National College of Ireland

Intrusion Detection System for Industrial Control Systems using Classification Techniques

MSc Research Project
MSc Cyber Security

Prabhjeet Singh Multani
Student ID: x20153449

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Prabhjeet Singh Multani

Student ID: X20153449

Programme: MSc in Cybersecurity **Year:** 2021-2022.

Module: Industry Internship

Supervisor: Vikas Sahni

Submission Due Date: 07/01/2022

Project Title: Intrusion Detection System in Industrial Control System using Classification Techniques

Word Count: **.....**6314..... **Page Count**.....21...

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....Prabhjeet Singh Multani.....

Date: .....06/01/2022.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□x
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Intrusion Detection System for Industrial Control Systems using Classification Techniques

Prabhjeet Singh Multani
x20153449@student.ncirl.ie

Abstract

The operational technology using intelligent devices and make all the machineries work automatically to increase the production and the attacks in these systems are increasing exponentially. Attackers can send the malicious data in the SCADA system to interrupt the function or cause severe damage. The researchers implemented several models to improve the intrusion detection system on this problem. The explanation in this paper is about the implementation of classification techniques using machine learning algorithms for building the model with the ICS dataset which is provided in the Mississippi university platform. Dataset has been cleaned, transformed, and analysed. In machine learning total four number of models are implemented such as Logistic Regression, Random Forest, Decision Tree, and XGB classifier. As mentioned earlier dataset is about industrial control system, where the bulk of units are allotted to natural and attack.

The proposed study focuses on machine learning techniques to detect attacks in the SCADA network. After implementing four models, the best detection model will be selected for IDS in ICS. The research was carried out to measure the accuracy, precision, recall, and f1-score and compared to get the best model for the detection of the attacks.

Keywords: - SCADA systems, Intrusion Detection System, Random Forest, Decision Tree, Logistic Regression, XGB classifier

1 Introduction

In the current scenario, the work style has become more advanced where humans utilize the latest technology to increase the production and achieve a high profit every year, which makes the business more profitable. This field includes like manufacturing, transportation systems, and many more. SCADA systems (supervisory control and data acquisition) are used to monitor and control a variety of industrial and infrastructural activities. SCADA is implemented in essential infrastructure assets such as electric power, water distribution networks, and a variety of other industries (Benisha and Raja Ratna, 2019).

This research will answer the following question, allowing future IDS systems to have a superior and more efficient analytical engine to extract actual alerts.

- Which model would be best suitable for classification techniques in the SCADA network?

For this research, intrusion detection system using classification techniques in ICS. Which basically detects the threats based on the dataset. The dataset is obtained from Mississippi university platform and in this dataset, the two dependent variables are a) attack and b)

natural. There are several other variables in this dataset that could be utilized to develop and analyse the machine learning model. There are several solutions available in the market but there are some limitations, and this research is trying to offer the best possible solution which can detect the signature-based attack using four algorithms that will be used in this research such as Logistic Regression, Random Forest, Decision Tree, and XGB classifier.

In this research, four separate models were trained using classification techniques. The next section would be related to previous work done by other researchers, methodology, design and specification, Implementation, Evaluation, and at last conclusion.

2 Related Work

In Industrial Control Systems, the research has done by other researchers in intrusion detection system using machine is going to be considered for our research. There is extensive work in Intrusion detection systems in SCADA. Our focus in this section is on previous research done by researchers detecting threats in the ICS networks, where many researchers have proposed work in a similar field with different strategies. So, their methodologies, frameworks, implementation techniques and future work will be explained. Mostly, the research is done using the machine learning technique. In Industrial Control System, the research is done by other researchers is going to be considered for our research.

2.1 Researched work on Machine Learning

The researchers (Zaki Khan and Serpen, 2019) implemented machine learning algorithms applying WEKA to develop or design to detect the threats on the SCADA system network of a gas pipeline infrastructure. Along with naive Bayes, two algorithms were trained as classifiers. Both the PART and Random-forest techniques produced good results (97.6%) in terms of performance for both binary and multi-class cases on the dataset, though the Random-forest classifier surpassed other algorithms in performance.

In this research, researchers (Tamy et al., 2019) evaluated the machine learning algorithms to detect the threats or attacks in SCADA where they have compared four ML algorithms results. Naïve Bayes, Random Forest, SVM and Tress J48. The data set used in this research was taken from a laboratory-scale gas pipeline. Their main purpose was to select the best possible algorithm to detect the attacks in the SCADA system. After measuring the precision, error rate, and accuracy in classification techniques, the best algorithm to detect the intrusion in SCADA is Random Forest with 99.30% of accuracy rate.

The researchers (Lopez Perez, Adamsky, Soua and Engel, 2018) performed research for reliable network attack detection in SCADA systems using a machine learning approach. The dataset was gathered from a public platform and relates to the gas pipeline system. They applied Random Forest and SVM to generate diversity in IDS, and they evaluated and compared both methods for the best accuracy rate, recall, and precision along with random hyper-parameter search results. The Random Forest algorithm has the best detection rate in SCADA, with 99.90 percent of benign data and 98.46 percent of assaults detected, and an overall detection rate of 99.58 percent.

This research done by researchers (Syamsul Arifin et al., 2021), the dataset including DoS attacks and running IEC protocol was used to identify DoS attacks on SCADA using a

machine learning technique. The IDS attack on the SCADA system was created using three machine learning models: SVM, Decision Tree, and Gaussian Nave Bayes. With a 99.99 percent accuracy rate, Decision Tree is the best approach for detecting DoS on both the testing and training datasets. The goal for the future is to build a comprehensive data attack in a dataset, especially for the SCADA IEC 60870- 5-104 system, where some packets can be recognized in the SCADA protocol.

The research (Alhaidari and AL-Dahasi, 2019) proposed an enhanced framework for identifying DDoS attacks using multiple machine learning methods such as Random Forest, J48, and Naive Bayes to discover attack patterns. KDDCup'99 dataset got trained and analyzed by these algorithms. Among all three approaches in machine learning the Random Forest classifier achieved the best performance (99.9998%) which just minimal difference with J48 and NB accuracies. Future work, a different dataset that would best fit SCADA systems using similar algorithms and other ML algorithms can also be used against different parameters selection.

The researchers (Rajesh and Satyanarayana, 2021) used a real-time SCADA test bed to investigate detecting malicious traffic in SCADA networks using machine learning methods. They developed their own dataset that included both normal and attack data in network traffic. Following that, feature extraction techniques such as Chi-Square, ANOVA, and LASSO were utilized to reduce the dimensionality of the feature dataset. Data plots were also used to balance out an unbalanced dataset. The algorithms RF, KNN, SVM, and NB are utilized to provide performance measures. With a receiver operating characteristic value of 99.96 percent, the SVM algorithm with filtering and SVMSMOTE approach exceeds the other three machine learning algorithms.

Researchers (Marsden, Moustafa, Sitnikova and Creech, 2017) suggested a study on probability risk identification-based intrusion detection system technique, which essentially analyses Modbus TCP/IP network packets for identifying replay attacks. The techniques used in this paper to evaluate the hybrid environment to configure a testbed which scalable, accurate and low-cost SCADA network. PRI approach is implemented and compared to three other machine learning algorithms, with KNN, Nave Bayes, and Random Forest showing lower DR and FPR than PRI. As a consequence of the configurable risk assignment to individual packets, the PRI-IDS outperforms rival algorithms for Modbus TCP, depending on the prospective process of each approach being tested. To improve performance, future development and reimplementation may be done in a real-time environment.

Researchers (Almseidin, Alzubi, Kovacs and Alkasassbeh, 2017) conducted many experiments and tests to measure the efficiency and productivity of machine learning algorithms in this study. More than six approaches were evaluated using the KDD intrusion detection dataset. They discovered that all machine learning algorithms are worthless for handling all forms of attacks, including RF, MLP, J48, and others, in this experiment. While detecting normal packets, Random Forest had the best accuracy rate and Bayes network classifiers had the best detecting accuracy rate. Other machine algorithms, unlike MLP, can design a training model in an acceptable length of time. True positive and average accuracy rates alone are insufficient to detect the incursion under this limitation, hence false positive and false negative should be considered.

Researchers (Khan et al., 2019) at IDS conducted a study in a SCADA system utilizing a hybrid multilevel approach, proposing HML as an anomaly detection approach in machine

learning. First, they accurately tested and trained the dataset model in this proposed model for detection. The main goal is to increase accuracy and plot the un-balanced dataset. The reliable insights between regular and irregular behaviour of the system were confirmed using DFR for extracting the features that are the result of the advised approach at DFR, as well as Bloom filter and KNN. In summary, the proposed IDS model can achieve high efficiency while maintaining a low cost. Future work will be done through deep learning approach to improve the DR.

The research carried-out by researchers (Duque Anton, Sinha and Dieter Schotten, 2019), Intrusion detection system in Operational technology using machine learning algorithms, Random Forest and SVM models were used to detect the attack where Random Forest detection rate were 90% and 95% on two datasets which was more than SVM, but algorithm capable of increasing the detection capabilities. For the future work, this model needs to be work on some improvements such as detecting the undetected attack.

The researchers (Perales Gomez et al., 2019) used ICS to perform anomaly identification on datasets, which involved 4 stages. The first two stages demonstrate how to attack the testbed and in next two stages capture the network traffic and extract the features. Two subsets were generated called Electra S7Comm and Electra Modbus. Support the implementation of ML and DL classifiers such as RF, SVM, and Neural Networks, both subsets demonstrated high precision and recall metrics. Future work, performance of the result needs some improvement on both ML and DL classifiers.

Researchers (V.Tomin, G.Kurbatsky, N.Sidorov and V.Zhukov, 2016) suggested a study on a unique semi-automated technique for assessing online security utilizing machine learning techniques such as ANNs, SVM, DT and more. The results of this proposed approach can identify the threat with high accuracy. This ML model could generate an alarm if any threats detected. This research basically provides online security, but the limitation of this model is on other OT sites need to use different datasets.

(Park, Li and Hong, 2018) in his research developed an IDS based model using principles of context awareness and machine learning. The model contributed to Ambient intelligence in a smart factory environment where the information is captured and shared in real-time using sensors. As in a smart factory, all the processes are automated there are chances of process hijacking, system malfunctioning, and leakage of critical production information. The proposed model in this paper solves these problems by effectively detecting an anomaly and process hijacking.

(Nkiruka Eke, Petrovski and Ahriz, 2019) have utilized artificial immune systems, LSTM, and RNN algorithm to create a model that automatically detects APT attacks. This model not only detects the threats and incoming attacks but also identifies illicit data flow and classifies it according to the kind of APT assault. The model was trained by researchers rigorously using UNSW training dataset in order to achieve excellent precision and accuracy rate. And LSTM had the highest accuracy rate 99.99%. The researchers further suggest creating a model which can detect dynamic behaviours at multiple stages of APT attacks.

The Industrial Control system cyber-attack detection model was created by (Mubarak, Hadi Habaebi, Islam and Khan, 2021) using deep learning algorithms like LSTM and RNN. They

evaluated the proposed model by incorporating and simulating various OT attacks scenarios and utilized industrial OT network-traffic dataset with ICS cyber test-kit. Further, they performed deep packet inspection (DPI) of data packet flow using metadata. DPI analysis gave them greater insight onto the details of OT traffic which are dependent on communicating protocols. The researchers state that the model initially consumes time but when reused achieves high accuracy when predicting attacks.

Researchers (TAMY, BELHADAOUI, RABBAH and RIFI, 2020) proposed a study based on the usage of UTM and machine learning approaches. They used SVM, OneR, K-NN, and RF in this model. These methods were tested and studied on a SCADA dataset. The generated simulation results demonstrate that RF is still the best classifier after PSO optimization of methods. Future research will concentrate on developing a hybrid model that uses many classifiers to improve intrusion detection.

2.2 Research Niche

References	Methodology	Algorithms for IDS	Future work
(Zaki Khan and Serpen, 2019)	Machine learning algorithms used to build the IDS for SCADA network	Random Forest, PART Classifier and naïve Bayes	Large dataset is required for better accuracy rate.
(Tamy et al., 2019)	Four Algorithms were used for IDS in machine learning and evaluate them in SCADA network	Naïve Bayes, SVM, J48 and RF	Need to trained other model for good accuracy except RF
(Lopez Perez, Adamsky, Soua and Engel, 2018)	In this, four steps were implemented to develop an ML-based for IDS. Data cleaning, data transformation, hyperparameter search and classification are the four steps.	SVM and RF	More machine learning algorithms will be used for implemented and developed hybrid model
(Syamsul Arifin et al., 2021)	The IDS model was built by data training and applies machine learning methods to identify DoS assaults. The 229 dataset was used for IDS model, and monitored the performance	Decision Tree, Gaussian Naïve Bayes and (SVM).	For detecting packet in protocol, dataset need to be created according to that for SCADA network
(Alhaidari and AL-Dahasi, 2019)	An enhanced framework for identifying DDoS attacks using multiple machine learning method.	J48, NB and RF	They do suggest doing an evaluation on these algorithms using different

			datasets including a dataset that fits with SCADA systems.
(Rajesh and Satyanarayana, 2021)	Detection of malicious traffic, machine learning algorithms were implemented and also evaluated.	SVM, KNN, RF and NB	Deep learning algorithms will be implemented on malicious traffic for better results.
(Marsden, Moustafa, Sitnikova and Creech, 2017)	Probability Risk Identification Based Intrusion Detection System for SCADA Systems	KNN, NB, RF and PRI	PRI models need to ne reimplemented into a real-time environment to increase the performance.
(Almseidin, Alzubi, Kovacs and Alkasassbeh, 2017)	Evaluation of Machine Learning Algorithms for Intrusion Detection System	J48, RF, DT, MLP, naïve Bayes and Bayes Network	False negative and false positive rates are also needed to be taken into consideration.
(Khan et al., 2019)	A Hybrid-Multilevel Anomaly Prediction Approach for Intrusion Detection in SCADA Systems	Hybrid Multilevel	Need to improve the DR values through deep learning
(Duque Anton, Sinha and Dieter Schotten, 2019)	Two data sets were analysed using SVM and RF algorithms for anomaly-based IDS	SVM and RF	NA
(Perales Gomez et al., 2019)	Four steps were used in the methodology, first selection of attacks, in next step attacks were launched, third step, capturing the network traffic and last step was computing the features from the network capture and generate the dataset	ML and DL classifier	Plan to improve the results using advanced model such as LSTM and CNN
(V.Tomin, G.Kurbatsky, N.Sidorov and V.Zhukov, 2016)	An innovative automated multi-model method based on machine learning for assessing internet security in power systems.	ANN, SVM, DT and etc	Want to implement on real-time scenario
(Park, Li and Hong, 2018)	Machine learning approach on smart factory-based ambient intelligence context-aware	ML	NA

	intrusion detection system.		
(Nkiruka Eke, Petrovski and Ahriz, 2019)	Four Major AIS Algorithms used in IDSs	LSTM, RNN	Further work will explore modelling combination of an optimised LSTM-RNN model and CNN on a time-series dataset -
(Mubarak, Hadi Habaebi, Islam and Khan, 2021)	Six steps taken for detecting ICS attack with ensembled ML and DPI using dataset	LR, KNN, NB, RF, ANN, SVM, ST, RNN	NA
TAMY, BELHADAOUI, RABBAH and RIFI, 2020)	Unified Threat Management methodology was proposed in this research	SVM, OneR, K-NN, and RF	Future research will concentrate on developing a hybrid model that uses many classifiers to improve intrusion detection.

Figure 1 Research Niche

To conclude in the related research work, many researchers used machine learning algorithms to detect the threats in operational technology, different methodologies were proposed and future work as well, but the focus was on accuracy rate, precision, recall and f1-score. In this research, different machine algorithms are going to be implemented on publicly available dataset for good accuracy, precision, recall and f1-score.

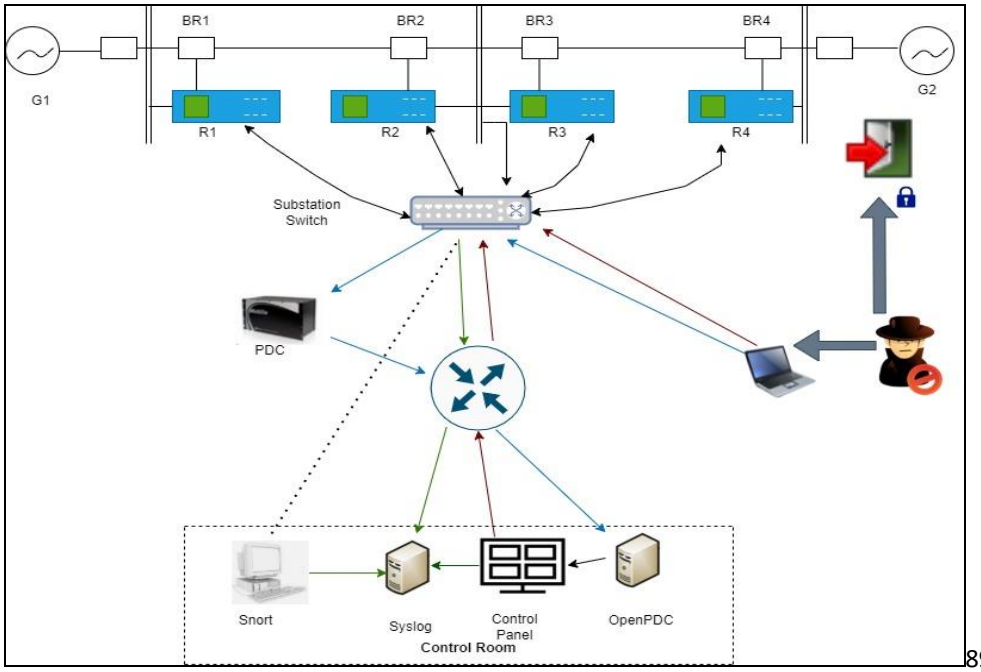
3 Research Methodology

3.1 Overview of Methodology

There are several methodologies used to detect the threats in IDS. Intrusion detection systems can reduce risks in operational technology while also providing additional security and protection in infrastructure. In this study, four algorithms are applied to achieve the best accuracy in IDS, and they are compared to determine the two best algorithms for detecting attacks in ICS, this is the main aim of this research to get the best accuracy rate in IDS. To implement the project first the dataset is required which is taken from a public platform called Mississippi university, the reason to choose this dataset as the research is basically on Industrial Control System. Out of three datasets, the triple- class dataset is taken and after that KDD technique is used for data mining on the dataset. The four algorithms Random Forest, Decision Tree Logistic Regression, and XGBC classifier have been used.

3.2 Justification of Dataset

The Mississippi university has provided an ICS dataset[1] on their website which is a public platform, there are three datasets that really are public platforms for industrial control systems. Binary, triple-class, and multiclass are the three types of datasets, and the triple-class dataset was chosen for our research. For a quick explanation, the below figure illustrates an ICS network diagram with several features or components. Starting with G1 and G2, as these are power generators, BR1 through BR4 are breakers that are divided into two portions. Breakers, often known as circuit breaker tripping, safeguard circuits from power overload. R1 to R4 are the four intelligent Electronic Devices, and each IED regulates one breaker, with R1 controlling BR1, R2 controlling BR2, and so on. When intelligent devices identify a breakdown, a distance protection technique is used to trip the breakers. There are numerous sorts of situations, some of which is natural and others of which is attack. For example, short-circuit fault and line maintenance are natural faults, whereas remote trip command injection, relay configuration modification, and data injection are attacks in the network.



89

Figure 2 Architecture diagram of Dataset

In the chosen dataset, there are three scenarios natural, no event, and attack. In the triple-class dataset 129 features or we can say parameter is given and each phasor measurement units have 29 types of measurements that calculate the electric waves using and, in our dataset, there are 4 PMU and a total of 116 PMU measurement columns. 29 columns in each PMU indicate different parts like R2-PA2: VH means phase A voltage phase angle measured by PMU R2. 12 more columns from control devices such as snort, Syslog, control panel, and OpenPDC.

3.3 KDD Methodology for data mining

In this paper the KDD technique is used for data mining on dataset.

[1] Tommy Morris - Industrial Control System (ICS) Cyber Attack Datasets (google.com)

1. Import a triple-class dataset into the data frame and choose a dependent variable.
2. Irrelevant data such as missing values or unwanted scrap variables are dropped by the dataset.
3. Null value values should be checked and deciding the algorithms.
4. After imputation, the dataset is examined for outliers and accurate values, as well as transformed to the required data type.
5. Identify the outlier and remove some outlier.
6. Use techniques like exploratory factorial analysis to choose variables.
7. The data set is divided into three sections: training, testing, and fitting to specific machine learning algorithms, with the results evaluated using specialized evaluation techniques

3.4 Justifications and support

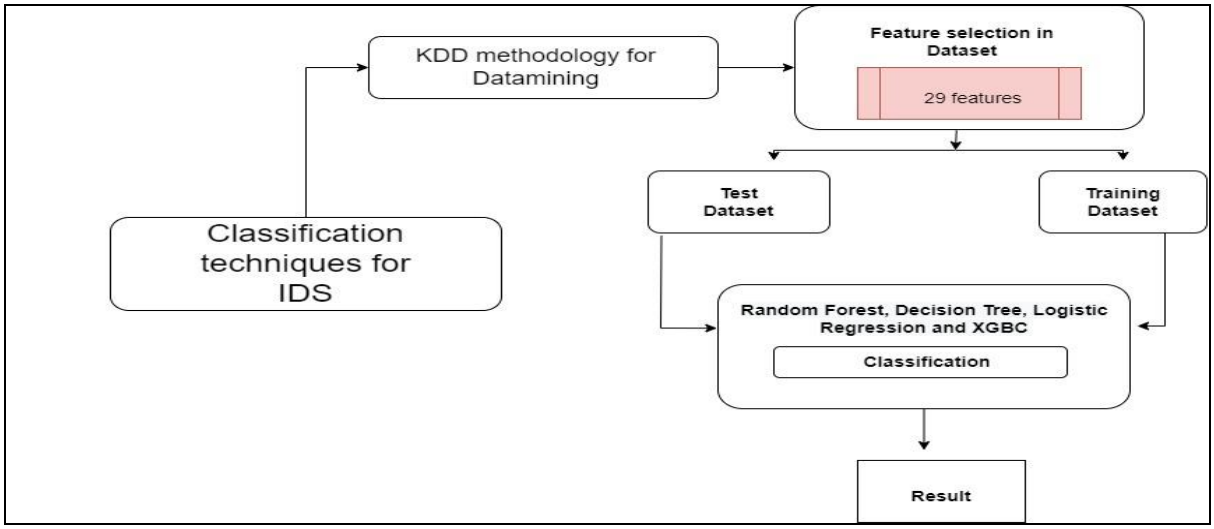


Figure 3 Flow Chart

The purpose of using this model is to meet the study's goal of improving detection accuracy. In ICS, after selecting a dataset, the KDD methodology is used for data mining, which is explained in figure 2 in seven steps, where 29 parameters are chosen for the experiment or implementation, where test and training datasets are used for implementation in classification techniques such as Random Forest, Decision Tree, Logistic Regression, and XGB. All four algorithms will have different outcomes and results where we can evaluate the best two algorithms.
Random Forest is a supervised machine learning technique that is most used in regression and classification problems. Its most essential characteristic is that it can certainly manage datasets with the variable in regression and categorical variables in classification. For classification difficulties, it produces superior results[2].

Implementation of Decision Tree mostly is used for resolving the errors in regression and classification because it is a member of the supervised learning algorithm family, the word "decision tree" refers to the flow of tree structure for predictions that results from such a

2 https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

sequence of feature-based splits. Decision Trees, like trees, begin with a root node and terminate with decisions made by leaves[3] .

Logistic Regression is another Machine Learning method that is used to classify data. The logistic function employs a set of parameters known as the sigmoid function. Predictions are converted to probabilities using this method (Saranya et al., 2020).

The XGBoost classifier is another Machine Learning technique that is used for structured and tabular data. XGBoost is a gradient-boosted decision tree implementation optimized for speed and performance. The XGBoost algorithm is an extreme gradient boost technique. That suggests it's a large Machine learning algorithm with a lot of moving elements. XGBoost is capable of handling huge, complex datasets. XGBoost is a strategy for ensemble modeling[4].

4 Design Specification

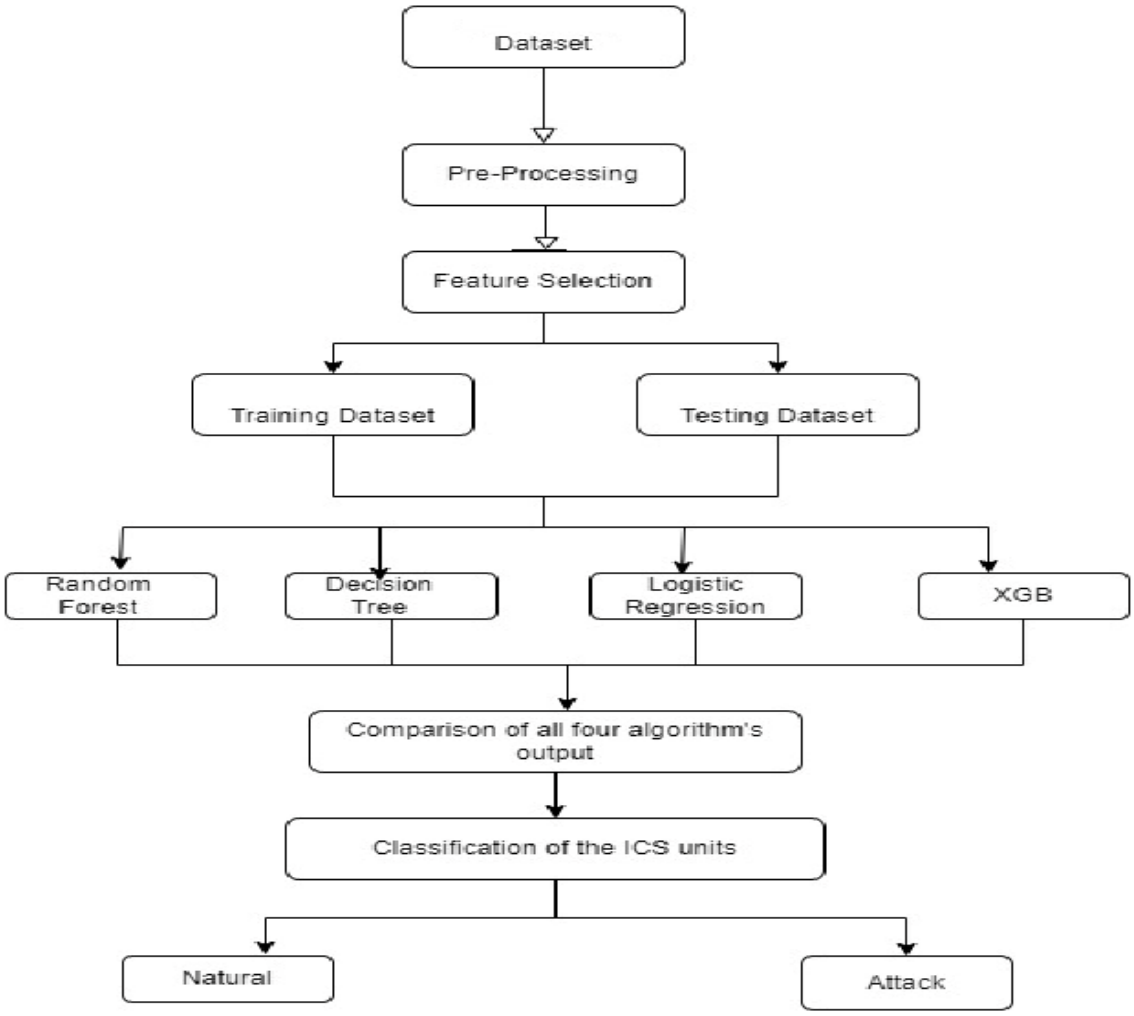


Figure 4 Framework

[3] https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/
[4] https://www.linkedin.com/pulse/xgboost-classifier-algorithm-machine-learning-kavya-kumar/

The main aim of this research is to design an experiment that will help to achieve the best accuracy rate to detect the intrusion in the ICS network using the machine learning classification technique. Above figure 5 is the framework that would assist in achieving the goal of the project. First, the triple-class Dataset is uploaded in google drive and Google colaboratory is chosen for the implementation, the dataset is loaded in the pre-processing section, and then feature selection is completed by doing some experiments where one of the Intelligent Electronic Devices is selected. Each IED has 29 parameters. In the next step, test and train data is split into 20–80 ratios respectively. Once the accuracy, precision, recall, and f1-score rate come out in the result, then compare all four algorithms accuracy rate on the dependent variables, dependent variables are natural and attack.

5 Implementation

This section illustrates the implementation of an intrusion detection system using classification techniques carried out in this research. Explanation of environment setup, dataset, target variables and model implementation are described below:

5.1 Environment setup:

Our implementation and model development are performed in Python. Google Colab is used for the development and execution of the model, which is a Google Research product. The reason for choosing Colab for this research is it runs on a Google server and provides us access to free TPU and GPU for faster processing.

5.2 Libraries/ Packages:

Below Libraries/Packages were installed in implementation

- **Matplotlib:** This library, also known as a visualization library, is utilized in our study to graphically represent a huge number of data but also the output results. Plotting graphs are possible.
- **NumPy:** This library is used in this research to handle and support the large and multi-dimensional arrays and matrices.
- **Pandas:** This library is based on the Python programming language and is extremely powerful and it is utilized for data manipulation and analysis.
- **Seaborn:** It is a data visualization library used to make statistical graphs.

5.3 Dataset

The dataset is taken from the Mississippi university platform, a dataset called Triple-class dataset in which total numbers of features are 129, 29 features were taken from the implementation and execution. The reason for the selection of these parameters is that all 29 selected features emerge from one intelligent device which is connected to the power generator, breakers, substation switch and indirectly provides the logs to snort and Syslog so, in short voltage of every phase, negative and positive voltage, a frequency for relay and some more measurement are given in each R1, R2, R3 and R4 (Intelligent Electronic Devices). Using this dataset for the research provides the best signature-based accuracy.

5.3.1 Data Overview:

This dataset has 239 columns and 83792 rows

5.3.2 Target Variable:

From this dataset 2 variables are taken

Dependent variable 1: Attack

Dependent variable 2: Natural

Against these two variables the experiment can be done to get the highest detection.

5.4 Model Implementation:

5.4.1 Data Pre-processing:

In this step, the imported or fetched data in Google Colab were identify the null, nan values and any missing values were cleaned for better output of this research. Using a heatmap, all the null values in the dataset can be shown. Imputation is finished after the process of finding missing or null values is done. The graph below shows that there is no null value.

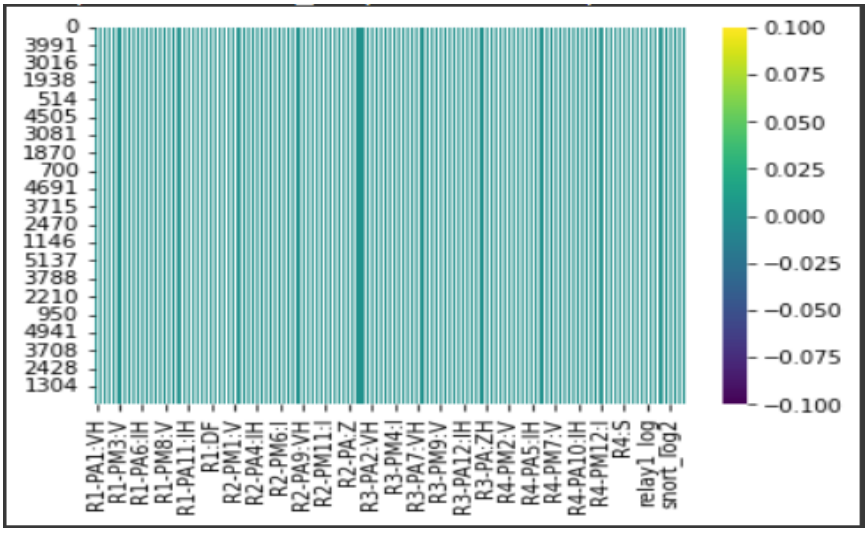


Figure 6 heatmap of missing values in dataset

The two dependent variables attacks and natural were encoded with the character to string values, where 0 string value assigned to natural, and 1 string value assigned to attack. Then checking the class imbalance, which means checking the un-even dataset where the attack rate has more in numbers in the dataset than natural, would have impacted the model not only in accuracy but also in the efficiency. So, to resolve this issue, the Imblearn method was utilized, with the SMOTE module generating an equal ratio of classes where efficiency and accuracy will not be impacted.

5.4.2 Feature Selection:

In feature selection, it is an extremely important part that needs to be completed in order to achieve efficiency and improve the score of the input features to improve the prediction of the dependent variable. This is made to reduce complexity and remove any irrelevant features that were not needed for the detection. Features are selecting from one of the R1 to R4 devices which is intelligent electronic devices in the ICS network. After experimenting with

all of the IEDs parameters, the following features were chosen for the research, the total number of features is 29 against dependent variables.

Features
R2-PA1:VH,R2-PM1:V,R2-PA2:VH,R2-PM2:V,R2-PA3:VH,R2-PM3:V,R2-PA4:I,R2-PM4:I,R2-PA5:IH,R2-PM5:I,R2-PA6:IH,R2-PM6:I,R2-PA7:VH,R2-PM7:V,R2-PA8:VH,R2-PM8:V,R2-PA9:VH,R2-PM9:V,R2-PA10:IH,R2-PM10:I,R2-PA11:IH,R2-PM11:I,R2-PA12:IH,R2-PM12:I,R2:F,R2:DF,R2-PA:ZH,R2:S,marker

All the parameters available in the dataset is either assigned to string value ‘0’ or ‘1’ which is natural and attack.

5.4.3 Data Splitting:

The triple-class dataset was divided into two parts for classification techniques where 80 percent is given to train and 20 percent is allocated to test data. This split data is used in all the algorithms for getting the best results.

5.4.4 Classification Model Training:

This model is using four separate machine learning algorithms and the performance of two variables (natural and attack) are compared. In this research paper, four machine learning algorithms are Logistic Regression, Decision Tree, Random Forest, and XGB. The data has been divided into two parts train and test, all algorithms using the same data for implementation in this model, where 29 features were selected to get the best output. After getting the result of each model, the output of each algorithm would be compared in accuracy, precision, recall, and f1-score.

6 Evaluation

In this section, the efficiency and outcomes of all models are evaluated and find the best accuracy of the proposed model, the classification techniques have been used which will detect the accuracy rate of attacks using the dataset. In machine learning, the confusion matrix depicts the performance of the model. For the evaluation of these classification models, various measures such as accuracy, precision, recall, and F1 score are considered. The dataset is divided into two parts 20-80 ratios given for testing and training purposes and the same data split method was applied in each model like Logistic Regression, Decision Tree, Random Forest, and XGB.

6.1 Model 1: Logistic Regression

Where 29 parameters have selected for performing this experiment and get the best accuracy on detecting the attack and natural dependent variables.

Accuracy result: The dataset was utilized for training and testing the algorithm. The dataset has 83792 rows and 129 columns, with just 29 parameters selected against dependent variables and below figure shown the result of accuracy 59 % using Logistic Regression.
Precision- 58%
Recall- 60%
F1-score- 59%

Classification report :

	precision	recall	f1-score	support
0	0.59	0.58	0.59	8392
1	0.58	0.60	0.59	8290
accuracy			0.59	16682
macro avg	0.59	0.59	0.59	16682
weighted avg	0.59	0.59	0.59	16682

Figure 7 output of logistic regression

Confusion Matrix: In general, it illustrates four possible combinations of predicted and actual values.

It is also excellent for analyzing variables like recall, precision, specificity, accuracy, and, mo st crucially, AUC-ROC curves.

- TP: indicates the intrusion as attacks correctly.
- TN: indicated the natural activity as non-attack correctly.
- FP: shows the natural activity as attack incorrectly.
- FN: shows the intrusion as natural activity incorrectly

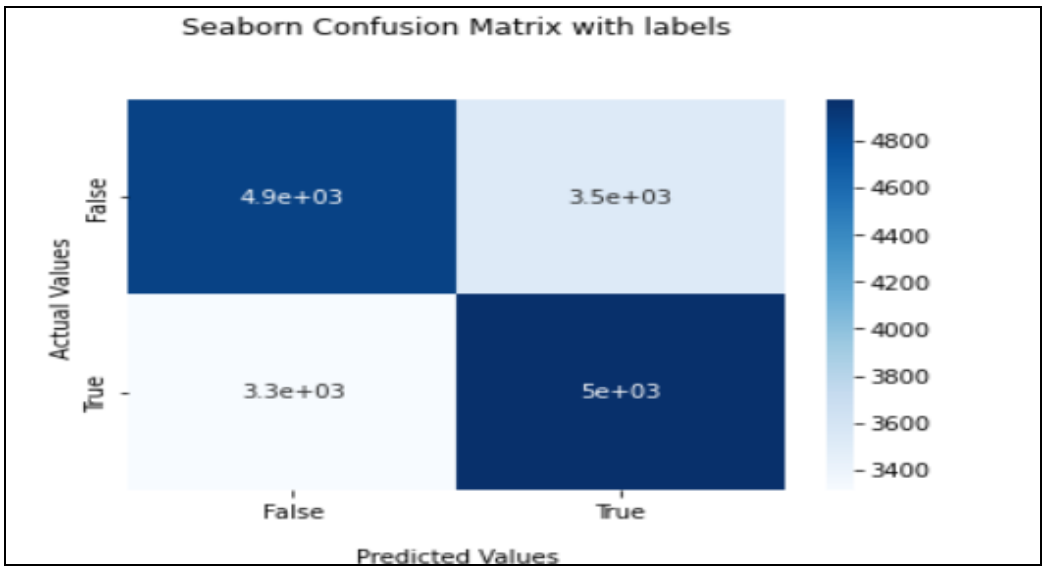


Figure 8 Confusion matrix of logistic regression

6.2 Model 2: Random Forest model

Another model where the same parameters have selected to get the best possible result, dependent variables are also same.

Accuracy result: Accuracy rate in this model is 90 %, same parameter and dependable parameter have been used in this algorithm. The percentage of attack is below mentioned.
Precision is 89%
Recall: 92%
F1 score: 90%

Classification report :

	precision	recall	f1-score	support
0	0.92	0.89	0.90	8392
1	0.89	0.92	0.90	8290
accuracy			0.90	16682
macro avg	0.90	0.90	0.90	16682
weighted avg	0.90	0.90	0.90	16682

Figure 9 output of random forest

Confusion Matrix: Below confusion matrix is after implementing on Random Forest model

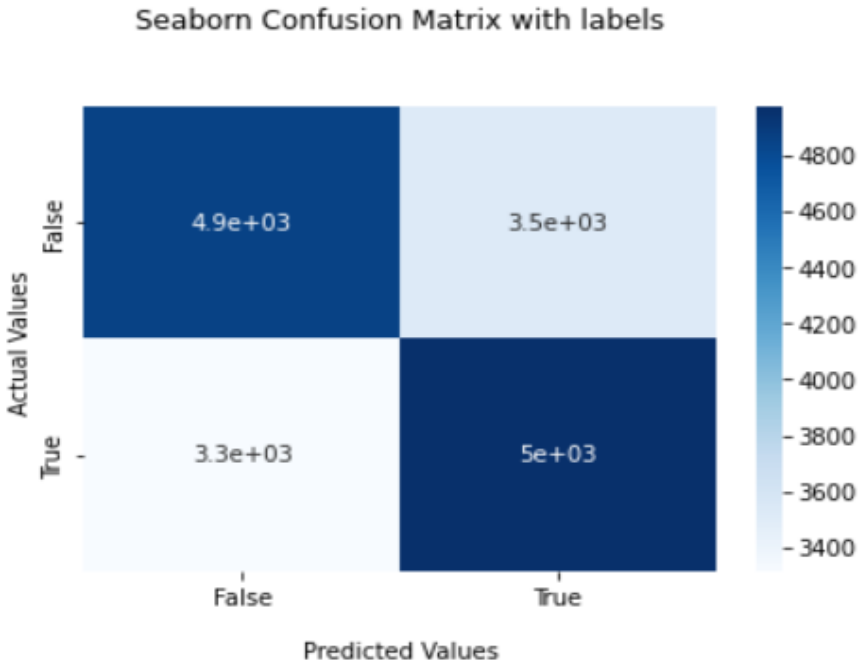


Figure 10 Confusion matrix of random forest

6.3 Model 3: Decision Tree

Third model is used in this classification techniques.

Accuracy rate: Accuracy rate in this model is 85 %, same parameter and dependable parameter have been used in this algorithm. The percentage of attack is below mentioned.

Precision is 86%

Recall: 85%

F1 score: 85%

Classification report :

	precision	recall	f1-score	support
0	0.85	0.86	0.86	8392
1	0.86	0.85	0.85	8290
accuracy			0.85	16682
macro avg	0.85	0.85	0.85	16682
weighted avg	0.85	0.85	0.85	16682

Figure 11 output of decision tree

Confusion Matrix: Below confusion matrix is the output is from Decision Tree Model

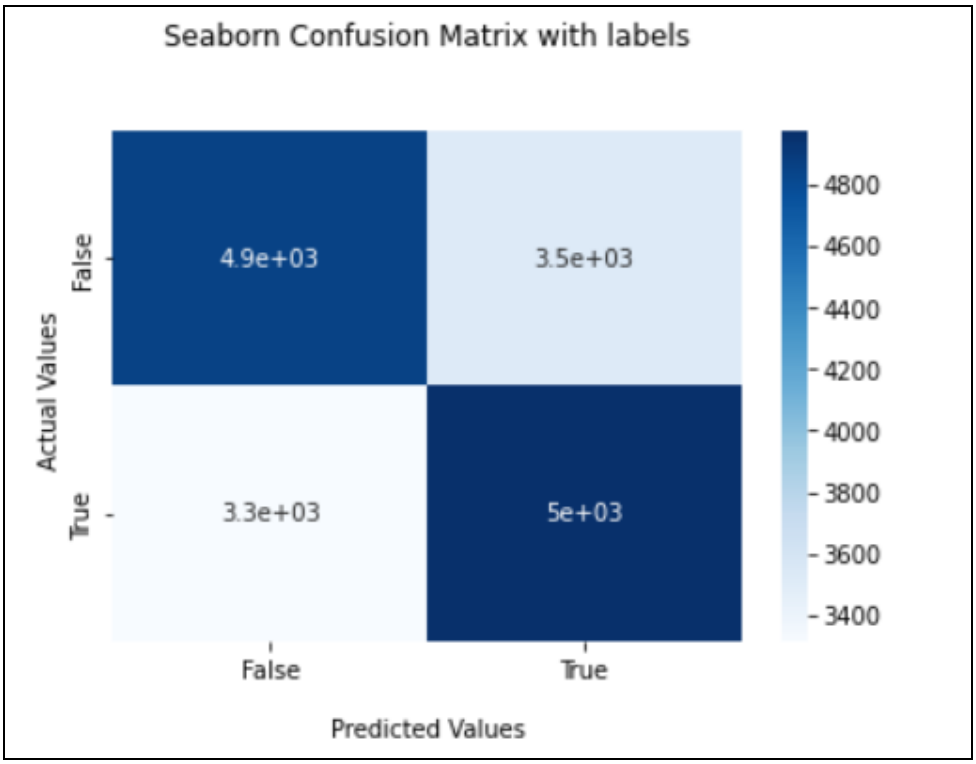


Figure 12 Confusion matrix of decision tree

6.4 Model 4: XGB classifier

The fourth and last model utilized to try to get the output is the XGB classifier model.
Accuracy: Accuracy rate in this model is 58 %, same parameter and dependable parameter have been used in this algorithm. The percentage of attack is below mentioned.
Precision is 55%
Recall: 76%
F1 score: 64%

Classification report :

	precision	recall	f1-score	support
0	0.62	0.40	0.48	8392
1	0.55	0.76	0.64	8290
accuracy			0.58	16682
macro avg	0.59	0.58	0.56	16682
weighted avg	0.59	0.58	0.56	16682

Figure 13 output of XGB classifier

Confusion Matrix: Below confusion matrix is from XGB Classifier model.

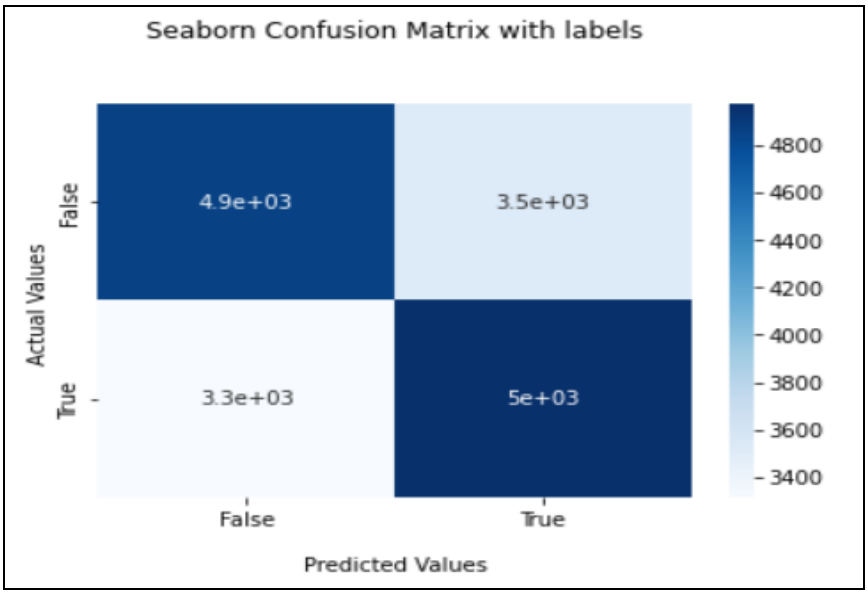


Figure 14 Confusion matrix of XGB classifier

6.5 Experiment 1/ Accuracy Comparison

In this experiment, the accuracy rate is compared, and find the best algorithm to detect the highest accuracy rate. Measuring accuracy is an effective approach for any Machine learning model. The graph illustrates the accuracy rate comparison on the same dataset with the same feature selection, with each algorithm having a different accuracy rate. Where the Random

Forest method outperforms the other three algorithms in terms of accuracy. Another best algorithm for a good accuracy rate is Decision Tree which is showing an 84 percent of accuracy rate just 6% lesser than Random Forest. Logistic Regression and XGB classifier algorithms have 59 and 58 percent of accuracy, respectively. So, in our first experiment, the Random Forest and Decision Tree outperform the other two algorithms in terms of accuracy rate.

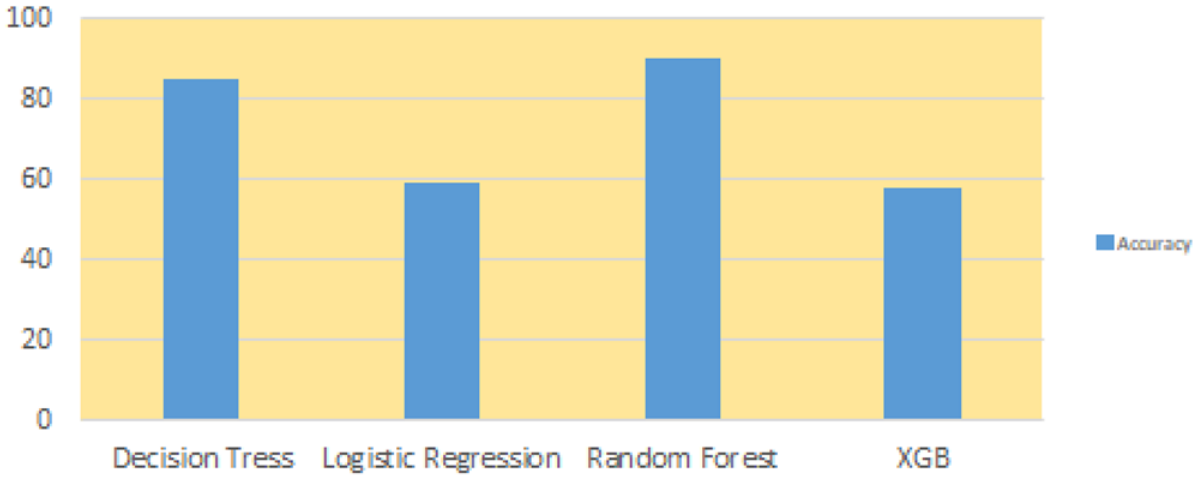


Figure 15 Comparison of Accuracy

6.6 Experiment 2 / Precision, Recall and F1-Score Comparison

Precision, recall, and f1-score are compared in the second experiment. Through PRF (Precision, Recall, and F1-score) metrics provide help with a better understanding of false positive and false negative rates. Similarly, in experiment 1, Random Forest has a good PRF rate, after that Decision Tree and the other two algorithms are comparatively average PRF rates

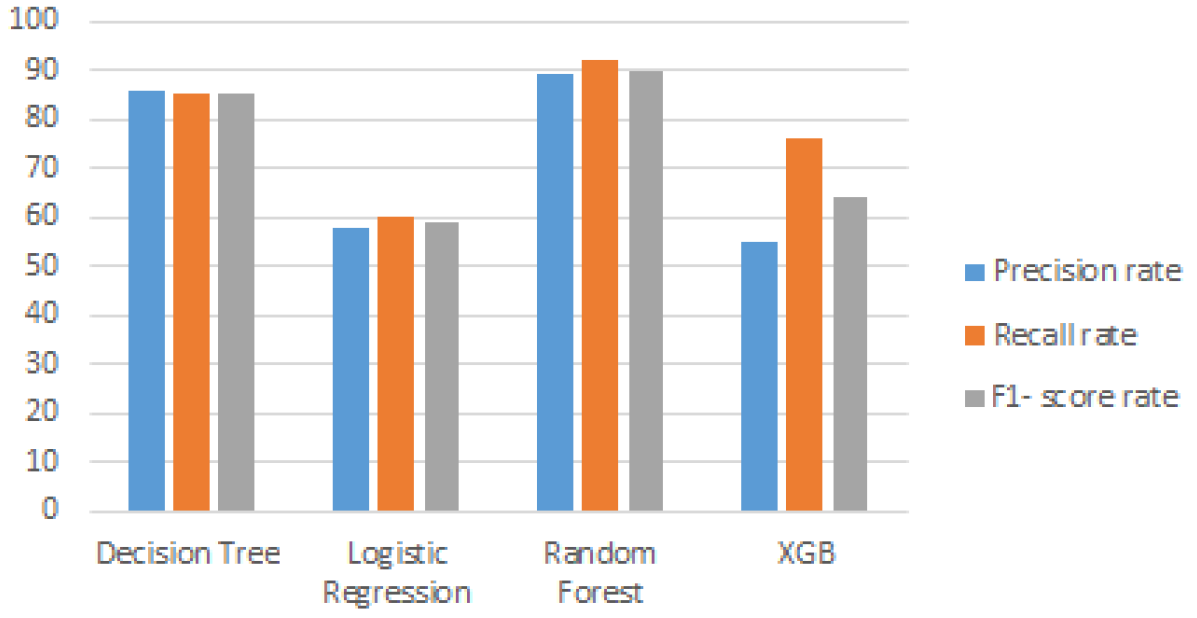


Figure 16 Comparison Precisions, Recalls and FI Scores

6.7 Discussion

In our analysis performing couple of experiments on different machine learning models. After analysing the findings and comparing each model in the below table, the random forest model has the highest accuracy and PRF rate, second Decision Tree model, which has the second-highest accuracy and PRF rate. Where Logistic Regression and XGB classifier have only average accuracy and PRF rate. The approach behind using classification techniques is to find a suitable model to detect the signature-based attack in Industrial control systems. Models were trained and tested to get the best possible output, to achieve that the dependent variable called natural considerably lower than another dependent variable called attack. Using SMOTE the dataset formed in balance. Implementing this model in the real world would provide a good impact while monitoring real-time intrusion detection. Dataset is quite crucial to get the highest accuracy rate, for improvement different simple datasets in SCADA network will provide the higher detection rate in the same model.

Model	Accuracy	Precision	Recall	F1- score
Logistic Regression	59%	58%	60%	59%
Random Forest	90%	89%	92%	90%
Decision Tree	85%	86%	85%	85%
XGB	58%	55%	76%	64%

Figure 17 Comparison of outputs

7 Conclusion and Future Work

The main objective of this research paper is to build an Intrusion Detection System using Machine Learning algorithms in Operational Technology and verify which algorithm has the best detection. Before being used in various models, dataset was first imported, analysed, and transformed. Also, this research got the result which can secure the ICS. When compared to XGB, Logistic Regression, and Decision Tree, the Random Forest algorithm provides the highest detection rate of attack using the Triple-Class dataset, with an accuracy of 90 percent. Also, the Decision Tree cannot be neglected as the accuracy shown in the implementation is 85 percent, where Logistic Regression is 59 percent of accuracy and XGB Classifier with the lowest accuracy of 58 percent. All models have different PRL rates as well and quite similar output is there.

Future work will include the implementation of classification techniques in real-time SCADA networks, as well as the development of an anomaly-based detection system for no-scope of vulnerabilities in ICS while detecting the attacks.

References

Benisha, R. and Raja Ratna, S., 2019. Design of Intrusion Detection and Prevention in SCADA System for the Detection of Bias Injection Attacks. *Security and Communication Networks*, 2019, pp.1-12.

Zaki Khan, A. and Serpen, G., 2019. Misuse Intrusion Detection Using Machine Learning for Gas Pipeline SCADA Networks. In: *Int'l Conf. Security and Management*. Ohio: IEEE, pp.1-8.

Tamy, S., Belhadaoui, H., Almostafa Rabbah, M., Rabbah, N. and Rifi, M., 2019. AN EVALUATION OF MACHINE LEARNING ALGORITHMS TO DETECT ATTACKS IN SCADA NETWORK. In: *2019 7th Mediterranean Congress of Telecommunications (CMT)*. Fez: IEEE, pp.1-5.

Lopez Perez, R., Adamsky, F., Soua, R. and Engel, T., 2018. Machine Learning for Reliable Network Attack Detection in SCADA Systems. In: *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. New York: IEEE, pp.1-6.

Syamsul Arifin, M., Stiawan, D., Rejito, J., Idris, M. and Budiarto, R., 2021. Denial of Service Attacks Detection on SCADA Network IEC 60870-5-104 using Machine Learning. In: *2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. Semarang: IEEE, pp.1-5.

Alhaidari, F. and AL-Dahasi, E., 2019. New Approach to Determine DDoS Attack Patterns on SCADA System Using Machine Learning. In: *2019 International Conference on Computer and Information Sciences (ICCIS)*. Sakaka: IEEE, pp.1-6.

Rajesh, L. and Satyanarayana, P., 2021. Evaluation of Machine Learning Algorithms for Detection of Malicious Traffic in SCADA Network. *Journal of Electrical Engineering & Technology*.

Marsden, T., Moustafa, N., Sitnikova, E. and Creech, G., 2017. Probability Risk Identification Based Intrusion Detection System for SCADA Systems. In: *International Conference on Mobile Networks and Management*. Canberra: IEEE, pp.1-11.

Almseidin, M., Alzubi, M., Kovacs, S. and Alkasassbeh, M., 2017. Evaluation of machine learning algorithms for intrusion detection system. In: *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*. Subotica: IEEE, pp.1-6.

Khan, I., Pi, D., Khan, Z., Hussain, Y. and Nawaz, A., 2019. HML-IDS: A Hybrid-Multilevel Anomaly Prediction Approach for Intrusion Detection in SCADA Systems. *IEEE Access*, 7, pp.89507-89521.

Duque Anton, S., Sinha, S. and Dieter Schotten, H., 2019. Anomaly-based Intrusion Detection in Industrial Data with SVM and Random Forests. In: *2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. Split: IEEE, pp.1-6.

Perales Gomez, A., Fernandez Maimo, L., Huertas Celdran, A., Garcia Clemente, F., Cadenas Sarmiento, C., Del Canto Masa, C. and Mendez Nistal, R., 2019. On the Generation of Anomaly Detection Datasets in Industrial Control Systems. *IEEE Access*, 7, pp.177460-177473.

V.Tomin, N., G.Kurbatsky, V., N.Sidorov, D. and V.Zhukov, A., 2016. MachineLearningTechniquesforPowerSystemSecurityAssessment. In: *IFAC-PapersOnLine 49-27 (2016) 445–450*. Irkutsk: IEEE, pp.1-6.

Park, S., Li, G. and Hong, J., 2018. A study on smart factory-based ambient intelligence context-aware intrusion detection system using machine learning. *Journal of Ambient Intelligence and Humanized Computing*, 11(4), pp.1405-1412.
Nkiruka Eke, H., Petrovski, A. and Ahriz, H., 2019. The Use of Machine Learning Algorithms for Detecting Advanced Persistent Threats. In: *SIN '19: Proceedings of the 12th International Conference on Security of Information and Networks*. Aberdeen: IEEE, pp.1-8.

Mubarak, S., Hadi Habaebi, M., Islam, M. and Khan, S., 2021. ICS Cyber Attack Detection with Ensemble Machine Learning and DPI using Cyber-kit Datasets. In: *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*. Kuala Lumpu: IEEE, pp.1-8.

TAMY, S., BELHADAOUI, H., RABBAH, N. and RIFI, M., 2020. CYBER SECURITY BASED MACHINE LEARNING ALGORITHMS APPLIED TO INDUSTRY 4.0 APPLICATION CASE: DEVELOPMENT OF NETWORK INTRUSION DETECTION SYSTEM USING HYBRID METHOD. *Journal of Theoretical and Applied Information Technology*, pp.1-14.

Saranya, T., Sridevi, S., Deisy, C., Duc Chung, T. and Khan, M., 2020. Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review. In: *Third International Conference on Computing and Network Communications (CoCoNet'19)*. KualaLumpur: IEEE, pp.1-10.