Aswan University Journal of Science and Technology Vol. 2, issue 2, December 2022

Deep Learning based Attacks Detection of DNP3 Protocol

Ahmed G. Yahia[a], Adly S. Tag El Dien[b], N. Abdel-Rahim [c]

[a]Faculty of Engineering, Helwan University, Cairo, Egypt,
AhmedG1990@yahoo.com
[b]Faculty of Engineering, Benha University, Cairo, Egypt,
adlytag@ feng.bu.edu.eg
[c]Faculty of Engineering, Future University, Cairo, Egypt,
naser.abdelrahim@fue.edu.eg

Abstract - SCADA systems contain many important components that communicate with each other through communication protocols designed for SCADA systems. This paper concerns distributed network protocol 3 (DNP3), which is considered a sufficient, trustworthy, and standard protocol for improving communications between multiple vendors. The vulnerabilities of this protocol form a disaster threat over the whole system, so this paper mentions these weakness points of this protocol. Also, the paper mentions the different types of attacks that exploit these vulnerabilities. So, it is necessary for researchers to continuously study mitigating these attacks without affecting the efficiency of the system. This goal is introduced in deep learning model algorithms dependent on neural networks. This paper introduces an ensemble deep learning algorithm (autoencoders) with decision tree (DT) multiple classification and support vector machine (SVM) multiple classification. After that, applying these two classifications models to a dataset to study the efficiency of each model and compares the results between each of them using performance metrics of deep learning algorithms and confusion matrixes which show the accuracy of each classifier.

Keywords: DNP3, Autoencoders, Decision tree (DT) classification, Support vector machine (SVM) classification.

1. INTRODUCTION

Supervisory Control and Data Acquisition (SCADA) systems monitor and control devices of critical infrastructures, such as power, telecommunication, transportation, pipelines, chemicals, and manufacturing plants. Legacy SCADA systems are isolated networks, that made them safe from outer threats. Now the increasing connection of SCADA systems to the Internet, as well as corporate networks, introduces serious security issues. Reports in [1] show security incidents are increased in SCADA infrastructure, so security concerns are propagate compared with common IT networks due to the impact of safety of society. In [2] discuss the components of Supervisory Control and Data Acquisition (SCADA) systems and Distributed Control Systems (DCSs) and configuration for the system. This control systems support real time monitor and control devices. These systems consist

Corresponding author E-mail: ashraf@aswu.edu.eg

Received November,25, 2022, received in revised form, December, 4, 2022, accepted December 4, 2022.

(ASWJST 2022/ printed ISSN: 2735-3087 and on-line ISSN: 2735-3095) https://journals.aswu.edu.eg/stjournal

of central controller and assign devices such as sensors and actuators the data transfer between the components utilized communication protocols used in industrial operations.in [3] represent logical analysis for threats, attacks, and how to mitigate these attacks in SCADA field. In [4] show the requirements, and the characteristics of communication system technologies. In [5] provide the security requirements and treat the vulnerabilities. SCADA communication frameworks and show security points and solutions in [6]. Provide in [7] comprehensive overview according to attackers point to view the threats and vulnerabilities of the systems of SCADA. in [8] This communications protocols must support the real time application and not suffer from delay that affect the industrial operation, that represent challenge for security application for these systems. In [9] another challenge is the high cost for securing the application of SCADA that make the industrial applications care well with security application. As a result, for these challenges, in [10] these make a disaster attack for the industrial application make it financial losses and losses in humanity souls. All the previous effects lead to secure the industrial system protocols [11], especially DNP 3 [12] protocol which using in most industrial environments as open and standard protocol. And for continuous changes in attacks there is need for training systems to detect the abnormal behavior using neural network [13] and deep learning [14] in industrial environments. So, this paper introduces an overview DNP3 protocol in **section .2**. After that in **section .3** Provides vulnerabilities of this protocol. These vulnerabilities result in attacks which mentioned in **section .4**. besides that, in **section .5**. the paper introduces a dataset to apply the deep learning algorithms. Then in **section .6** the paper provides model algorithm ensemble auto encoder for unbalanced data depends on neural networks with decision tree multiple classifications. In other hand **section.7**. Mentions the same algorithm depends on SVM multiple classifications. After that, in **section .8** compares the results between these two algorithms by the metrics and confusion matrices for each classification which measure the efficiency of the model algorithms.

2. DNP3 OVERVIEW

DNP3 Distributed Network Protocol is produced in 1990s which standards-based communication protocol developed to enhance communications among systems, this layered protocol offers higher data-transfer [15] integrity than most communication protocols. These layers utilize OSI model as in DNP3 layers application layer, transport layer, and lower layers are considered as datalink and physical layer. Transport layer divides the message of application to segments compatible with data link frame size. Application layer: determine the role of message request message from master to slave or replay message from slave to master may be solicited message (which be determined request from master), or unsolicited message (that is update or alarm from slave). Figure 1. show the format of application layer.

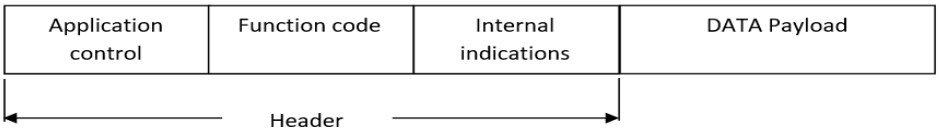


FIGURE 1. Application layer format

Application control: divide the data size exceed the limit size of transport layer into packets by two flags show the first or last packet at the application layer, the second flag show the sequence

number of packets. Function code: the purpose of message replay message or request message. Internal indication contains some information about outstation in the replay message.

3. Vulnerabilities of DNP3 Protocol

A study in [16] makes classification of attacks according to the target and threat type summarized vulnerabilities of protocol in points:

DNP3 is open protocol make devices connect with different protocols such as TCP, UDP, HTTP [17] this causes many ports to attack the protocol.

DNP3 introduce an architecture allow remote access to depend on logical point to implement encryption and authentication, [18] this makes the data transfer in plain text. DNP3 does not support security model, the only remote security is the username and password for authentication in DNP3 secure authentication version. In addition to all this device come with factory default for username and password this make easily penetrate the system using dictionary [19].

The protocol supports remote access to download configuration files to devices, reconfiguration, and restarting the devices. so, if the attacker penetrates the system may cause damage to the system [20].

4. ATTACKS ON DNP3

Man-in-the-middle Attack: This attack snoops on or captures the traffic transfer between the master and slave devices. The attacker can also modify the packet and transmit it to the respective devices.

Packet Modification and Injection Attack: The attacker can pick up the packet transferring between the master and slave device and may change its contents.

Denial-of-Service (DOS) Attack: The attacker attempts to make a service or a network resource unavailable to its intended users, or it can temporarily interrupt or suspend the services of a host connected to the network by randomly sending unexpected messages.

Replay Attack: The attacker can maliciously repeat or delay the valid message.

Spoofing: the attacker gains an illegitimate advantage, which can distort the data.

5. KDD Cup 1999 Dataset

To test deep learning algorithms, we should implement this model algorithms and test them in real industrial control systems (ICS), but it is more difficult to stop real systems in an industrial environment, so we will try these algorithms on a real dataset and study the results. Kddcup99 is held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. Which considered as a network intrusion detector, a predictive model capable of defining attacks, and ` normal connections [21]. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

This dataset contains 42 features with a total of 494021 observations and we will define the nature of the attack using autoencoders algorithms with DT multiple classifications and SVM multiple classifications.

In this paper we will use the Python 3.0 programming language for processing and applying analysis on dataset with Google Collab editor [22].

After processing and removing the inconsistent data from the dataset we get 42 attributes with 145586 instances.

Table 1. shows the count of Normal, and attacks categorizes.

category	Nature of category	count
Normal	Normal	87832
DOS (Denial of service)	Attack	54572
Prob	Attack	2131
R2L (Root to local)	Attack	999
U2R (User to root attack)	Attack	52

6. Auto encoder algorithm with DT multiple classification

A deep learning model that works with raw imbalanced datasets [23], which is considered a challenge to determine the attack from normal data [24]. The suggested solution makes a new balanced form from a raw dataset and passes it to an ensemble deep learning model for classification. The deep learning model consists of multiple unsupervised stacked auto encoders (SAE), which are considered a neural network to get a compressed representation of raw data. Then the encoders play a role in data preparation by being used to train the model [25]. So, we apply multiple autoencoders (AE) to extract a new form from unlabeled data to gather distinct patterns. Then, the result from each SAE is passed to a deep neural network (DNN) via a super vector and concatenated using a fusion activation vector. Lastly, a DT is used, as a binary classifier to identify attacks from the newly merged forms. The schematic of the proposed model is presented in Fig.2.

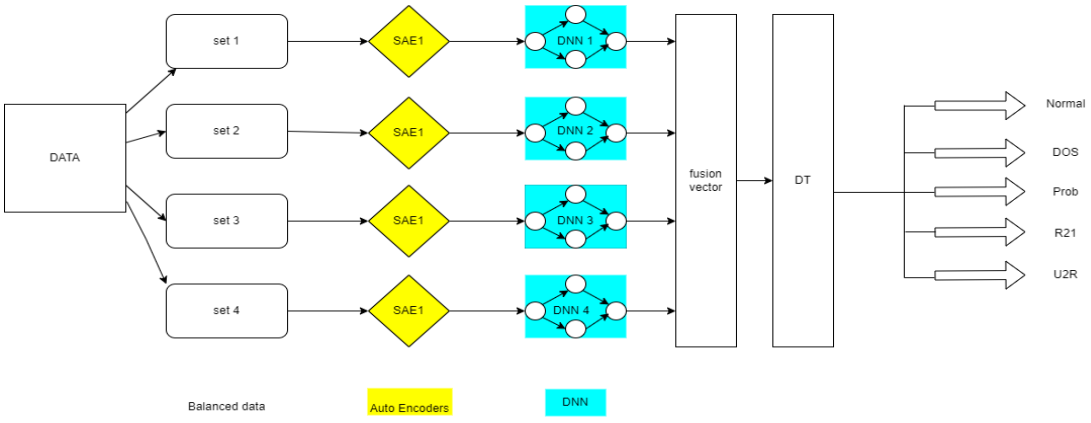


FIGURE 2. deep learning model with decision tree multiple classifications.

An ensemble deep representation-learning model based on SAE to enhance the overall performance of the model. This is achieved by extracting an equal and balanced set and transferring it to multiple AEs. The input sample x in a sample set X corresponding to the hidden layer is represented in the following equation

$$f(x_i) = u\,(r_1 x_i + c_1) \quad (1)$$

r and c represent the weight matrix of neurons and the bias vector of all neurons between the input and hidden layers, respectively [26], u is a function of the hidden layer used after beginning the training process by updating the next input layer to construct a set of stacked multi-layer AEs. To enhance the performance of each AE, a layer is added to develop the generalization of the model by reducing the dependence of the output on a specific set of parameters. Also, the number of nodes and layers was selected through cross-validation of various networks with critical analysis of loss history and validation accuracy. Binary Cross-Entropy (BCE) is used as the cost function.

$$j = -\frac{1}{m}\sum_{i=1}^{m} y_i \log\left(p(y_i)\right) + (1 - y_i).\log\left(1 - p(y_i)\right) \quad (2)$$

Where m is the number of samples, y_I is the attack value, $p(y_I)$ the probability of an attack sample. When the new representations are generated from the imbalanced dataset, they are passed to an ensemble of data neural networks (DNN) to detect normal from abnormal behaviors. The results from each DNN are then concatenated via a super vector using a fusion activation function. The fusion activation function of the sigmoid layer is

$$L = \sum_{i=1}^{m} y_i \log(q_i).w_s + (1 - y_i)\ \log(1 - q_i).w_l \quad (3)$$

Where y_I is the label of the i_th sample, q_I is the prediction of i_th, w_s is the weight of the unstable sample, w_l is the weight of stable sample. And passed on to a decision tree (DT) to detect the nature of the newly collected data.

7. Autoencoder with SVM Multiple classification

SVM in [27], and [28] is a large margin classifier of capacity (n) are divided by an (n + 1) dimensional hyperplane in such a way that each item has the maximal possible distance from the grouping hyperplane. The instances are noted as groups.

$$(A_i, B_i), i = 1, \ldots, n, B \in \{-1,1\} \quad (4)$$

A is a vector describing an instance of data in an n-dimensional feature space. B describes the attribution of the instance as belonging to one of two classes, while n is the number of instances. First, the SVM is trained with a labelled set of instances. It is a supervised classification method, meaning the training set needs to contain information about the correct classes. After training, the attribution of the test and productive data is produced by the signum function, as shown in (5). w is the normal vector of the separator hyperplane; d is the offset from the hyperplane.

$$B_i = \mathrm{sgn}(w, A_i - d) \quad (5)$$

When applying SVMs, obtaining a linear hyperplane to split the data set is desirable.
Figure 3 shows the previous model when using SVM classification instead of a decision tree.

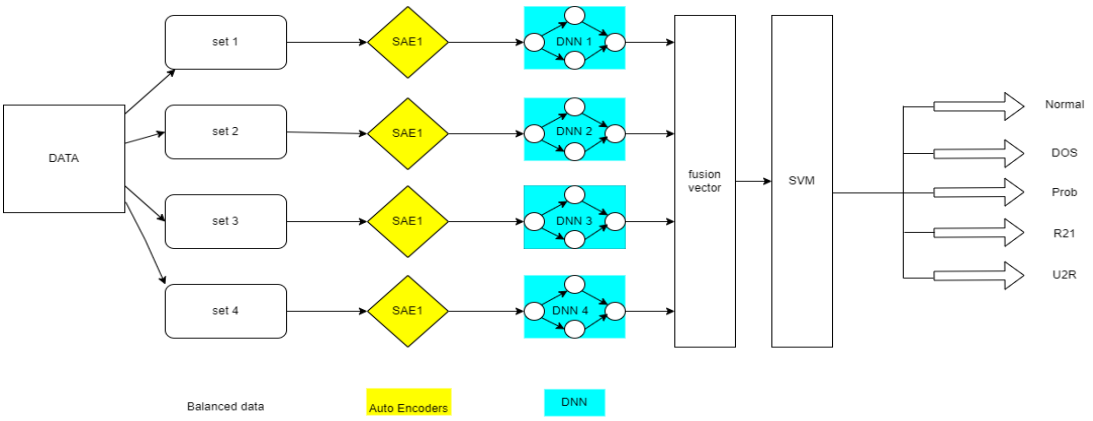


FIGURE 3. **deep learning model with SVM multiple classifications.**

8. comparison results

To test these algorithms, we should implement this model and test them in real industrial control systems (ICS), but it is more difficult to stop real systems in an industrial environment, so we will try these algorithms using the Python 3.0 programming language and Google Collab editor [22] on the dataset by dividing it into a training portion to train the system and a testing portion to test the accuracy of the system. The results of the metrics [29] and the confusion matrices will determine the performance of each algorithm.

When it comes to the security of ICSs, the concern revolves around detecting cyber-attacks while achieving high scores on imbalanced datasets, thereby minimizing the rate of false alarms. As with standard machine learning metrics, which are measured by their values, TP represents the number of correct attack instances, TN is the number of correct analyses of normal instances, FP is the number of incorrect analyses of normal instances as attacks, and FN is the number of incorrect classifications of attacks as normal instances. The performance of machine learning is determined by some of metrics:

Accuracy(A): the Ratio of samples analyzed correctly over the dataset

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Precision(P): the percentage of analyzed positive samples.

$$P = \frac{TP}{TP + FP} \quad (7)$$

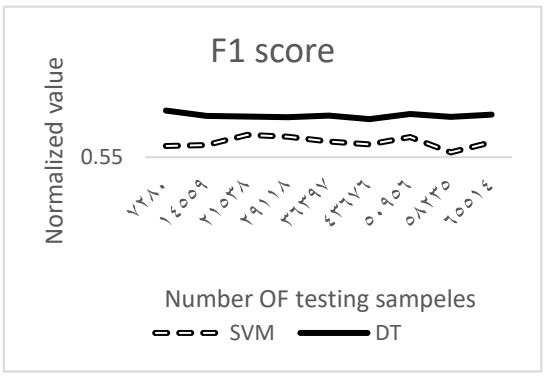
Recall(R): the ratio of predicted positive samples over the total

$$R = \frac{TP}{TP + FN} \quad (8)$$

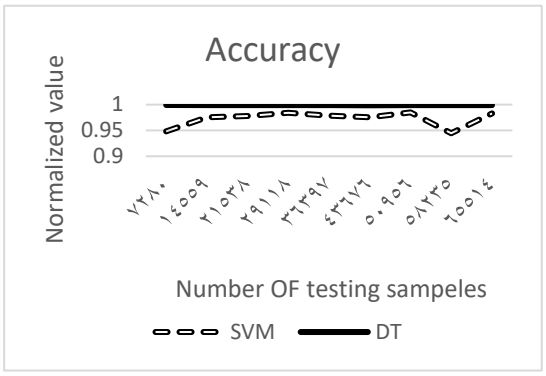
F1 Score (F1): the Harmonic mean of precision and recall represent equal balance between R, P which important for imbalanced data.

$$F1 = \frac{2 * TP}{2 * TP + FN + FP} \quad (9)$$

Figure 4 shows f1 score, accuracy, precision, and recall of the auto encoder algorithm with decision tree classification rather than SVM classification by changing the size of the training and testing datasets.

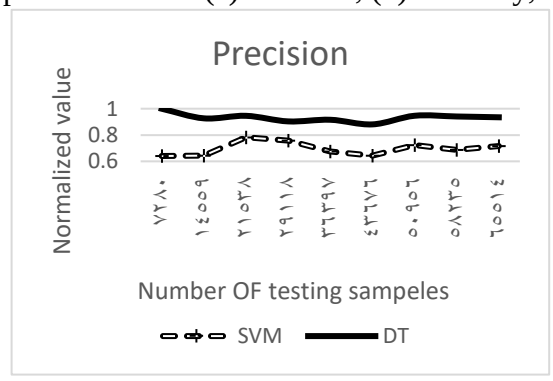


(a)

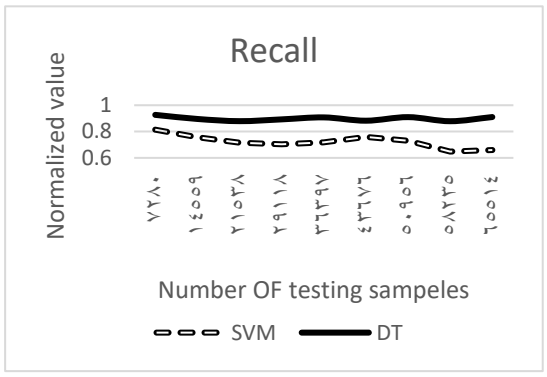


(b)

FIGURE 4. shows the characteristics for performance in (a) F1 score, (b) accuracy, precision, and (d) recall.

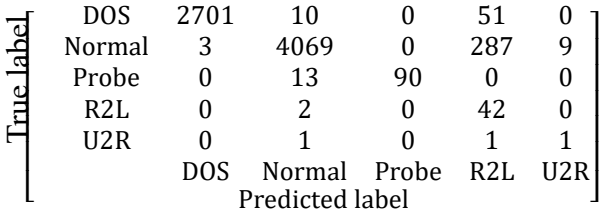


(c)

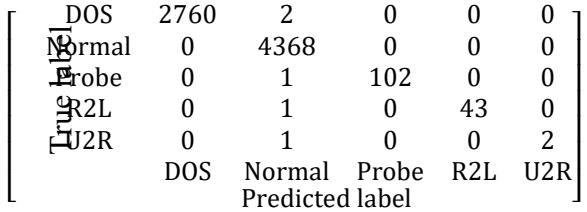


(d)

In the other hand we need to know the count of samples which were classified correctly by these models was showed in the confusion matrixes in figures 5,6,7,8,9,10,11,12,13 for each algorithm which provide the number of testing samples which predicted correctly for each classification by changing the size of training dataset and testing set.



(a)



(b)

Figure 5. shows confusion matrixes for (a) SVM, and (b) DT with 5% testing samples.

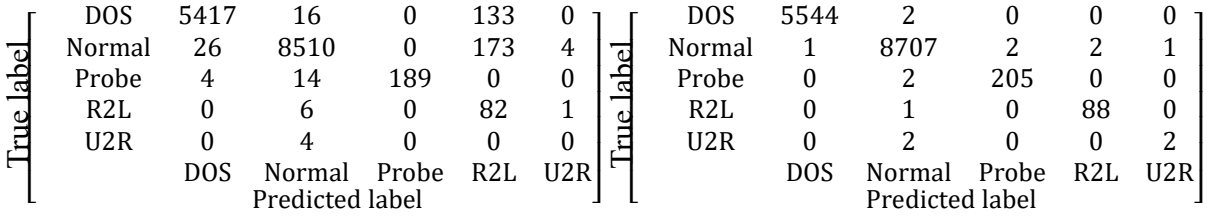


Figure 6. shows confusion matrixes for (a) SVM, and (b) DT with 10 % testing samples.

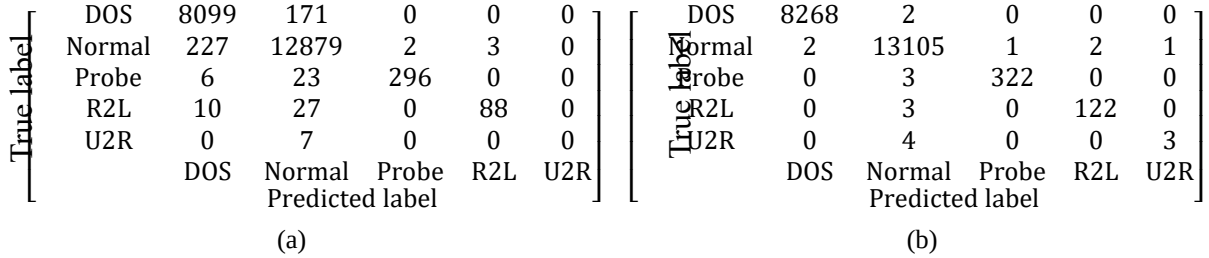


Figure 7. shows confusion matrixes for (a) SVM, and (b) DT with 15 % testing samples.

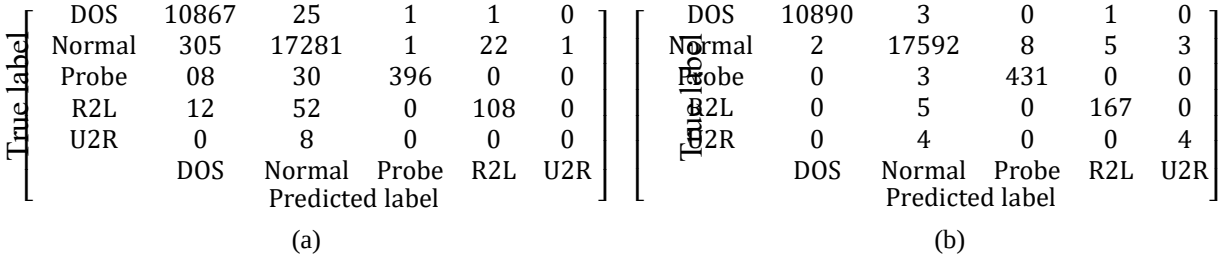


Figure 8. shows confusion matrixes for (a) SVM, and (b) DT with 20 % testing samples.

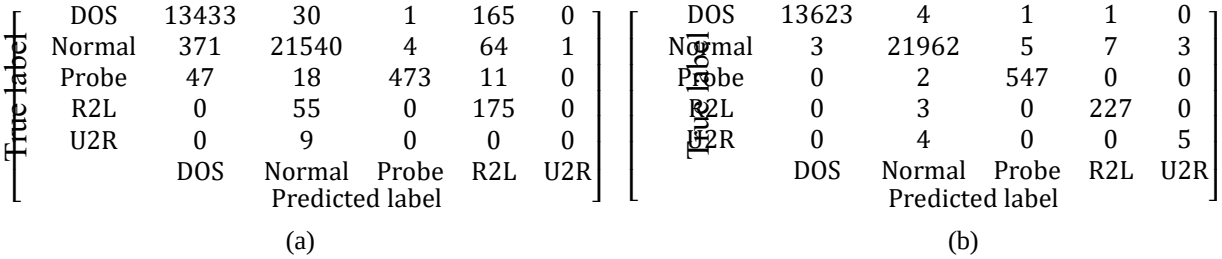


Figure 9. shows confusion matrixes for (a) SVM, and (b) DT with 25 % testing samples.

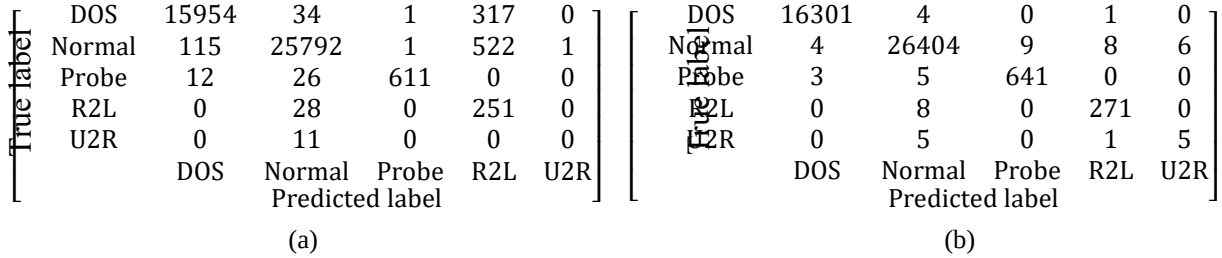


Figure 10. shows confusion matrixes for (a) SVM, and (b) DT with 30 % testing samples.

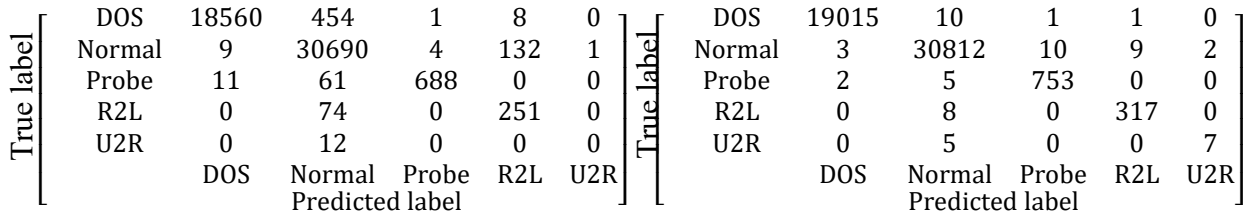


Figure 11. shows confusion matrixes for (a) SVM, and (b) DT with 35 % testing samples.

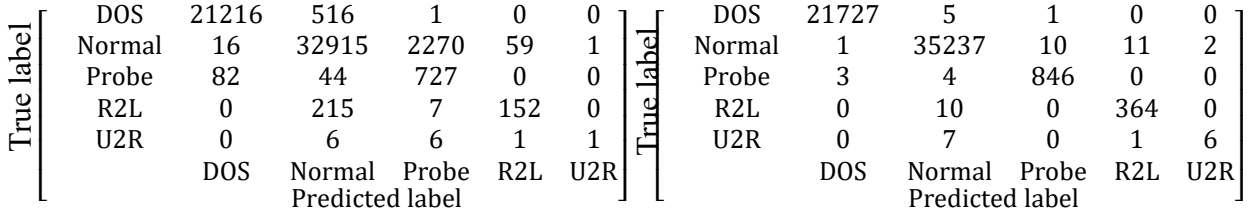


Figure 12. shows confusion matrixes for (a) SVM, and (b) DT with 40 % testing samples.

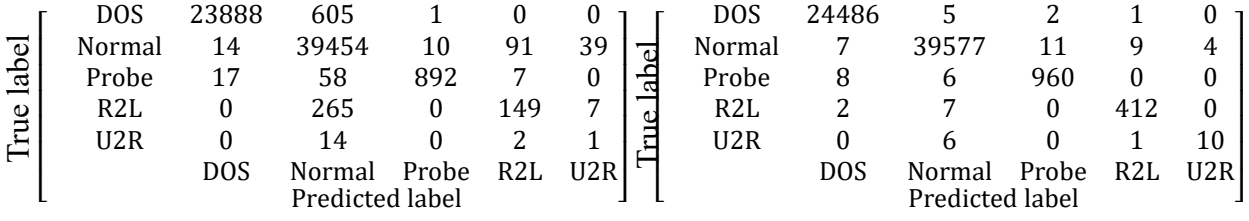


Figure 13. shows confusion matrixes for (a) SVM, and (b) DT with 45 % testing samples.

The confusion matrices show that DT classification predicted samples correctly more than SVM classification. In addition to that DT classification take less time than SVM which is needed in industrial environments.

9. Conclusion

In SCADA systems, we must be concerned with the data transferring without any change or delay. And DNP3 is considered a standard protocol for communication in SCADA systems where speed and future propagation are concerned, so this paper tries to mitigate the weaknesses of this protocol without affecting the time delay related to this treatment. Also, it must face new attacks each day with the raw data in the SCADA systems, which direct the researchers to the deep learning machine. So, this paper introduces deep learning models with multiple classification methods and shows that DT classification has higher performance than SVM classification. These multiple classifications provide more information with attack samples which lead to determine the nature of the attack. With the systems being trained continuously, this will lead to improve the system security performance.

References

1. Pliatsios, D., Sarigiannidis, P., Lagkas, T., and Sarigiannidis, A. G. "A Survey on SCADA Systems: Secure Protocols, Incidents, Threats and Tactics". IEEE Communications Surveys & Tutorials,2020. vol.22.NO.3.
2. Upadhyay, D., and Sampalli, S. "SCADA (Supervisory Control and Data Acquisition) Systems: Vulnerability Assessment and Security Recommendations". Computers and Security,2019. doi: 10.1016/j.cose.2019.101666.
3. Hilal, H., and Nangim, A. "Network security analysis SCADA system automation on industrial process". International Conference on Broadband Communication, Wireless Sensors, and Powering (BCWSP) ,2017. doi:10.1109/bcwsp.2017.8272569.
4. Yan, Y., Qian, Y., Sharif, H., and Tipper, D. "A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements and Challenges". IEEE Communications Surveys & Tutorials,2013. doi:10.1109/surv.2012.021312.00034.
5. W. Wang and Z. Lu. "Cyber security in the smart grid: Survey and challenges". Computer networks, 2013. vol. 57, no. 5, pp.1344–1371.
6. J. Gao, J. Liu, B. Rajan, R. Nori, B. Fu, Y. Xiao, W. Liang, and C. Philip Chen. "SCADA communication and security issues". Security and Communication Networks, 2014. vol. 7, no. 1, pp. 175–194.
7. S. Nazir, S. Patel, and D. Patel, "Assessing and augmenting SCADA cyber security: A survey of techniques". Computers & Security,2017. vol. 70, pp. 436–454.
8. Pidikiti, D. S., Kalluri, R., Kumar, R. K. S., and Bindhumadhava, B. S. "SCADA communication protocols: vulnerabilities, attacks and possible mitigations". CSI Transactions on ICT, 2013. 1(2), 135–141. doi:10.1007/s40012-013-0013-5.
9. Iqbal, A., and Iqbal, M. T. "Low-Cost and Secure Communication System for SCADA System of Remote Microgrids". Journal of Electrical and Computer Engineering, 2019.1–12.doi:10.1155/2019/1986325.
10. Korman, M., Välja, M., Björkman, G., Ekstedt, M., Vernotte, A., and Lagerström, R. "Analyzing the Effectiveness of Attack Countermeasures in a SCADA System". Proceedings of the 2nd Workshop on Cyber-Physical Security and Resilience in Smart Grids - CPSR-SG'17, 2017.doi:10.1145/3055386.3055393.
11. Jingran, W., Mingzhe, L., Aidong, X., Bo, H., Xiaojia, H., and Xiufang, Z. "Research and Implementation of Secure Industrial Communication Protocols". IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS),2020.
12. Radoglou-Grammatikis, P., Sarigiannidis, P., Efstathopoulos, G., Karypidis, P.-A., and Sarigiannidis, A. " DIDEROT: An Intrusion Detection and Prevention System for DNP3-based SCADA Systems". Proceedings of the 15th International Conference on Availability, Reliability and Security,2020.
13. Staar, B., Lütjen, M., and Freitag, M. "Anomaly detection with convolutional neural networks for industrial surface inspection". Procedia CIRP,2019. vol. 79, pp. 484–489.
14. Gupta, C., and Farahat, A. "Deep Learning for Industrial AI: Challenges, New Methods and Best Practices". Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and DataMining,2020. doi:10.1145/3394486.3406482.
15. Hoogenboezem, J. "Distributed Network Protocols—The Old and the New DNP3, IEC 60870-5 and IEC 61850". IDC Technologies Industrial Automation Conference,2012.

16. Drias, Z., Serhrouchni, A., & Vogel, O. “Taxonomy of attacks on industrial control protocols”. International Conference on Protocol Engineering (ICPE)and International Conference on New Technologies of Distributed Systems (NTDS),2015.
17. Njova, D., Ogudo, K., and Umenne, P. “Packet Analysis of DNP3 protocol over TCP/IP at an Electrical Substation Grid modelled in OPNET”. IEEE PES/IAS PowerAfrica ,2020.
18. Pham, B., Huff, C., Nick Vendittis, P., Smit, A., Stinskiy, A., and Chanda, S.” Implementing Distributed Intelligence by Utilizing DNP3 Protocol for Distribution Automation Application”. IEEE/PES Transmission and Distribution Conference and Exposition (T&D), 2018.
19. Jain, P., and Tripathi, P.” SCADA security: a review and enhancement for DNP3 based systems”.CSI Transactions on ICT, 2013. 1(4), pp.301–308.
20. Amoah, R., Camtepe, S., and Foo, E. “Formal modelling and analysis of DNP3 secure authentication”. Journal of Network and Computer Applications, 2016.vol. 59, pp. 345–360. doi: 10.1016/j.jnca.2015.05.015.
21. http://kdd.ics.uci.edu/.
22. https://colab.research.google.com/
23. Al-Abassi, A., Karimipour, H., Dehghantanha, A., and Parizi, R. M. “An Ensemble Deep Learning-based Cyber-Attack Detection in Industrial Control System”. IEEE Access,2020.
24. Mohit Sewak* Sanjay K. Sahay and Hemant Rathore “An Overview of Deep Learning Architecture of Deep Neural Networks and Autoencoders”. Journal of Computational, 2020.
25. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. “Adaptive Computation and Machine Learning series”, 2017.
26. Lei, N., An, D., Guo, Y., Su, K., Liu, S., Luo, Z., … Gu, X. “A Geometric Understanding of Deep Learning. Engineering”. arxiv, 2020.
27. D. Shalyga, P. Filonov, A. Lavrentyev, "Anomaly detection for water treatment system based on neural network with automatic architecture optimization”. arxive, 2018.
28. Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C.” High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning”. Pattern Recognition, 2016.vol. 58, pp. 121–134.
29. I. P. Turnipseed, "A new SCADA dataset for intrusion detection System research" Mississippi State University,2015.