Enhanced Intrusion Detection Using RNA Encoding and Frequent Pattern Mining on CICIDS2017 Dataset

Omar F. Rashid[1], Marwan I. Shukur[2], Humam Al-Shahwani[3]
[1]Department of Geology, College of Science, University of Baghdad, Baghdad, Iraq.
[2]Department of Electronic Engineering, College of Electrical Engineering, University of Technology, Baghdad, Iraq.
[3]Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq.

Article Info

Article history:

Received June, 12, 2025
Revised July, 15, 2025
Accepted Aug., 10, 2025

Keywords:

Intrusion Detection System
Frequent Pattern Mining
RNA Encoding
Network Security
Misuse Detection

ABSTRACT

Intrusion Detection System is a security system that serves as a shield to the infrastructure in place. Over the years, IDS technology has expanded significantly in order to meet the demands of new computer crime. Starting from the middle of the eighties to the current century, efforts have been made to improve its ability to identify attacks without compromising the performance of the network. This paper proposes a new method for a misuse IDS model based on RNA encoding and Frequent Pattern Mining for improving pattern extraction and classification, where Frequent Pattern Mining can deal with large amounts of data, therefore fast extractions of patterns and can detect intrusion in real-time. In this work, the characteristic patterns related to the malicious activity are identified using the CICIDS2017 dataset containing different types of network attacks. The proposed system comprises four stages: Data Selection and Preprocessing, RNA encoding, Frequent Pattern Mining and lastly classification. The experiments showed that the proposed IDS model has high accuracy and lower computational complexity as well as better scalability to be applied to real-time network intrusion detection.

Corresponding Author:

Marwan I. Shukur
Department of Electronic Engineering, College of Electrical Engineering, University of Technology.
AlSinna’a street, Baghdad, Iraq
Email: Marwan.I.Shukur@uotechnology.edu.iq

1. INTRODUCTION

The Internet is considered an essential part of our life and one of the most important tools. It occurs in human life in most fields, including business, education, and entertainment. In other words, when technology progresses, usage of the network prevails in any sphere of our existence. With this usage of the network, let to get the dangers of an attack on the network [1]. Computer network security has emerged as one of the greatest challenges of the recent past. One of the most effective security measures is the use of a security system in a network. Out of the two Firewall is one of the mechanisms used. Still, it is not very effective in protecting the network from attacks since the firewall is only capable of identifying attacks that originate from outside the network. More so, in the last couple of years, the rate of attacks related to networks has rampantly risen [2 - 3]. Consequently, as people are looking for an Intrusion detection system (IDS), an alternative security method, there has been growing interest among researchers. Intrusion detection is the practice of observing a computer system or a network with the intention of identifying unauthorized use or improper usage of resources in the system. IDs intend to identify such action and then take appropriate action to prevent additional loss or data surrender [4]. The most common IDS technologies, methodologies and approaches in the literature are given in Figure 1 [5].

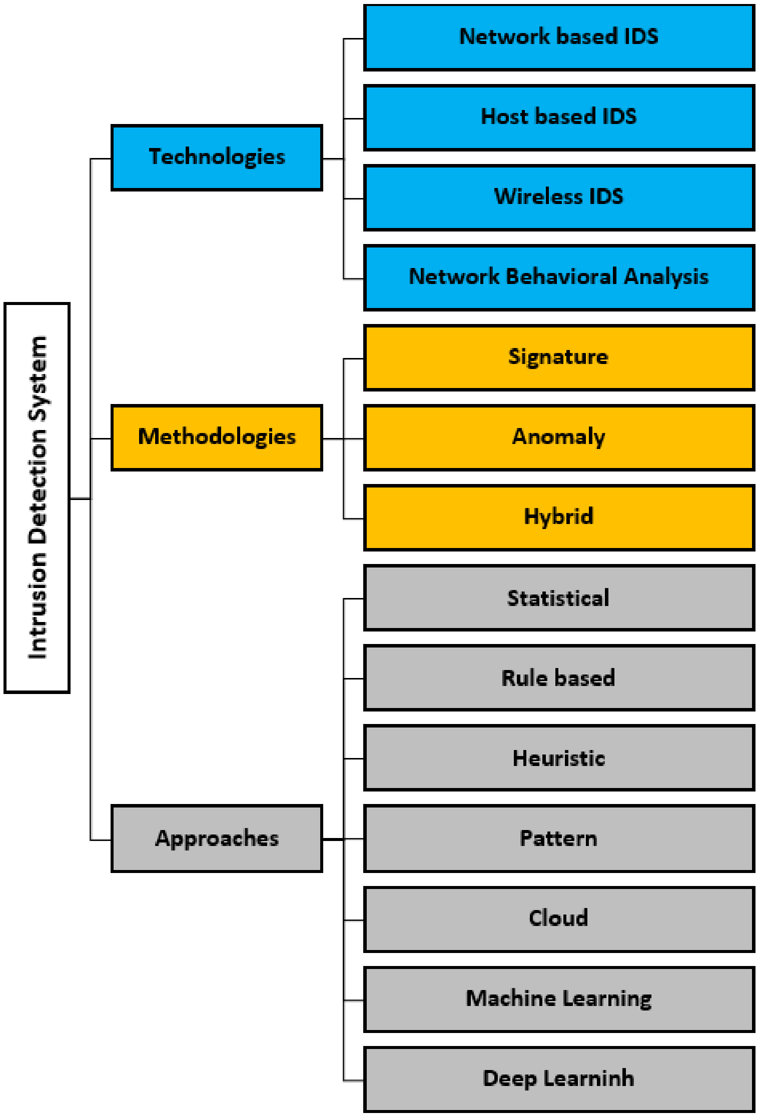

Figure 1. IDS Classification

Different types of IDS techniques have been described in the security context literature. Gama et al. [6] built a new method to enhance IDS in drone communications by simulating real network traffic, where this method is based on the Recurrent Neural Network and Long Short-Term Memory Network. A novel intelligent system is proposed by [7] based on feature selection with Rough set theory and the Bayes theorem, where the suggested feature selection method is done by computing the core features and grouping them using estimated probability. A new detection method in real-time by using the least squares method and conditional entropy is proposed [8], where two functions enhance the penalty term of least squares. Meanwhile, Thakur et al. [9] suggested a new model for selection features. Then, they utilized the deep learning method for classification, where the data points are from two different distributions - one generic to all network intrusions and the other specific to the domain. An anomaly IDS is suggested [10], where this system is bone by utilizing RNA encoding and the ResNet50 Model. The encoding is done by dividing records into groups, and RNA characters represent every group; then, the ResNet architecture addresses training challenges. A lightweight architecture for IDS is proposed [11], where this architecture is based on Parallel Deep Auto-Encoder to utilise local and surrounding information in the feature vector, where this method is evaluated based on KDDCup99, CICIDS2017, and UNSW-NB15 datasets. A novel class imbalance processing method for a large-scale dataset is suggested [12]. The proposed method is done by combining the Synthetic Minority Over-Sampling Technique and Gaussian Mixture Model, where IDS integrates imbalanced class processing with a convolutional neural network, and this model is verified based on the UNSW-NB15 and CICIDS2017 datasets.

A novel IDS method using a convolutional neural network and gated recurrent unit in IoT is presented [13], where this model can capture intricate features and learn relational aspects crucial to IoT security. Alqahtani [14] proposed an active IDS based on a Deep Convolutional Neural Network to protect information from intrusion using a Pascal triangle, where this method consists of five convolutional layers. The performance evaluation of this method is done based on the CICIDS2018 dataset. A novel IDS method with information encryption is suggested by [15], where this system is applied to detect intrusion and protect information from intruders. This method has three phases: Intrusion identification using hybrid machine learning and deep learning method, advanced deep learning method for attack detection, and advanced encryption model to secure the information.

A distributed federated IDS method is presented by [16], where the suggested method uses the labelled data information as the prior knowledge to detect new attacks and uses the blockchain method for the federated learning process for the consensus all framework. Qi et al. [17] proposed a novel IDS based on quantum particle swarm optimization and extreme learning machine, and this method starts by suggesting a feature selection method based on partitioned gains, where the achieved results showed that the proposed system has achieved speed detection time and high accuracy. A new features selection method is presented for IDS [18], this method is done based on three steps. The first step of this method is extracting the keys and their positions, then selecting the features using the extracted key positions, and finally, classifying records as normal or an attack. A new online learning method for IDS is presented [19], where the proposed method learns and practices to work within an online environment, the random forest algorithm used to train samples by inserting a new tree to train.

2. METHOD

In this study, a new misuse IDS is presented which incorporates RNA encoding as well as FPM to improve the detection of network intrusions. This approach employs the CICIDS2017 dataset, which is well recognized for its inclusion of current and modern attack types. The proposed method consists of four steps and these steps are illustrated in Figure 2.

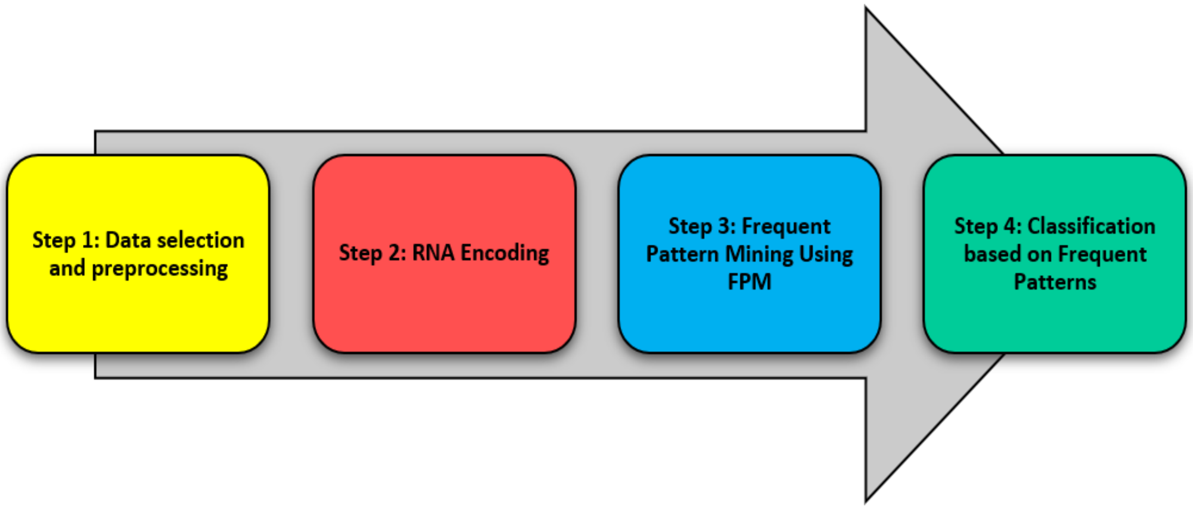
Step 1: Data selection and preprocessing
Step 2: RNA Encoding
Step 3: Frequent Pattern Mining Using FPM
Step 4: Classification based on Frequent Patterns

Figure 2. The proposed IDS method steps

2.1 Data Selection and Preprocessing

The CICIDS2017 dataset is selected for this research because it covers not only DDoS and DoS attacks, FTP-Patator, and Web Attacks among others. Such diversity guarantees the model’s efficiency as an answer to different intrusion schemes. To enhance fairness and achieve a balanced set of attack and benign traffic, a random sample from the dataset is selected. As for the chosen dataset, 70% of the data are used for model training, while the remaining 30% is used for testing.

2.2 RNA Encoding

RNA encoding is employed on the dataset to convert the numerical attribute values into RNA sequences that can be used in pattern extraction. Two RNA characters represent every attribute, and it is shown that all values in the dataset can be represented in such a way. This encoding is useful for the subsequent steps of pattern recognition because the RNA sequences distinguish the most frequently occurring patterns and sequences associated with network attacks.

2.3 Frequent Pattern Mining Using FPM

For extracting the patterns, related to each type of attack, Frequent Pattern Mining (FPM) methods are used. FPM observes periodicity in the RNA-encoded sequence and finds periodicity measures that characterize attack traffic from normal traffic. In this study, the FP-Growth algorithm is adopted due to its ability to cope with large databases and the disadvantage of having to generate candidates, as in the case of the Apriori algorithm. FP-Growth results in a condensed form of the data known as an FP-tree and extracts patterns of a given size, in this case, sequences of six RNA characters.

The nature of the relationships that FPM helps to discover, based on frequent patterns specific to each type of attack, enables the identification of meaningful associations that define malicious activity. These patterns are then analyzed to decide the relative frequency at which they appear in attack traffic and this forms the basis of classification. The frequency and relevancy of the patterns are high in order to employ them for the identification of these patterns in the testing data.

2.3.1 Frequent Pattern Mining Steps for Pattern Extraction

1. **Data Preparation**:

   - First, need to prepare the used dataset, which in this research is RNA-encoded, where all features from network traffic are in a proper format. The encoding process has made numerical values to RNA sequences which can be considered as categorical data in FPM.
   - This step may also involve data sampling so that only such features with high variability within the data set are used in subsequent steps with an assumption that they are regularly involved in attacks.

2. **Pattern Mining Algorithms**:

   - **Apriori Algorithm**: This algorithm is useful for the mining of frequent item sets (sets of feature-value pairs that are frequently conjoint) using the approach of extension of frequent subsets. For instance, in DDoS attacks, if specific features are encoded in RNA, Apriori will recognize these as frequent patterns because they are always associated.
   - **FP-Growth Algorithm**: This is more efficient than the Apriori technique since the various candidate sets are not generated, but instead a data structure referred to as an FP-tree is used. FP-Growth reduces the data and discovers frequent patterns without the need for candidate generation, which is beneficial for large datasets such as CICIDS2017.

3. **Sequence Pattern Mining**:

   - If IDS has to detect sequential behaviours (as many attacks are), then you can use techniques such as Sequential Pattern Mining, which includes Prefix Span. This approach is for finding the most often repeated sub-sequences in the traffic data encoded by RNA, and that is why it is good for detecting multiple-stage attacks, including APTs that consist of a set of different actions.
   - You can concentrate on consecutive RNA sequences that relate to specific sorts of attacks. Thus, it will enable IDS to identify not only the sort of attacks but also the process of attack.

4. **Pattern Filtering and Ranking**:

- After identifying frequent patterns, the results need to be refined to eliminate patterns that are common to both normal and attack traffic, as they are unlikely to help in identifying threats.
- Then, it is able to sort and rank the extracted patterns using support, which is the number of times a pattern has been extracted and confidence, which is the probability of an attack given a pattern. High confidence patterns are good to use for detection because they show a high correlation with attack traffic.

2.4 Classification Based on Frequent Patterns

After the extraction of the frequent patterns, the frequent patterns are used for the classification of the testing subset. Each record within the test data is then analyzed for the occurrence of such frequent patterns in the RNA-encoded sequence. That is why if the sequence contains patterns connected with a definite type of attack, it is considered to be an intrusion. Otherwise, it is benign. This classification approach uses the frequent patterns that are obtained from the training phase, through which the IDS is able to quickly detect known attack behaviours using RNA sequences.

3. RESULTS AND DISCUSSION

A new IDS model is implemented based on frequent pattern mining, where the validation was carried out by using the CICIDS2017 dataset, with assessment criteria being the detection rate, the false positive rate, and the accuracy. Where the obtained DR results for different attack types that appear within the used dataset are shown in Table 1 and Figure 2.

Table 1. Obtained DR results for different attack types

Attack name	Detection Rate
Bot	99
DDoS	82
DoS GoldenEye	99
DoS Hulk	99
DoS Slowhttptest	82
DoS slowloris	71
FTP-Patator	99
Heartbleed	100
Infiltration	88
PortScan	66
SSH-Patator	94
Brute Force	73
Sql Injection	91
XSS	72

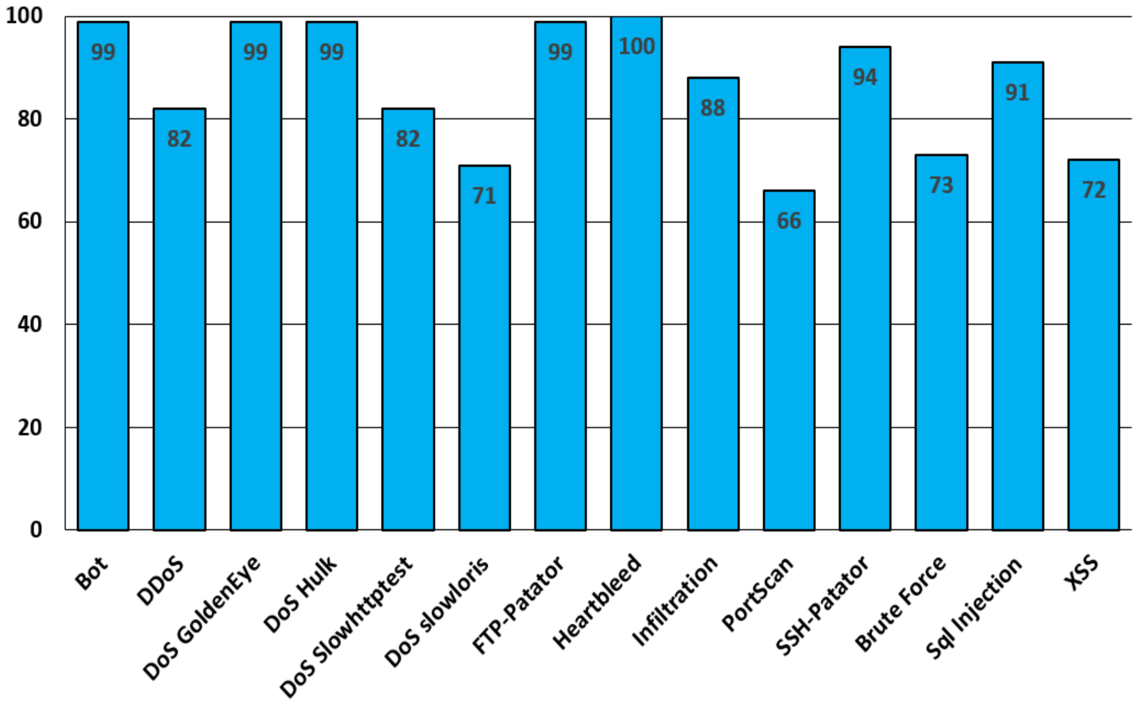

Figure 3. Obtained DR results for different attack types

From Table 1 and Figure 3, the achieved DR results for different attack types are high, where the highest achieved DR result is equal to 100% for Heartbleed. At the same time, the obtained DR, FAR, and accuracy are shown in Table 2 and Figure 4.

Table 2. Achieved results based on the proposed IDS method

IDS Method	Results
Detection Rate	87.3%
False alarm rate	10.1%
Accuracy	88.6%

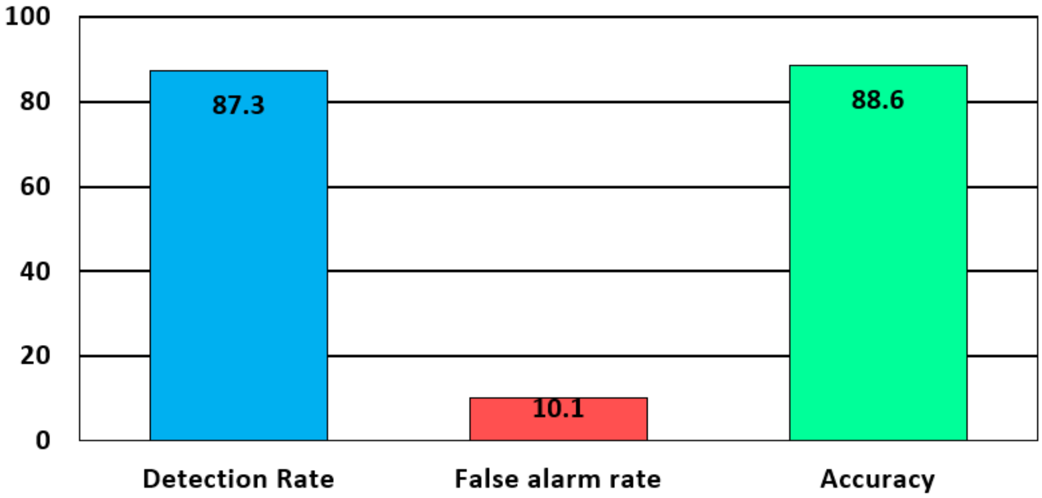

Figure 4. Achieved results based on the proposed IDS method

As shown in Table 2 and Fig. 3, the achieved DR, FAR, and accuracy results are equal to 87.3%, 10.1%, and 88.6% respectively.

4. CONCLUSION

This work presents a new misuse IDS by using RNA encoding and Frequent Pattern Mining to identify network attacks. Where the proposed method is applied based on the CICIDS2017 dataset to detect frequent patterns specific to malicious traffic. The results prove that FPM achieved a high detection rate result, and the ability to be applied in real-time making FPM a feasible solution for detecting multiple threats in a network. For future work, this system can train to use in real time for detecting new threats. Also can used FPM with other data mining methods, such as association rule mining or clustering, that may help to analyse more complicated attack behaviours.

REFERENCES

[1] Malik R., Raza H., and Saleem M., Towards A Blockchain enabled integrated library managment system using hyperledger fabric: using hyperledger fabric, International Journal of Computational and Innovative Sciences. (2022) 1, no. 3, 17–24.

[2] Rashid, O.F.,Othman, Z.A.,Zainudin, S.. A novel DNA sequence approach for network intrusion detection system based on cryptography encoding method. International Journal on Advanced Science, Engineering and Information Technology, 2017, 7(1), pp. 183–189

[3] A. Khraisat, I. Gondal, P. Vamplew and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques datasets and challenges", Cybersecurity, vol. 2, no. 1, pp. 1-22, Dec. 2019.

[4] Malik J. A. and Saleem M., Blockchain and cyber-physical system for security engineering in the smart industry, Security Engineering for Embedded and Cyber-Physical Systems, 2022, CRC Press, Boca Raton, FL, USA, 51–70.

[5] M. Ozkan-Okay, R. Samet, Ö. Aslan and D. Gupta, "A Comprehensive Systematic Literature Review on Intrusion Detection Systems," in IEEE Access, vol. 9, pp. 157727-157760, 2021, doi: 10.1109/ACCESS.2021.3129336.

[6] Menna Gamal, Mohamed Elhamahmy, Sanaa Taha, Hesham Elmahdy, Improving intrusion detection using LSTM-RNN to protect drones' networks, Egyptian Informatics Journal, Volume 27, 2024, https://doi.org/10.1016/j.eij.2024.100501.

[7] Mahendra Prasad, Sachin Tripathi, Keshav Dahal, An efficient feature selection based Bayesian and Rough set approach for intrusion detection, Applied Soft Computing, Volume 87, 2020, https://doi.org/10.1016/j.asoc.2019.105980.

[8] Jing Zhang, Yige Yuan, Jiahong Zhang, Yang Yang, Wenjin Xie, Anomaly detection method based on penalty least squares algorithm and time window entropy for Cyber–Physical Systems, Journal of King Saud University - Computer and Information Sciences, Volume 35, Issue 10, 2023, https://doi.org/10.1016/j.jksuci.2023.101860.

[9] Soumyadeep Thakur, Anuran Chakraborty, Rajonya De, Neeraj Kumar, Ram Sarkar, Intrusion detection in cyber-physical systems using a generic and domain specific deep autoencoder model, Computers & Electrical Engineering, Volume 91, 2021, https://doi.org/10.1016/j.compeleceng.2021.107044.

[10] Subhi, M., Rashid , O. F., Abdulsahib , S. A., Hussein , M. K., & Mohammed , S. M. (2024). Anomaly Intrusion Detection Method based on RNA Encoding and ResNet50 Model. Mesopotamian Journal of CyberSecurity, 4(2), 120–128. https://doi.org/10.58496/MJCS/2024/011.

[11] Amir Basati, Mohammad Mehdi Faghih, PDAE: Efficient network intrusion detection in IoT using parallel deep auto-encoders, Information Sciences, Volume 598, 2022, Pages 57-74, https://doi.org/10.1016/j.ins.2022.03.065.

[12] Hongpo Zhang, Lulu Huang, Chase Q. Wu, Zhanbo Li, An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset, Computer Networks, Volume 177, 2020, https://doi.org/10.1016/j.comnet.2020.107315.

[13] Qaddos, A., Yaseen, M.U., Al-Shamayleh, A.S. et al. A novel intrusion detection framework for optimizing IoT security. Scientific Reports, vol. 14, (2024). https://doi.org/10.1038/s41598-024-72049-z.

[14] Alqahtani, A.S. Deep Convolutional Neural Network for Active Intrusion Detection and Protect data from Passive Intrusion by Pascal Triangle. Wireless Personal Communications (2024). https://doi.org/10.1007/s11277-023-10846-x.

[15] Pradeepthi, C., Maheswari, B.U. Network intrusion detection and prevention strategy with data encryption using hybrid detection classifier. Multimedia Tools and Applications, vol. 83, 40147–40178 (2024). https://doi.org/10.1007/s11042-023-16853-1.

[16] Sun, N., Wang, W., Tong, Y. et al. Blockchain based federated learning for intrusion detection for Internet of Things. Frontiers of Computer Science, vol. 18, 185328 (2024). https://doi.org/10.1007/s11704-023-3026-8.

[17] Qi, H., Liu, X., Gani, A. et al. Quantum particle Swarm optimized extreme learning machine for intrusion detection. The Journal of Supercomputing, vol. 80, 14622–14644 (2024). https://doi.org/10.1007/s11227-024-06022-y.

[18] Rashid O.F., Othman Z.A., Zainudin S. Features selection for intrusion detection system based on DNA encoding. (2019) Lecture Notes in Networks and Systems, 67, pp. 323 - 335, DOI: 10.1007/978-981-13-6031-2_23.

[19] Chuang, PJ., Huang, PY. Enhancing network intrusion detection by lifelong active online learning. The Journal of Supercomputing, vol. 80, 16428–16451 (2024). https://doi.org/10.1007/s11227-024-06070-4.

[20] Zhang, Z., Wang, L., Zhu, J. et al. MIM: A multiple integration model for intrusion detection on imbalanced samples. World Wide Web, vol. 27, (2024). https://doi.org/10.1007/s11280-024-01285-0.