MINAR International Journal of Applied Sciences and Technology
ISSN: 2717-8234
Article type: Research Article

Received: 12/01/2024 **Accepted:** 09/02/2024 **Published:** 01/03/2024

DETECTING ANOMALIES USING UNSUPERVISED MACHINE LEARNING ALGORITHMS

Raya Jasim EASA[1]

University of Mosul, Iraq

Asma Salim YAHYA[2]

University of Mosul, Iraq

Abstract

Unsupervised machine learning is considered more challenging compared to supervised machine learning. This is due to dealing with uncategorized and unlabeled data. Many unsupervised algorithms are available, but the problem is in the sensitivity of the parameters of the algorithms. Moreover, unsupervised machine learning algorithms struggled with time constraints, especially with systems that need real-time processing such as anomaly detection systems. This article proposes a method for anomaly detection in real time. The proposed method used concepts inspired by the DBSCAN algorithm. Modifications were performed on the original version to improve the performance in terms of accuracy and time. The dataset used was MAWI, which is considered a standard in the network security area. The performance of the proposed algorithm was compared to other clustering algorithms such as KNN, K-Means, and PCA as well as the original version of the DBSCAN. The results showed promising aspects of the proposed algorithm because it provides efficient performance in terms of accuracy and time**.**

Keywords: *Anomaly Detection, K-means algorithm, KNN algorithm, Network Security, Unsupervised Machine Learning.*

http://dx.doi.org/10.47832/2717-8234.18.15

[1] raya.jassim@uomosul.edu.iq

[2] asma_alkhairi@uomosul.edu.iq

1. Introduction

1.1. Overview

Networks are currently everywhere and become an important part of life's social activities. These activities are managed by applications that work mostly online. This means data are exposed to unauthorized access by anonymous users and such a case is considered a risk [1]. Therefore, security architects try to design efficient Intrusion Detection (ID) models that can differentiate between normal and abnormal activities within the network [2] as well as hold the main security features of Confidentiality, Integrity, and availability. Abnormal activities are also called anomalies, which may cause risk.

To detect anomalies in a network, several methods can be used. For instance, statistical methods such as regression approaches may contribute to detecting anomalies by building statistical models for that purpose [3-5]. Moreover, machine learning, including deep learning, can also be used to detect anomalous activities within a network [6-8].

The most frequent methods used in designing ID systems are the ones that depend on machine learning and deep learning. These methods can be either supervised or unsupervised. The supervised learning methods are mainly performing classification tasks using labelled data [9]. On the other hand, unsupervised methods perform clustering tasks using unlabeled data [10]. This research focuses on the unsupervised machine learning approaches. In unsupervised learning, data are clustered to be normal or abnormal (anomaly), which makes it easier to detect anomalies (see Figure 1).

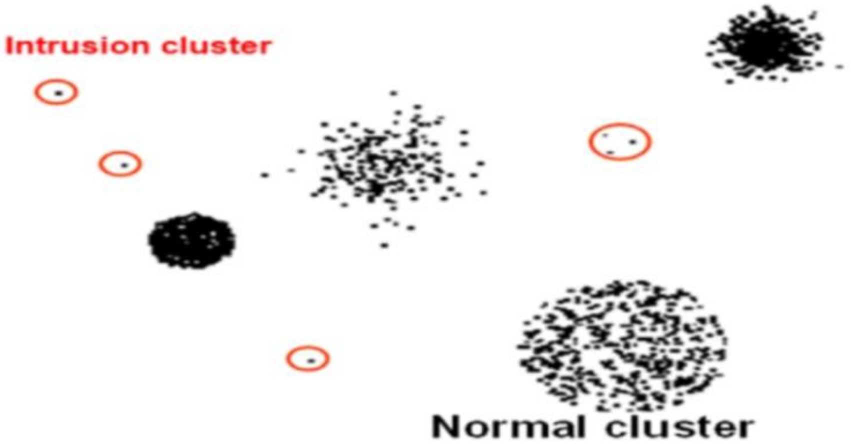

Figure 1: Anomaly detection using unsupervised methods (clustering).

The literature presents many unsupervised research methods. One of the early studies is the study of [11], who suggested a method to detect outliers in big data. The authors used many methods in the detection process. They used six datasets in their work; UCI, Diabetic, Phishing, Banknot, Forest CoverType, NSL-KDD, and Spambase datasets. They used many evaluation metrics such as Purity, Mirkin, and F-measure, which represent the quality of the generated clusters. The algorithms they used in clustering the data were k-means and the proposed batch-based algorithm. They were able to obtain approximately 96% of accuracy.

Another proposed in [12] a novel method that was efficiently able to detect anomalies. The authors used deep learning concepts in designing the method. They merged both CNN and clustering loss methods and called it Top-K DCCA. They evaluate their proposed method with some methods in the literature using the F-score metric. The comparison showed the superior performance of the proposed method when involving datasets of industrial processes.

[13] proposed a method that combines supervised and unsupervised methods to detect anomalies in a network. Their method was supported by concepts inspired by parallel computing. The dataset used was NASA logs with the XGBoost model. The results showed the efficiency of their proposed method in detecting anomalies. The normal events within the network were efficiently clustered.

Furthermore, the selection of the algorithm depends on the dataset and the nature of its attributes. Therefore, some studies tried to compare different methods on different datasets. The study of [14] compared the performance of the k-means clustering when using both; the MAWI dataset and the NSL-KDD dataset. The results showed that the selection of features, the training set, and the quality of the dataset used played a crucial role in the accuracy of the clustering algorithm used.

Dealing with high-dimensional data is considered one of the key factors that affect the performance of the process of detecting anomalies in the field of network security. Therefore, researchers try to reduce the dimensionality of data to deal with the data more adequately and eventually obtain a more accurate detection rate. [15] tried to enhance unsupervised learning in detecting anomalies. They utilized the deep autoencoder (DAE) along with some clustering methods to produce low-dimensional data. Their proposed method can neglect redundant data. The algorithms used with the DAE were DBSCAN, K-Means, and Mean-Shift. The datasets used were Thyroid, Arrhythmia, and Pen_global. The results showed that the proposed update to the mentioned algorithms provided efficient results in terms of detecting anomalies.

Another study performed by [16] proposed a real-time method called (RE-ADTS) for detecting anomalies in time series data. They used 52 datasets to test the proposed method. The metrics used in evaluating the performance of the proposed method were recall, precision, and F-measure.

In a study performed by [17] ,various data stream methods were utilized on the CICIDS2017 datasets, encompassing multiple novel forms of attacks. The optimal algorithm that meets the requirements of high accuracy and short computation time was selected after the results. Moreover, the UNSW-NB15 dataset and CNN are applied by [18] to create a supervised network to save time and money, Recursive Feature Elimination (RFE) and Extreme Gradient Boosting (XGB). Also, bias toward the dataset's majority class is lessened via the Bayesian Gaussian Mixture Model (BGMM) and Synthetic Minority Oversampling

Technique (SMOTE) with a 98.80% accuracy rate for binary classification and a 96.49% accuracy rate for classification into multiple categories, including the data demonstrate that this model outperforms existing techniques.

Based on the above literature, it can be seen there are many methods available to deal with different applications and data types. Also, no stable or standard method can be adopted by researchers. Moreover, most of the methods in the literature struggled with the complexity issue, which means they need time and processing abilities to perform a detection process. However, most of the applications currently work online and are in touch with several networks and it is required to have real-time processing. As a result, it is needed to minimize the complexity as much as possible. Moreover, one of the issues in clustering algorithms is that most of them are sensitive to parameters. Hence, the contribution of this work is to propose a new version of DBSCAN that includes a parameter that can reduce the sensitivity of other parameters and can be used for anomaly detection in real-time. The proposed method used concepts inspired by DBSCAN. The dataset of this work is MAWI, which is considered a standard in the network security area.

The rest of this document is divided as follows: Section 2 presents the proposed method as well as the description of the dataset used and the evaluation metrics. Section 3 demonstrates the experimental results and discussion about them. Section 4 will provide the conclusions of this work.

2. Research Methodology

This section provides a detailed description of the dataset used in this work and the method followed as well as the evaluation metrics. The general workflow of this article is summarized in Figure 2.

2.1. MAWI dataset

This dataset originated by MAWILab [19] and is widely used for testing and validating anomaly detection approaches. It includes network traffic data for the years 2007 to 2023 and each year includes the monthly traffic data. This dataset includes all the information needed to perform the process of anomaly detection as described in [20]. For instance, Figure 3 shows the anomalies in the dataset for the years 2007 to 2023. The figure shows the fluctuations of anomalous behaviour in the dataset.

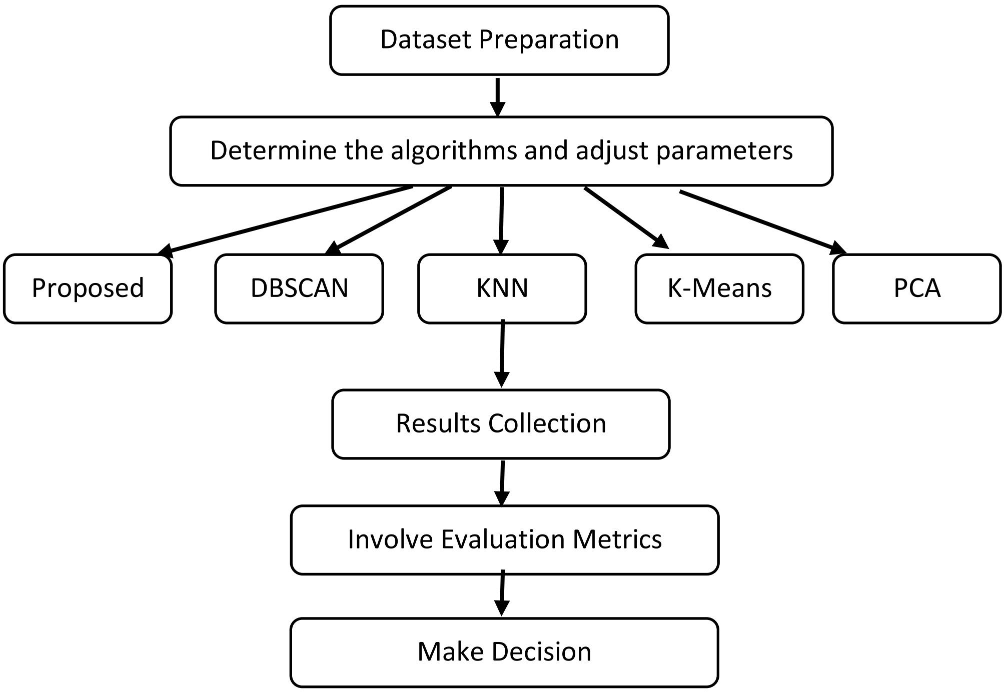

Figure 2: Workflow diagram.

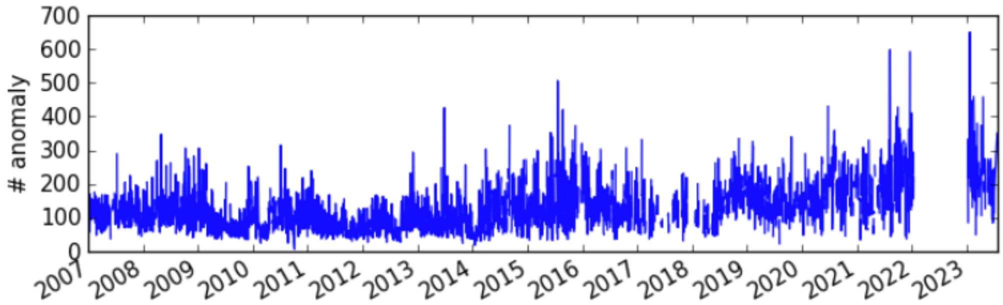

Figure 3: Anomalies in MAWI dataset (2007-2023).

2.2. Proposed Method

The proposed method is based on the DBSCAN algorithm [20], which is a popular clustering algorithm that is classified as unsupervised. The reason behind choosing this algorithm is that DBSCAN behaves similarly to the human way of thinking and can deal with unlabeled data. The DBSCAN works based on two parameters:

- ϵ, controls the decision of a data point to be in a particular cluster or core point (core points are selected earlier). The distance between a point and the core point will be compared to the value of ϵ, if the difference is less than or equal to ϵ then the point is considered a neighbour otherwise it will be discarded or considered an outlier.

- MinPts represents the minimum number of neighbours and depends on the dimensions of the data where MinPts is less or equal to (dimension +1), which is most likely to be at least 3.

The pseudocode of the DBSCAN can be illustrated in the following pseudocode, and Figure 4 demonstrates the procedure of the algorithm.

```
DBSCAN (dataset, eps, MinPts)
{
 # cluster index
 C=1
 For each unvisited point p in the dataset {
 Mark p as visited
 # find neighbours
 Neighbours N = find the neighbouring points of p
 If |N| >= MinPts:
 N = N U N'
 If p' is not a member of any cluster:
 Add p' to cluster C }
}
```

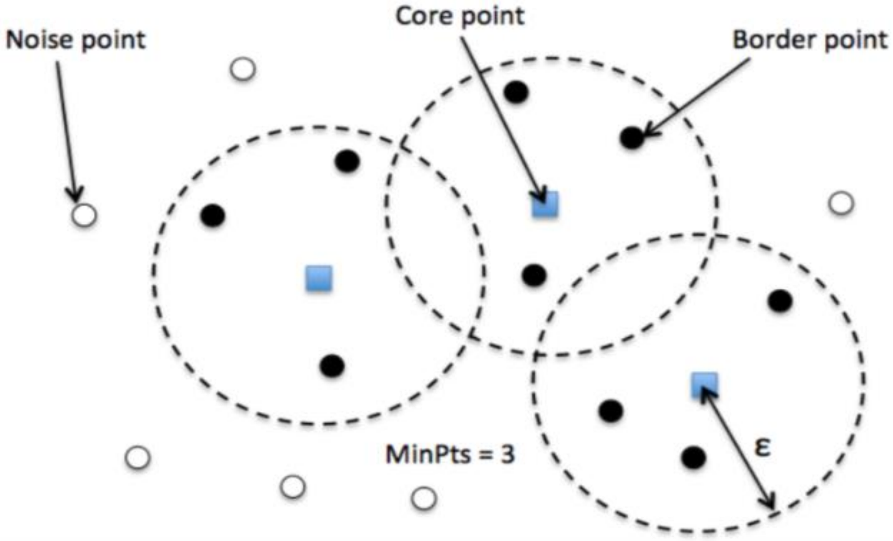

Figure 4: The procedures followed in the DBSCAN algorithm.

After describing the DBSCAN algorithm, we present the proposed modification of it. The new version of the DBSCAN is similar to the original with one main difference, that is, we add a parameter that can reduce the sensitivity of the other two parameters. This issue in DBSCAN is frequently faced by researchers. Therefore, we propose to add a parameter called *estimated core points* (*ECP*) that will represent the number of estimated clusters, which is in our case 3 (normal, anomalous, and outlier) as shown in Eq. 1.

$$ecp = \begin{Bmatrix} Normal\,Activity \\ Anomalous\ Activity \\ Outlier \end{Bmatrix} (1)$$

For more accuracy, we propose that the new version of DBSCAN will include several core points equal to *ecp*^{2}, which in this case is 9. This means the number of stages for clustering (*STC*) will be:

$$stc = \frac{ecp^2}{ecp} \qquad (2)$$

This modification was tested for a variety number of *ECP* and different test datasets and the results were promising. Therefore, this modification is considered powerful in terms of improving the accuracy of the generated clusters.

The proposed algorithm is benchmarked with some similar algorithms in the literature such as K-means [21], KNN [22], and PCA [23] as well as the original DBSCAN algorithm.

2.3. Assessment Metrics

To evaluate the performance of the proposed and the benchmarking algorithms, metrics were used as follows:

1- Accuracy: it measures the accuracy of the clustering algorithm and can be calculated as follows:

$$Accuracy = \frac{DP}{N}\ X\ 100\ (3)$$

Where DP denotes the detected points, and N is the total number of points.

2- Time: represents the time consumed in performing the clustering process.

3. Results and Discussions

The results of implementing the proposed algorithm and the benchmarking can be divided into two portions accuracy and time.

3.1. Accuracy Assessment:

The accuracy of the algorithms implemented in this article is demonstrated in Figure 5. The performance shows that there is an enhancement in the performance of the DBSCAN compared to the proposed algorithm. However, the difference is too significant but noticeable. It is also clear that the proposed algorithm outperforms the other clustering algorithms in terms of accuracy.

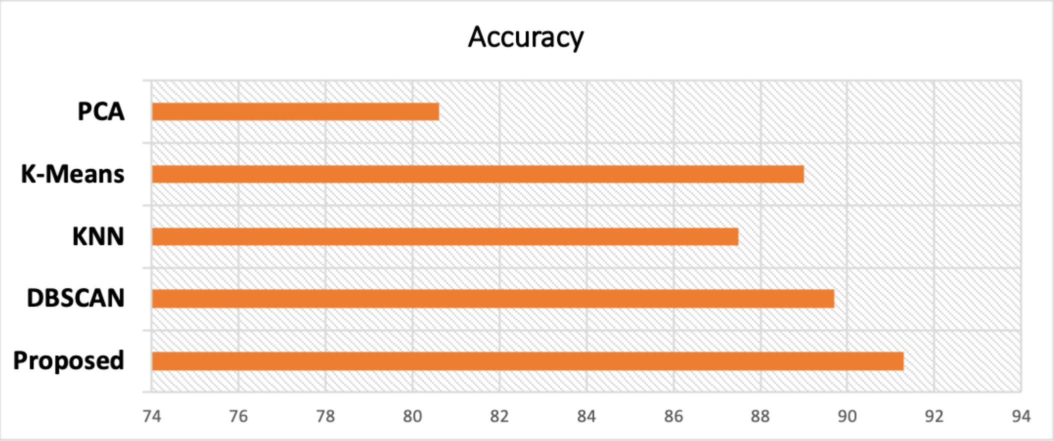

Figure 5: The performance in terms of accuracy of the algorithms implemented in this work.

3.2. Time:

The performance of the proposed algorithm and the other implemented algorithms in this work in terms of time consumption is shown in Figure 6. The figure shows that the proposed algorithm slightly underperformed the original version of the DBSCAN algorithm. This is considered a side effect of the newly added parameter, which makes it a little more complex and costs more time during the processing. However, compared to the other algorithms in this work, the proposed one showed efficient performance.

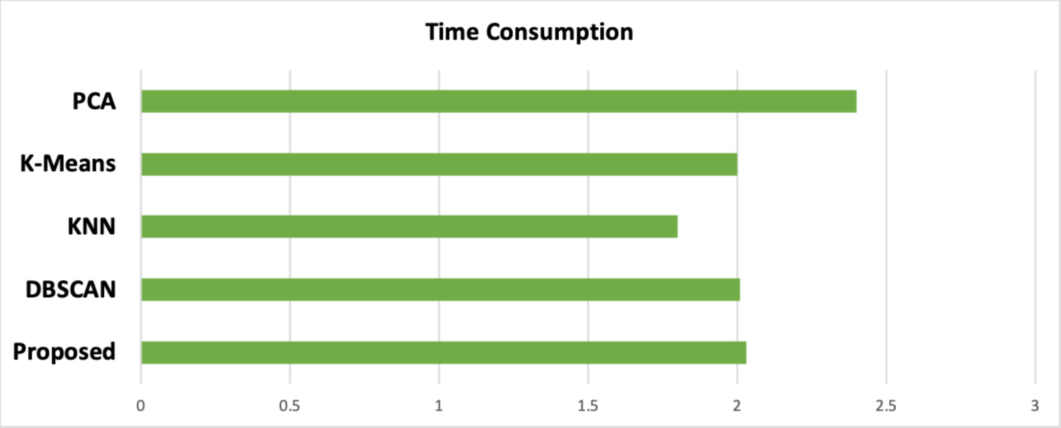

Figure 6: The performance in terms of time consumption of the algorithms implemented in this work.

A comparison between the accuracy of the algorithms implemented and demonstrated in Figure 5, and the performance of the proposed algorithm and other algorithms regarding the time consumption as demonstrated in Figure 6. The performance highlighted that there is an enhancement in the performance of the DBSCAN compared to the proposed algorithm. However, the difference is too significant but noticeable. It is also clear that the proposed algorithm outperforms the other clustering algorithms in terms of accuracy. In contrast, the performance of the proposed algorithm and the other implemented algorithms in this work in terms of time consumption is shown in Figure 6. The figure shows that the proposed algorithm slightly underperformed the original version of the DBSCAN algorithm. This is considered a side effect of the newly added parameter, which makes it a little more complex and costs more time during the processing. However, compared to the other algorithms in this work, the proposed one showed efficient performance.

As can be seen in the obtained results (accuracy and time), the total performance of the proposed algorithm showed efficient performance. However, more investigation is required to assure the efficiency of the proposed algorithm such as using more datasets and investigating the parameters more deeply. Finally, the investigation should include more evaluation metrics aiming to have more concrete results and make the proposed algorithm reliable when used by researchers and developers.

4. Conclusions

In this article, it has proved that a modified version of the DBSCAN algorithm for anomaly detection in real-time. The modifications on the original version of the DBSCAN were performed to enhance the efficiency of accuracy and time. The dataset used was MAWI, which is considered a standard in the network security area. Thus, the results showed promising aspects of the proposed algorithm because it provides efficient performance in terms of accuracy and time compared to the original version of the DBSCAN, KNN, K-Means, and PCA. Accordingly, the results presented effective work based on the proposed algorithm as long as it offers efficient performance in accuracy and time compared to the DBSCAN, KNN, K-Means, and PCA. Future work is planned to involve more datasets to more accurately test the proposed algorithm. Also, more modifications can be performed on the current version of the DBSCAN algorithm.

References

[1] K. DeMedeiros, A. Hendawi and M. Alvarez, “Survey of AI-based anomaly detection in IoT and sensor networks,” *Sensors*, vol. 23, no. 3, pp. 1352, Jan. 2023, doi: 10.3390/s23031352.

[2] L. Cerdá-Alabern and G Iuhasz, “Dataset for Anomaly Detection in a Production Wireless Mesh Community Network,” Data in Brief, vol. 49, pp. 109342, June 2023. doi: 10.1016/j.dib.2023.109342.

[3] N. Moustafa and J. Slay, “The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set,” Information Security Journal: A Global Perspective, vol. 25, no. 1-3, pp. 18-31, 2016, doi: 10.1080/19393555.2015.1125974.

[4] M. Nooribakhsh and M. Mollamotalebi, “A review on statistical approaches for anomaly detection in DDoS attacks,” Information Security Journal: A Global Perspective, vol. 29, no. 3, pp. 118-133, 2020, doi: 10.3390/s21206886.

[5] W. Jia, R. M. Shukla and S. Sengupta, “Anomaly detection using supervised learning and multiple statistical methods,” in 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1291-1297, 2019, doi: 10.1109/ICMLA.2019.00211.

[6] Y. Xu, L. Zhang, B. Du and L. Zhang, “Hyperspectral anomaly detection based on machine learning: An overview,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 3351-3364, 2022, doi: 10.1109/JSTARS.2022.3167830.

[7] B. Pranto, H. A. Ratul, M. Rahman, I. Jahan and Z. B. Zahir, “Performance of machine learning techniques in anomaly detection with basic feature selection strategy - a network intrusion detection system,” Journal of Advances in Information Technology, vol. 13, no. 1, pp. 36-44, 2022, doi: 10.12720/jait.13.1.36-44.

[8] X. Hu, C. Xie, Z. Fan, Q. Duan, D. Zhang, L. Jiang, ... and J. Chanussot, “Hyperspectral anomaly detection using deep learning: A review,” Remote Sensing, vol. 14, no. 9, pp. 1973, 2022, doi: 10.3390/rs14091973.

[9] A. B. Nassif, M. A. Talib, Q. Nasir and F. M. Dakalbab, “Machine learning for anomaly detection: A systematic review,” IEEE Access, vol. 9, pp. 99, 2021, doi: 10.1109/ACCESS.2021.3083060.

[10] Y. Chang, Z. Tu, W. Xie and J. Yuan, “Clustering driven deep autoencoder for video anomaly detection,” in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV, pp. 329-345, Springer International Publishing, 2020, doi: 10.1007/978-3-030-58555-6_20.

[11] R. Alguliyev, R. Alguliyev and L. Sukhostat, “Anomaly detection in Big data based on clustering,” Statistics, Optimization & Information Computing, vol. 5, no. 4, pp. 325-340, 2017, doi: 10.19139/soic.v5i4.365.

[12] G. S. Chadha, I. Islam, A. Schwung and S. X. Ding, “Deep convolutional clustering-based time series anomaly detection,” Sensors, vol. 21, no. 16, pp. 5488, 2021, doi: 10.3390/s21165488.

[13] J. Henriques, F. Caldeira, T. Cruz and P. Simões, “Combining k-means and xgboost models for anomaly detection using log datasets,” Electronics, vol. 9, no. 7, pp. 1164, 2020, doi: 10.3390/electronics9071164.

[14] O. I. Al-Sanjary, M. A. Roslan, R. A. Helmi and A. A. Ahmed, “Comparison and detection analysis of network traffic datasets using K-means clustering algorithm,” Journal of Information & Knowledge Management, vol. 19, no. 3, pp. 2050026, 2020, doi: 10.1142/S0219649220500264.

[15] C. Zhang, J. Liu, W. Chen, J. Shi, M. Yao, X. Yan, ... and D. Chen, “Unsupervised anomaly detection based on deep autoencoding and clustering,” Security and Communication Networks, vol. 2021, pp. 1-8, 2021, doi: 10.1155/2021/7389943.

[16] T. Amarbayasgalan, V. H. Pham, N. Theera-Umpon, and K. H. Ryu, “Unsupervised anomaly detection approach for time-series in multi-domains using deep reconstruction error,” Symmetry, vol. 12, no. 8, pp. 1251, 2020, doi: 10.3390/sym12081251.

[17] A. A. Abdulrahman and M. Kh. Ibrahim, “Intrusion Detection System Using Data Stream Classification,” Iraqi Journal of Science, vol. 62, no. 1, pp. 319-328, 2021, doi: 10.24996/ijs.2021.62.1.30.

[18] W. F. Kamil and I. J. Mohammed, “Adapted CNN-SMOTE-BGMM Deep Learning Framework for Network Intrusion Detection using Unbalanced Dataset,” Iraqi Journal of Science, vol. 64, no. 9, pp. 4846-4864, 2023, dio: 10.24996/ijs.2023.64.9.43.

[19] R. Fontugne, B. Borgnat, P. Abry. and K. Fukuda, "MAWILab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking," in Proceedings of the 2010 ACM Conference on Emerging Networking Experiments and Technology, CoNEXT 2010, Philadelphia, PA, USA, 2010.

[20] K. Khan, S. U. Rehman, K. Aziz, S. Fong and S. Sarasvady, “DBSCAN: Past, present and future,” in the fifth international conference on the applications of digital information and web technologies (ICADIWT 2014), pp. 232-238, IEEE, May 2014.

[21] J. A. Hartigan, and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” Journal of the Royal Statistical Society. series c (applied statistics), vol. 28, no. 1, pp. 100-108, 1979.

[22] G. Guo, H. Wang, D. Bell, Y. Bi and K. Greer, “KNN model-based approach in classification,” in On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, Proceedings Springer Berlin Heidelbergm, pp. 986-996, 2003.

[23] K. Y. Yeung, and W. L. Ruzzo, ”Principal component analysis for clustering gene expression data,” Bioinformatics, vol. 17, no. 9, pp. 763-774, 2001.