Computer Systems Science & Engineering
DOI:10.32604/csse.2021.014384
Article

Tech Science Press

Anomaly Detection in ICS Datasets with Machine Learning Algorithms

Sinil Mubarak[1], Mohamed Hadi Habaebi[1,*], Md Rafiqul Islam[1], Farah Diyana Abdul Rahman and Mohammad Tahir[2]

[1]International Islamic University Malaysia, Jalan Gombak, 53100, Malaysia
[2]Sunway University, Selangor, 47500, Malaysia
*Corresponding Author: Mohamed Hadi Habaebi. Email: habaebi@iium.edu.my
Received: 17 September 2020; Accepted: 14 December 2020

Abstract: An Intrusion Detection System (IDS) provides a front-line defense mechanism for the Industrial Control System (ICS) dedicated to keeping the process operations running continuously for 24 hours in a day and 7 days in a week. A well-known ICS is the Supervisory Control and Data Acquisition (SCADA) system. It supervises the physical process from sensor data and performs remote monitoring control and diagnostic functions in critical infrastructures. The ICS cyber threats are growing at an alarming rate on industrial automation applications. Detection techniques with machine learning algorithms on public datasets, suitable for intrusion detection of cyber-attacks in SCADA systems, as the first line of defense, have been detailed. The machine learning algorithms have been performed with labeled output for prediction classification. The activity traffic between ICS components is analyzed and packet inspection of the dataset is performed for the ICS network. The features of flow-based network traffic are extracted for behavior analysis with port-wise profiling based on the data baseline, and anomaly detection classification and prediction using machine learning algorithms are performed.

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

1 Introduction

Control system is defined as the hardware and software component of an Industrial Automation and Control System (IACS). The key components of the industrial control system (ICS) include Supervisory Control and Data Acquisition (SCADA), Human Machine Interface (HMI), Programmable Logic Controllers (PLC), Remote Terminal Unit (RTU), and Distributed Control System (DCS). A SCADA system helps to collect data from field sensors that enable us to control the system through a human-machine interface (HMI) software.

Cybersecurity solutions for Information Technology (IT) are well established and secured but less work has been done on cybersecurity for operational technology (OT) (Sentryo, 2019). In the IT sector, the confidentiality of information has the highest priority, whereas, in OT-ICS, the highest priority is for the

availability of information. The OT systems lack the cybersecurity culture and with increased digitization, more cyber-attacks surface. IT risks are sources of fraud, financial losses, privacy, and data leaks, wherein OT risks are sources of health, safety, and environmental casualties. At present, OT networks have an inconsistent deployment of security policies and standards wherein IT networks have strong security policies [1]. Applications and protocols in the OT domain are customized in SCADA, HMI, and DCS, whereas for the IT domain it is already standardized in email, internet, video, etc.

Intrusion detection systems have proved to be a reliable security process for anomaly detection in traditional IT, which identifies all inbound and outbound network traffic for security breach and check the traffic for matching signatures. Then, it signals an alarm when the matching is not found. Network-based IDS (NIDS) scans entire networks and detects malicious traffic activity, whereas Host-based IDS (HIDS) scans for a specific host and monitors each system event.

Intrusion detection systems can work conjointly with IT security systems, but unfortunately, IT systems do not meet the industrial requirements. However, the ICS cyber threats are growing at an alarming rate on industrial automation applications. The continuity of services with the safe operation is of great importance since many ICSs are in a position where a failure can result in a threat to human lives, environmental safety, or production output.

Some of the main challenges faced by OT ICS are [2] the lack of asset visibility for brownfield control systems, ongoing modifications, and upgradations in process plants. Multiple Original Equipment Manufacturers (OEM) in single plant operation are using different communication protocols. ICS vendors are not familiar with IT cybersecurity protocols or technology, and they do not have hands-on experience with ICS devices due to a shortage of experienced cybersecurity personnel.

Furthermore, many universities have difficulties to build their own OT ICS Cyber Range lab facilities dedicated to industrial use-case scenarios to carry out the research activities due to financial constraints. Currently, many researchers utilize publicly available ICS datasets for analysis of detection techniques with machine learning algorithms, as the industrial entities are reluctant to disclose the operational datasets to the public due to the sensitivity and criticality of industrial assets.

In recent years, cyber-attacks on industrial control systems had been increased many-fold due to the digitization of the industrial sector. The prime examples of notable recent industrial control system cyber-attack incidents include- Stuxnet attack on Iran nuclear facility, the Duqu & Flame attack on Iran offshore facility, the Havex remote access trojan, the Shamoon attack on Saudi Aramco, the Petya Ransomware attack in India, and the Triton-Triconex Safety Instrumented System attack on Saudi Aramco [2].

Little research had been carried to identify the advantages of using machine learning in ICS SCADA systems with real network traffic data testbed simulation and its behavior analysis for anomaly detection. The architecture of a typical modern SCADA reference model is shown in Fig. 1 consists of the following layers [3].

The root causes of cyber vulnerabilities in ICS SCADA systems are due to poorly secured legacy systems, delayed patch updates of software vulnerabilities, lack of cyber-security situational awareness, remote access for maintenance, large deployment areas, distributed operating mode, growing interconnectivity, and lack of built-in security with SCADA protocols.

The contribution of this paper is to highlight the machine learning techniques, for attack detection with SCADA public dataset and introduce innovative data profiling with flow-based behavior analysis using packet inspection of network traffic data. The dataset is processed and profiled for modeling for the abnormal prediction detection with the anomaly-based machine learning algorithms for intrusion detection in ICS systems.

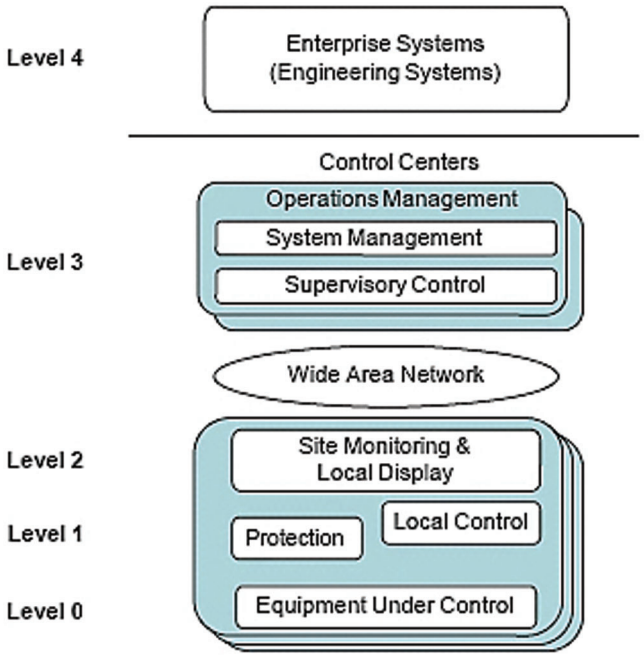

Figure 1: Typical SCADA reference model [3]

The rest of the paper is subdivided as follows. Section 2 deals with the literature review of Machine learning techniques for ICS SCADA intrusion detection systems and different types of public SCADA datasets. It also explicates the ML-based SCADA IDS steps and the performance metrics criteria for the evaluation of algorithms. Section 3 provides ML analysis of the public dataset and its performance evaluation comparison along with validation. Section 4 illustrates the network traffic analysis by data profiling and the baseline is determined by the traffic flow activities of the network with packet inspection. The feature extracted processed dataset is modeled to predict the abnormality classification in the network traffic data with the comparison of different machine learning algorithms and the paper concludes with future works for behavior analysis with multiple port-based protocol analysis and multiple anomaly criteria with hybrid machine learning algorithms. The paper concludes in section 5.

2 Related Work

This section discusses the Machine learning applications which are predominantly deployed across various industries and their applications due to their computing power, data collection, and storage capabilities. An intrusion detection system (IDS) integrated with machine learning (supervised and unsupervised techniques) can improve the detection rates of attacks for SCADA systems [4].

Machine learning algorithms are widely implemented in the intrusion detection system (IDS) to overcome the high false positives issue in prediction. Different machine learning techniques- such as supervised and unsupervised, which uses statistical techniques to learn, classify and predict the outcome methods, can be analyzed as mentioned in Tabs. 1 and 2 [5].

In supervised methods, the pre-labeled dataset feature is required (classification/regression) whereas unsupervised methods do not need pre-labeled data (dimensional reduction, clustering) for analysis. Clustering is mainly applied for forensic analysis, regression for network packet parameters prediction, and comparison with the normal ones, whereas classification is applied to identify different classes of network attacks such as scanning and spoofing [6]. An anomaly detection method for deception attacks in

the industrial control system is introduced by investigating the behavior of normal, attack-free activities [7]. Different existing intrusion detection systems using artificial neural networks (ANN) for detecting malicious network activity for different datasets have been reviewed [8]. A novel dataset focused on IoT combined network, power features and attacks utilizing WEKA application to train, test, and cross-validate the dataset for classification of detection with Naive Bayes (NB), support vector machines (SVMs), multilayer perceptron (MLP), Random Forest (RF), ZeroR ML classifiers has been introduced [9].

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

Table 2: Machine learning—Unsupervised methods

Algorithm	Technique
k-means	Clustering–Iteration process
K-medoid	Robust clustering, less sensitive to noise
Principal component analysis	Dimensional reduction methods

2.1 Public SCADA Datasets

A framework for security testbed for Modbus/TCP-based has been introduced in [9] for SCADA security evaluation and testing environment. The analysis of machine learning algorithms for SCADA systems can be performed with industrial public datasets, mathematical modeling of the system, and using ICS cyber test kit with OT network traffic simulation. The building of detection models using SCADA data is performed using manual definition- which is time-consuming and expensive or using the machine learning strategies that automatically build a detection model based on the training data set [10].

In [11], SCADA dataset features were extracted from captured network traffic and the performance of different supervised ML algorithms such as Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbor (KNN) compared. In [12], the multi-class power system datasets include 37 scenarios of both normal and attack instances, and the three ML techniques of KNN, NB, RF methods were analyzed.

In [13], a testbed was designed for supervised ML approach for anomaly detection of energy monitoring-based water supply system, and the three different datasets obtained from the testbed were analyzed with Random forest, KNN, and SVM algorithms. A real-time dataset which includes normal traffic along with 35 types of cyber-attacks, is utilized to train and test the ML classifiers intrusion detection system but shows a high false positive rate for the algorithms [14].

To overcome the present IDS drawbacks, the implementation of Machine learning techniques with IDS integrated along with real operational technology traffic data has become a vital and innovative concept.

Furthermore, a cyber-physical ICS testbed can provide a hands-on simulation platform with real-time network data with various types of cyber-attacks for security evaluation and testing environment for research purposes.

2.2 Machine Learning Algorithms–ICS SCADA

Machine Learning (ML) based intrusion detection in SCADA systems follow the steps below as represented in Fig. 2 [15].

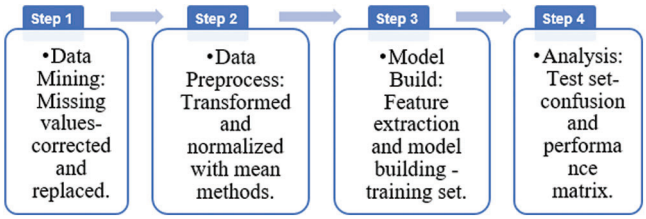

Figure 2: ML-based SCADA IDS steps

In the data cleaning and mining stage, missing values in the SCADA dataset are corrected, split randomly into training and test sets for better results. Then data normalization is followed where the improper features are replaced with mean and normalized values. Then the dataset features are extracted, and the model is built based on the machine learning algorithms to detect anomalies in the dataset. Finally, the IDS performance can be analyzed with parameters, such as Accuracy, precision, sensitivity (recall), receiver operating curve (ROC), F-score, etc.

The common IDS, like Snort and Bro, alerts can be classified as mentioned in Tab. 3, with True Negative (TN) for no attack-no alert, False Positive (FP) for no attack-alert, False Negative (FN) for attack-no alert, and True Positive (TP) for attack-alert.

Table 3: IDS evaluation matrix

Actual data (Attack)	Predicted data (Alarm)	
	Negative (0)	Positive (1)
Negative (0)	True Negative (TN-00)	False Positive (FP-01)
Positive (1)	False Negative (FN-10)	True Positive (TP-11)

3 Methodology and Analysis with Public Datasets

This section describes machine learning analysis of public datasets. SCADA datasets with attack vectors are used for the evaluation of different machine learning algorithms' performances. Most of the datasets such as KDD 1999, DARPA, Gao's dataset is outdated and are associated with information technology systems, which are also unsuitable for SCADA IDS research. An improved Cyber-physical SCADA dataset from Mississippi state university's in-house SCADA lab which contains both normal and attacks traffic is used to evaluate the ML algorithms performance for SCADA IDS. The dataset contains network traffic data with 274,628 instances having normal activity along with 35 cyber-attacks class subtypes of data flow.

Each instance of Modbus RTU packets contains 20 features with Man in the middle (MITM) attacks, 214580 normal instances (78%), and 60048 (22%) attack instances are represented in below Tab. 4.

Table 4: Modbus RTU packets instance features

S.n	Features	Features Description
1/2	Address/Function	Modbus slave device address/Modbus function
3/4	Length/Set point	Modbus packet length/Pressure set point–Auto
5/6	Reset rate	PID gain/PID reset rate
7/8	Deadband/Cycle time	PID dead band/PID cycle time
9/10	Rate/System mode	PID rate/Automatic (2), manual (1), or off
11	Control scheme	Either pump (0) or solenoid (1)
12	Pump	Pump control; on (1) or off (0).
13	Solenoid	Relief valve; opened (1), closed (0)- manual
14/15	Pressure/Command	Pressure measurement/ Command (1) or response
16/17	CRC rate/Time	Cyclic redundancy check rate/Timestamp
18	Binary	Attack (1) or normal (0)
19	Categorized	Category of attack (0–7)
20	Specific	Specific attack

The dataset is randomly split into 80% training sets for modeling and the rest 20% test sets for ML algorithm evaluation. The dataset includes 274628 instances with a training set of 219,702 and a test set with 54,926 observation instances.

3.1 Machine Learning with R-Studio

The binary classification is evaluated with the programming tool R Studio, which is an open-source environment for statistical computing for data analysis. The dataset (in .csv format) fetched by the R-studio program, is corrected and split into training and test sets. The data features are normalized and modeled with a training set for ML algorithms. The supervised methods (logistic regression/KNN) are used to model and train the dataset and binary classification of pre-labeled output label feature number#18 as shown in Fig. 3, is predicted and compared with the test dataset and its performance is evaluated with the confusion matrix.

S. N	1	2	3	4	5	6	7		8	9	10
Features	A	F	L	P	PID Parameters						S
Features Value	4	16	90	115	0.2	0.5	1		0	0	0

11	12	13	14	15	16	17	18	19	20
C	P	S	P	C	CRC	Time Stamp	Output Label		
1	0	0	?	1	17219	141862164.995592	1	1	1

Figure 3: Modbus RTU packets instance [16]

Logistic regression function is used to model with the training set and probability for binary attribute classification is predicted on the test set which detects the normal/attack status. The logistic regression (LR) confusion matrix evaluation is performed with the R-studio platform for the test dataset and provides a prediction accuracy of 99.99%, for total observations of 54926 instances, as mentioned in below Tab. 5.

Table 5: Logistic regression confusion matrix evaluation

SCADA_Binary Test	SCADA_ Prediction (0)	SCADA_ Prediction (1)	Row Total
0	42917	0	42917
1	0	12009	12009
Column Total	42917	12009	54926

The data analysis with the KNN ML technique mentioned in Tab. 6, shows an accuracy parameter of 83.72% for the total observations of 7500 instances.

Table 6: KNN confusion matrix evaluation

SCADA_Binary Test	SCADA_ Prediction (0)	SCADA_ Prediction (1)	Row Total
0	6078	0	6078
1	1221	201	1422
Total	7299	201	7500

3.2 Machine Learning Algorithms-WEKA Platform

This section describes the machine learning analysis of public datasets. The WEKA platform has pre-processing tools called filters for attribute selection, normalization purposes, and classifier models for predicting nominal or numeric quantities, such as support vector machines, logistic regression, BayesNet, decision trees–J48 method, Meta-classifiers: bagging, boosting, and voting stacking algorithms. The dataset in ARFF format is fetched by the application, which is then pre-processed. And relevant features are filtered based on a ranking method for training the dataset with base learner Decision tree-J48 classifier ML algorithm. The classification performance is evaluated with another BayesNet ML classifier. The hybrid (J48 and BayesNet) ML classifier is applied with the base learner–J48 decision tree and meta classifier–BayesNetwork to obtain the best prediction capabilities. Once the dataset is loaded with all features, pre-processed, and the feature extraction method is built, the dataset attributes are ranked based on the information gain parameter. Then the decision tree-J48 classifier is used to train the feature filtered dataset, as the base learner, while the BayesNet classifier is used for the hybrid model for classification performance evaluation as shown in Fig. 4.

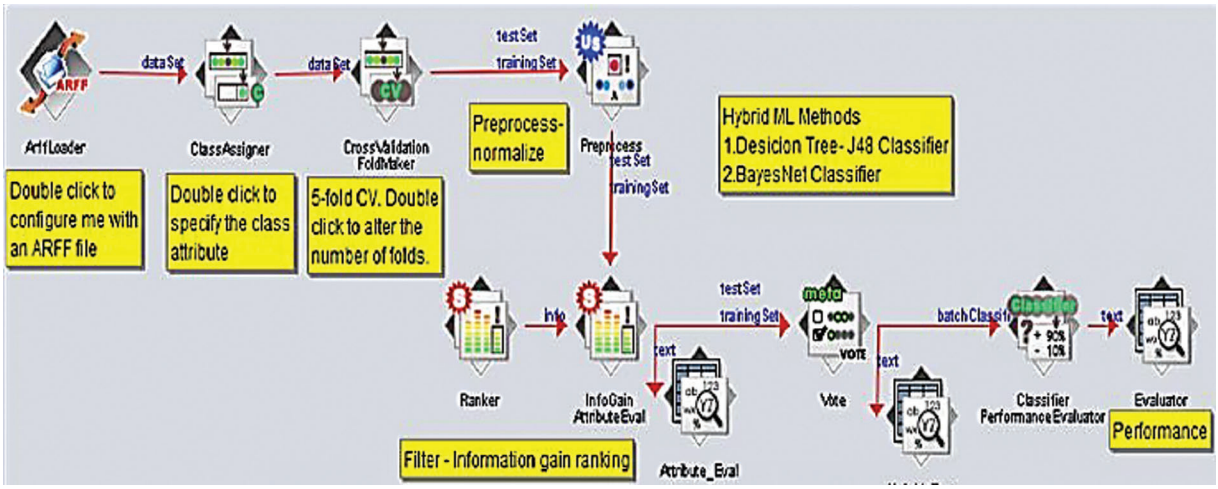

Figure 4: Hybrid ML algorithm classifier

The following five feature attributes are extracted, filtered, and ranked based on the information gain ranking attribute selection criteria method, as mentioned in Tab. 7.

Table 7: Feature extraction

Rank	Attribute (No.)	Gain
1	Time (17)	0.3077
2	Pressure (14)	0.1622
3	CRC rate (15)	0.1408
4	Length (3)	0.0885
5	Function (2)	0.0809

The ML performance metrics of hybrid classifier algorithm with instances–25000 and 274628, 5-fold cross-validation, for the five attributes are evaluated in Tab. 8.

Table 8: Hybrid classifier algorithm for 25000 instances

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.996	0.028	0.992	0.996	0.994	0.996	0
0.972	0.004	0.988	0.972	0.978	0.996	1
0.991	0.023	0.991	0.991	0.991	0.996	

The ML classification performance of different algorithms has been compared in Tab. 9.

The machine learning algorithms performances evaluated are benchmarked in Tab. 10, with the following results on SCADA and KDD datasets.

Table 9: ML classification algorithms performance

ML Algorithm	Instance	Attributes	Accuracy
Logistic	274628	18	99.99%
KNN	25000	18	83.72%
Logistic	25000	18	83.84%
SVM	25000	18	86.25%
J48	25000	18	97.79%
Bayes Network	25000	18	91.73%
Hybrid: LR + BayesNetwork	25000	05	89.37%
Hybrid: J48 + BayesNetwork	25000	05	99.09%
Hybrid: J48 + BayesNetwork	274628	05	100.00%

Table 10: Benchmark ML performance on datasets

ML Algorithm on SCADA dataset	Instances	Accuracy
SVM	54927	94.36%
BLSTM	54927	98.40%
Random Forest	55251	99.41%
ML Algorithm on KDD dataset	Instances	Accuracy
J48	6000	93.10%
Bayes Network	6000	90.73%

4 Methodology and Results with ICS Network Dataset

The goal of this section is to analyze the network traffic data which is encapsulated in network packets as a .pcap file format. The .pcap file is taken from Wireshark and converted into .CSV file with the Spyder platform for machine learning prediction analysis.

The Wireshark is used to capture, analyze signals, and data traffic over the communication channel. Such a channel varies from a local computer bus to a satellite link, that provides a means of communication using a standard communication protocol (networked or point-to-point). The network traffic data (pcap file) is used for prediction analysis and convert the .pcap file to .csv format for ease of use and analysis.

The activity between components is a set of traffic between two components/devices. The traffic dataset is imported with Spyder python and the initial observation is as in Tab. 11.

Table 11: Dataset with column description

Column	Description
Time	The timestamp of each captured packet
Source	Source IP of the TCP communication
Destination	Destination IP of the TCP communication
Protocol	The protocol used to communicate between source and destination
Length	Data packet Length
Info	Wireshark packet of the summarized packet that specific communication

The dataset has 86799 instances with 06 data columns, without any null values and 'NA' character values, as highlighted in Fig. 5.

4.1 Profiling of Network Traffic Data

The packet inspection of network data traffic is performed with both pre-processing and post-processing techniques. The traffic flow-based intrusion detection serves as an anomaly-based intrusion detection system where the baseline is determined by the flow of the network. Pre-processing of network traffic data is performed with the Spyder platform [17]. The Scientific Python Development Environment (Spyder) is an integrated development environment (IDE) that has libraries: such as Regular expression to filter and remove the unwanted expressions: =, [ ] < > from the dataset.

```
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Time         86798 non-null  float64
 1   Source       86798 non-null  object
 2   Destination  86798 non-null  object
 3   Protocol     86798 non-null  object
 4   Length       86798 non-null  int64
 5   Info         86798 non-null  object
dtypes: float64(1), int64(1), object(4)
memory usage: 4.6+ MB
```

Figure 5: Dataset with data column and datatype

The traffic flows are a port-wise set of packets and have different protocols such as TCP, UDP, ICMP has different flow properties. The Transmission Control Protocol (TCP) operates at transport layer of OSI layer for services/applications with dedicated destination port: HTTP (80), FTP (20, 21), Telnet (23), SMTP (25), DNS (53), HTTPS (443), Modbus (502), ISO-on-TCP–Siemens S7 Communication (S7comm) Protocol (RFC 1006)—(102), whereas the User Datagram Protocol (UDP) includes: SNMP (161), Syslog (514) [18].

The pre-processing of data is based on the communication protocols which is "TCP" and is assigned to discrete output value: 1, for further analysis and classification. The dataset is post-processed with a column named Info, which is then further extracted and assigned to each separate column for prediction and classification analysis. This profiling is done with functions with the Spyder python platform.

The feature extraction of the dataset is obtained by splitting the Info column of network traffic data which is a critical part of data analysis and each of the relevant features from Info columns such as source port, destination port, Ack, Seq, Len Packet, Window is extracted by filtering unwanted characters, which is vital for training and testing is shown in Tab. 12.

Table 12: Post-processed dataset with feature extraction

	Source	Destination	Protocol	Length	Source_ Port	dest _Port	Ack	Seq	Len Packet	Window
0	219.216.128.25	192.168.68.130	1	1506	80	43624	1	1	1452	64240
1	192.168.68.130	219.216.128.25	1	54	43624	80	1453	1	0	65535
2	219.216.128.25	192.168.68.130	1	1506	80	43624	1	1453	1452	64240
3	192.168.68.130	219.216.128.25	1	54	43624	80	2905	1	0	65535
4	219.216.128.25	192.168.68.130	1	1506	80	43624	1	2905	1452	64240
5	192.168.68.130	219.216.128.25	1	54	43624	80	4357	1	0	65535

The behavior analysis of traffic data is performed with packet inspection, the data flow is analyzed, and classification output is identified based on the column Info data baseline, as shown in Fig. 6. In normal scenario result, an output of '1' is assigned for classification, whereas in anomaly scenario result, an output of '0' is assigned based on keywords "Dup ACK," "Previous segment not captured."

[TCP Dup ACK 400#3] 43624 > 80 [ACK] Seq=1 Ack=310012 Win=65535 Len=0

TCP Dup ACK 86798#1 32586 502 ACK Seq=1141 Ack=2186 Win=64768 Len=0

The output has 44814 counts for result value '0' where-as the result value '1' has 41220 counts. The network traffic dataset having labels and relevant features are trained and modeled with different machine learning algorithms and result classification is predicted with machine learning analysis.

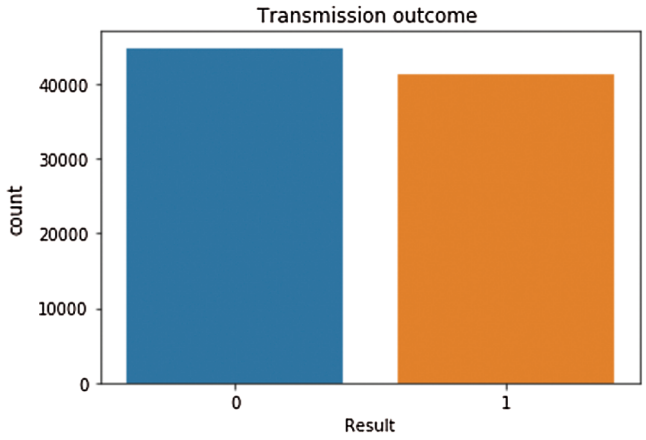

Figure 6: Profiled classification result

4.2 Results of Anomaly Prediction with Machine Learning Algorithms

The dataset which is processed for model evaluation as mentioned in Fig. 7, is split into 80% train data and 20% test data and the training data is utilized to model with independent variables for different Machine Learning Algorithms (MLA) and accuracy classification is predicted for test and train data with confusion matrix parameters, mentioned in Fig. 8.

Source	Destination	Protocol	Length	Source_Port	destination_Port	Ack	Seq	Len_Packet	Window	Result
219.216.128.25	192.168.68.130	1	1514	80	43638	140	229963	1460	64240	1
219.216.128.25	192.168.68.130	1	1514	80	43642	160	1359209	1460	64240	1
192.168.68.130	219.216.128.25	1	54	43628	80	119113	160	0	64240	0
192.168.68.130	219.216.128.25	1	54	43642	80	573175	160	0	65535	0
219.216.128.25	192.168.68.130	1	1506	80	43638	140	145201	1452	64240	1
219.216.128.25	192.168.68.130	1	1514	80	43624	1	1184541	1460	64240	0
192.168.68.130	219.216.128.25	1	54	43630	80	69697	155	0	65340	1

Figure 7: A processed dataset for model evaluation

The logistic regression model is evaluated, and the predicted binary outputs are represented as a probability function that is converted to discrete '0' and '1'. The training accuracy is at 65.23% where-as maximum test accuracy is at 65.15%.

K-Nearest Neighbor (KNN) model is applied with nearest neighbors algorithm and the K-value which is the threshold point at which the performance of train/test accuracy start to dip or decrease is determined for train and test dataset. K-value is the odd increment value. Each instance and the training dataset has a K-value of 9 for the 5th element while the test dataset has a K-value of 7 for the 4th element as identified in below Fig. 9. The training accuracy is at 71.43% where-as maximum test accuracy is at 69.75%.

The Naïve Bayes model has two options for independent variables. The Gaussian method has more accuracy for continuous variables, whereas the Multinomial method has higher accuracy for categorical discrete independent variables. The training accuracy is at 61.69% where-as maximum test accuracy is at 61.30%.

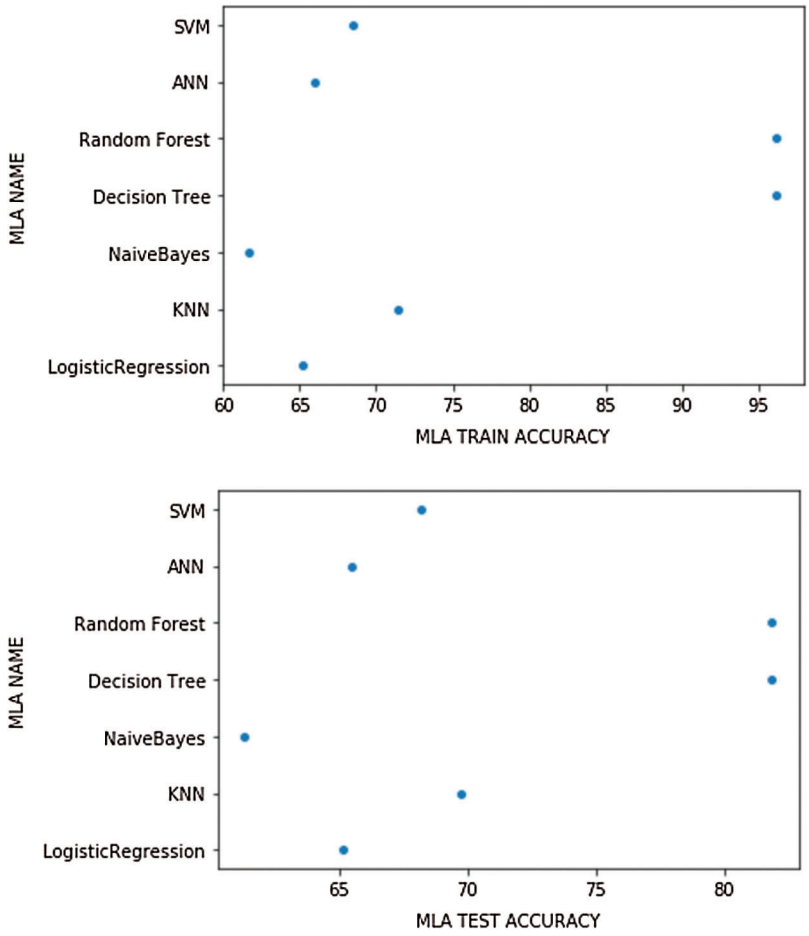

Figure 8: MLA train and test dataset accuracies

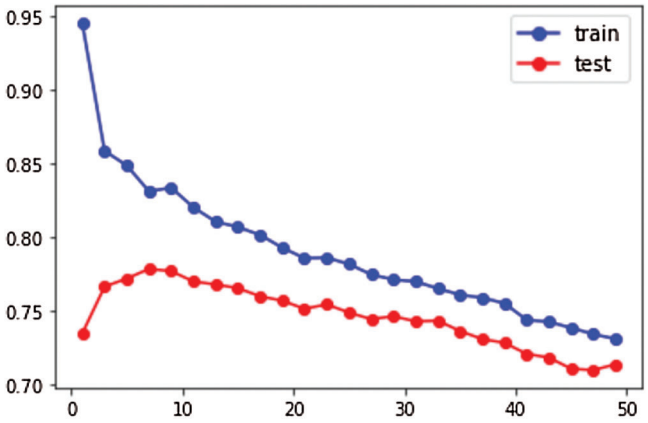

Figure 9: KNN K-value for train and test dataset

The decision tree model has two options—the entropy method for information gain where the root node is identified, and the other option is the Gini method for the impurity measurement. The Decision tree with entropy method exhibits the highest train accuracy parameter with 96.18 whereas test accuracy Random

Forest model averages the multiple decision trees and provides better accuracy. The training accuracy is at 61.69% where-as maximum test accuracy is at 61.30%.

Artificial Neural Network (ANN) is a hybrid model with the black box technique, where each layer can have 100 networks. The training accuracy is at 65.94% where-as maximum test accuracy is at 65.45%. Support Vector Machine (SVM) uses a hyperplane (linear boundary) method and has different kernel types-rbf, poly, sigmoid which can reduce overfitting.

The comparison of machine learning algorithms performance is mentioned in Tab. 13.

Table 13: MLA train and test dataset accuracy performances

Name	Train accuracy	Test accuracy	Difference
Logistic Regression	65.23	65.15	−0.08
KNN	71.43	69.75	−1.68
Naïve Bayes	61.69	61.30	−0.39
Decision Tree	96.18	81.85	−14.33
Random Forest	96.16	81.95	−14.21
ANN	65.94	65.45	−0.49
SVM	68.48	68.20	−0.28

5 Conclusion and Future Works

The signature-based detection for ICS OT cyber-attacks using R-studio and WEKA platform utilizing public datasets has been analyzed. It is noted that the public datasets are not accurate and do not suit industrial use-case scenarios. An innovative behavior analysis of network traffic in ICS with a baseline model is performed with the Spyder python platform. The flow-based network traffic data is profiled with single communication protocol-based behavior analysis for the normal scenario of the ICS network traffic with packet inspection and classification is predicted with different machine learning algorithms for anomaly detection. The Cyber Security Management Systems (CSMS) provide well-established methods with high accuracy for protecting the control system assets from cyber-attacks which includes the development of the basic cybersecurity policies, and its compliance with ISA/IEC 62443 standards.

In future work, the real-time ICS network traffic data will be extracted and a completely generic anomaly detection system without the need for prior knowledge of variables will be proposed to be developed, as in [19,20]. The Packet Capture (PCAP) files for real-time network data analysis can be performed with industrial sensors to obtain the relevant metadata from the OT network. The behavior analysis with multiple port-based protocol analysis and multiple anomaly criteria with hybrid machine learning algorithms using real-time industrial control system integrating cyber-attack test cases with portable ICS cyber kit will be implemented in future works. Advanced cyber-attacks such as reconnaissance, interruption (DoS), interception (MITM), firmware analysis can also be simulated with penetration test tools.

Funding Statement: This work was conducted at the IoT and wireless communication protocols laboratory, International Islamic University Malaysia and is partially sponsored by the Publication-Research initiative grant scheme no. P-RIGS18-003-0003.

Conflicts of Interest: The authors of this article declare no conflict of interest.

References

[1] ISA, "Security for industrial automation and control systems, Part 3-3: *System Security Requirements and Security Levels*," 2013.

[2] D. McMillen, "Security attacks on industrial control systems," 2016. [Online]. Available: https://securityintelligence.com/attacks-targeting-industrial-control-systems-ics-up-110-percent/.

[3] L. Van, "Sequential detection and isolation of cyber-physical attacks on SCADA systems," Ph.D. Thesis. University of Technology of Troyes, 2015.

[4] 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, 1–6, pp. 2017.

[5] L. A. Maglaras and J. Jiang, "Intrusion detection in SCADA systems using machine learning techniques," Ph.D. Thesis. University of Huddersfield, UK, 2018.

[6] Q. Qassim, "An anomaly detection technique for deception attacks in industrial control systems," in *IEEE 5th Intl. Conf. on Big Data Security on Cloud*. Washington, DC, USA, 267–272, 2019.

[7] R. L. Perez, F. Adamsky, S. Ridha and E. Thomas, "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, 2018.

[8] I. Solomon, A. Jatain and B. Shalini, "Neural network-based intrusion detection: State of the art." India: International Conf. on Sustainable Computing in Science, Technology and Management (SUSCOM-2019), Amity University Rajasthan, 2019.

[9] J. Foley, N. Moradpoor and H. Ochen, "Employing a machine learning approach to detect combined internet of things attacks against two objective functions using a novel dataset," *Security and Communication Networks*, vol. 2020, no. 2, pp. 1–17, 2020.

[10] 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, pp. 94–110, 2014.

[11] N. Moradpoor and A. Hall, "Insider threat detection using supervised machine learning algorithms on an extremely imbalanced dataset," *International Journal of Cyber Warfare and Terrorism*, vol. 10, no. 2, pp. 1–26, 2020.

[12] 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. 76, pp. 1–15, 2018.

[13] A. Robles-Durazno, N. Moradpoor, J. McWhinnie and G. Russell, "A supervised energy monitoring-based machine learning approach for anomaly detection in a clean water supply system," in *Proceedings of the IEEE Int. Conf. on Cyber Security and Protection of Digital Services*, Glasgow, UK, 2018.

[14] I. Turnipseed, "A new SCADA dataset for intrusion detection research," Ph.D. Thesis. Mississippi State University, USA, 2015.

[15] A. Polyakov, "Machine learning for cybersecurity 101, *Dzone*, *AI Zone*, 2018. [Online]. Available: https://dzone.com/articles/machine-learning-for-cybersecurity-101.

[16] R. Vinayakumar, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525–41550, 2019.

[17] P. Singh, "Machine learning with PySpark," Apress, Springer Nature Publishing Co., NY, USA, 2018.

[18] A. Almehmadi, "SCADA networks anomaly-based intrusion detection system," in *SIN'18: Proceedings of the 11th Int. Conf. on Security of Information and Networks*, Cardiff, UK, pp. 1–4, 2018.

[19] R. Malaiya, D. Kwon, C. Suh, H. Kim and J. Kim, "An empirical evaluation of deep learning for network anomaly detection," *IEEE Access*, vol. 7, pp. 140806–140817, 2019.

[20] S. D. Anton, L. Ahrens, D. Fraunholz and H. D. Schotten, "Time is of the essence: Machine learning-based intrusion detection in industrial time series data," IEEE Int. Conf. on Data Mining Workshops (ICDMW), pp. 1–6, 2018.