Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 8, Issue. 3, March 2019, pg.193 – 203

CYBER ATTACK DETECTION IN REMOTE TERMINAL UNIT OF SCADA SYSTEMS

Ali Hasan Dakheel[1]; Osman Nuri Ucan[2,*]; Oguz Bayat[3]; Hamzah Hameed Jasim[4]

ali.dakeel19715@gmail.com, Osman.ucan@altinbas.edu.tr

ABSTRACT: Supervisory Control and Data Acquisition (SCADA) systems are widely used in critical infrastructures such as water distribution networks, electricity generation and distribution plants, oil refineries, nuclear plants, and public transportation systems. Every communication is done through encrypted messages to protect the pipeline from any intrusion from outside, it is almost impossible to interpret the observed payload. SCADA systems typically do require a high throughput but are much more tolerant of delays and outside intrusion. In addition, many SCADA systems may have much greater resource constraints than would be found in traditional IT systems to protect the industrial systems from any kind of intrusion or Cyber Attack. This lack of computing resources along with performance constraints can make it difficult or impossible to apply standard security technologies. The results from this proposed system were validated with a realistic text files with malicious data provided by the network operators online. Using SCADA systems, unauthorized access to network and switches could be more tightly controlled while keeping a human in the loop; that is, human supervision and interaction were, and still are, part of SCADA systems. However, technological advances and the maturation of SCADA systems has pushed more of the supervisory function onto the computer systems that make up modern SCADA systems. In the early development of SCADA systems attention was given to physical security, but virtually no attention was given to electronic or cyber security. The systems were obscure and the skills and technology needed to interact with the systems and update network security were simply not readily available; security of this type is often referred to as security through obscurity. This pattern has continued and today, most dedicated SCADA applications have not included built-in security.

Keywords: Platform, SCADA, Modbus, Cyber Attack Detection, Critical Infrastructure Machine learning

INTRODUCTION

Supervisory Control and Data Acquisition (SCADA) systems are widely distributed systems used to continuously supervise and monitor critical infrastructures such as water distribution networks, electricity generation and distribution plants, oil refineries, nuclear plants, and public transportation systems [1]. By automating data collection from field devices in remote sites, SCADA systems provide operators with a real-time view of the whole process. SCADA systems and the processes they manage are critical to any nation’s defense and economy.

For the purpose of reducing costs and increasing efficiency, the SCADA technology migrated from isolated (monolithic) systems to networked architectures that communicate with corporate network and the internet [3]. This enabled the replacement of proprietary protocols by open SCADA protocols; it also allowed interoperability between different Remote Terminal Unit (RTU) vendors. However, this improvement has been implemented on top of most existing SCADA systems without considering its impact towards cyber security [2]. Security breaches against SCADA systems can disrupt and damage the operation of critical infrastructures, contaminate the ecological environment, cause huge economic losses and even more dangerously human lives loss [4]. Upon this filtering, the flows will be aggregated in order to give the program a wider and more granular view of what is happening, as it will be explained further ahead. Having the flows preprocessed and aggregated, the two machine learning techniques that were previously mentioned will come into play. For the unsupervised learning, a clustering algorithm will be used, which will group together flows that share similar patterns. Having the clusters been generated, there will be a need for an expert intervention, in order to label the clusters. This is necessary for the next step, which applies supervised learning, in which Support Vector Machines (SVM) will be used. The SVM will then receive the labeled flows, and will be able to indicate whether the input flows are perceived by the system as being malicious, or benign behavior, therefore highlighting the intrusions present in system. Similarly, This approach will allow the detection of the source of the attacks, rather than to identify which kind of attack it is, i.e. it detects the devices (be them desktops, laptops, mobile phones or compromised servers) that are behaving maliciously. Also, this system focusesff on detecting volumetric attacks, i.e. it allows to unveil relevant patterns in large scale and intensity network attacks, therefore not allowing to discover attacks such as Buffer Overflows, SQL Injections, Phishing, Cross Site Scripting, and so on [5].

BACKGROUND

In June 2010, Stuxnet worm, the new cyber weapon of cyber-warfare, had struck the Iranian nuclear facility. The Stuxnet worm is the first malware to specifically target industrial programmable logic control systems. The worm uses Siemens' default passwords and gradually modifies the code running in Programmable Logic Controllers to behave them different from their primary purpose. More specifically, Stuxnet alters the frequency of electric current; as result, the drives switch between high and low speeds for which they were not designed [5, 7–9].

After all those incidents, SCADA security has become a concern for researchers, industries and governments. In the past few years, a lot of effort has been made to improve the security of SCADA system. Access control, authentication and intrusion detection are some of the most commonly used security mechanisms to secure critical infrastructures. Since most of the critical infrastructures with SCADA system operate 24/7 and require huge investment, it is not practical to make research on actual SCADA system. Furthermore, for the reason of confidentiality, it is very challenging to collect network traffic signatures on real CADA system for evaluating various SCADA specific security solutions. To address these problems, it would be feasible to develop SCADA platform for critical infrastructures security research. In the subsequent paragraphs of this section, we will summarize some of the existing SCADA platforms.

Some countries established government sponsored SCADA platforms at national level. To facilitate security research and development in the energy sector the National SCADA Test Bed (NSTB) has been established since 2003 by the U.S. Department of Energy [10]. Similarly, in Japan Control System Security Centre Tohoku Tagajo Headquarter Platform has been established since 2012 [11].

For academic and research purpose Morris et al. [12] proposed a small scale real SCADA platform for various industrial applications. On the other hand, Davis et al. [13] proposed simulation based SCADA platform for power grid. They used network client, PowerWorld Server and RINSE network emulator to represent the control station, power station and communication networks respectively. Chabukswar et al. [14] Used C2WindTunnel to simulate the controllers and DDoS attacks on a well-known chemical processing system called Tennessee Eastman Control Challenge.

PROBLEM STATEMENT

The gas pipeline industry face many attacks from the intruders that could dismantle the stakes of many industries for their personal benefit on different architectural platforms besides these platforms, several virtual or simulation based SCADA platforms have also been proposed in recent years [6]. However, most of these platforms are not freely available. Even those platforms, which are developed on open software, are sector and application specific. Moreover, these platforms lack re-configurability for different attack scenarios. In this paper, to address these limitations, an open SCADA platform is proposed, which can be easily extended to various critical infrastructure sectors. More importantly, different from previously developed platforms, the proposed SCADA platform is user friendly and easily reconfigurable for different types of Cyber Attacks. Modbus TCP/IP protocol is used for communication between the SCADA components. The platform can be used for cyber security assessment and vulnerability investigation on SCADA systems. With our platform, one can also generate benchmark dataset to develop and evaluate attack detection and protection technologies for SCADA systems. With the current version of the platform, we have already collected network and control data for SCADA security research and development. Though the platform is primarily developed for SCADA security research, it can also be used for educational purpose.

METHODOLGY

The techniques of machine learning were used to predict the potential attack in the industry of gas pipeline. Initially SCADA protocols did not include security features, which resulted in vulnerabilities to message modification, spoofing, and sniffing attacks. The protocol vulnerabilities are really outside the RTU security perimeter, at least in their specification, but it is important to mention them, and keep them in mind in considering lower layers. One reason for keeping these in mind is the principle of easiest penetration [14]. If an RTU supports an insecure SCADA protocol that can easily be attacked and used to control or damage the connected physical processes, it will be impossible for a lower level prevention mechanism to protect the RTU since it has no way of differentiating between authentic and un-authentic SCADA communications. This approach parallels the definition of a physical security perimeter that is a standard approach in securing physical places. SCADA systems are large distributed systems, and in developing a layered approach to security for them it is important to identify security boundaries for different components.

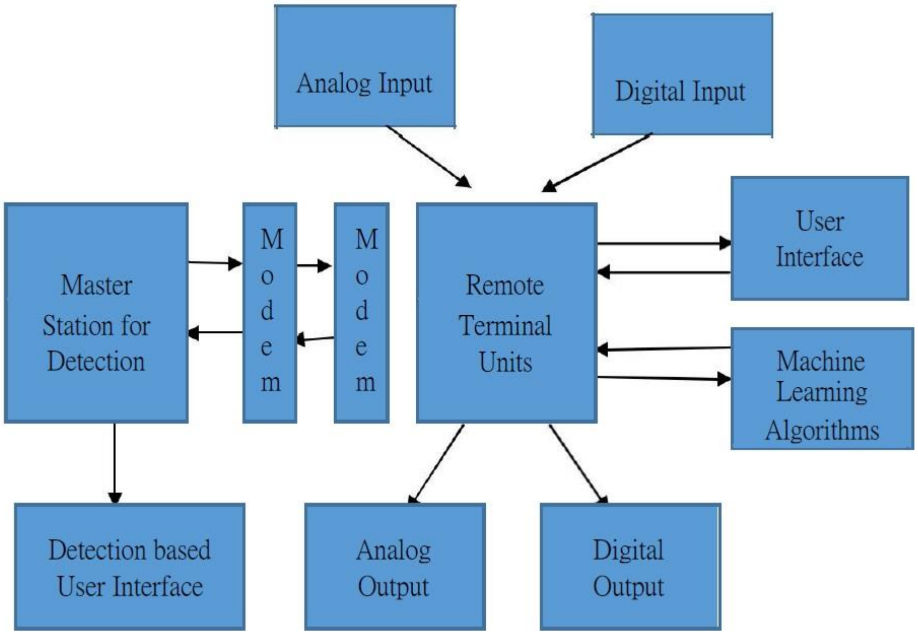

Figure 1: *Typical Supervisory System using Machine Learning.*

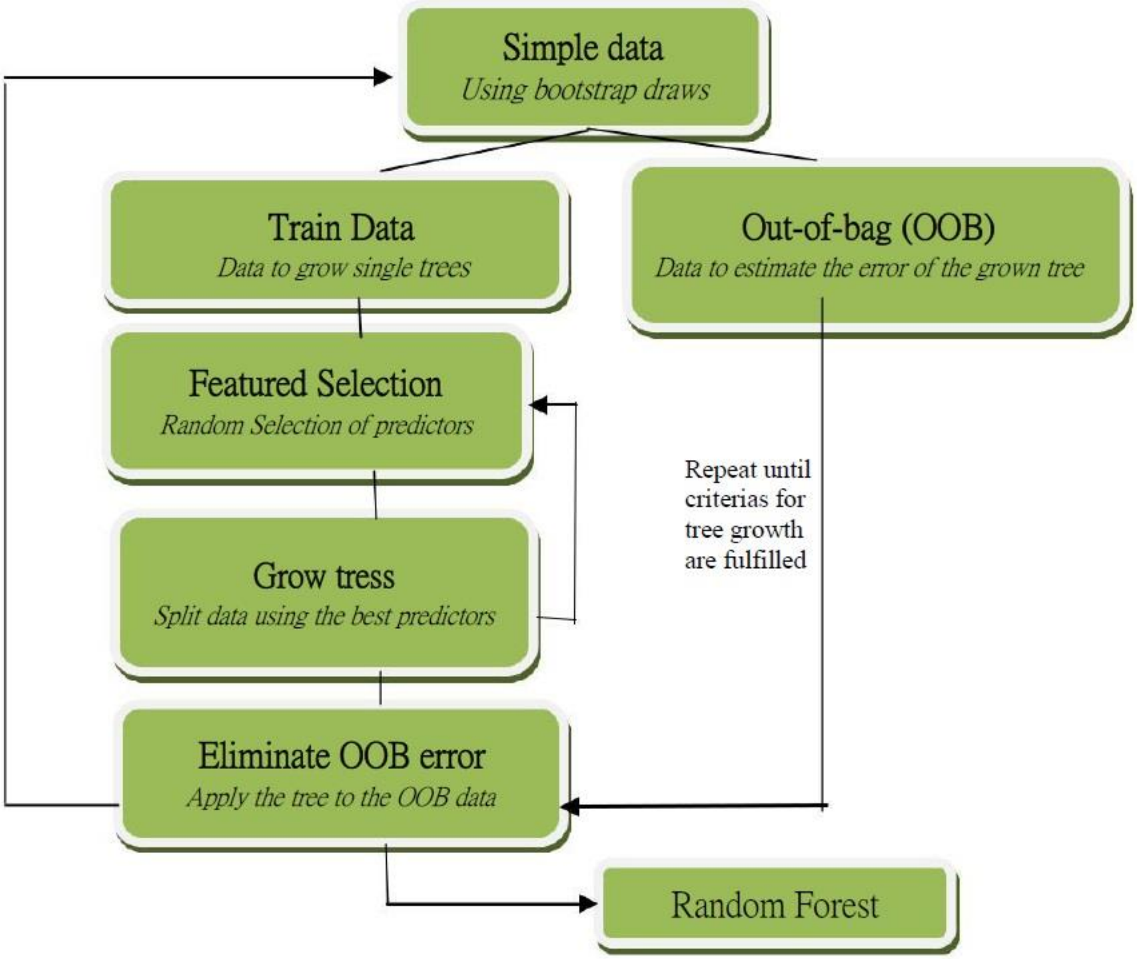
Simple data
Using bootstrap draws
Train Data
Data to grow single trees
Out-of-bag (OOB)
Data to estimate the error of the grown tree
Featured Selection
Random Selection of predictors
Repeat until criterias for tree growth are fulfilled
Grow tress
Split data using the best predictors
Eliminate OOB error
Apply the tree to the OOB data
Random Forest

Figure 2: *Random Forest Algorithm used for classification of Cyber Attack.*

The first option provides the user with a brief description of the data being analyzed, featuring the number of entries of the dataset, and, for each feature, it also presents its minimum and maximum value, and also the mean value and standard deviation. The Precision and accuracy is also measured and attained using random forest approach.

This way, we get a simple overview of the data, statistically. The fourth option presents, for each cluster and features, its mean value and standard deviation. As for the fifth option, window is opened in the browser, that allows to choose which cluster to analyze, and provides an interactive visualization of the dataset it contains. The control station includes Master Terminal Unit (MTU) and Human Machine Interface (HMI). The MTU is the heart of the SCADA system that supervises and controls the activities of various RTUs and sends the control commands to the physical process [15]. The MTU uses polling technique to collect data from RTU. Based on the collected data the HMI provides a visual representation of the remote process for operators. The SCADA system through its HMI also provides flexibility for operators to change plant configurations. These clustering algorithms converge when the variation of the distance between the data points and the clusters centers start converging. With this in mind, the Encryption Method starts by computing the error function that is used as a stopping criterion in the algorithm, known as: Within Sum of Squares (WSS), which is mathematically defined as such:

$$WSS = \mathrm{P}\Sigma_{\mathrm{i=0}}^{k}\ x\in^{p} ci\ dist(x, ci)^{2}$$

The measurement data collected by sensors will be forwarded to the control layer through the communication network. On the other hand, the commands coming from the control station will be sent to actuators through the communication networks. SCADA systems use various communication media including telephone, fiber optic, radio, and satellites to connect the remote terminal units with the control station. Nowadays there are many types of SCADA communication protocols. Modbus TCP [16], Distributed Network Protocol version 3 (DNP3) and Profibus are some of the most widely used protocols.

In this work Modbus protocol is used for communication between the components of the proposed SCADA platform. Modbus protocol supports request response messaging between master and slave. Modbus RTU and Modbus TCP/IP are the two variants of Modbus protocol. In this study the Modbus TCP/IP version of Modbus protocol is considered. In Modbus TCP/IP, the Modbus packet is embedded in the TCP segment and TCP port is dedicated to the protocol.

IMPLEMENTATION

To implement the Cyber Attack detection system that could detect and recognize any potential attack in the gas pipeline system we designed a machine learning based system that could learn and predict the situation and specificity of attack. An excellent example of such a SCADA system is the distribution system used by electric utilities, which is one of the oldest and most familiar SCADA systems. In gas pipeline, gas distribution SCADA is used to collect information from remote parts of a power distribution grid; for example the volts, amps or phase angle of a particular line in a substation, and provide it to a central control installation. In addition, SCADA allows an operator at the centralized control station to trip breakers at remote substations in response to conditions reported by the SCADA system. Other well-known industries that use

SCADA are the gas and oil utilities and nuclear power production.

Having the data filtered, the system proceeds to execute the machine learning algorithms. This consists in two distinct parts, as the name suggests: the machine phase and the learning phases. The goal of the former is to aggregate values by a specific key; the goal of the latter is to perform some sort of operation on the aggregated keys generated by the Machine phase. The libraries are import and data is loaded based on different algorithms of machine learning.

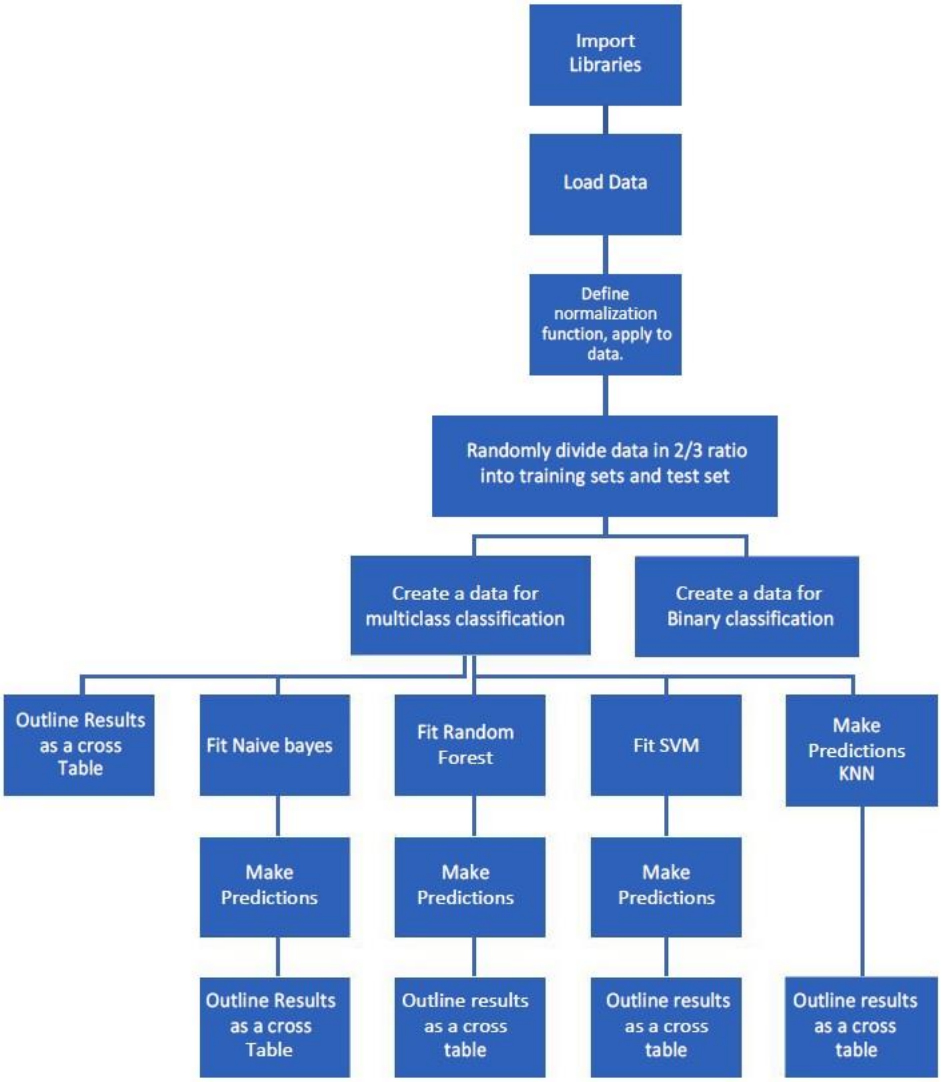
Import Libraries
Load Data
Define normalization function, apply to data.
Randomly divide data in 2/3 ratio into training sets and test set
Create a data for multiclass classification
Create a data for Binary classification
Outline Results as a cross Table
Fit Naive bayes
Fit Random Forest
Fit SVM
Make Predictions KNN
Make Predictions
Make Predictions
Make Predictions
Outline Results as a cross Table
Outline results as a cross table
Outline results as a cross table
Outline results as a cross table

Figure 3: *Model 1 depicts the whole distinctive algorithms used for detecting and analyzing the type of attack.*

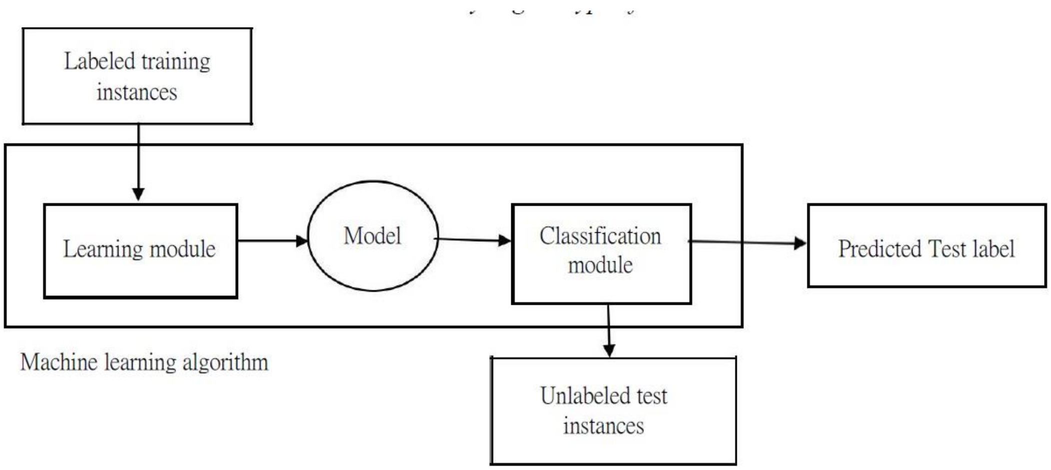

Figure 4: Training model for the labeled and unlabeled test for predicting the unusual Cyber Attack or intrusion from external network.

The model that predicts test labels from the dataset which knowingly classifies the instance of attacks which could probably destroy the normal operations of industry. That's the reason an autonomous system using machine learning is designed for the early detection of pipeline attacks near in future. Nowadays the user must run the unsupervised learning module. Once the flows are aggregated, the unsupervised learning module may take place, in order to cluster the flows to predict the detection of any potential attack.

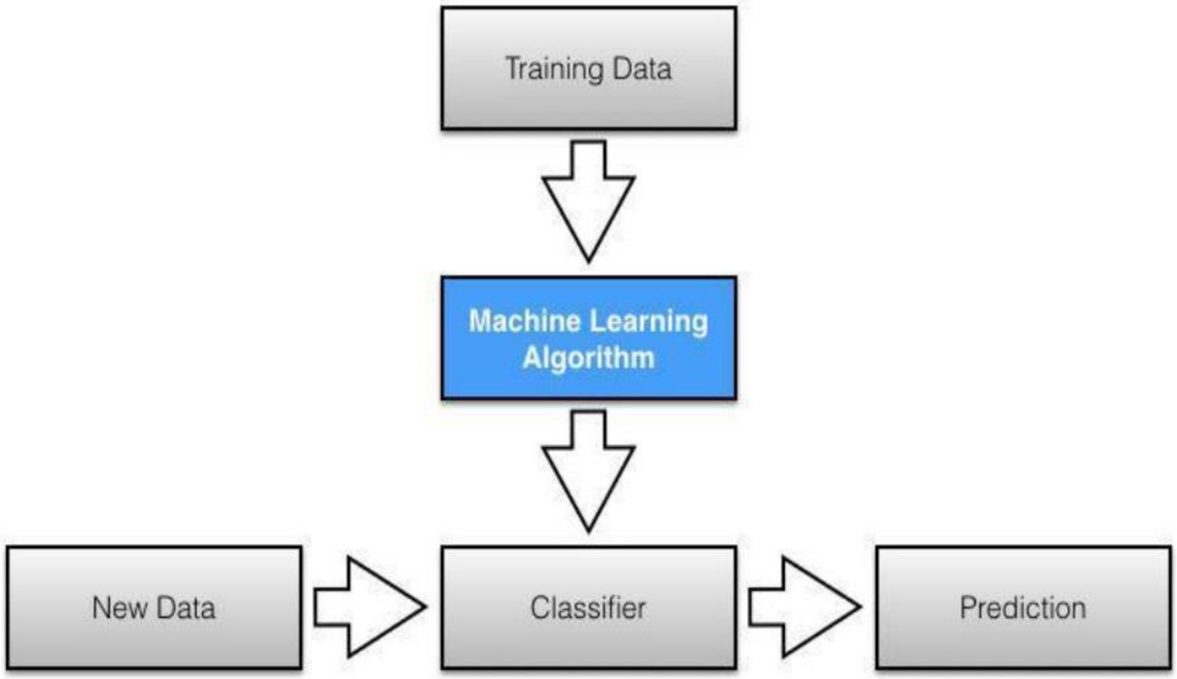

Figure 5: *Terminal user interface for the unsupervised learning module [17].*

RESULTS

When the data is described in more than three dimensions, it becomes a difficult task to represent it visually. For this reason, a -Dimensional representation of the dataset is also provided. Instead of dropping several features until only two are left, the dataset is rearranged such that there are now two features that describe the original ones. To compile the results for the feasible detection of the classification part comes. We used 'scikit-learn' to import classification algorithms. We classified our dataset from "Tommy Morris Dataset for Industrial Control System (ICS)" of Gas Pipeline System with many classifiers like SVM, Random Forest and k-NN. Some of the classifiers were quite efficient but some other were not up-to the mark. The Recall and Accuracy are represented in the result analysis part below.

k-NN: We examined K-NN classifier with different values of K. But 2-nearest neighbors earned the best result among them. We also set the weight of the algorithm according to the distance between two points. The leaf size was restricted to 30. Setting all these parameters, we got an accuracy of 87.37%.

RANDOM FOREST: This classifier worked with less efficiency for our corpus. We set the criterion of the algorithm to 'entropy' and got the result of 85.50% accuracy.

SVM: SVM is one of the best fitted classifier in our work. There are several kernels used in SVM but we implemented 'k-NN' and 'RF'. Linear SVM gained an accuracy of 86.04% setting the gamma value of the algorithm to 10, SVM with RF kernel gave us the best result of 85.50%.

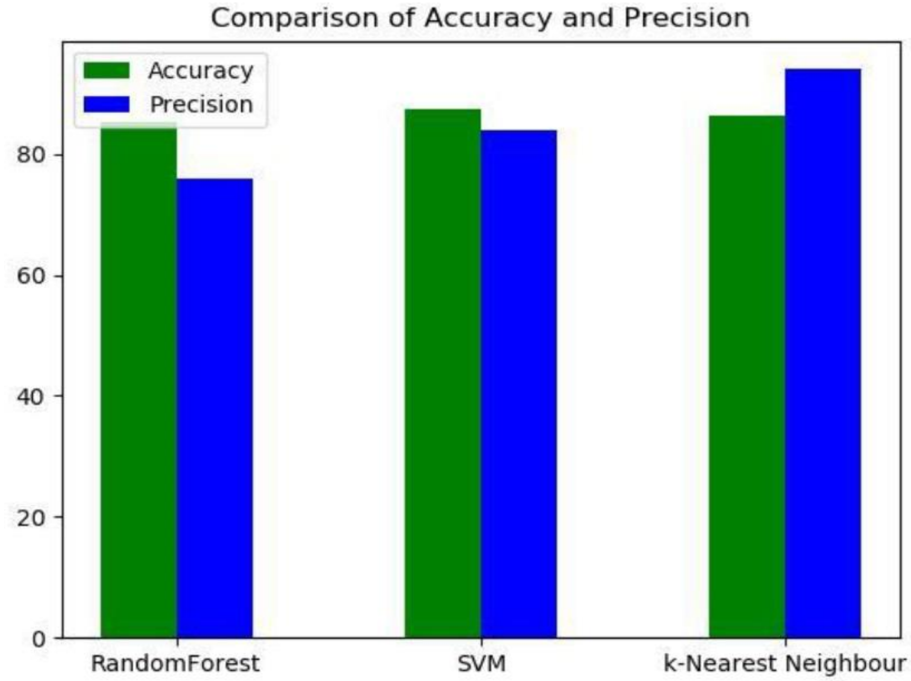

Figure 6: *The accuracy and precision of all three supervised learning algorithm for detecting the Cyber Attack in remote terminal units of SCADA system.*

These machine learning algorithms find the way in which all the attacks are identifiable by tracking the data provided by the source, and not the action will be taken against its results based on the algorithms, since we aim to find the malicious hosts; by analyzing the destination aggregation key, in this case, we would find the victims of the attacks, rather than the attackers, which is not the goal of this work. The reason that some attack were not identified, is that this system's focus is on large volumetric attacks, i.e. attacks that occur in large volumes, that exhaust the bandwidth of a network, and with feature values that tend to inflate. In the case of the attack, the system was only able to identify the one host that had the largest attack volume, as the rest of the host were producing a silent attack, that the system was not able to detect.

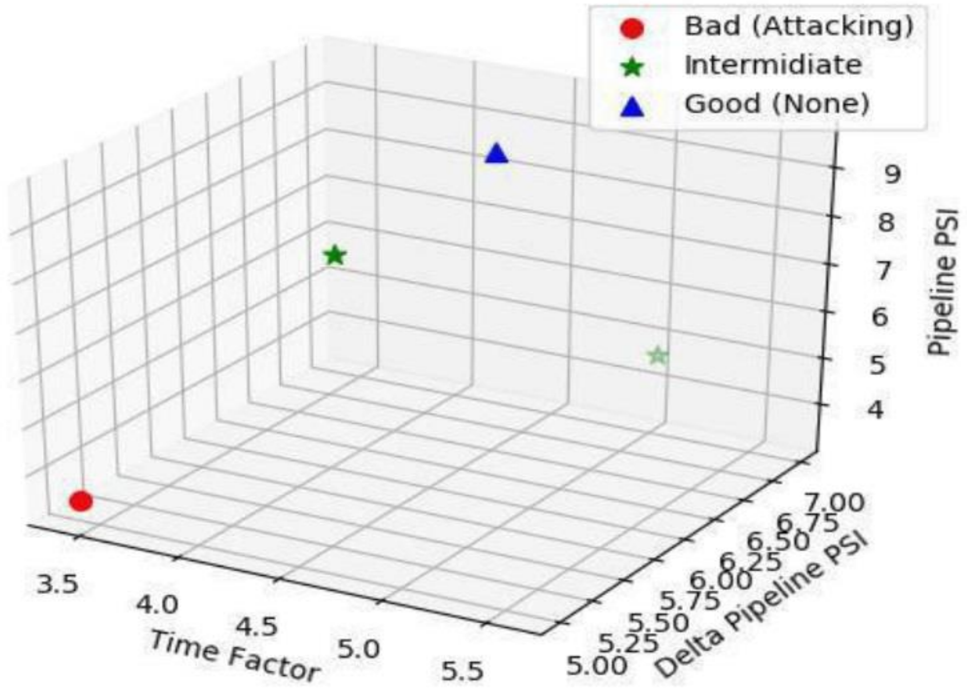

Figure 7: *The Pipeline PSI clusters is in y-axis and the Delta Pipeline represented on x-axis with time factor cluster on z-axis of detection graph. The Dataset is dividing into four clusters for detection of attack.*

SVM		k-NN		RF	
Detection Accuracy	**Detection Precision**	**Detection Accuracy**	**Detection Precision**	**Detection Accuracy**	**Detection Precision**
86.04%	94.04%	87.37%	83.90%	85.50%	75.89%

Table 1: *The precision and accuracy of all three algorithms extracted from the dataset of gas pipeline in remote terminal units of SCADA system.*

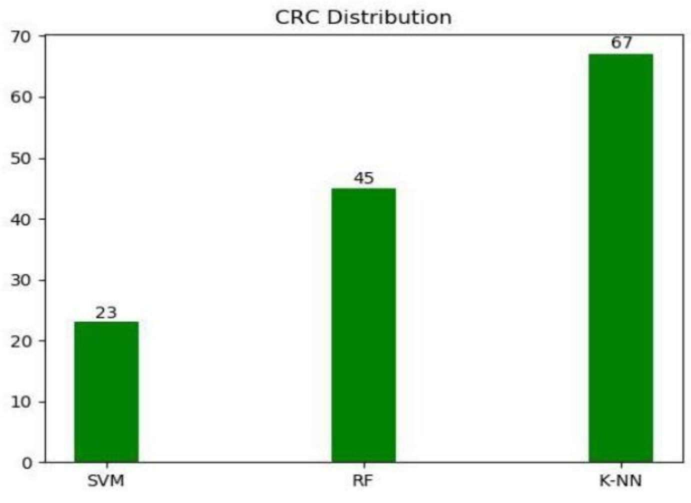

Figure 8: *A command packet captured by all the supervised algorithms in the RTU of normal operation of the platform.*

CONCLUSION

In this paper, we presented a simple but scalable SCADA security platform, which contains SCADA components that mimic real SCADA components. The developed platform is user friendly and easily reconfigurable for different types of Cyber Attacks. In the platform, simultaneous attacks can be initiated. Different nodes in the platform. For our thesis work, to detect Cyber Attack in the gas pipeline industry specially using the balanced approaches of supervised machine learning algorithms, we chose a quite balanced data that had reasonable amount of data; cleaned and preprocessed it. We used model to fit and transform our data into machine recognizable form to detect and recognize the effect of appearing attack in future. Extracted the best features by using suitable methods. Split the data and applied various supervised machine learning classifying algorithms such as Support Vector Machine, Logistic, K-Nearest Neighbor and Random Forest to detect performance. We tweaked and experimented with our data and model to achieve best accuracy possible. Eventually, we got the best result using Support Vector Machine using kernel for detecting the Cyber Attacks in the giant gas pipeline industry based on introductory technology of SCADA systems.

FUTURE WORK

In the future, we would like to evaluate the platform with several SCADA communication protocols. Furthermore, we will setup an arrangement to make the platform to be remotely accessible for researchers in the area of SCADA security. Using the data collected from the platform, we also plan to develop and evaluate a new SCADA specific intrusion detection system.

REFERENCES

[1] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, “Adversarial Machine Learning,” in ACM workshop on Security and artificial intelligence, 2011.

[2] A. Perrig, R. Szewczyk, J. Tygar, V. Wen, and D. Culler. SPINS: Security protocols for sensor networks. Wireless Networks 8:521–534, 2002.

[3] A. Perrig, R. Canetti, D. Song, and J. Tygar. Efficient and secure source authentication for multicast. In: Network and Distributed System Security Symposium, NDSS’01, 2001.

[4] A. Perrig, R. Canetti, J. Tygar, and D. Song. Efficient authentication and signing of multicast streams over lossy channels. In: IEEE Symposium on Security and Privacy, 2000.

[5] Washington Post. Dissertation could be security threat. http://www.washingtonpost.com/ac2/wp-dyn/ A23689-2003Jul7.

[6] F. Pierazzi, G. Apruzzese, M. Colajanni, A. Guido, and M. Marchetti, “Scalable architecture for online prioritization of cyber threats,” in International Conference on Cyber Conflict (CyCon), 2017.

[7] J. Kim, J. Kim, H. L. T. Thu, and H. Kim, “Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection,” in IEEE International Conference on Platform Technology and Service (PlatCon), 2016.

[8] P. Torres, C. Catania, S. Garcia, and C. G. Garino, “An analysis of Recurrent Neural Networks for Botnet detection behavior,” in IEEE Biennial Congress of Argentina (ARGENCON), 2016.

[9] G. E. Dahl, J. W. Stokes, L. Deng, and D. Yu, “Large-scale malware classification using random projections and neural networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

[10] G. D. Hill and X. J. Bellekens, “Deep Learning Based Cryptographic Primitive Classification,” arXiv preprint, 2017.

[11] R. Pascanu, J. W. Stokes, H. Sanossian, M. Marinescu, and A. Thomas, “Malware classification with recurrent networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[12] M. Z. Alom, V. Bontupalli, and T. M. Taha, “Intrusion detection using deep belief networks,” in IEEE National Aerospace and Electronics Conference (NAECON), 2015.

[13] W. Hardy, L. Chen, S. Hou, Y. Ye, and X. Li, “DL4MD: A Deep Learning Framework for Intelligent

Malware Detection," in International Conference on Data Mining (DMIN), 2016.

[14] G. Tzortzis and A. Likas, "Deep belief networks for spam filtering," in IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2007.

[15] G. Mi, Y. Gao, and Y. Tan, "Apply stacked auto-encoder to spam detection," in International Conference in Swarm Intelligence, 2015.

[16] M. Stevanovic and J. M. Pedersen, "An efficient flow-based botnet detection using supervised machine learning," in IEEE International Conference on Computing, Networking and Communications (ICNC), 2014.

[17] S. Ranjan, Machine learning based botnet detection using real-time extracted traffic features, Google Patents, 2014.