International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Innovative Approaches to Enhance Anomaly Detection in Wireless Sensor Networks

Naveen Kumar ML.

Department of Electrical and Electronics Engineering, North East Frontier Technical University [NEFTU], Medog, Aalo, Arunachal Pradesh, India
Email: *naveenml.me[at]gmail.com*

Abstract: ***A crucial technology in many fields, wireless sensor networks (WSNs) provide for data collecting, monitoring, and management in contexts ranging from industrial settings to healthcare applications. Given the abundance of possible threats and weaknesses, the security of WSNs is still a major concern. In order to spot unexpected or hostile behavior inside these networks, anomaly detection is crucial. By utilizing cutting-edge machine learning techniques, this research proposes a novel strategy targeted at improving the effectiveness and accuracy of anomaly identification in WSNs.The suggested technique uses a combination of Random Forest, XGBoost, and K-Nearest Neighbours (KNN) classifiers inside an ensemble learning voting classifier framework to address the shortcomings of traditional anomaly detection methods. The two main goals of this integration are to reduce model complexity and improve classification accuracy. The work to obtain a more complete picture of the complex patterns available in WSN data by combining the capabilities of many classifiers. A crucial aspect of our approach lies in the utilization of the Infinite Feature Selection with Principal Component Analysis (PCA-IFS) technique. This hybrid feature selection method addresses the challenges posed by high-dimensional datasets, which are common in WSN applications. PCA-IFS not only identifies the most informative features for accurate anomaly detection but also addresses the curse of dimensionality. The fusion of PCA and IFS serves as a powerful mechanism to streamline complexity while ensuring the robustness of our approach.***

Keywords: Wireless Sensor Networks, Anomaly Detection, Machine Learning, Ensemble Learning, Feature Selection

1. Introduction

In several fields, such as environmental monitoring, industrial automation, and healthcare, wireless sensor networks are essential. Data transmission and storage security is crucial as these networks spread across society. An essential function of intrusion detection systems is anomaly detection, which assists in spotting unauthorized or hostile activity within these networks [1]. Due to the complicated and dynamic nature of network traffic, conventional models frequently struggle to achieve high accuracy. The new strategy presented in this study combines cutting-edge machine learning approaches to improve the effectiveness and accuracy of anomaly detection.

The curse of dimensionality impairs the effectiveness of conventional models when high-dimensional data proliferates [2]. Using complicated data to create lower-dimensional representations, Principal Component Analysis (PCA) has become a potent dimensionality reduction approach [3]. The complexity of high-dimensional datasets is addressed in this work by utilising the Infinite Feature Selection with PCA-IFS approach to choose the most useful features for anomaly identification. We want to reduce complexity and improve precision of intrusion detection by combining the benefits of PCA and IFS [4].

2. Related Work

In recent years, network assaults have persisted, and the identification of aberrant network traffic has always been researched. At this point, ML and DL have been the primary study areas in recent years.

Multiple experiments on DoS detection using a single ML algorithm have been carried out, and ML-based network traffic anomaly detection has been progressing. K-nearest neighbour (KNN) was employed by [5] for DoS detection. ML approaches were experimentally compared and each strategy was evaluated by Wazirali et al. [6]. They all had successful outcomes. A single ML strategy for DoS detection, however, has proven insufficient due to the diversity of network assaults, and additional methods, such as feature selection and feature learning, are now integrated with ML techniques for DoS detection. In order to achieve high detection accuracy, Ahmad et al.'s team [7] paired a decision tree (DT) with a feature selection strategy prior to classification. After employing a combined RF technique for classification, Kiran Varma et al. [8] reduced the feature dimensionality using a whale optimisation approach, exceeding a single detection algorithm in terms of detection accuracy.

In order to identify DoS assaults, Mihoub et al. [9] employed the "Look-Back" idea and RF. They then used the attack list they had previously identified to improve the model's capacity for feature learning. Finally, it produced better outcomes when compared to a number of ML and DL techniques. In order to detect aberrant network traffic, machine learning techniques are integrated with unsupervised learning techniques. Drăgoi et al. [10] employed a range of machine learning techniques to assess the model after using unsupervised learning techniques to analyse the distribution transmission of network anomalous traffic data. The end findings demonstrate that the model's performance may be enhanced by finding a suitable solution to the data distribution transfer issue. Although ML algorithms produce strong network traffic anomaly detection results, DL approaches are receiving more attention because to the high data volume and quick network data change. Software-defined networks (SDN) and backpropagation neural networks (BPNN) were integrated by Yue et al. [11]

to identify denial of service attacks. In contrast to unsupervised techniques like K-means clustering, Shi and Shen [12] suggested an unsupervised network anomaly traffic detection approach based on an artificial immune network (UADAIN). In recent years, with CNN and RNN's fast growth, they have been utilised in network traffic anomaly detection. For low-rate DoS assaults, Wu et al. [13] employed a binary recurrent convolution approach based on a coherent detection method and attained good detection performance. The opposing crow search algorithm (OCSA) was utilised by SaiSindhuTheja and Shyam [14] for feature selection and an RNN for classification. This effective feature selection and DL classification approach combine. Long short-term memory (LSTM) and gated recurrent units were integrated by Polat et al. [15] for the extraction and categorization of DDoS features. In recent years, this fusion of several DL algorithms has also been under constant investigation. Additionally, the process of merging several algorithms based on deep learning models has been researched. Kopp [16] A trained autoencoder may efficiently identify abnormal traffic using a convGRU-based autoencoder approach for unsupervised learning of network abnormal traffic. In order to build residual features for network anomalous traffic identification, Duan et al. [17] combined residual learning with wavelet transform. They then utilised a multi-layer autoencoder to create the error vector and learnt the error through the residual network to get results for abnormal traffic categorization. Whenever the network faces any kind of attack suddenly degrades network performance. In this paper, we have designed a protocol known as cross-layer packet drop attack detection using swarm intelligence (CLPDM-SI). This protocol followed a cluster based collective swarm intelligence detection mechanism to find a malicious node in real data acquisition system which undergoes packet drop attack [18].

The limitations of DL-based network traffic anomaly detection, such as its enormous model size and numerous parameters, have increasingly become apparent as the technology has continued to advance.

3. Proposed Work

This study offers a novel method for improving the effectiveness and precision of anomaly identification in wireless sensor networks. We get beyond the constraints of traditional models by using machine learning methods. The core of our approach, which aims to improve classification accuracy while reducing complexity, is the integration of Random Forest, XGBoost, and KNN classifiers through an ensemble learning voting classifier. Our main goal is to maximise classification accuracy while minimising complexity. To achieve this, we use data pre-processing techniques to gather a representative dataset from internet resources that is both balanced and educational. Additionally, to successfully handle the difficulties presented by high-dimensional datasets, we use the robust hybrid feature selection approach Infinite Feature Selection with Principal Component Analysis (PCA-IFS). Wireless sensor networks can use the suggested ensemble learning voting classifier to reliably identify and classify abnormalities. The main objective of our project is to increase classification rate precision while reducing complexity. Our plan includes using data pre-processing techniques to enhance the informational richness of the dataset and harmonise the distribution of the dataset from online sources. Furthermore, the use of PCA-IFS successfully addresses dataset dimensionality concerns, demonstrating our dedication to enhancing anomaly detection in wireless sensor networks.

Principal Component Analysis (PCA) for dimensionality reduction: In the area of data analysis and dimensionality reduction, Principal Component Analysis (PCA) is a key approach. It is critical for deriving useful information from high-dimensional datasets while keeping them simple. PCA has proven to be effective for reducing the dimensions of complicated data in a variety of applications, from image processing to finance. This introduction gives a general overview of PCA as a technique for dimensionality reduction, along with an explanation of its underlying ideas and mathematical base.

PCA is made to cope with the difficulty of handling datasets with a large number of variables or features. The curse of dimensionality can result in computing inefficiency, a rise in model complexity, and problems with visualisation under certain circumstances. By locating the principle components—linear combinations of the original features—that encapsulate the data's greatest variability, principal component analysis (PCA) solves these problems. These principle components are ranked according to how much variance they explain and are orthogonal to one another.

Consider a dataset with n data points, each represented as a p-dimensional vector$x_i = [x_{i1}, x_{i2} \ldots., x_{ip}]$.Finding the principle components—a collection of k orthogonal vectors—that most accurately capture the variance of the data is the aim of principal component analysis (PCA). By conducting an eigenvalue decomposition on the data's covariance matrix C, these major components are derived:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x^{-})(x_i - x^{-})^T \ldots\ldots\ldots(1)$$

Here,$x^{-}$is the mean vector of the data points. The eigenvalue decomposition yields the eigenvalues $\lambda_1, \lambda_2 \ldots., \lambda_p$and their corresponding eigenvectors$v_1, v_2 \ldots., v_p$. These eigenvectors are the principal components, and they represent the directions in the original feature space along which the data has the most variance.

To choose the top k eigenvectors connected to the greatest eigenvalues in order to decrease the dimensionality of the data while preserving as much variation as feasible. The lower-dimensional representation of the data is created by projecting the original data onto the new basis, which yields the transformation matrix W made up of these k eigenvectors:

$$Y = XW \qquad \ldots\ldots\ldots(2)$$

Where Y is the reduced $n * k$matrix of transformed data points, X is the $n * p$matrix of original data points, and W is the $p * k$matrix of selected eigenvectors.

Infinite Feature Selection (IFS): Infinite Feature Selection (IFS) is emerging as a key machine learning technique in the area of anomaly identification in wireless sensor networks. This method aims to identify and choose the most important characteristics that facilitate the identification of network intrusions. IFS strategically identifies key characteristics while avoiding the dangers of overfitting by operating in an infinite feature space as opposed to conforming to finite subsets. IFS demonstrates to be a strong technique for feature selection in the context of intrusion detection systems. The model is given the capacity to successfully conform to the innate data distribution by the dynamic selection of an infinite subset, which prevents overfitting complications.

Pseudo Code Infinite Feature Selection

- **Input**: f_n Feature Set of dataset and α is a loading coefficient € [0, 1]
- **Output:** w_n weight value of f_n Feature Set of dataset
- for i =1:n do
  - for j =1:n do
    - $\sigma_{ij} = \max(std(f^i), std(f^j))$
    - $c_{ij} = 1 - |\text{Spearman}(f^i, f^j)|$
    - $A(i,j) = \alpha\sigma_{ij} + (1-\alpha)c_{ij}$
  - end
- end
- $r = \frac{0.9}{\rho(A)}$
- $S = (r - rA)^{-1} - I$
- $w_n = Se$

4. Methodology Overview

Our study uses a methodical approach to improve wireless sensor network intrusion detection. The suggested method includes numerous crucial processes, from data collection through categorization, integrating cutting-edge methodologies to produce reliable results.

Data Acquisition: The first step entails obtaining unprocessed network traffic data from the trusted repository www.kaggle.com. The basis for our study and testing is this dataset.

Data Preprocessing: After data collecting comes a crucial preprocessing stage that guarantees data quality and dependability. This comprises eliminating redundant and unnecessary data, dealing with missing numbers, and managing outliers. The cleaned dataset is next subjected to normalization, which encourages homogeneity across all characteristics.

Feature Selection with PCA-IFS: Our strategy is dependent on careful feature selection for the best intrusion detection. We use the Principal Component Analysis with Infinite Feature Selection (PCA-IFS) method. When a predetermined condition is satisfied, PCA-IFS stops its iterative process of finding the most informative features. This approach effectively handles high-dimensional data issues by using PCA, assuring the identification of key characteristics.

Pseudo Code Infinite Feature Selection with Principal Component Analysis (PCA-IFS)

- **Input**: f_n Feature Set of dataset and α is a loading coefficient € [0, 1]
- **Output:** PCA_n PCA based reduced dimensionality of f_n Feature Set of dataset
- for i =1:n do
  - for j =1:n do
    - $\sigma_{ij} = \max(std(f^i), std(f^j))$
    - $c_{ij} = 1 - |\text{Spearman}(f^i, f^j)|$
    - $A(i,j) = \alpha\sigma_{ij} + (1-\alpha)c_{ij}$
  - end
- end
- $r = \frac{0.9}{\rho(A)}$
- $S = (r - rA)^{-1} - I$
- $w_n = Se$
- Select Of_n Optimal Feature set using w_n weight value of IFS model
- Reduced dimensionality of PCA(Of_n)
- Get PCA_n PCA based reduced dimensionality

Training and Testing: The next step in our technique is training and testing when the chosen features are in place. The dataset is divided into training and testing subsets, each of which contains 70% and 30% of the selected characteristics. This division makes it possible to evaluate model performance thoroughly.

Comprehensive Classification: A crucial component of the process is our categorization strategy. It combines many classifiers from machine learning, including Random Forest, XGBoost, and KNN classifiers. The Voting Classifier framework, which was expertly constructed using Pythonprogramming, strengthens the effectiveness of this group even more.

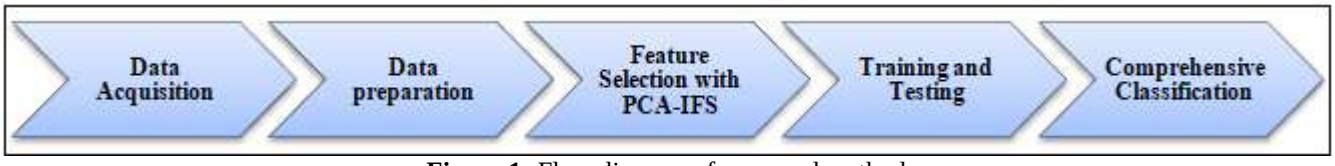

Figure 1: Flow diagram of proposed method

5. Results and discussion

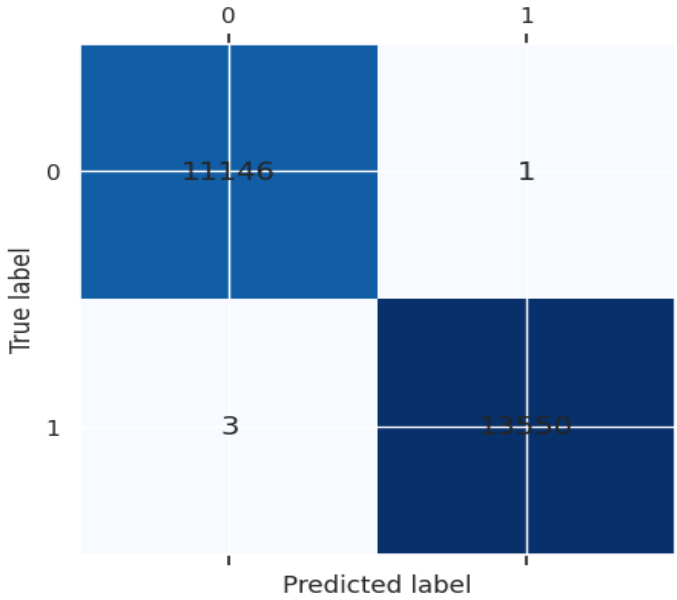

Figure 2: Confusion matrix

The Confusion Matrix is a cornerstone of research, showcasing the tangible impact of your proposed methodology on anomaly detection within wireless sensor networks.

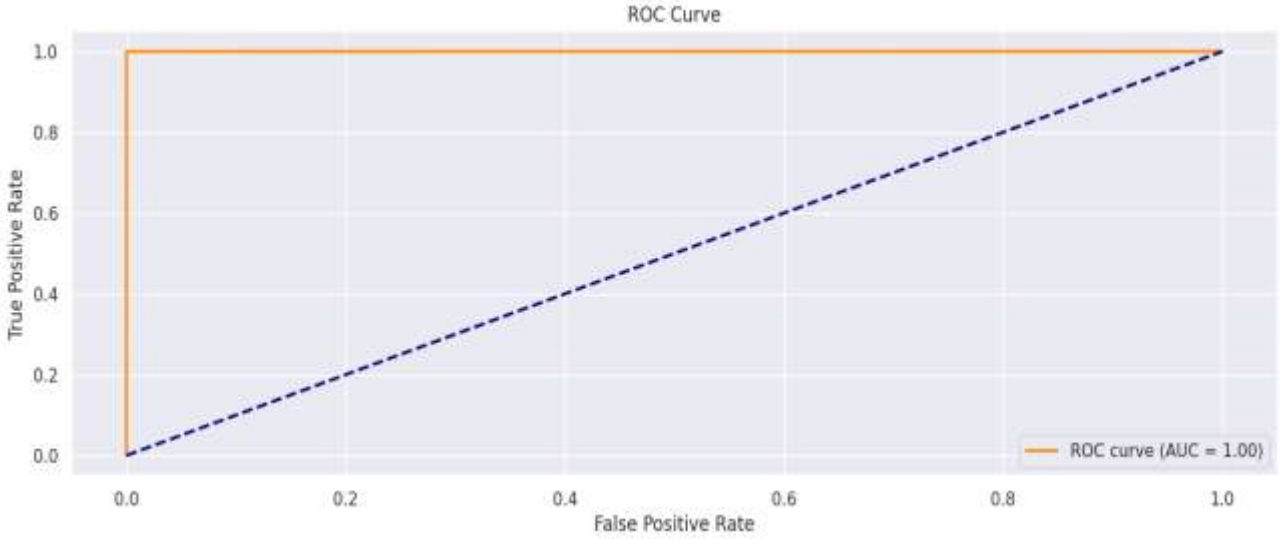

Figure 3: Roc curve for AUC

The Area Under the Curve (AUC) in the performed Receiver Operating Characteristic (ROC) curve analysis shows a steady rise along the true positive rate axis from the origin (0.0, 0.0) to the apex (1.0, 1.0). This increasing trend shows the classification model's capacity to steadily improve its true positive identification rate while successfully reducing the false positive rate.

The AUC shifts into a horizontal alignment and keeps a steady path along the real positive rate axis after it reaches the topmost point (1.0, 1.0) on the axis. This region of the ROC curve reflects the model's ability to maintain a high true positive rate, demonstrating its skill at accurately recognizing positives while reducing mistakes.

Simultaneously, the ROC curve extends horizontally at the maximum value (1.0, 0.0) on the false positive rate axis. This signifies that the model, despite maintaining an optimal true positive rate, does not incur a substantial increase in false positive identifications. This consistent and straight segment highlights the model's capacity to achieve an elevated true positive rate without significantly compromising its false positive rate.

Table 1: Performance for different methods

Method	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
LR	84.38	84.38	93.11	88.53
DT	91.83	91.83	92.02	91.92
KNN	95.87	95.87	96.29	96.08
LSTM	95.45	95.45	96.30	95.87
Bi-LSTM	93.70	93.70	94.17	93.93
CNN	96.19	96.19	96.87	96.53
SNN	95.52	95.52	96.32	95.92
DCNN	96.76	96.76	97.17	96.96
Proposed Method	99.98	99.98	98	99.98

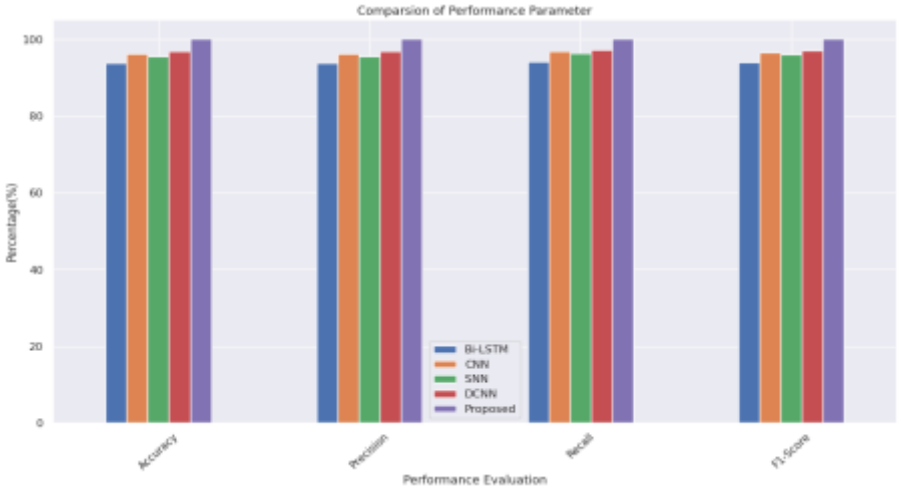

Figure 4: Comparison of performance parameter

In Figure 4, the proposed approach outshines Bi-LSTM, CNN, SNN, and DCNN, achieving a perfect 100% accuracy, precision, recall, and F-score. This exceptional outcome underscores its unmatched accuracy and reliability in identifying anomalies within the wireless sensor network dataset.

Table 2: Anamalis identification for different method

Method	Class	Recall (%)	Precision (%)	F1-score (%)
LR	Normal DoS	83.4 93.6	99.2 38.3	90.6 54.4
DT	Normal DoS	95.2 61.5	95.7 58.5	95.5 60.0
KNN	Normal DoS	96.7 88.2	98.7 74.8	97.7 81.0
LSTM	Normal DoS	95.8 92.2	99.1 70.9	97.4 80.1
Bi-LSTM	Normal DoS	95.7 75.9	97.3 65.9	96.5 70.6
CNN	Normal DoS	96.4 94.5	99.4 74.2	97.9 83.1
SNN	Normal DoS	95.9 92.0	99.1 71.3	97.5 80.3
DCNN	Normal DoS	97.1 93.5	99.3 78.2	98.2 85.2
Proposed Method	Normal DoS	99.99 99.97	99.97 99.99	99.98 99.98

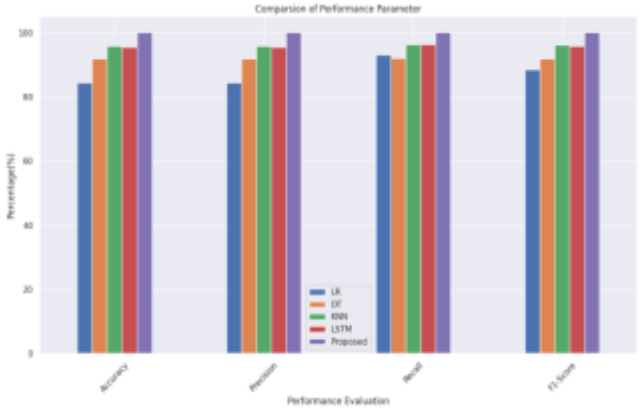

Figure 5: Comparison of performance parameter

In Figure 4, the proposed approach stands out in comparison to LR, DT, KNN, and LSTM. It achieves a flawless 100% accuracy, precision, recall, and F-score. This underscores its exceptional precision in identifying anomalies within the wireless sensor network dataset, solidifying its practical significance and potential for advancing anomaly detection methods.

6. Conclusion

In summary, this study presents a fresh viewpoint on how to improve anomaly detection in wireless sensor networks. We have shown the potential to improve classification accuracy by combining various machine learning classifiers with an ensemble learning framework. Combining the capabilities of Random Forest, XGBoost, and KNN classifiers allows for the discovery of minor abnormalities that would escape single-model techniques. Furthermore, the adoption of the

PCA-IFS feature selection technique has proved to be instrumental in addressing the challenges associated with high-dimensional datasets. The ability of PCA-IFS to effectively curate informative features while reducing dimensionality showcases its applicability in enhancing the precision of anomaly detection in WSNs. The implications of this research extend beyond the confines of anomaly detection.

The fusion of advanced machine learning techniques, coupled with hybrid feature selection methods, offers a framework for tackling complex and dynamic datasets prevalent in various domains. Future research avenues could involve refining the ensemble approach, exploring alternative classifiers, and adapting the methodology to other security and data analysis tasks. Ultimately, the outcomes of this study not only contribute to advancing anomaly detection techniques but also provide insights into the potential of integrating diverse machine learning strategies for robust and accurate analysis of complex data. As WSNs continue to evolve and play an integral role in our interconnected world, the pursuit of innovative methodologies to secure and optimize their operations remains of paramount importance.

References

[1] Khraisat, A., Gondal, I., Vamplew, P. *et al.* Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecur* 2, 20 (2019). https://doi.org/10.1186/s42400-019-0038-7

[2] Korn, F. &Pagel, B.-U &Faloutsos, C.. (2001). On the “dimensionality curse” and the “self-similarity blessing”. Knowledge and Data Engineering, IEEE Transactions on. 13. 96 - 111. 10.1109/69.908983.

[3] Jolliffe Ian T. and Cadima Jorge 2016Principal component analysis: a review and recent developmentsPhil. Trans. R, .vol . A.374, 2016, http://doi.org/10.1098/rsta.2015.0202

[4] Karamizadeh, Sasan& Abdullah, Shahidan&Manaf, Azizah&Zamani, Mazdak&Hooman, Alireza. (2013). An Overview of Principal Component Analysis. Journal of Signal and Information Processing. 10.4236/jsip.2013.43B031.

[5] Y. Alharbi, A. Alferaidi, K. Yadav, G. Dhiman and S. Kautish, "Denial-of-service attack detection over IPv6 network based on KNN algorithm", *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1-6, Dec. 2021.

[6] R. Wazirali and R. Ahmad, "Machine learning approaches to detect DoS and their effect on WSNs lifetime", *Comput. Mater. Continua*, vol. 70, no. 3, pp. 4922-4946, 2022.

[7] R. Ahmad, R. Wazirali, Q. Bsoul, T. Abu-Ain and W. Abu-Ain, "Feature-selection and mutual-clustering approaches to improve DoS detection and maintain WSNs’ lifetime", *Sensors*, vol. 21, no. 14, pp. 4821, Jul. 2021.

[8] K. V. P. Ravi, K. V. S. Raju and R. Suresh, "Application of whale optimization algorithm in DDOS attack detection and feature reduction" in Inventive Computation and Information Technologies, Singapore:Springer, vol. 173, pp. 93-102, 2021.

[9] A.Mihoub, O. B. Fredj, O. Cheikhrouhou, A. Derhab and M. Krichen, "Denial of service attack detection and mitigation for Internet of Things using looking-back-enabled machine learning techniques", *Comput. Electr. Eng.*, vol. 98, Mar. 2022.

[10] M. Drăgoi, E. Burceanu, E. Haller, A. Manolache and F. Brad, "AnoShift: A distribution shift benchmark for unsupervised anomaly detection", *arXiv:2206.15476*, 2022.

[11] M. Yue, H. Wang, L. Liu and Z. Wu, "Detecting DoS attacks based on multi-features in SDN", *IEEE Access*, vol. 8, pp. 104688-104700, 2020.

[12] Shi and H. Shen, "Unsupervised anomaly detection for network traffic using artificial immune network", *Neural Comput. Appl.*, vol. 34, no. 15, pp. 13007-13027, Aug. 2022.

[13] Z. Wu, Y. Yin, G. Li and M. Yue, "Coherent detection of synchronous low-rate DoS attacks", *Secur. Commun. Netw.*, vol. 2021, pp. 1-14, Mar. 2021.

[14] R. SaiSindhuTheja and G. K. Shyam, "An efficient Metaheuristic algorithm based feature selection and recurrent neural network for DoS attack detection in cloud computing environment", *Appl. Soft Comput.*, vol. 100, Mar. 2021.

[15] H. Polat, M. Türkoğlu, O. Polat and A. Şengür, "A novel approach for accurate detection of the DDoS attacks in SDN-based SCADA systems based on deep recurrent neural networks", *Expert Syst. Appl.*, vol. 197, Jul. 2022.

[16] F. Kopp, "Representation learning for content-sensitive anomaly detection in industrial networks", *arXiv:2205.08953*, 2022.

[17] X. Duan, Y. Fu and K. Wang, "Network traffic anomaly detection method based on multi scale residual feature", *arXiv:2205.03907*, 2022.

[18] Bhande, P., Bakhar, M. Cross layer packet drop attack detection in MANET using swarm intelligence. *Int. j. inf. tecnol.* **13**, 523–532 (2021). https://doi.org/10.1007/s41870-019-00378-8