovery of ICS components. All the protocol data are combined and sorted in the form of dataframe for analysis (insecure HTTP, Telnet, port scan, ARP duplicate MITM for Modbus TCP) and the abnormalities in the cyber kit datasets have been identified, as mentioned in below [Figs. 18a–18d](#).

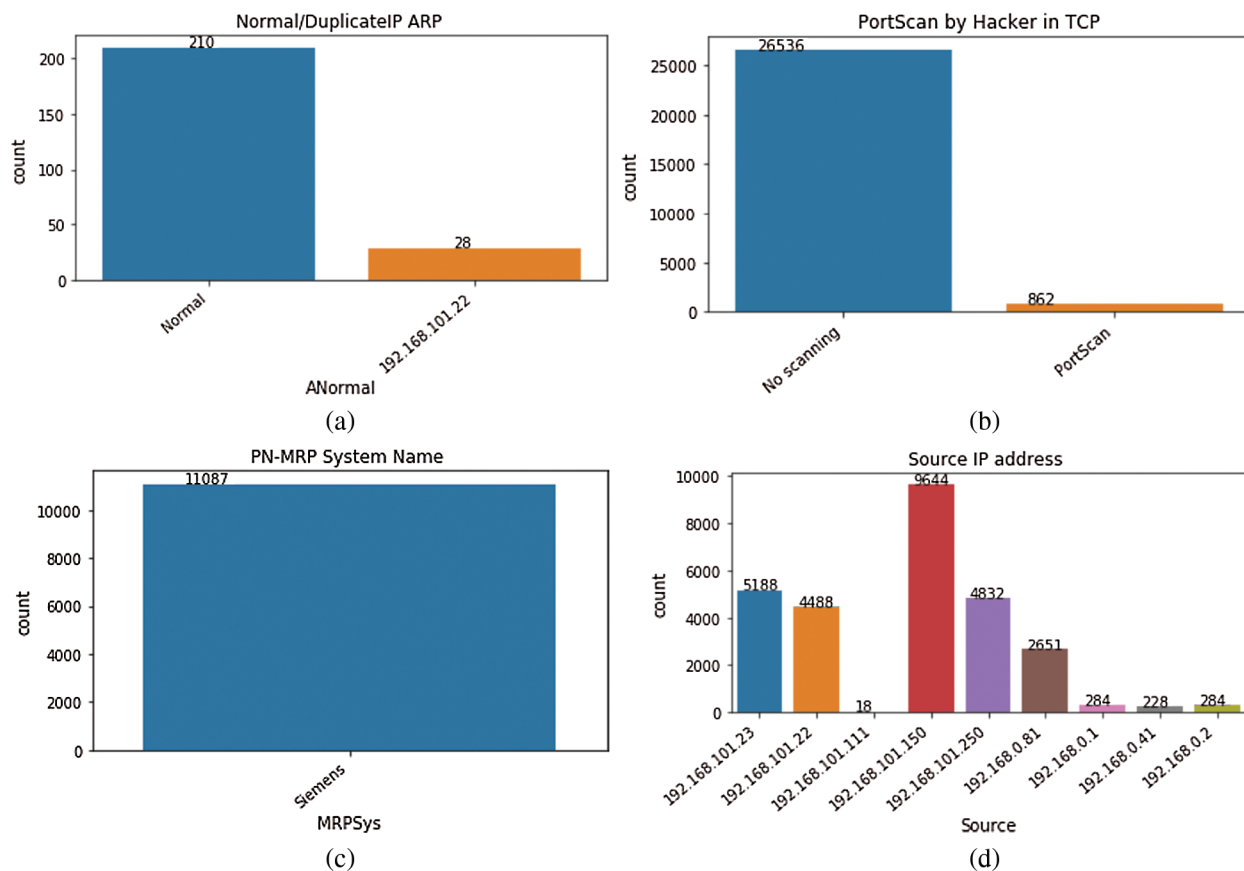


Figure 18: (a) ARP analysis, (b) Port scan, (c) Asset discovery, (d) Source IP address

5 Conclusion and Future Works

The modern control systems are vulnerable to many attack vectors, and intrusion detection methods with ML algorithms integrated are implemented to counteract and protect the system. It is worth noting that not all threats are from outside the network and the public datasets are not accurate and does not suit industrial operational scenarios.

The present drawback on the availability of real-time datasets motivated us to design an ICS portable cyber kit for real-time network traffic in industrial control systems [20]. The behavior analysis for the normal scenario of ICS network traffic is studied. Deep packet inspection is carried out to extract the metadata, where the important information is identified and extracted from “info” column based on ICS OT protocols and the capability of machine learning algorithms for anomaly detection is identified. Decision tree ML technique is a top-down recursively partitioned method based on training set attributes and faster in response. Its performance does not affect by missing values in the data and is more stable and accurate.

Ensemble learning is a technique which creates multiple models and combines them to produce improved results. Deep learning (DL) RNN (recurrent neural network) method has memory storage along with complex data handling capability. LSTM (Long short-term memory) is one of technique of RNN, which are very good at learning from sequential data can be integrated for abnormal detection classification with the test kit [21,22]. The cyber test kit package can be utilized to simulate any OT scenarios for industrial datasets, which can be utilized for research & training, and machine learning

software development. Advanced cyber-attacks such as reconnaissance, interruption (DoS), interception is simulated with penetration test tools and can be detected using deep packet inspection technologies, which are efficient way to detect zero-day attacks.

Funding Statement: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Research Group Program under Grant Number (R.G.P.1/219/42), Grant received by E. A. A. E and F. M. S. <https://www.kku.edu.sa/en/taxonomy/term/3226>.

Conflicts of Interest: The authors of this article declare no conflicts of interest to report regarding the present study.

References

- [1] W. Schwab and M. Poujol, "The state of industrial cybersecurity 2018," *Kaspersky Lab*, 2018. [Online]. Available: <https://ics.kaspersky.com/media/2018-Kaspersky-ICS-Whitepaper.pdf>.
- [2] S. Hakak, W. Z. Khan, M. Imran, K. R. Choo and M. Shoaib, "Have you been a victim of covid-19-related cyber incidents? Survey, taxonomy, and mitigation strategies," *IEEE Access*, vol. 8, pp. 124134–124144, 2020.
- [3] M. Keshk, N. Moustafa, E. Sitnikova and G. Creech, "Privacy preservation intrusion detection technique for SCADA systems," in *Military Communications and Information Systems Conf. (MilCIS)*, Canberra, Australia, pp. 1–6, 2017.
- [4] A. Almalawi, X. Yu, Z. Tari, A. Fahad and I. Khalil, "An unsupervised anomaly-based detection approach for integrity attacks on SCADA systems," *Computers & Security*, vol. 46, no. 3, pp. 94–110, 2014.
- [5] N. V. Tomin, V. G. Kurbatsky, D. N. Sidorov and A. V. Zhukov, "Machine learning techniques for power system security assessment," in *IFAC Workshop on Control of Transmission and Distribution Smart Grids*. Prague, Czech Republic, 445–450, 2016.
- [6] M. Zaman and C. Lung, "Evaluation of machine learning techniques for network intrusion detection," in *IEEE Network Operations and Management Sym. (NOMS)*, Taipei, Taiwan, pp. 1–5, 2018.
- [7] M. A. Teixeira, T. Salman, M. Zolanvari, R. Jain, N. Meskin *et al.*, "SCADA system testbed for cybersecurity research using machine learning approach," *Future Internet*, vol. 10, no. 8, pp. 76, 2018.
- [8] M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," in *IEEE 15th Int. Sym. on Intelligent Systems and Informatics (SISY)*, Subotica, Serbia, pp. 277–282, 2017.
- [9] A. Mathur and N. Tippenhauer, "SWaT: A water treatment testbed for research and training on ICS security," in *Int. Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*. Vienna, Austria, pp. 31–36, 2016.
- [10] R. L. Perez, F. Adamsky, R. Soua and T. Engel, "Machine learning for reliable network attack detection in SCADA systems," in *17th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-18)*, New York, USA, pp. 633–638, 2018.
- [11] A. Jicha, M. Patton and H. Chen, "SCADA honeypots: An in-depth analysis of Conpot," in *IEEE Conf. on Intelligence and Security Informatics (ISI)*, Tucson, USA, pp. 196–198, 2016.
- [12] A. V. Serbanescu, S. Obermeier and D. Yu, "ICS threat analysis using a large-scale honeynet," in *3rd Int. Sym. for ICS & SCADA Cyber Security Research*, Germany, 2015.
- [13] S. Mubarak, "Cyber-attacks analysis and mitigation with machine learning techniques in ICS SCADA systems," *Control Systems*, vol. 11, no. 1, pp. 9, 2019.
- [14] Siemens, "Cybersecurity: How to keep industrial control systems safe," 2021. [Online]. Available: <https://ae.webinar.siemens.com/cybersecurity-how-to-keep/3eef037e9e75aa2ee7ae>.
- [15] S. Mubarak, M. H. Habaebi, M. R. Islam, F. Diyana and M. Tahir, "Anomaly detection in ICS datasets with machine learning algorithms," *Computer Systems Science and Engineering*, vol. 37, no. 1, pp. 33–46, 2021.

- [16] J. J. Diaz, "Using snort for intrusion detection in MODBUS TCP/IP communications," *Sans*, 2011. [Online]. Available: <https://www.sans.org/reading-room/whitepapers/detection/paper/33844>.
- [17] P. Sweep, "Detect/analyze scanning traffic using Wireshark," *Pentest*, 2013. [Online]. Available: <https://www.koenig-solutions.com/documents/PenTestExtra-06-2013.pdf>.
- [18] R. Abdulhammed, H. MUSAFAER, A. ALESSA, M. Faezipour and A. Abuzneid, "Features dimensionality reduction approaches for machine learning based network intrusion detection," *Electronics*, vol. 8, no. 3, pp. 322, 2019.
- [19] X. Gao, C. Shan, C. Hu, Z. Niu and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," *IEEE Access*, vol. 7, pp. 82512–82521, 2019.
- [20] S. Mubarak and M. H. Habaebi, "Real-time ICS SCADA system cyber kit testbed with industrial hacking scenarios," *Mendeley Data*, V1, 2021. [Online]. Available: <https://data.mendeley.com/datasets/k76xhm22yj/1>.
- [21] Q. S. Qassim, A. R. Ahmad, R. Ismail, A. Abu Bakar, F. Abdul Rahim *et al.*, "An anomaly detection technique for deception attacks in industrial control systems," in *IEEE 5th Int. Conf. on Big Data Security on Cloud (BigDataSecurity)*, Washington, USA, pp. 267–272, 2019.
- [22] J. Kim, J. Kim, H. L. Thi Thu and H. Kim, "Long short-term memory recurrent neural network classifier for intrusion detection," in *Int. Conf. on Platform Technology and Service (PlatCon)*, Jeju, South Korea, pp. 1–5, 2016.