Article

JOURNAL OF UNIVERSITY OF BABYLON

For Pure and Applied Sciences (JUBPAS)

Vol.32; No.4.| 2024

Strengthening SCADA System Security through a Novel Intrusion Detection Method Using artificial intelligence Algorithm

Asmaa Dheyaa Saed Mashtah[1*]

College of Engineering and Engineering Techniques, Al-Mustaqbal University, 51001, Babylon, Iraq

*Corresponding author: E-mail: asmaadheyaa1996@gmail.com

Accepted: 8/11/2024 **Published:** 31/12/2024

ABSTRACT

Background Intrusion detection in SCADA systems is essential for ensuring operational security and preventing unauthorized access. Traditional methods often face challenges related to high dimensionality and inefficiencies in accurately identifying threats.

Materials& Methods: This paper introduces a novel methodology that optimizes intrusion detection performance through a series of systematic steps, beginning with data preprocessing to eliminate redundant features, reducing the initial dataset from 35 to 17 features while maintaining critical information integrity. The approach utilizes Principal Component Analysis (PCA) to reduce dimensionality, transforming variables that are related into unrelated main components and retaining 99% of the original data variance. For classification, a Radial Basis Network (RBN) is employed, with parameters such as spread value and the number of hidden layer neurons carefully selected to enhance model performance. To avoid the problem of vanishing gradient, Emperor Penguin Optimization is applied to the training process. It should be noted that the population size and maximum iterations are carefully adjusted for an optimized training procedure. The obtained performance is 99.22% for accuracy, 99.31% for precision, 99.09% for recall, and 99.20% for F1 score.

Conclusion: These results confirm the model's appropriate performance for the detection of intrusions while avoiding any false alarms and validate it as a very efficient solution for strengthening security on SCADA systems.

Keywords: SCADA system security; Intrusion detection (ID); Radial Basis Network (RBN); Emperor Penguin Optimization (EPO); Principal Component Analysis (PCA).

ISSN: 2312-8135 | Print ISSN: 1992-0652 | www.journalofbabylon.com | jub@itnet.uobabylon.edu.iq | info@journalofbabylon.com

Article
JOURNAL OF UNIVERSITY OF BABYLON
For Pure and Applied Sciences (JUBPAS)
Vol.32; No.4.| 2024

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

INTRODUCTION

SCADA has become one of the key elements of any modern industrial system, serving to monitor and regulate critical infrastructures such as power supply grids, water supply networks, and transportation systems. While digital communications technologies are increasingly enabled to allow connectivity between systems, so too do they increasingly expose those same systems to an ever-growing threat of cyber-attack [1]. Due to the critical nature of SCADA systems—under incessant, different kinds of hostile attacks—actually surviving without the elaboration of robust intrusion detection mechanisms is impossible. Traditional methods cannot cope with the increasing complexity and volume of cyber-attacks from SCADA systems, and it is urgent that new, more advanced effective methods be developed [2].

It remains challenging how to properly identify malicious activities in SCADA networks, which are usually masked in a sea of normal operation data. These challenges are further complicated by the fact that SCADA systems also must be designed for minimal downtime and a high degree of accuracy, ruling out most conventional security solutions that could introduce performance bottlenecks or false alarms. Therefore, there is a dire need to develop some innovative methodologies for intrusion detection that ensures high accuracy with very low false positive rates and process the large-sized data with a high speed efficiently [3]. Addressing this problem requires a combination of advanced machine learning techniques, effective feature reduction, and optimization methods tailored to the unique characteristics of SCADA networks.

In introduces a deep-learning-based omni-intrusion detection system (IDS) for SCADA networks, focusing on detecting both temporally uncorrelated and correlated attacks. This method provides a comprehensive security solution that can handle various types of cyber threats. reference proposes a machine learning approach specifically for detecting and classifying attack threats in cyber-physical systems, which also includes SCADA systems. The study emphasizes the need for a robust security structure in SCADA systems, which are often highly vulnerable to cyber-attacks [4]. Paper presents a novel method that integrates machine learning classifiers like Random Forest and k-nearest neighbors (KNN) to detect intrusions in SCADA systems that are connected with IoT devices. It analyzes how well these classifiers can identify anomalies in network traffic [5]. focuses on supervised learning methods to identify and classifying intrusions in SCADA systems. The study also provides a thorough review of existing methodologies, available datasets, and optimization strategies to secure critical infrastructures (CIs) [6]. Reference introduces a novel approach that uses deep reinforcement learning (DRL) for anomaly detection in SCADA networks. This method addresses the limitations of previous deep learning and reinforcement learning approaches, specifically in the domain of industrial control systems (ICS) [7]. highlights a machine learning-based method for intrusion detection in ICS and SCADA networks. It incorporates Defense-In-Depth architecture and energy consumption forecasting models, aiming to enhance security and ensure data availability [8]. presents a machine learning-based approach for detecting malicious traffic in IPCS-SCADA networks. It uses feature selection methods and SVMSMOTE to enhance intrusion detection

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652 | www.journalofbabylon.com | jub@itnet.uobabylon.edu.iq | info@journalofbabylon.com

performance. proposes a stacking ensemble model for SCADA intrusion detection. The model combines Random Forest, LightGBM, and XGBoost, with a Multilayer Perceptron (MLP) network as the meta-classifier, improving detection accuracy [9]. introduces an Autoencoder-based IDS (AE-IDS) for SCADA networks, focusing on the DNP3 communication protocol. The unsupervised AE-IDS model outperforms other supervised learning algorithms [10]. analyzes the performance of deep learning architectures (CNN, autoencoders, DBN, and LSTM) for intrusion detection based on anomalies in SCADA devices, with a focus on enhancing cybersecurity [11]. presents an integrated framework for smart grid intrusion detection. This framework combines feature engineering with machine learning classifiers, using Gradient Boosting Feature Selection (GBFS) to optimize detection rates and execution speed [12]. proposes a few-shot learning method (FS-IDS) for identifying intrusions in SCADA systems, effectively detecting cyber-attacks with limited examples by improving data quality through advanced preprocessing techniques [13]. introduces an interconnected structure for intrusion detection in SCADA-based power grids, utilizing RFE-XGBoost for feature selection and a majority-vote ensemble of nine classifiers to improve detection performance [14]. focuses on the Modbus protocol, employing machine learning algorithms like SVMs, decision trees, and k-nearest neighbors for detecting security threats in SCADA systems [15]. presents a hybrid method to tackle class imbalance in SCADA IDS, using SMOTE_SVM and CSL, achieving improved precision and recall metrics [16]. uses autoencoders for anomaly detection in SCADA networks, focusing on reconstruction loss in unsupervised learning to detect cyber threats, particularly in a gas pipeline dataset [17]. proposes an innovative approach using Artificial Intelligence (AI) to develop an intrusion detection system (IDS) for SCADA systems. The study evaluates various deep learning and machine learning algorithms to identify the most effective method for detecting cyber-attacks. The research findings indicate that the Bi-LSTM (Bidirectional Long Short-Term Memory) IDS technique demonstrates superior performance compared to previous approaches in securing SCADA systems against intrusions [18]. introduces a three-stage classifier model for SCADA intrusion detection, classifying attacks into seven classes and 35 sub-attacks, significantly enhancing detection accuracy [19]. proposes a method to detect Denial of Service (DoS) attacks in interconnected SCADA systems, showing that the Decision Tree algorithm achieves 99.99% detection accuracy [20]. develops a deep learning-based IDS for SCADA systems focusing on the DNP3 communication protocol, addressing critical data injection and modification attacks [21]. introduces a fused feature selection approach in a three-phase IDS framework for SCADA networks, improving computational efficiency and performance over existing methods [22]. applies supervised learning algorithms for SCADA intrusion detection, showing that the SMO-SVM model outperforms other classifiers, particularly with data preprocessed via principal component analysis (PCA) [23]. uses a Random Forest (RF) algorithm with Adaptive Boosting (AdaBoost) to enhance the detection rate of network intrusion in SCADA systems, addressing the vulnerabilities posed by interconnectivity with corporate networks [24]. presents an organized review of supervised learning approaches to SCADA intrusion detection, analyzing existing methodologies, datasets, and recommendations for improving security [25]. reviews the challenges faced by IDS in response to growing network sizes and novel attack vectors. It

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

emphasizes the potential of machine learning and deep learning to enhance rate of accuracy detection and reduce false alarms.

Although there are some developments in research for intrusion detection systems, there are a few research gaps regarding the security of SCADA systems. Most of the existing approaches depend on conventional machine learning algorithms, which may fail to detect the complex patterns and dynamics in the operation data for such SCADA networks. In addition, high-dimensional data processing without losing significant information is one of the major challenges. While some studies focus on dimensionality reduction techniques, the ways in which such techniques may be integrated with advanced models of classifications are only scantily explored. Moreover, there is a lack of understanding in the case of optimization algorithms and how such algorithms can improve neural network-based training processes for intrusion detection systems, specifically for overcoming problems like the vanishing gradient problem in the case of Radial Basis Networks.
These research lacunas are targeted in this paper by proposing a novel methodology to optimizing intrusion detection in SCADA networks. In the proposed approach, extensive data preprocessing is combined with dimensionality reduction using the Principal Components Analysis method and classification by applying RBN, trained using Emperor Penguin Optimization-EPO. These elements, in this framework, interact for improving intrusion detection accuracy and efficiency in SCADA. Contributions of this research are:

- Developing a custom RBN architecture tailored for intrusion detection in SCADA systems, optimizing variables like as spread value and the number of hidden neurons to balance complexity and performance.
- Introducing the EPO algorithm to effectively train the RBN, addressing the vanishing gradient problem and improving the ability of the system to acquire complex patterns.
- Applying PCA to decrease the feature space while preserving 99% of the original variance, enhancing the efficiency of the classification process.

BASIC CONCEPTS

This section gives a general overview of the critical underlying principles and fundamental concepts that are required in the development of the proposed method.

A. *Radial Basis Networks:*

Radial basis networks (RBN) are a specific family of artificial neural networks which derives inspiration from the theory of function approximation. They are capable of learning complex patterns based on a radial basis function (RBF) — typically a Gaussian function. In RBNs, the idea is to calculate the Euclidean distance between the input vector and each neuron's center; then apply radial basis functions to adjust weights or activation level of each neuron for output. RBNs have a three-layer structure networks that include input layer, a hidden layer and an output layer as seen in Figure 1. The input layer is designed to accept one or more predictor variables, and each input predictor variable must be associated with an independent neuron. The input layer

ISSN: 2312-8135 | Print ISSN: 1992-0652 | www.journalofbabylon.com | jub@itnet.uobabylon.edu.iq | info@journalofbabylon.com

is present to pass this information vector into the hidden layer in which it undergoes transformation.

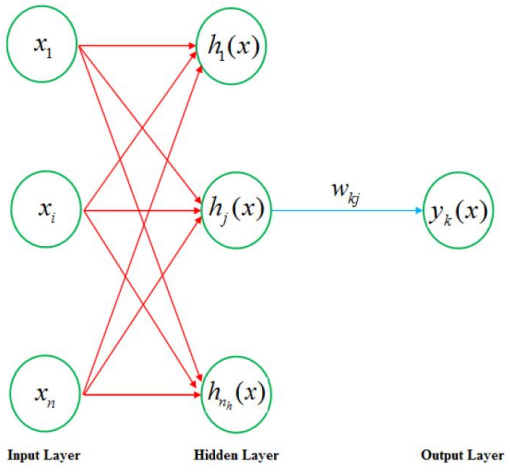

Figure 1. Architecture of a radial basis network [26]

The neurons in the hidden layer will compute the degree of influence of each input using radial basis functions. Each hidden neuron is related with a center c_j and a width σ_j (both parameters of the radial basis function). The result of a single neuron in the hidden layer is equal to the nonlinear transformation of the input vector x:

$$h_j(x) = \exp(-\frac{\|x-c_j\|^2}{\sigma_j^2}) \qquad (1)$$

In this equation, $h_j(x)$ represents the output of the *j-th* hidden neuron, x is the input vector, and c_j is the center of the radial basis function. The Euclidean distance $\|x - c_j\|^2$ measures how far the input is from the center of the radial basis function. If this distance is zero (i.e., the input is exactly at the center), the output is 1, indicating maximum influence. Conversely, as the distance increases, the output tends to 0, implying diminishing influence.

The output layer performs a linear combination of the hidden layer outputs. For each output neuron $y_k(x)$, the value is calculated by summing the weighted outputs of all hidden neurons:

$$y_k(x) = \sum_{j=1}^{n_h} w_{kj} h_j(x) + b_k \qquad (2)$$

Here, $y_k(x)$ is the output of the *k-th* neuron, w_{kj} is the weight connecting the *j-th* hidden neuron to the *k-th* output neuron, and b_k is a bias term added to adjust the final output. The weighted outputs from the hidden layer are summed and passed to the output layer, producing the network's final result [26], [27].

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

RBNs are particularly effective for problems where the relationship between input and output is nonlinear, making them suitable for tasks such as classification and regression. RBNs can handle complex data pattern effectively, due to a local sensitivity of input data for radial basis function.

B. *Emperor Penguins Optimization Algorithm*

The Emperor Penguin Optimization (EPO) algorithm discussed in this paper is modeled after the actions and movement patterns of emperor penguins in their colony. This, resulting from the key stages which are performed by the algorithm that repeat as penguins' spiral-like movements and heat distribution patterns [28]:

Initial Population of Emperor Penguins: The algorithm starts by generating a population of emperor penguins (initial solution) arbitrarily in the search space. These penguins form a cluster, or "huddle "which is the primary component of the optimization process.

Huddle Boundary Calculation: Emperor penguins are more likely to be located on the boundary of polygonal-shape formation when huddling. At some point, they choose two adjacent penguins, and the flow of wind around the huddle influences the boundary's shape, which resembles a polygon. Given that wind moves faster than the penguins, mixed-variable concepts are used to simulate this random boundary. If φ represents the wind speed and Ψ denotes its gradient according to the following Equation:

$$\Psi = \nabla \Phi \qquad (3)$$

The vector Ω, combined with Φ, forms a complex value:

$$F = \Phi + i\Omega \qquad (4)$$

The region has an L-shape, with the best cost value (C-value) generally located in the area where the penguins adjust their positions within the boundary of the huddle, using a jiggling mechanism with complex numbers i (where i is the imaginary constant).

Temperature Characteristics Calculation: The penguins aim to conserve energy by insulating the huddle and maintaining temperature. Mathematically, if the radius of the polygon R>1, the temperature T=0, and if R<1, T=1. These temperature traits help guide the penguins' movement and resource utilization. The temperature T′ is calculated by the following formula:

$$T' = \left(T - \frac{Max_{iteration}}{x - Max_{iteration}}\right)$$

$$T = \begin{cases} 0, & if\ R > 1 \\ 1, & if\ R < 1 \end{cases} \qquad (5)$$

Where x is the current iteration, $Max_{iteration}$ is maximum iteration allowed, R is the polygon radius and T is time required to search solution into search space.

ISSN: 2312-8135 | Print ISSN: 1992-0652 | www.journalofbabylon.com | jub@itnet.uobabylon.edu.iq | info@journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652 | www.journalofbabylon.com | jub@itnet.uobabylon.edu.iq | info@journalofbabylon.com

Emperor Penguins Distance Calculation: Once the huddle boundary is formed then the distance between each emperor and best-found solution will be calculated. The solution along with the best fitness value is said to be optimal. The other penguins recalculate their new location based on this solution by the following formula:

$$\vec{D}_{ep} = Abs\left(S\left(\vec{A}\right).\vec{P}(x) - \vec{C}.\vec{P}_{ep}(x)\right) \quad (6)$$

ere, $\vec{D}_{ep}$ is the distance between the penguins and the optimal solution, $\vec{A}$ and $\vec{C}$ act as parameters to avoid collision among the penguin, where $\vec{P}$ signifies the best solution (or the position of the optimal penguin). The function $S(\vec{A})$ represents the social force helping to move each pair of penguins towards the closest penguin in their collision group. The terms $\vec{A}$, $\vec{C}$ and $S(\vec{A})$ are given by the following relations:

$$\vec{A} = \left(M \times (T' + P_{grid}(Accuracy)) \times Rand()\right) - T' \quad (7)$$

$$P_{grid}(Accuracy) = Abs\left(\vec{P} - \vec{P}_{ep}\right) \quad (8)$$

$$\vec{C} = Rand() \quad (9)$$

$$S\left(\vec{A}\right) = \left(\sqrt{f.e^{-x/l} - e^{-x}}\right)^2 \quad (10)$$

Where M is a movement parameter set to 2 to prevent collisions, T' is the temperature characteristic, $P_{grid}(Accuracy)$ evaluates the accuracy of the polygonal grid, Rand () is a random function, and e is Euler's number. Parameters f and 1 used to control exploration and exploitation, with suggested values between [2-3] and [1.5-2], respectively.

Termination and Population Update: After each complete iteration of the algorithm, it is decided if the maximum number of iterations has been reached. If yes, the solution where penguin has highest fitness value is sort and this is our final answer. If not, the algorithm keeps adjusting positions of the emperor penguins until it reaches to optimal solution.

I.METHODOLOGY

The suggested approach is focused on optimizing the effectiveness of intrusion detection of SCADA systems whereby it continues through a series of steps starting with data preprocessing. Originally there were 35 features in the dataset, but removing 4 of those due to the redundant information results in theremaining 31 features. This dimensionality reduction happens after doing all the preprocessing steps which contain normalization, data cleaning

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

from NAN or NULL values, and converting the characters to numbers. After normalization, the duplicated features become unusable as they result in a fraction of 0/0 due to the absence of variability between their maximum and minimum values. These features are excluded to maintain the dataset's integrity.

Post-preprocessing, the data can be further subject to dimensionality reduction using Principal Component Analysis (PCA) and fetching themost important features. PCA is used to turn the correlated variables into uncorrelated ones (principal components) which results in less complexity while keeping information. The feature matrix is transformed into a lower-dimensional space by calculating the eigenvectors and eigenvalues of the covariance matrix, where eigenvectors represent the directions of greatest variance in the data. To retain a high degree of information, eigenvectors contributing to a cumulative sum of eigenvalues of at least 0.99 are selected. This ensures that 99% of the variance in the original data is preserved. As a result, we were able to reduce the dimensionality from 31 features to just 17 while preserving information that is crucial and important for further analysis.

A radial basis network (RBN) is chosen for classification. The RBN is defined by an input layer, several hidden layers of radial basis functions, and an output layer. The spread value and the number of neurons in the hidden layer are two crucial parameters in RBN architecture. The spread value affects the smoothness of the radial basis functions, helping the network captures the complexity of the intrusion detection task. In this paper, a spread value of 1 is chosen to reflect the variations in the data. The number of neurons in the hidden layer affects the network's ability to model complex patterns. To ensure the network has enough capacity to correctly solve this problem, 10 neurons are used, which are larger enough number of neurons while not causing over fitting. Figure 2 shows the designed RBN architecture. It should be noted that since the problem is binary classification, instead of using the SoftMax output layer, a linear one is chosen and the binary classification is done using a threshold equal to zero. This threshold is considered in the customized training procedure of the network to properly adapt the network's weights and biases.

To train the network, a novel training approach is replaced with the conventional back propagation approach to reduce the vanishing gradient problem which is a prevalent issue in RBN. The derivative of Gaussian RBFs with respect to their inputs can become very small, especially when the input is not in the vicinity of the center. This small derivative reduces the gradient signal passed back through the network, exacerbating the vanishing gradient issue. Moreover, the parameters defining the RBFs (centers and widths) are highly sensitive. Small gradients mean that these parameters may adjust very slowly, making it difficult for the network to fine-tune the RBFs to better fit the data. Hence, the Emperor Penguin Optimization (EPO) algorithm is used to train the designed RBN better. For this purpose, EPO parameters should be adjusted first. Two important parameters in the EPO algorithm, the highest possible number of iterations and size of the population were adjusted to improve training performance. The

maximum number of iterations is the length of time that the algorithm continues to run, and should be adjusted depending on a trade-off between getting a robust solution and spending computation time. To guarantee an optimal solution without excessive computation, 150 iterations were set as the maximum. The population size refers to the number of candidate solutions considered in every generation. A bigger population is better for the algorithm in the sense that the space of target solutions is explored more fully, but this obviously brings an extra computational load. In our work, a population size of 100 was chosen to strike an appropriate balance between exploration and computational efficiency. By setting the parameters of the EPO algorithm, the RBN can be trained using our new approach. In this regard, the learnable parameters (weights and biases) are considered decision variables, and the error (1-accuracy) of the network is considered as the cost function. By adopting this approach, the proposed method is completed for detection of intrusion in SCADA systems.

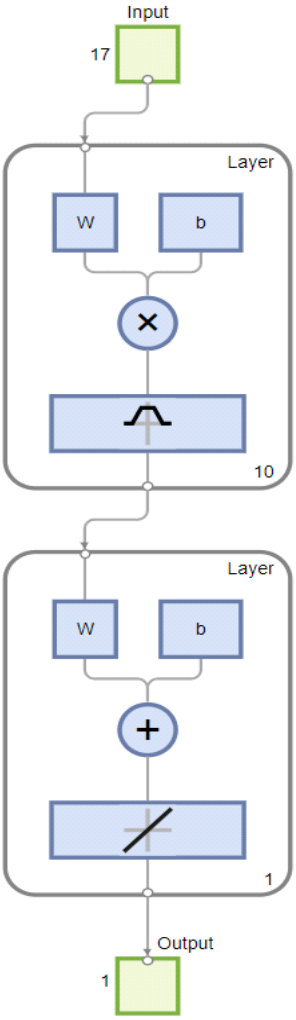

Figure 2. The structure of the designed RBN network [27]

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

II. DATASET

In this paper, three datasets are utilized for evaluating the proposed method: an industrial control system dataset focused on gas transmission, the NSL-KDD, and the UNSW-NB15 datasets. The industrial control system dataset, specifically collected for intrusion detection in SCADA systems, is used for the primary implementation and evaluation of the method. The NSL-KDD and UNSW-NB15 datasets are just included to assess the method's generalizability across different datasets and to enable result comparison.

A. Industrial control systems in gas transmission

It should be noticed that for the implementation of the proposed method, we use a dataset of an industrial control system used in a gas transmission setup. This dataset includes 31 instances of cyber-attacks and can efficiently be used for training and testing classifiers. Consequently, this dataset is made up of a total of 22,538 samples; each sample has 35 different features of which some are:

- Packet length of sent/received data.
- The connectivity address, if any, associated with the system.
- Control type and layout information.
- Duration of the relationship.
- Electrical valve specifications.
- Operational mode of pumps.
- Nature of compliance to directions.
- Pressure scales and measurements.
- Control system state settings.

These features aggregately contribute to the completeness of the dataset, which enables training and evaluation of the proposed approach in the context of SCADA systems.

B. NSL-KDD

The NSL-KDD dataset is an improved version of the KDD Cup 99 dataset, widely used for training and evaluating intrusion detection systems. It consists of 125,973 samples in the training data and 22,544 samples in the testing data. In this dataset, samples are labeled into four main categories of attacks: DoS (Denial of Service, disrupting network resources), R2L (Remote to Local, gaining remote access to the system), U2R (User to Root, escalating privileges from a normal account), and Probe (scanning the network for vulnerabilities). Each sample has 41 features that describe various aspects of network traffic and behaviors, such as protocol type, number of failed connections, and the number of packets sent. These features are either continuous or categorical, and each sample is labeled as either "normal" or "attack."

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

C. *UNSW-NB15*

The UNSW-NB15 dataset is a benchmark dataset developed to evaluate the performance of intrusion detection systems (IDS) with a more realistic representation of modern network traffic. Created by the Cyber Range Lab of the Australian Centre for Cyber Security in 2015, this dataset includes 100,000 network connection records for training and testing, representing both normal and malicious activities. UNSW-NB15 simulates various types of attacks, organized into nine categories: Fuzzers, Analysis, Backdoors, DoS (Denial of Service), Exploits, Generic, Reconnaissance, Shellcode**,** and Worms. Each record has 49 features covering a range of network characteristics, including flow information, basic TCP/IP header details, content, time, and additional statistical properties of the network traffic. The dataset includes labeling for both binary classification (normal or attack) and multi-class classification (attack type), making it suitable for various machine learning and deep learning approaches in intrusion detection research

III. EVALUATION METRICS

The primary concern on classification tasks, is the evaluation of model performance to guarantee that the predictions are reliable and accurate. The most popular evaluation measures are accuracy, precision, recall and F1 score. These metrics give insights into effectiveness of the classifier in various aspects, particularly for imbalanced datasets or multi-class problems.

1. Accuracy: This is one of the simplest metrics for evaluating classifiers. It simply calculates how many correct classifications were made among the all. It is a great marker of model performance over all, but it might not be very reflective for imbalanced dataset where one class outnumbers the other. It is calculated as equation 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where:

- TP refers to rightly forecasted positive situations.
- TN refers to rightly forecasted negative situations.
- FP refers to wrongly forecasted positive situations.
- FN refers to wrongly forecasted negative situations.

2. Precision: Precision, often referred to as Positive Predictive Value, is the proportion of true positive predictions out of the total number of predicted positives. It is necessary in cases when the false positive predictions are high. It is defined as:

$$Precision = \frac{TP}{TP+FP}$$

3. Recall: Recall which is also referred to as Sensitivity or True Positive Rate, The true positive rate is the proportion of actual positives that are correctly identified. This is important in cases

where false negatives are costly, e.g. disease detection (missing a person with a positive case can have severe consequences). It is calculated as follows:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

4. **F1 Score**: The F1 score represents the harmonic average of precision and recall. It offers a balanced statistic when precision and recall are equally significant, which is very beneficial when working with data sets that are imbalanced. A high F1 score indicates both high precision and recall, meaning the classifier is both accurate in predicting positive instances and able to capture a large proportion of actual positives. It is calculated as:

$$F_1 Score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \quad (4)$$

IV. SIMULATION RESULTS

This section provides the detailed analysis of results derived based on the evaluation of the suggested methodology in intrusion detection in SCADA systems. In this regard, the dimensionality reduction and training results are presented first; next, the trained model will be evaluated. Simulations will be performed on MATLAB 2024a using an Intel Core i7 13650HX CPU, with 32 GB of RAM and an NVIDIA RTX 4060 GPU with 8GB GDDR6 memory.

D. Model Optimization Results

In this section, we will look at the results from the model optimization process, focusing on how preprocessing and training the Radial Basis Network (RBN) made a difference. we will evaluate how well the different classes can be distinguished and kept an eye on the training error of the RBN as it improves.

Figure 3 illustrates how various classes can be distinguished using the first three features, both before and after applying the Principal Component Analysis (PCA) algorithm. From the figure, you can see that PCA really helped to enhance the separation between the classes, which in turn improved the overall structure of our dataset. The results show that this dimensionality reduction not only keeps the important information but also helps the model better distinguish between different types of intrusions. While there is still some overlap between classes in a few samples, it is worth mentioning that we have 14 other features to help boost the performance of intrusion detection in SCADA systems.

ISSN: 2312-8135 | Print ISSN: 1992-0652
www.journalofbabylon.com | jub@itnet.uobabylon.edu.iq | info@journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

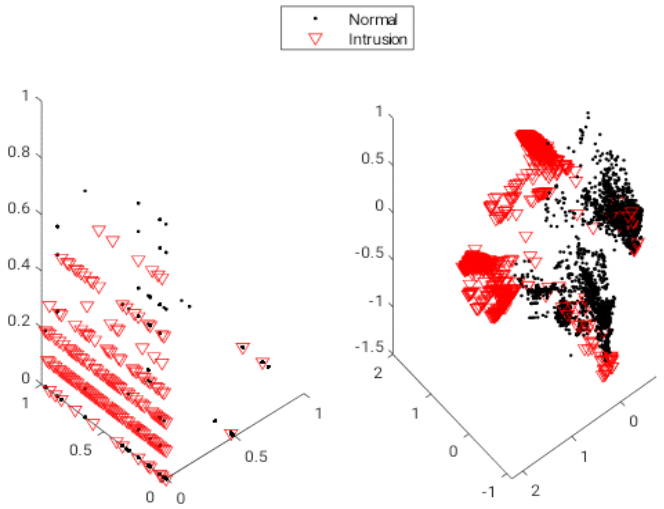

Figure 3. Discriminability of classes, before and after applying PCA [in the current study]

Moreover, Figure 4 reveals the error of the RBN network during its training iterations using EPO. This chart gives information on the network's convergence pattern. As can be seen from the curve, it converges quickly to the global best value. Finally, at convergence, the final trained error is 0.0071, which testifies to the EPO algorithm's fine performance on RBN training in terms of just how well it is doing this intrusion detection business. These results confirm both the effectiveness of reducing dimensionality by PCA and training with EPO.

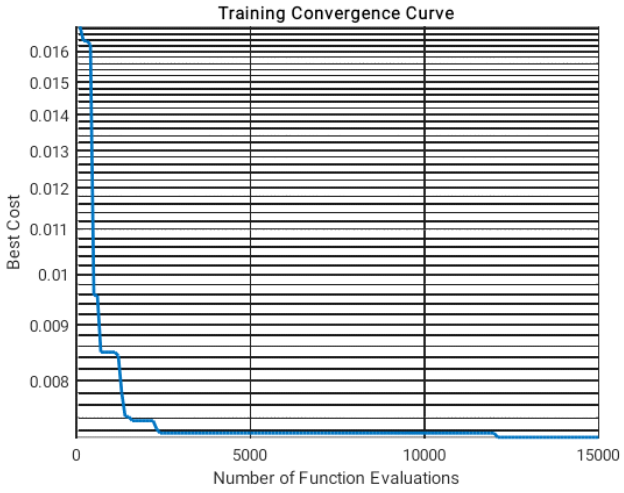

Figure 4. The convergence curve of EPO in training the RBN network [in the current study]

E. *Normal Traffic Classification Results*

In this section, we will go into the performance detail of our intrusion detection model using a

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

mix of evaluation metrics, confusion matrix, and ROC curve, through which this detailed overview gives us an insight into how the model identifies intrusions to belong either to normal activities or to attacks. Key indicators of effectiveness are accuracy, precision, recall, and the F1 score. Precision informs the ratio of how often the model is right overall, while precision will show how many of the flagged intrusions were correct. Recall-sensitivity calculates the model's performance of catching all the real threats. The F1 score balances precision and recall, which could be useful when dealing with data that is imbalanced.

Table 1 shows the values of these different evaluation metrics for both the training and testing datasets in this regard. The network provides an accuracy of 99.29% with precision of 99.37%, recall of 99.18%, and F1 score of 99.27% for the training dataset. The results confirm that the proposed approach of intrusion detection shows strong performance in a variety of evaluation criteria while ensuring dependable detection in the security of the SCADA system. Especially, the high precision value indicates the proper ratio of correctly identified positive cases in the proposed method. The testing values are really close to the training values as can be shown in Table 1 with accuracy, precision, recall, and F1-score equal to 99.22%, 99.31%, 99.09%, and 99.20%, respectively. These close results indicate the proper generalization ability of the proposed method for unseen data.

Table 1. The overall performance evaluation of the proposed method using introduced evaluation metrics

	Accuracy	Precision	Recall	F1-Score
Training Dataset	99.29%	99.37%	99.18%	99.27%
Testing Dataset	99.22%	99.31%	99.09%	99.20%

The confusion matrix is used to do additional evaluation. A confusion matrix is a kind of performance measurement in classification models that describes the true positive, true negative, false positive, and false negative. As a metric, this is capable of providing detailed insights into the accuracy of each class classification; it is also very effective in the identification of the patterns of misclassifications to allow further model fine-tuning.

Figure 5 illustrates the suggested intrusion detection model's confusion matrix. This matrix informs that the model correctly classified 2787 intrusion instances and 3921 normal instances, while there were also 5 false alarms and 48 missed detections. These results show that most of

the intrusion and normal traffic instances are correctly differentiated by the model, with only a limited number misclassified.

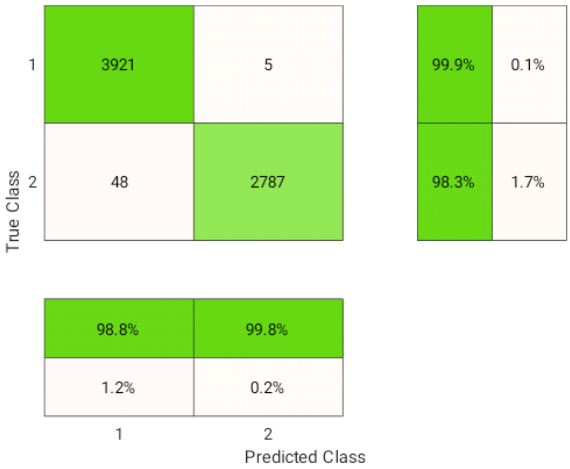

Figure 5. The confusion matrix result of the proposed method [in the current study]

Another great ally in the assessment of performance in the classification is the ROC curve. That would plot the rate of true positive, better known as sensitivity, against the rate of false positive, giving a visual idea of the compromise among sensitivity and specificity. The region surrounded by the curve-ROC, or simply AUC, is normally taken a summarizing metric for general effectiveness of any method, where a higher value serves as increased indication for better discriminative power.

Figure 6 represents the ROC curve for the proposed intrusion detection system, where an AUC value of over 0.99 shows excellent discriminative capability of the model in balancing sensitivity and specificity. This is confirmation that the proposed approach effectively can realize an accurate intrusion detection with a low false alarm rate, further validating the effectiveness of the proposed method for SCADA security. These results demonstrate the overall reliability and robustness of the proposed method, confirming its ability to enhance SCADA system security through accurate intrusion detection.

ISSN: 2312-8135 | Print ISSN: 1992-0652 | www.journalofbabylon.com | jub@itnet.uobabylon.edu.iq | info@journalofbabylon.com

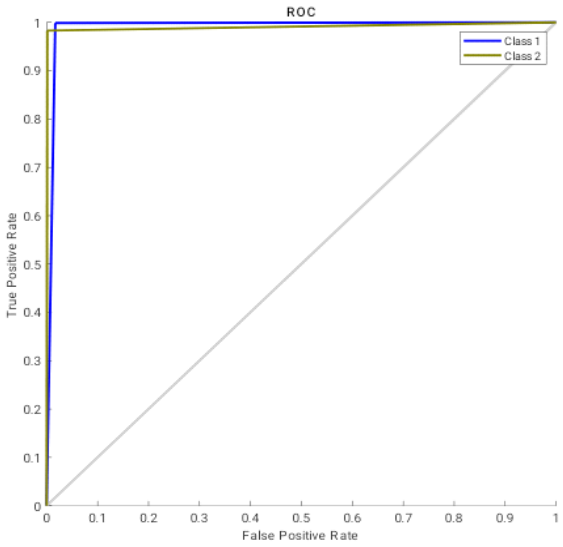

Figure 6. The ROC curve of the proposed method [in the current study]

V. COMPARE RESULTS OF EVALUATING DIFFERENT DATASETS

In this section, we present a comparative analysis of the proposed method's accuracy on three widely used datasets in the field of intrusion detection: UNSW-NB15**,** NSL-KDD, and Industrial Control Systems in Gas Transmission. The proposed approach leverages an optimized Radial Basis Function (RBF) neural network, fine-tuned with the Penguin Optimization Algorithm, specifically designed for accurate intrusion detection in SCADA (Supervisory Control and Data Acquisition) systems. As illustrated in Figure 7, the accuracy results indicate that the proposed method achieves exceptional performance across all three datasets, with an accuracy of 99.81% on the Industrial Control Systems in Gas Transmission dataset, 99.27% on the NSL-KDD dataset, and 99.02% on the UNSW-NB15 dataset. These results underscore the robustness and adaptability of the optimized RBF neural network in detecting intrusions across different environments, highlighting its potential for practical applications in SCADA systems and other industrial control settings.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

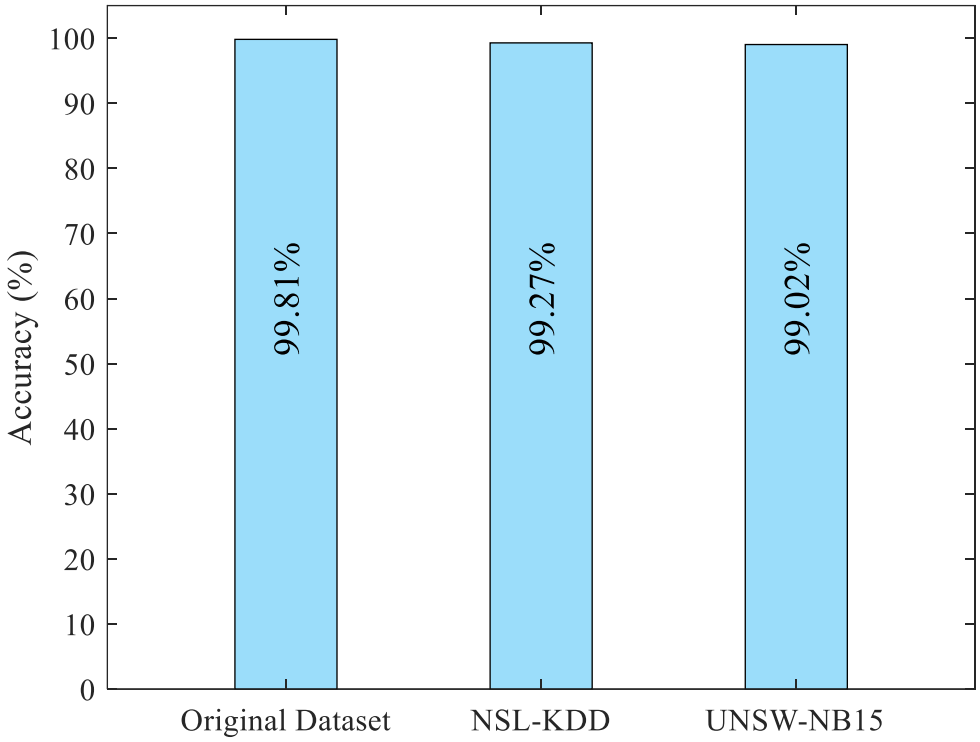

Figure 7. Comparison of Accuracy of the Proposed Method Across Different Datasets [in the current study]

VI. COMPARISON

This section presents an extensive contrast between the suggested approach and other techniques reviewed in the literature. It describes each approach, then summarizes the main characteristics of each method in one table in order to make the comparison easier [29].

presents an intrusion detection approach for ICS using Deep Convolutional Neural Networks (DCNN) and Long Short-Term Memory (LSTM) networks. The models were trained on ICS and gas pipeline data from Mississippi State University (MSU) without needing prior knowledge of network architecture. The LSTM model achieved over 99% accuracy and an AUC-ROC value of 99.50% on the ICS data [30]. focuses on securing Industrial Internet of Things (IIoT) systems, which face unique threats compared to traditional IT networks. The paper evaluates multiple IDS models using six machine learning models on the WUSTL-IIoT-2021 dataset. The RF model stood out with an accuracy of 97.97%, outperforming previous models [31]. discusses the rise in cyber-attacks on ICS alongside improvements in production efficiency. A hybrid intrusion detection method combining CNN and BiLSTM is proposed, with SMOTE-ENN used to balance imbalanced data. The method achieved 97.7% accuracy on the CICIDS-2017 dataset and 85.5% accuracy on the gas pipeline dataset [32]. explores the constant threat of viruses and malware in

computer networks and the role of intrusion detection in securing IoT-driven cyber-physical systems. The study employs deep learning, including Generative Adversarial Networks (GANs), to enhance accuracy and reduce false positives. The hybrid model of CNN and LSTM reached an accuracy of 99% on the KDD-Cup dataset.

All the reviewed methods show a very good intrusion detection performance; however, the proposed approach outperforms the others owed to its optimal adjustments and efficient training process. An overview of the comparative analysis is presented in Table 2.

Table 2. The comparison of the proposed method with other intrusion detection approaches

Reference	Methods used	Accuracy
[29]	DCNN and LSTM	95%
[30]	Multiple IDS models utilizing six machine learning techniques	98.88%
[31]	CNN and BiLSTM	98.79%
[32]	CNN and LSTM	96.7%
Proposed method	PCA and RBN trained with EPO	99.81%

CONCLUSIONS

The proposed approach improves SCADA intrusion detection through preprocessing, dimensionality reduction, and advanced classification. PCA retains essential data, while the RBN network ensures accurate pattern recognition. The EPO algorithm optimizes training and resolves gradient issues. Evaluation shows high accuracy, confirming the model's reliability for industrial security.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Conflict of interests:

There are non-conflicts of interest.

References

[1] T. Gao, J., Gan, L., Buschendorf, F., Zhang, L., Liu, H., Li, P.,& Lu, "Omni SCADA intrusion detection using deep learning algorithms," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 951–961, 2020, doi: 10.1109/JIOT.2020.3009180.

[2] C. K. T Öztürk, Z Turgut, G Akgün, "Machine learning-based intrusion detection for SCADA systems in healthcare," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 11, 2022, doi: https://doi.org/10.1007/s13721-022-00390-2.

[3] K. Chakrabarti, "Intrusion detection of SCADA system using machine learning techniques: A study," *IoT Secur. Paradig. Appl.*, pp. 135–164, 2020, https://doi.org/10.1145/3374135.3385282.

[4] K. O. A. Alimi, O. A., Ouahada, K., Abu-Mahfouz, A. M., Rimer, S., & Alimi, "A review of research works on supervised learning algorithms for SCADA intrusion detection and classification," *Sustainability*, 2021, doi: https://doi.org/10.3390/su13179597.

[5] A. C. F Mesadieu, D Torre, "Leveraging Deep Reinforcement Learning Technique for Intrusion Detection in SCADA Infrastructure," *IEEE Access.*, pp. 63381–63399, 2024, doi: 10.1109/ACCESS.2024.3390722.

[6] H. Gaiceanu, M., Stanculescu, M., Andrei, P. C., Solcanu, V., Gaiceanu, T., & Andrei, "Intrusion detection on ics and scada networks," *Recent Dev. Ind. Control Syst. Resil.*, 2020, doi: https://doi.org/10.1007/978-3-030-31328-9_10.

[7] P. Rajesh, L., & Satyanarayana, "Evaluation of machine learning algorithms for detection of malicious traffic in scada network," *J. Electr. Eng. Technol.*, vol. 17, pp. 913–928, 2022, doi: https://doi.org/10.1007/s42835-021-00931-1.

[8] T. L. Nguyen, D. D., Le, M. T., & Cung, "Improving intrusion detection in SCADA systems using stacking ensemble of tree-based models," *Bull. Electr. Eng. Informatics*, 2022, doi: https://doi.org/10.11591/eei.v11i1.3334.

[9] S. H. M Altaha, JM Lee, M Aslam, "An Autoencoder-Based Network Intrusion Detection System for the SCADA System," *J. Commun*, 2021, doi:10.12720/jcm.16.6.210-216.

[10] M. C. HC Altunay, Z Albayrak, AN Özalp, "Analysis of anomaly detection approaches performed through deep learning methods in SCADA systems," *2021 3rd Int. Congr. Human-Computer Interact. Optim. Robot. Appl.*, 2021, doi: 10.1109/HORA52670.2021.9461273.

[11] S. S. D Upadhyay, J Manero, M Zaman, "Gradient boosting feature selection with machine learning classifiers for intrusion detection on power grids," *IEEE Trans. Netw. Serv. Manag.*, pp. 1104–1116, 2020, doi: 10.1109/TNSM.2020.3032618.

[12] T. L. Y Ouyang, B Li, Q Kong, H Song, "FS-IDS: a novel few-shot learning based intrusion detection system for scada networks," *ICC 2021-IEEE Int. Conf. Commun.*, 2021, doi: https://doi.org/10.1109/ICC42927.2021.9500667.

[13] D. Upadhyay, J. Manero, M. Zaman, and S. Sampalli, "Intrusion Detection in SCADA Based Power Grids : Recursive Feature Elimination Model With Majority Vote Ensemble Algorithm," vol. 8, no. 3, pp. 2559–2574, 2021, doi: 10.1109/TNSE.2021.3099371.

[14] B. Phillips and E. Gamess, "An Evaluation of Machine Learning-based Anomaly Detection in a SCADA System Using the Modbus Protocol," pp. 188–196, 2020, https://doi.org/10.1145/3374135.3385282.

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

[15] Mahalakshmi M; Ramkumar M P; Emil Selvan G S R, "Scada intrusion detection system using cost sensitive machine learning and smote-svm," *2022 4th Int. Conf. Adv. Comput. Commun. Control Netw.*, 2022, doi: 10.1109/ICAC3N56670.2022.10074251.

[16] D. P. S Nazir, S Patel, "Autoencoder based anomaly detection for SCADA networks," *Int. J. Artif. Intell. Mach. Learn.*, p. 17, 2021, doi: 10.4018/IJAIML.20210701.oa6.

[17] M. A. Umer, "Machine Learning for Intrusion Detection in Industrial Control Systems : Applications , Challenges , and Recommendations, https://doi.org/10.1016/j.ijcip.2022.100516."

[18] A. A. LA Aldossary, M Ali, "Securing SCADA systems against cyber-attacks using artificial intelligence," *2021 Int. Conf. Innov. Intell. Informatics, Comput. Technol.*, 2021, doi: 10.1109/3ICT53449.2021.9581394.

[19] M. A. S. Arifin and J. Rejito, "Denial of Service Attacks Detection on SCADA Network IEC 60870-5-104 using Machine Learning," no. October, pp. 20–21, 2021, doi**:** 10.23919/EECSI53397.2021.9624255.

[20] S. H. M Altaha, JM Lee, M Aslam, "Network Intrusion Detection based on Deep Neural Networks for the SCADA system," *J. Phys. Conf. Ser.*, 2020, doi: 10.1088/1742-6596/1585/1/012038.

[21] D. K. LAC Ahakonye, CI Nwakanma, JM Lee, "SCADA intrusion detection scheme exploiting the fusion of modified decision tree and Chi-square feature selection," *Internet of Things*, vol. 21, 2023, doi: https://doi.org/10.1016/j.iot.2022.100676.

[22] O. A. Alimi, "Supervised learning based intrusion detection for SCADA systems," pp. 4–8, 2022, doi: 10.1109/NIGERCON54645.2022.9803101.

[23] U. A. S Sivakumar, BS Ananthanarayanan, "Intrusion Detection System for Securing the SCADA Industrial Control System," *Proc. Int. Conf. Comput. Intell. Data Sci. Cloud Comput.*, pp. 645–658, 2021, doi: https://doi.org/10.1007/978-981-33-4968-1_50.

[24] J. Suaboot *et al.*, "A Taxonomy of Supervised Learning for IDSs in SCADA Environments, https://doi.org/10.1145/3379499."

[25] F. A. Z Ahmad, A Shahid Khan, C Wai Shiang, J Abdullah, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 1, 2021, doi: https://doi.org/10.1002/ett.4150.

[26] H. R. Gholam Ali Montazer, Davar Giveki, Maryam Karami, "Radial Basis Function Neural Networks: A Review Gholam," *Comput. Rev. J*, vol. 1, no. 1, pp. 52–74, 2018.

[27] M. Albahar, A. Alharbi, M. Alsuwat, and H. Aljuaid, "A Hybrid Model based on Radial basis Function Neural Network for Intrusion Detection," vol. 11, no. 8, pp. 781–791, 2020.

[28] S. Harifi, M. Khalilian, J. Mohammadzadeh, and S. Ebrahimnejad, "Emperor Penguins Colony : a new metaheuristic algorithm for optimization," *Evol. Intell.*, vol. 0, no. 0, p. 0, 2019, doi: 10.1007/s12065-019-00212-x.

[29] S. C. YK Saheed, S Misra, "Autoencoder via DCNN and LSTM Models for Intrusion Detection in Industrial Control Systems of Critical Infrastructures," *2023 IEEE/ACM 4th Int. Work. Eng. Cybersecurity Crit. Syst.*, 2023, doi: 10.1109/EnCyCriS59249.2023.00006.

[30] Abdulrahman Mahmoud Eid; Ali Bou Nassif; Bassel Soudan; Mohammad Noor Injadat, "IIoT network intrusion detection using machine learning," *2023 6th Int. Conf. Intell. Robot. Control Eng.*, 2023, doi: 10.1109/IRCE59430.2023.10255088.

[31] J. Wang, C. Si, Z. Wang, and Q. Fu, "A New Industrial Intrusion Detection Method Based on CNN-BiLSTM," 2024, doi: 10.32604/cmc.2024.050223.

[32] V. Gunnam, S. R., Vepuri, S. K., & Nallarasan, "Detection of Real Time Malicious Intrusions Using GAN (Generative Adversarial Networks) in Cyber Physical System," *2024 5th Int. Conf. Emerg. Technol.*, 2024, doi: 10.1109/INCET61516.2024.10593381.

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652