International Journal of Web-Based Learning and Teaching Technologies
Volume 17 • Issue 2

Enhanced SCADA IDS Security by Using MSOM Hybrid Unsupervised Algorithm

Sangeetha K., Kebri Dehar University, Kebri Dehar, Ethiopia
Shitharth S., Kebri Dehar University, Kebri Dehar, Ethiopia
Gouse Baig Mohammed, Vardhaman College of Engineering, India

ABSTRACT

Self-organizing maps (SOM) are unsupervised neural networks that cluster high dimensional data and transform complex inputs into easily understandable inputs. To find the closest distance and weight factor, it maps high dimensional input space to low dimensional input space. The closest node to data point is denoted as a neuron. It classifies the input data based on these neurons. The reduction of dimensionality and grid clustering using neurons makes to observe similarities between the data. In the proposed mutated self-organizing maps (MSOM) approach, the authors have two intentions. One is to eliminate the learning rate and to decrease the neighborhood size, and the next one is to find out the outliers in the network. The first one is by calculating the median distance (MD) between each node with its neighbor nodes. Then those median values are compared with one another. If any of the MD values significantly varies from the rest, they are declared as anomaly nodes. In the second phase, they find out the quantization error (QE) in each instance from the cluster center.

KEYWORDS

Internet Security, Intrusion Detection System (IDS), Mutated Self-Organizing Maps (MSOM), Quantization Error (QE), Self-Organizing Maps (SOM), Supervisory Control and Data Acquisition (SCADA)

1.INTRODUCTION

Supervisory control and data acquisition system are such an integral part of the latest automation industries. This receives data from various sources like sensors, RTU (Remote Terminal Units) and smart meters. The major tasks performed by SCADA (Rakas et al., 2020; Tamy et al., n.d.) is to monitor the connected data fetching sources. SCADA systems are mainly used to control and monitoring purposes in various industrial applications. It can be used for a small office building to monitor environmental conditions also used to monitor complex conditions in a nuclear power plant SCADA (Ferrag et al., 2020; Khan et al., 2019; Waagsnes & Ulltveit-Moe, n.d.) . To protect control systems, systems are evaluated before being deployed in production. So the operators have a good understanding of what types of vulnerability those systems may be introducing into their environment. One of the challenges of control systems is that many of them have been developed in an environment that works very well in operations, but they don't have all of the cybersecurity safeguards built into them (Shitharth et al., 2021; Suaboot et al., 2020). Sensor nodes which sense physical phenomenon that

DOI: 10.4018/IJWLTT.20220301.oa2

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

occur around them. These sensor nodes are majorly used for medical purposes, agriculture, industrial purposes, and so on. SCADA system uses wired or wireless sensor networks to transport the data from the master station. SCADA systems mainly use wireless sensor networks due to their frequent changing topology nature and the possibility of reconfiguration of networks. Using the wireless sensor networks this information or data is transmitted through a router, firewall, and switches. The first layer of protection is a router. The router should be configured with a VPN tunnel on the router side. Firewalls in control systems are used to protect unauthorized access (Priyanga et al., 2019; Teixeira et al., 2018). Data should be encrypted, to increase the level of information security, while accessing information through the internet. Various security threats (Gao et al., n.d.; Gao et al., 2020; Shitharth & Winston, 2016) are evolving every day like unauthorized access to the control software, virus infection and one more major threat is intruders sending malicious packets to host devices. By sending these packets anyone can control the SCADA devices.

2. RELATED WORK

SCADA network system has a high level of flexibility and adaptability, yet it consumes notable resources (Gao et al., 2020). Feature sets of these techniques are easily accessible by the attacker. SCADA systems are vulnerable to LOWSPAN or IP based networks. Lowspan architecture for SCADA systems is used for various applications (Shitharth & Winston, 2016). When SCADA systems used in industrial environments, it requires high receptivity, high acceptance, stability, and measurements. Yang, et al (n.d.) suggested a deep learning model for detecting the intrusions from the SCADA system based on the hand-crafted features. The main intention of this paper was to develop a retaining scheme for handling new threats by characterizing the salient temporal patterns. Ghosh, et al (2019) conducted a detailed survey on various issues and challenges related to the SCADA security. Here, the different types of threats could affect the normal operations of SCADA systems have been investigated, which includes masquerade, virus and norms, eavesdrop, Denial of Service (DoS), and Trojan horse. Qassim, et al (2017) intended to analyze various security vulnerabilities in the SCADA systems for ensuring the increased security of data transmission. Here, various testbed approaches have been validated based on the parameters of scalability, reliability, accuracy, safety, and repeatability. Almalawi, et al (2015) developed a data drivel clustering approach for detecting the normal and precarious stages of SCADA systems by using the proximity based detection rules. The main purpose of this work was to minimize the false positives by estimating the Euclidean distance for relabeling the critical states. In addition to that, the rank based precision measure was also computed for analyzing the efficiency of this IDS framework. Mo, et al (2013) presented a comprehensive survey for identifying the integrity attacks of SCADA networks. The aim of this review was to provide the possible countermeasures for detecting the harmful intrusions against the SCADA networks.Gumaei, et al (2020) implemented a cyberattacks detection framework for enhancing the security of SCADA based on the optimal selection of features. Here, the Correlation based Feature Selection (CFS) model was utilized to enhance the detection accuracy of network by eliminating the irrelevant features. In addition to that, the KNN algorithm was deployed to accurately predict the intrusions based on the optimal set of features. Kalech, et al (2019) suggested a temporal pattern recognition approach incorporated with the Hidden Markov Models (HMM) for ensuring the security and reliable communication in SCADA networks. Here, the feature extraction was mainly peformed to analyze the normal behavior of temporal patterns. Also, it utilized the HMM for categorizing the types of intrusions according to the generated patterns of the given dataset. Lai, et al (2019) employed a CNN model for predicting the anomalies with increased accuracy and reduced misclassification results. Here, the correlation between the features have been analyzed with reduced cost of error for identifying the anomalies. Also, the different types of classification techniques such as HMM, SVM ensemble, DT and CNN models have been validated and compared based on the accuracy. From the analysis, it was observed that the CNN model outperforms the other techniques with high performance

values. Yet, the computational complexity of this mechanism is need to be reduced for ensuring the increased efficiency of this prediction system. Yang, et al (2016) constructed a multi-dimensional IDS for improving the secruity of SCADA systems against the vulnerabilities. Also, the Protocol Whitelisting Detection (PWD) approach was used for increased the QoS of network with reduced traffic. In this paper Davidson et al., (n.d.), the author proposed Zigbee feature set supports data encryption that determines the changes in key distribution encryption SCADA networks have been incorporated with the internet which in turn has significantly increased threats to critical infrastructure. In this case, Scada systems should implement an architecture with various data frames that protect the architecture from external attacks.

3. EXISTING SYSTEM

Before classifying any unsupervised data, first and foremost the data has to be clustered. In a neural network, for classifying any high dimensional data SOM (Self Organizing Maps) is highly preferred. This architecture usually includes three layers such as the input layer, set of weights and Kohonen layer. The Kohonen layer is fully connected with neurons that have weights for each input. These weights get trained over time based on the difference between the input and the weight values.

3.1 SOM Algorithm

1) Calculate the number of neuron nodes and initialize weight for every neuron.
2) Traverse each vector from the training data set.
3) The best Matching Unit (BMU) is calculated based on the weighted similarity of each neuron that is closely associated in distance with the input vector using the Euclidean distance formula.
4) Track those BMU nodes and update them in the neighborhood of input nodes.
5) Then BMU is calculated for the updated neighborhood nodes.
6) Over a while, the number of neighbors gets gradually decreased.
7) Finally, the winning BMU's and neighbor nodes are alike sample nodes. The closer the distance, the more the weight is modified and vice versa if more the distance.
8) Repeat step 2 for N iterations.
   N - Number of iterations

$$Wn(t+1) = Wn(t) + \theta\,(v,n)\,\alpha(t)(D(t)) - Wn(t)) \quad (1)$$

Wn is the updated weight node $\theta(n,t)$ is the neighborhood function constraining w.r.t BMU

Where $\alpha(t)$ is a learning coefficient monotonically downsizing that and D(t) is the input vector.

In this figure, though the neurons in the lattice seem to be in the same size they aren't. The hexagonal lattice consists of neighbors that belongs to the smallest neighborhood of the neuron. The number of neurons should be affixed so that the results can be granular. Irregular neuron scaling may affect the accuracy of the prediction model.

3.2 Problem With the Existing SOM

Though the existing algorithm has many pros, it also has its cons. The major drawback of the existing SOM is its slow-paced learning rate and detecting anomaly nodes. Both these issues are inter-related because the learning rate is highly dependent on the neighborhood size MSOM requires neuron weights be necessary and sufficient to cluster inputs. When an SOM is provided too little information or too much extraneous information in the weights, the groupings found in the map may not be entirely accurate or informative.

Figure 1:Decision diagram of lattice output

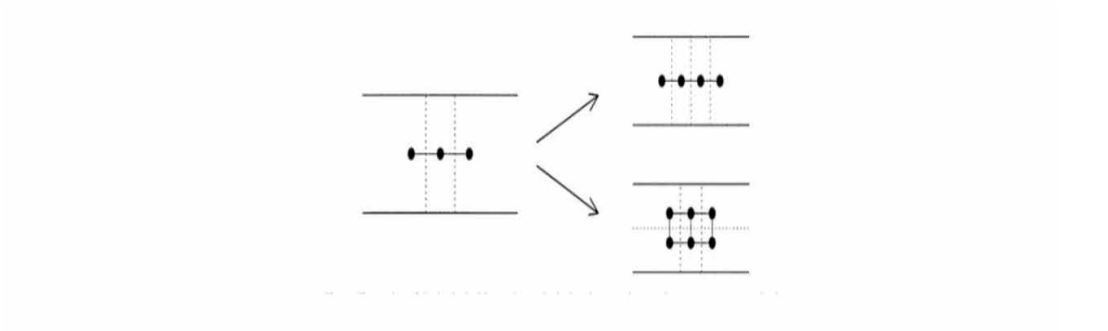
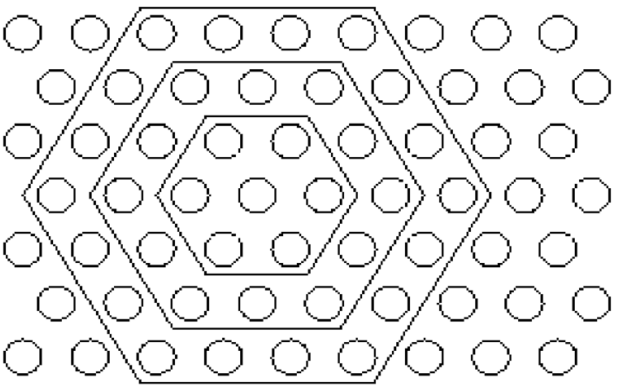


Figure 2: Neighborhood of a given winner unit



As the training progress, the learning rate parameters and neighborhood size vary according to the external variables. Here the major issue is such external variables are not always genuine. There are a lot of outliers also called anomalies which tend to decide the learning rate. So this paper's main motto is to get rid of those anomalies with the use of distributed agents which is used to attain fault tolerance and resistance. In addition to that, an enhanced learning rate is attained which depends only on internal variables, not on any outliers.

3.3 Mutated SOM Algorithm with Distributed Agents

As the SOM algorithm is concerned, the basic idea falls into multidimensional projection into one dimensional or two-dimensional neuronal structures. As compared to other estimation models like parzen windows, vector quantization and kernel density estimator that predict cluster density, SOM is much faster and its fault tolerance is relatively better than the aforementioned techniques. As mentioned in fig 3, the weights are mapped with the neurons based on their input layer levels. In existing Self-Organising Maps (SOM), there is a problem of dependence on the learning rate, the size of the neighborhood function and the decrease of these parameters as training progresses.

4. PROPOSED WORK

Our proposed system majorly targets to get rid of the learning rate from the system which eventually results in decreased neighborhood size. The prime intention of this paper is to accurately detect intrusions from the SCADA sensor network data through unsupervised learning. Usually, in unsupervised learning, the data collected has no labels and the output is always uncertain. As a result, an unsupervised algorithm has to understand the patterns in the data and then further process the desired output. This research proposes a modified neural network algorithm called MUSOM (mutated Self Organizing Maps). This algorithm has two major uses. One is it is used to accomplish fault tolerance, error correction and dimensionality reduction concerning the unknown anomalies. The next one is to find the outlying anomalies in the network.

4.1 Dimensionality Reduction

A novel learning methodology is introduced which makes the memory size of the neighborhood dependent on inlying node variables rather depending upon the outlying values. The algorithm is as follows:

$$Wn(t+1) = Wn(t) + \theta\,(n,t)\,\alpha(t)\,(D(t)) - Wn(t)) + \theta\,(n,t)\,(t/T) \tag{2}$$

Wn is the modified weight node θ(n,t) is the neighborhood function constraining w.r.t BMU

Where α(t) is a learning coefficient monotonically downsizing that and D(t) is the input vector. The lattice distance between the BMU and neuron n decides the value of neighborhood function θ (n,t).

Our MSOM main motto is to eliminate or reduce the learning rate completely. This is achieved through learning continuously from the environment nodes and only the initialization of the node is enough. As it does this, the training time will get reduced automatically. It also doesn't require additional memory as it updates the nodes on the flow.

4.2 Implemented MSOM Algorithm

The MSOM algorithm is implemented using pandas and also by using specific library functions like minisom and sompy. Initialize Grid size so that the number of clusters is fixed. The central grid node size is not affixed. The N values will vary gradually as the algorithm progress. Once N is defined, initialization for each cluster starts randomly.

A random data instance v(t) is chosen from the training data. Then every cluster center is examined thoroughly and the center is nearest to the chosen instance selected as BMU.

Formula to calculate the radius of the BMU:

The BMU radius is given byμ

$$\mu\,(t) = \mu 0\,(\exp(-\lambda)),\ t = 1, 2, 3... \tag{3}$$

where μ0 - initial radius, λ - time constant, and t - current iteration. At the initial stage μ0, is high with whole grid covered and eventually gets lower w.r.t time.

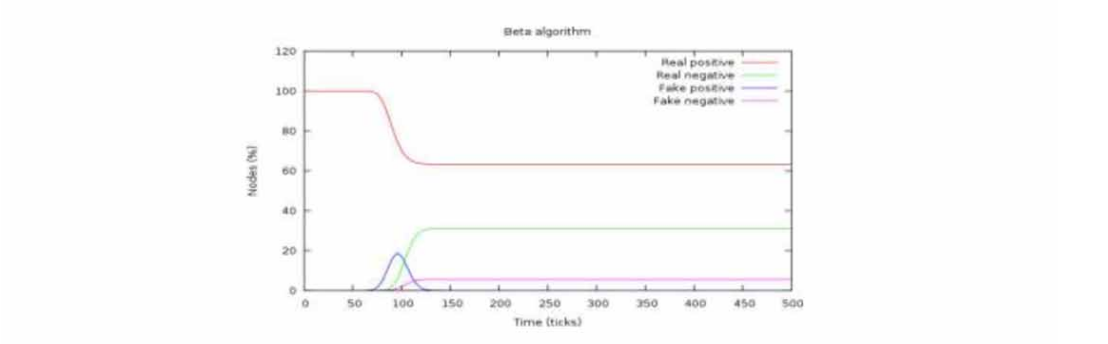
The cluster center is adjusted periodically since it lies in the BMU neighborhood. In instance n(t), the n-gram value is varied subjected to its presence in the node.

$$t = (t+1) + (L(t)((t)\ n(t)) \tag{4}$$

where L(t) - learning time

$$L(t) = L0\ \exp(-\lambda),\ t = 1, 2, 3.. \tag{5}$$

Figure 3: Weight Matrix Mapping



Whereas the node distance from the BMU is highly influenced by dist(t) and it is calculated by Formula 4:

$$\text{dist}(t) = \exp(-2\ \mu\ 2(t)),\ t=1,2\ldots.n. \tag{6}$$

where dis is the node distance from the BMU and μ (t) – radius function.

The n-gram is added to the cluster center, only in the absence of the instance v(t). The basic principle of this algorithm is to modify the node weightage w.r.t the present instance. Based on this all the neighborhood nodes get closer to the instance. On every iteration, the node distance would be updated. Gradually every node's comparison time would get shorter and shorter as the cluster size decreases periodically. This makes the classifier more stable and less time complexity.

5. EXPERIMENTAL RESULTS

The proposed algorithm has been implemented using python libraries. To make sure that the performance is better than the previous algorithm we have made a comparison graph with a beta algorithm, linear algorithm, and MSOM. The beta algorithm is used for risk assessment in WSN whereas a linear algorithm is used for node placement and secure routing in the sensor network. Certain calculated parameters that let us measure how effective the proposed system performance is when compared with the existing work:

Real Positive (RP)- accounts for the case in which a truly positive instance is categorized positive itself

$$RPRate \approx \frac{Positivescorrectlyclassified}{TotalPositives} \tag{7}$$

Real Negative (RN)- interprets for the case in which a truly a positive instance is categorized negative

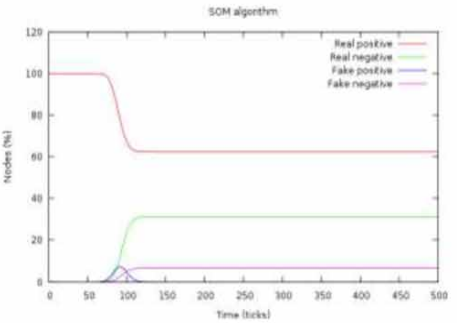
Fake Negative (TN)- accounts for the case in which a truly negative instance is categorized negative

Fake Positive (FP)- infers for the case in which a truly the negative instance is categorized positive

$$FPRate \approx \frac{\text{Negativesincorrectlyclassified}}{\text{TotalNegatives}} \tag{8}$$

One of the major threats in this scenario is a clone attack. Here the attackers try to capture a good amount of packets and re-architect it then finally replicate into larger numbers. Those packets will try to take over the entire network .

Figure 4.Occurrence of the real/fake positives/negatives for a clone attack.



As mentioned in Fig 4, SCADA is more vulnerable to the node replication attack (clone attack). Attackers through compromising one sensor node replicate many clones having the same identity (ID) from the compromised node, and place these clones in various places of network. Moreover, Self Organizing Map(SOM) provides a data visualization technique which helps to understand high dimensional data by reducing the dimensions of data to a map. SOM also represents clustering concept by grouping similar data together.

6. CONCLUSION

Our proposed algorithm deploys an agent that examines every node's communication in the promiscuous mode to trace the symptoms of malicious behavior. The SOM agents have a reputation system analyzing the trust value of every node. This trust value is calculated based on the node behavior. Our proposed system will give a higher reputation value for the trusted node and lower value for the malicious node which basically cannot be trusted at all. The major advantage is that the reputation values are chosen from the relative nodes which make the system more reliable from the outliers. In the future, this MSOM algorithm is about to get tested in a real-time testbed environment with the help of SCADA. Then as a further improvement, it would be tested with the SCADA firewall for improved intrusion detection.

REFERENCES

Almalawi, A., Fahad, A., Tari, Z., Alamri, A., AlGhamdi, R., & Zomaya, A. Y. (2015). An efficient data-driven clustering technique to detect attacks in SCADA systems. *IEEE Transactions on Information Forensics and Security*, *11*(5), 893–906.

Davidson, Andel, Yampolskiy, McDonald, Glisson, & Thomas. (n.d.). *On scada plc and fieldbus cyber-security.* Academic Press.

Ferrag, M. A., Babaghayou, M., & Yazici, M. A. (2020). Cyber security for fog-based smart grid SCADA systems: Solutions and challenges. *Journal of Information Security and Applications*, *52*, 102500.

Gao, Gan, Buschendorf, Zhang, Liu, Li, Dong, & Lu. (n.d.). *LSTM for SCADA intrusion detection.* Academic Press.

Gao, J., Gan, L., Buschendorf, F., Zhang, L., Liu, H., Li, P., Dong, X., & Lu, T. (2020). Omni SCADA intrusion detection using deep learning algorithms. *IEEE Internet of Things Journal*, *8*(2), 951–961.

Ghosh, S., & Sampalli, S. (2019). A survey of security in SCADA networks: Current issues and future challenges. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 135812–135831.

Gumaei, A., Hassan, M. M., Huda, S., Hassan, M. R., Camacho, D., Del Ser, J., & Fortino, G. (2020). A robust cyberattack detection approach using optimal features of SCADA power systems in smart grids. *Applied Soft Computing*, *96*, 106658.

Kalech, M. (2019). Cyber-attack detection in SCADA systems using temporal pattern recognition techniques. *Computers & Security*, *84*, 225–238.

Khan, I. A., Pi, D., Khan, Z. U., Hussain, Y., & Nawaz, A. (2019). HML-IDS: A hybrid-multilevel anomaly prediction approach for intrusion detection in SCADA systems. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 89507–89521.

Lai, Y., Zhang, J., & Liu, Z. (2019). *Industrial anomaly detection and attack classification method based on convolutional neural network* (Vol. 2019). Security and Communication Networks.

Mo, Y., Chabukswar, R., & Sinopoli, B. (2013). Detecting integrity attacks on SCADA systems. *IEEE Transactions on Control Systems Technology*, *22*(4), 1396–1407.

Priyanga, S., Gauthama Raman, M., Jagtap, S. S., Aswin, N., Kirthivasan, K., & Shankar Sriram, V. (2019). An improved rough set theory based feature selection approach for intrusion detection in SCADA systems. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 3993–4003.

Qassim, Q., Jamil, N., Abidin, I. Z., Rusli, M. E., Yussof, S., Ismail, R., Abdullah, F., Ja'afar, N., Hasan, H. C., & Daud, M. (2017). A survey of scada testbed implementation approaches. *Indian Journal of Science and Technology*, *10*(26), 1–8.

Rakas, S. V. B., Stojanović, M. D., & Marković-Petrović, J. D. (2020). A review of research work on network-based scada intrusion detection systems. *IEEE Access : Practical Innovations, Open Solutions*, *8*, 93083–93108.

Shitharth, S., Satheesh, N., Kumar, B. P., & Sangeetha, K. (2021). *IDS Detection Based on Optimization Based on WI-CS and GNN Algorithm in SCADA Network. In Architectural Wireless Networks Solutions and Security Issues*. Springer.

Shitharth, S., & Winston, D. P. (2016). A New Probablistic Relavancy Classification (PRC) based Intrusion Detection System (IDS) for SCADA network. *Journal of Electrical Engineering*, *16*(3), 278–288.

Suaboot, J., Fahad, A., Tari, Z., Grundy, J., Mahmood, A. N., Almalawi, A., Zomaya, A. Y., & Drira, K. (2020). A taxonomy of supervised learning for idss in scada environments. *ACM Computing Surveys*, *53*(2), 1–37.

Tamy, Belhadaoui, Rabbah, Rabbah, & Rifi. (n.d.). *An evaluation of machine learning algorithms to detect attacks in SCADA network*. Academic Press.

Teixeira, M. A., Salman, T., Zolanvari, M., Jain, R., Meskin, N., & Samaka, M. (2018). SCADA system testbed for cybersecurity research using machine learning approach. *Future Internet*, *10*(8), 76.

Waagsnes & Ulltveit-Moe. (n.d.). *Intrusion Detection System Test Framework for SCADA Systems.* Academic Press.

Yang, Cheng, & Chuah. (n.d.). *Deep-learning-based network intrusion detection for SCADA systems.* Academic Press.

Yang, Y., Xu, H.-Q., Gao, L., Yuan, Y.-B., McLaughlin, K., & Sezer, S. (2016). Multidimensional intrusion detection system for IEC 61850-based SCADA networks. *IEEE Transactions on Power Delivery*, *32*(2), 1068–1078.

K. Sangeetha received her B.E. degree in Computer Science from Ramakrishna College of Engineering, Coimbatore, India, in affiliation with Anna University, Chennai, India in 2012; and her M.E. degree in Computer Science & Engineering from SNS College of Engineering, Coimbatore, India, in affiliation with Anna University, Chennai, India in 2014. She is pursuing her Ph.D. degree in the Department of Computers Science &Engineering, Sri SatyaSai University of Technology of Medical Sciences. She has published six International Journals along with many International & National conferences. Her current research interests include Cyber Security, Cryptography and Network Security.

S. Shitharth received his B.Tech. degree in Information Technology from Kgisl Institute of Technology, Coimbatore,India, in affiliation with Anna University, Chennai, India in 2012; and his M.E. degree in Computer Science & Engineering from Thiagaraja College of Engineering, Madurai, India, in affiliation with Anna University, Chennai, India in 2014. He completed his Ph.D. degree in the Department of Computers Science & Engineering, Anna University. He is currently working as Assistant Professor in Vardhaman College of Engineering, Hyderabad. He has published more than 10 International Journals along with 12 International & National conferences. He has even published 3 patents in IPR. He is also an active member in IEEE Computer society and in 5 more professional bodies. His current research interests include Cyber Security, Critical Infrastructure & Systems, Network Security & Ethical Hacking. He is an active researcher, reviewer and editor for many international journals.

Gouse Baig Mohammad completed (Ph. D-CSE) Acharya Nagarjuna University in February 2020 and M.Tech-CSE from Jawaharlal Nehru Technological University, Hyderabad in 2010. He has attended many workshops and published more than 10 International Journals.His current research areas are IoT and Machine learning.