International Journal of Future Generation Communication and Networking
Vol. 13, No. 1, (2020), pp. 85-98

Detection of Intrusion using Enhanced Machine Learning Model in SCADA Wireless Network

R. B. Benisha[1], Dr. S. Raja Ratna[2]
Full-Time Research Scholar[1], Associate Professor[2]
Department of computer science and Engineering,
V V College of Engineering, Tisaiyanvilai, India
beni.rb53@gmail.com[1], gracelinrr@yahoo.com[2]

Abstract: *The remote communication and its control purposes are highly integrated and the control through the wireless network is monitored by Supervisory Control and Data Acquisition (SCADA) systems. The attacks are isolated based on the optimal feature selection that is extracted from the sensor data. The cluster between the matrix and the optimal features are extracted and labeled. During clustering, the initial processing of attacks is removed and it is performed by the Mean shift clustering algorithm. The irrelevant features from the clustered data are concealed using the Intrusion detection system based Enhanced Cuckoo Search optimization algorithm (IDS-ECSO) and it is used to select the best features. The relevancy vector is used to classify the attacks and it is performed using Genetic Machine Learning based Neural Network (GML-NN). The classification results of the SCADA data set with the proposed method are analyzed and compared with the other methods.*

Keywords: *Supervisory Control and Data Acquisition; Optimization; Intrusion; Clustering*

1.Introduction

Supervisory Control and Data Acquisition (SCADA) systems should be monitored to function correctly. In earlier days SCADA and the network are independent and isolated; so that controlling actions are performed with the computer systems. In recent days when large amount of data are used, the SCADA transmission is done through the network path. So security is very less during transmission and an additional system should be used to control the databases. The intruder aims to target the industrial control systems and degrade their growth of production. Hieb [1] described the security system called the intrusion detection system. Figure.1 shows the detection of intruders in the SCADA network.

Intrusion detection system monitors the data flow by detecting the attacks happening in the network. IDS is placed before and after firewall to monitor the data flow and it is deployed with a alarm to indicate whether any attacks occur. In this paper IDS with enhanced model is used in the SCADA network to protect the database. Yang et.al implemented a state space factor and fuzzy evaluation method to predict the cautions in the network of SCADA [2]. Zhang et.al implemented the cyber system for SCADA to determine the vulnerabilities in distribution of power. The CB game theory is used for evaluating the functionality of the resources [3]. The probability of successive rate is

calculated for 24 sub stations in distributing power [4]. The security is improved in the SCADA network using enhanced model of data clustering. To identify the attack from the SCADA data set, this model is used. The clustered data are separated from the attack features.

The paper is organized as follows; Section 2 describes the existing works; Section.3 describes the proposed method; Section.4 describes the performance analysis and Section 5 concludes paper.

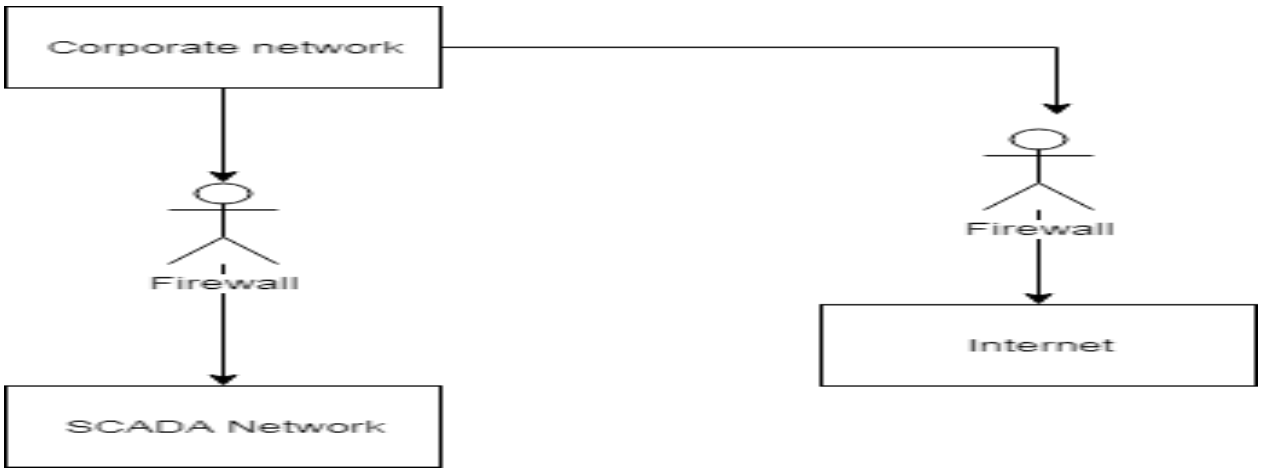

Fig.1 Detection of intruders in SCADA Network

2.Existing work

This section predicts the existing works of clustering and classification of intrusion detection system. Three forms of attack prediction are obtained 1) Network classification 2) Optimization of optimal features 3) Isolation.

In the initial process the attacks are logically checked whether any wrong data are present in the network [5]. Zhang et.al [6] analyzed the power distribution with the IDS by using Markov process. Also information pattern is used to identify different types of attacks in the SCADA [7]. Cherifi and Hamami implemented a data spread analysis to incorporate a prevention layer between physical and data link layer. The performance of the intrusion detection system is improved by the protocol layers [8]. The Denial of Service (DoS) attack is detected using GOOSE algorithm and the features are extracted using Message specification protocol [9]. The MMS protocol fails to encrypt the message through the network; so the enhanced machine learning technique is used for the isolation of DoS attack. In the field of power applications database should be protected from the intruders from modifying the data. Chen et.al [10] developed a prevention model for identifying the defensive attack. The Markov decision process is used to calculate the varied time instant parameter. It can be done through three forms clustering, optimization and classification.

In the clustering step, preprocessing is the major work to detect the different types of attack from the input data base. In the wind turbine power system deap learning based neural network is implemented to classify the different types of attacks [11]. The optimal

features are determined using fuzzy logic with different isolation model. To enhance the prediction process Koopan analysis is used with eigen vectors [12]. The input database is arranged in a clustered pattern by calculating the features and attacks with the eigen vectors [13]. Also in wind system, correlation is used for analyzing the operations and separated by the K-means clustering algorithm [14]. The SCADA is checked each and every instant to obtain the correlation matrix. The automated IDS is used in the network to report the different types of attacks from one system to another system during communication. The sensor location details of the attack are clearly identified by the IDS [14]. Alternatively cyber system is implemented in wireless sensor network for the protection by a digital framework [15]. In the recent years IoT technology is well developed in the SCADA network that monitors the performance of the control systems. The SVM is the major technique used in the SCADA to control and monitor the machine to identify the attacks from the database. Since large amount of data are transmitted through the network big data processing is used to manage and determine the malicious activities in the network [16]. The NN technique [17] is used to reduce the processing time during the classification.

To perform the improved efficiency and accuracy of the performance, this paper proposes an enhanced

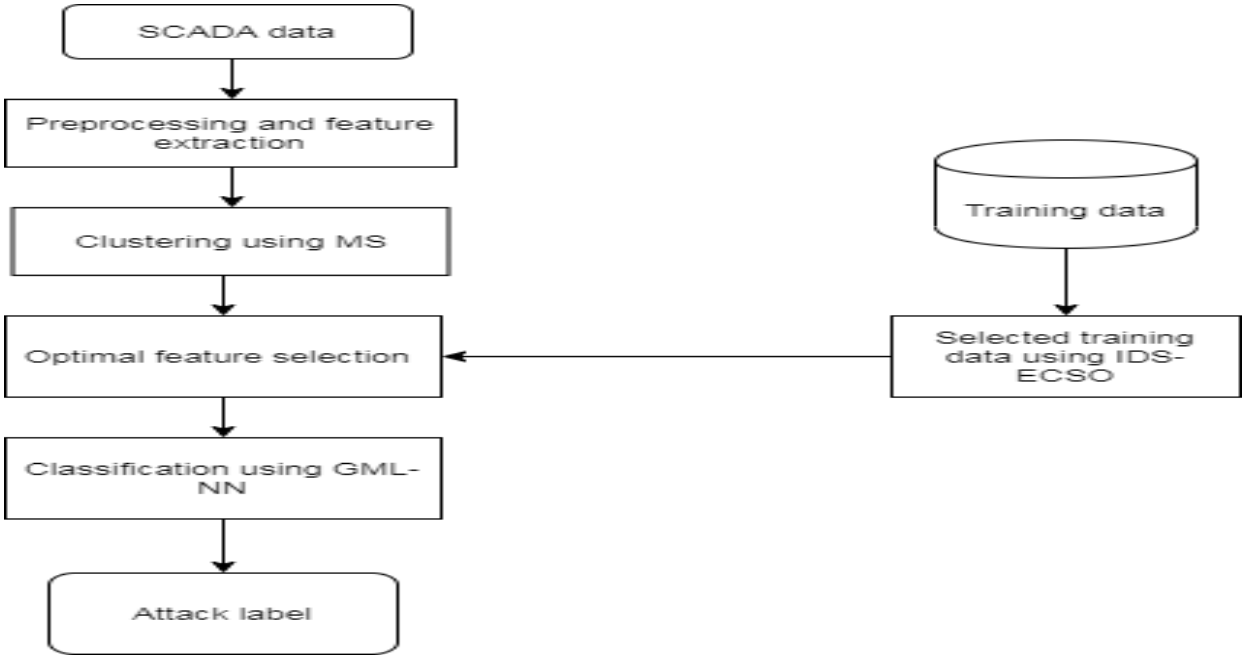

Figure.2 Block diagram of the proposed architecture

model for clustering and IDS based enhanced cuckoo search optimization for feature selection and for classification Genetic Machine learning based Neural Network is used. The major objective of the proposed work is given below;

- Grouping the data base in to a organized form by means of Mean shifting clustering.
- To select optimal features, IDS based enhanced cuckoo search optimization is used and the supervised classification is obtained by GML-NN algorithm.

- The attacks are predicted and labeled with its alarm entry which gives alert to the SCADA system.

The description of the proposed work is detailed in the following sections with the comparative analysis.

3. Proposed Scheme:

In this section clustering and selection of optimal features and classification in the SCADA network are described. The major intention of this work is to enhance the attack detection by selecting best features. In order to enhance the process data are clustered based on their mean distance between the features. The mean shifting algorithm is used to label feature data. The overall data set is applied with the enhanced learning process to predict the different types of attack. Only the classifier cannot predict the different types of attacks, optimization process should be used followed by the clustered data. To select optimal features IDS-ECSO is used to predict the best location. Finally GML-NN algorithm is used to classify the attacks from overall features. Figure.2 shows the overall proposed system architecture.

3.1 Preprocessing of data and extraction of features:

The preprocessing of data in the SCADA network contains several sensor parameters to monitor the functionalities. In this work network traffic determines the SCADA data matrix. The first step used is clustering and it is applied using mean shift clustering algorithm. This algorithm extracts the needed data and labels it to predict the feature that is needed for segmentation. An enhanced model is applied in the mean shifting algorithm to improve the relevant data. The dimensional space and the density using MS algorithm; so that clusters are labeled and segmented. The function $F(N)$ is denoted for n number of data points as;

Algorithm.1 clustering using MS algorithm

Input: Dataset from the SCADA Z
Output: Feature data label $F(N)$
Step.1 Initialize $Y_i = \{y_1, y_2, ...y_n\}$
Step.2 Set $F(1) = 0$; $\nabla F^{'}(N) = 1$
Step.3 if $j = \{1...n\}$;
$d = \{y - y_i\}$;// *distance estimation*
Step.4 Apply MS model
Step.5 If $\nabla F(N) < \nabla^{'} F(N)$;//*gradient calculation*
$F(N) = Y_j$
Step.6 End process

$$F(N) = \frac{1}{nR^d} \sum_{j=1}^{n} L(\frac{y - y_j}{2}) \quad (1)$$

Where R represents radius and L denotes enhanced model function. The enhanced model using the mean shifting algorithm is represented as;

$$L(y) = \exp(\frac{-(y - y_j)^3}{2N} \quad (2)$$

Where N is the normalization of the data point y . The gradient of density is represented as;

$$\nabla F(N) = \frac{2N}{nR^{2d}} \sum_{j=1}^{n} (y - y_j) g(\frac{y - y_j}{2})^2 \quad (3)$$

Where $g(N) = -L^{'}(Y)$ (4)

The exponential form of the enhanced model is increased while comparing with the other models of distance estimation method. It is explained in the algorithm 1.

The clustered data are arranged and segmented to indicate the network traffic. From the overall dataset the data points of traffic are indicated. The extracted features are classified based on the optimal feature selection.

3.2 Optimization of features:

The extracted feature data cannot classify and label the attacks accurately since the attacks types are same as the normal data and thus it leads to misclassification. In order to overcome this issue an IDS based Enhanced CSO algorithm is used to select the optimal features needed for classification. Many research works uses optimization techniques but it is only designed for a fixed data set. In the proposed scheme, optimization deals with the evaluation of separate data that are highly relevant to each class. The best location of the features is estimated by using the enhanced model. The cuckoo lays the eggs at different places but the best location is predicted by this algorithm. The propagation distance is evaluated by;

$$D_P = d_P \times rand(2) \times l_1 + d_{2P} \times rand(2) \times l_2 \quad (5)$$

Where l_1 and l_2 determines the searching coefficients. Using this optimization the process involved is the selection and spreading behavior. The spreading stage by its position is given as;

$$P_P = rand(2) \times d_i \times 2 + d_i \quad (6)$$

While the positions are updated the distance of standard deviation and its weights are identified as;

$$D_P^r = \sqrt{\frac{\sum (P_P - P_P^{'})^2}{2N+2}} \quad (7)$$

The updated positions are calculated by;

$$P_{P,m}^{'} = D_{P,m} + P_p \quad (8)$$

Where $D_{P,m}$ represents the random values obtained by the Gaussian distribution and the best features are selected using the probability function as;

$$P = \sqrt{\frac{F(P_{P,m}^{'})}{N \times F_{\max}}} \times M_x^P \quad (9)$$

Algorithm.2 Optimal feature selection using IDS-ECSO

Input: Selected data features $F(N)$
Output: Best attributes F_s
Step.1 Initialize cuckoo particles
$d_i = rand(size(F(N)))$ *; // positions updated*
$d_P = d_i \times 2 - d$ *; //position of cuckoo particles*
Step.2 If $j = \{1,.....n\}$ *;*
$M * N$ *; for* $i = 1$ *;// crossover*
Step.3 Fitness calculation
If $P < rand$ *; //probability estimation*
$F_s(N) = probability$
Step.4 if $j+1$ *; //mutation*
$d_P = d_{P-1}$ *; //replacement*
Step.5 End process

Where M_x^P represents the maximum values used for the selection process. Algorithm.2 explains the steps involved in selecting optimal features using IDS-ECSO. The best selected features are selected and given to the classifier that generates automatic label of attacks in which the malicious types of data are isolated.

3.3 Classification process:

In this process, both the testing, training data and the selected features are given to the classifier which uses GML-NN algorithm. Using this algorithm each node is interconnected by the training feature parameters. This forms the composite form of multi class pattern of data points. In the process of training, weighted node values are generated randomly as W_P. The automation rule factor is used for calculating the weight values and the probability of the nodes. It is given by the following equation as below;

$$AF_C = \sum_n P(n/H_1)L(H_1 - n) + \sum_n P(n/H_0)L(H_0 - n) \quad (10)$$

The change in the automation factor can be given as;

$$\Delta AF_C = (1 - 2H_i)(W_P + \sum_{i \neq j}(H_j W_P) \quad (11)$$

Where n indicates the training feature samples. The relationship between the nodes are obtained by

$$(1 - 2H_i) = \begin{Bmatrix} 1; falseH_i \\ 0; trueH_i \end{Bmatrix} \quad (12)$$

Algorithm.3 Classification using GML-NN

Input: Best attributes F_S
Output: Attack label A_L
Step.1 Initialize the weight of testing
Set T *; //parameter*
Step.2 If $j \leq N$ *then*
Set $R = rand(j)$ *; random value generated*
Step.3 Rule change using AF_C
If $R < AF_C$ *; then*
Step.4 $A_L = 1$*; // probability calculation*
else $A_L = 0$
$T = 0.98 * T$ *; label*
Step.5 End process

The probability that changes the state H_i is given by;

$$AF_C(F) = \frac{1}{1 - \exp(\Delta AF(i))F} \quad (13)$$

Where F indicates the parameters used for controlling the rule updation. Algorithm 3 describes the steps involved in testing process malicious data are classified from the normal data.

4. Performance Evaluation:

In the performance analysis, a data sets such as traffic data set is used for implementing with the proposed scheme. The data set contains traffic attacks of water storage system. Table.1 describes the different types of attacks and the percentage of number of samples.

Table.1 Attacks and samples for testing dataset

Attack	Sample (%)
Normal	55
Integrity	4
DoS	18
Bias Injection	1
Replay	10
Sinkhole	2
Reconnaissance	2
Shealthy	8

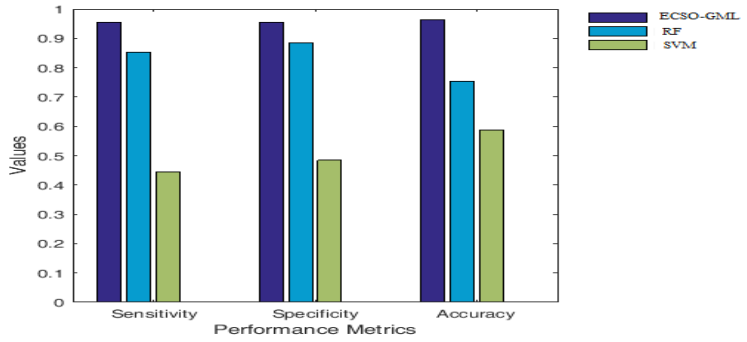

Figure.3 Overall performance metrics of the parameters Sensitivity, Specificity and accuracy with the existing methods

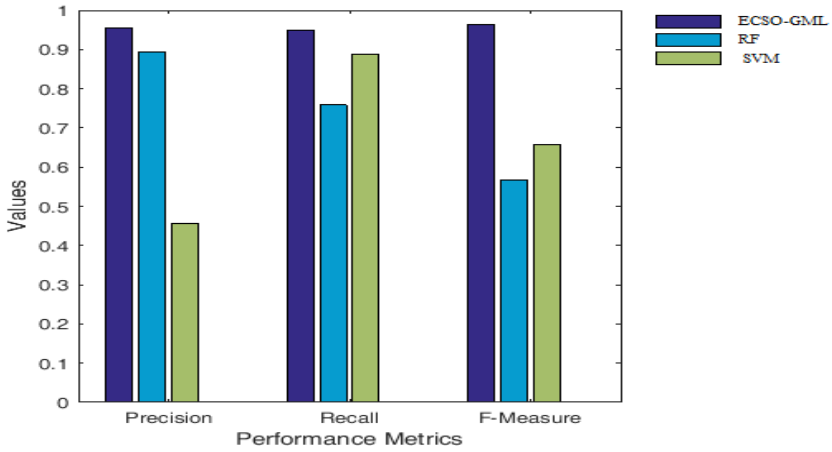

Figure.4 Overall performance metrics of the parameters Precision, Recall and F-measure with the existing methods

In this analysis 55% normal data and 45% of attacked data is present. To analyze the performance 60-90% of data samples are used and they are compared with the existing models. The proposed work is implemented in the MATLAB tool. The parameters used for the evaluation of performance are sensitivity, specificity, precision, recall and F-measure. The sensitivity and specificity is defined as the ratio of true positive to the true positive and false measure. The sensitivity can also be called as recall. FPR is calculated by the probability of false alarm rate. Precision refers to the fraction between actual positive and negative values. It defines that the data retrieved by the relevant number of

samples. False rate is defined as the average of precision and recall to determine the performance that denotes the accuracy. Figure.3 and 4 describes the overall performance metrics of the parameters with the existing methods.

4.1 Output results:

The proposed work performance is analyzed with these parameters and it is represented in graphs and tables. The proposed ECSO-GML scheme is compared with the existing algorithms of SVM, and RF.. An experimental result shows that the accuracy in the data set attains 96%. The comparison with the different methods using dataset is described in table 2. The true positive rate increases accuracy with the reduction in the false positive rate. This reduction in the FPR resembles less error rate with the proposed method. This increases the precision value that is separated for each class. Table 3 describes the false rate values of the proposed method and the existing method. The proposed method obtains 1.2% for the single class classification and 2.1% for multiclass. With the proposed method the average value of the precision is 98.8% while dealing with eight different types of attacks.

Table.2 Detection accuracy for the dataset

Models	Accuracy (%)
SVM	91%
HNA-NN	92.7%
NB	92.9%
Decision tree	93.4%
CSO	94.7%
GA	94.9%
DBN	95.7%
ECSO-GML	98.7%

Table.3 F-measure for the dataset

Model	F-measure	
	Single class	Multi class
SVM	90.7	87
NB-CSO	63.2	56.3
DBN-SVM	58.2	42.3
RF	97.5	97.2
RF-CSO	97.3	95.5
RF-SVM	98.3	97.6
ECSO-GML	99.8	99.56

Table.4 Comparing precision values of RF and ECSO-GM for dataset

Attacks	RF	ECSO-GML
A1	96.3	99.6
A2	94.5	98.5
A3	95.3	98.6
A4	96.8	99.4
A5	97.4	99.2
A6	96.2	98.5

A7	95.3	99.3
A8	95.3	99.5

Table.5 Comparing F-measure values of RF and ECSO-GM for dataset

Attacks	RF	ECSO-GML
A1	98.3	99.4
A2	97.5	98.6
A3	95.4	99.6
A4	94.8	99.5
A5	95.4	98.2
A6	96.6	98.6
A7	96.3	99.4
A8	95.5	99.5

Recall is calculated to determine the relevancy classified feature set. With the eight different types of attack the proposed method forms the increased range of 90-10% in dataset 2. The F-measure is compared with the existing methods which obtains 1.6% greater output than the RF classifier. Table.4 and 5 describes the comparison of precision and recall values of RF and ECSO-GM for dataset. Table.7 describes the overall comparison parameters of RF and ECSO-GML.

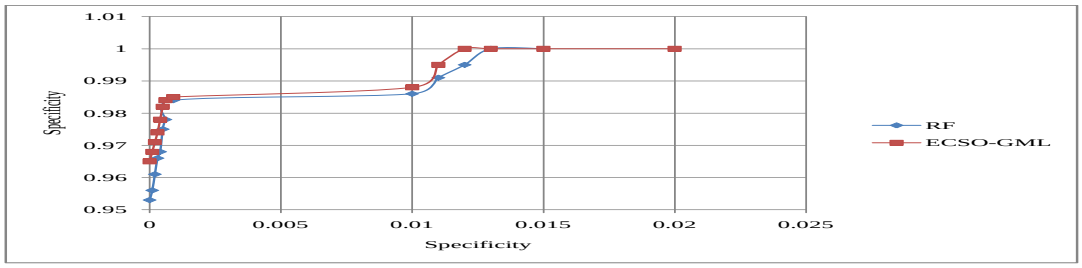

Figure.5 ROC curve of RF and ECSO-GML for the overall dataset

Table.7 Comparison of dataset with the parameters

Parameters	RF	ECSO-GML
Precision	98.75	99.75
Recall	98.75	99.75
Detection accuracy	98.3	99.5
F-measure	98.3	99.23
Sensitivity	98.6	99.64
Specificity	98.4	99.87

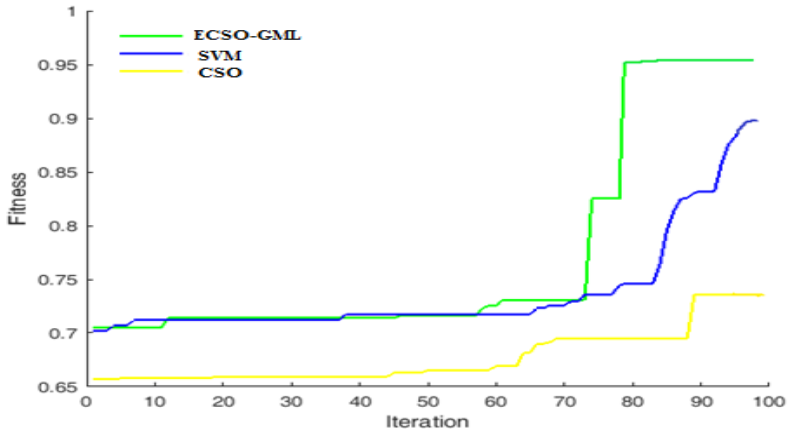

Figure.6 Fitness value under different iterations

Table.4 Overall comparison parameters of RF and ECSO-GML

PARAMETER	RF	ECSO-GML
TP	6826	6835
TN	46754	48579
FP	40	36
FN	40	36
SENSITIVITY	0.951	0.986
SPECIFICITY	0.946	0.985
ACCURACY	0.946	0.976
PRECISION	0.955	00.981
RECALL	0.86	0.9758
F-MEASURE	0.0036	0.0018

The relevancy between the normal and the attack label are represented from the classified output which performs better performance rate in multi labeled feature with low level of training features. Table 4 describes the average values of the parameters of the proposed scheme. Hence overall accuracy of the proposed method obtains 0.5% higher than the existing methods. The proposed method achieves 1.8%greater than the existing method in overall output. The ROC curve determines the classification performance with different thresholds as shown in figure.5. Figure.6 shows the fitness value under different iterations. While comparing the datasets the proposed method obtains better performance.

5. Conclusion:

In this paper, a novel proposed method is used for the detection and classification of malicious data in the SCADA network. This is done by clustering and IDS based ECSO algorithm for the selection of optimal features and GML-NN algorithm for the classification. To identify the attack from the predicted data certain conditions are evaluated. The conditions refers to that the values are limited for testing vector to match the prediction. Clustering is done to select all the features. This process selects the relevant features that are needed for selecting best features. The IDS based ECSO

algorithm is used to select the best optimal features needed for classification. The accuracy with the novel method is improved with the minimum duration of time. The combination of optimized features is given to the classifier. In the classification process GML-NN method is used to make the training classified output better than the existing method. Two different datasets are used to compare the performance with the existing methods. The performance analysis achieves better clustering, optimization and classification output than the traditional algorithms.

Future work:

The proposed IDS based ECSO and GML-NN algorithm is developed with the supervised feature pattern and also with the parallel processing method to analyze bulky data with different features.

Availability of data and materials:

"Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Competing interests:

R. B. Benisha and Dr. S. Raja Ratna declares that they have no Competing interests.

Animal Rights:
This article does not contain any studies with animal subjects performed by the any of the authors.

Funding:
The data needed for the research work is taken from Gamesa Wind turbines. Since in Wind Turbines traffic based attacks and time delay based interruption attacks cause huge damage which leads to the growth in production, Real time data stored in SCADA are taken and analyzed.

Consent for publication:
The article is original, has not already been published in a journal, and is not currently under consideration by another journal

6. References

[1] Hieb, J.: 'Security hardened remote terminal units for SCADA networks',2008.
[2] Yang, L., Cao, X., Li, J.: 'A new cyber security risk evaluation method for oil and gas SCADA based on factor state space', *J. Solitons Fractals*, 2016, **89**,pp. 203–209.
[3] Zhang, Y., Wang, L., Xiang, Y., *et al.*: 'Inclusion of SCADA cyber vulnerability in power system reliability assessment considering optimal resources allocation', *IEEE Trans. Power Syst.*, 2016, **31**, (6), pp. 4379–4394.
[4] Almalawi, A., Fahad, A., Tari, Z., *et al.*: 'An efficient data-driven clustering technique to detect attacks in SCADA systems', *IEEE Trans. Inf. Forensics Sec.*, 2016, **11**, (5), pp. 893–906.

[5] Li, W., Xie, L., Deng, Z., *et al.*: 'False sequential logic attack on SCADA system and its physical impact analysis', *Comput. Secur.*, 2016, **58**, pp. 149–159.
[6] Zhang, Y., Wang, L., Xiang, Y.: 'Power system reliability analysis with intrusion tolerance in SCADA systems', *IEEE Trans. Smart Grid*, 2016, **7**,(2), pp. 669–683.
[7] Finogeev, A.G., Finogeev, A.A.: 'Information attacks and security in wireless sensor networks of industrial SCADA systems', *J. Ind. Inf. Integr.*, 2017, **5**, pp. 6–16.
[8] Cherifi, T., Hamami, L.: 'A practical implementation of unconditional security for the IEC 60780-5-101 SCADA protocol', *Int. J. Crit. Infrastruct. Prot.*, 2018, **20**, pp. 68–84.
[9] Lahza, H., Radke, K., Foo, E.: 'Applying domain-specific knowledge to construct features for detecting distributed denial-of-service attacks on the GOOSE and MMS protocols', *Int. J. Crit. Infrastruct. Prot.*, 2018, **20**, pp. 48–67.
[10] Chen, Y., Hong, J., Liu, C.C..: 'Modeling of intrusion and defense for assessment of cyber security at power substations', *IEEE Trans. Smart Grid*, 2018, **9**, (4), pp. 2541–2552.
[11] Sun, P., Li, J., Wang, C., *et al.*: 'A generalized model for wind turbine anomaly identification based on SCADA data', *Appl. Energy*, 2016, **168**, pp.550–567.
[12] Raak, F., Susuki, Y., Hikihara, T.: 'Data-driven partitioning of power networks via Koopman mode analysis', *IEEE Trans. Power Syst.*, 2016, **31**,(4), pp. 2799–2808.
[13] Fang, R., Shang, R., Wu, M., *et al.*: 'Application of gray relational analysis to *k*-means clustering for dynamic equivalent modeling of wind farm', *Int. J. Hydrog. Energy*, 2017, **42**, (31), pp. 20154–20163.
[14] Khan, R., Khan, S.U.: 'Design and implementation of an automated network monitoring and reporting back system', *J. Ind. Inf. Integr.*, 2018, **9**, pp. 24–34.
[15] Elhoseny, M., Hosny, A., Hassanien, A.E., *et al.*: 'Secure automated forensic investigation for sustainable critical infrastructures compliant with green computing requirements', *IEEE Trans. Sustain. Comput.*, 2017.
[16] Parwez, M.S., Rawat, D.B., Garuba, M.: 'Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network', *IEEE Trans.Ind. Inf.*, 2017, **13**, (4), pp. 2058–2065.
[17] Raman, M.G., Somu, N., Kirthivasan, K., *et al.*: 'A hypergraph and arithmetic residue-based probabilistic neural network for classification in intrusion detection systems', *Neural Netw.*, 2017, **92**, pp. 89–97.

Author Details:

R.B.Benisha, (Corresponding author) completed B.E degree in Electronics and communication Engineering in Vins Christian college of Engineering in the year 2011. She completed her M.E degree in Applied Electronics in Velammal Engineering college Chennai in the year 2013. Currently she is pursuing her Ph.D (full-time) under Anna University, Chennai from 2018. The research topic is "Network security in SCADA".


Dr. S. Raja Ratna completed her B.E degree in Electrical and Electronics Engineering from The Indian Engineering College, Tirunelveli in 2000, and the M. Tech degree in Computer and Information Technology from Manonmanium Sundaranar University, Tirunelveli in 2005. She completes her Ph. D degree at the Information and Communication Engineering at Anna University, Chennai in the year 2015. Currently she is working in V V college of Engineering, Tisaiyanvilai.

Author's Contribution:

R.B.Benisha has published four International conference papers and one International journal paper. Dr.S.Raja Ratna has published Four annexure I journals and five annexure II journals. Her research interests include denial-of-service attacks, jamming attacks, secure routing algorithm and security in networks.