International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Mar-24)

http://dx.doi.org/10.12785/ijcds/150177

Applying Hybrid Clustering with Evaluation by AUC Classification Metrics

Ali Fattah Dakhil[1], Waffaa M. Ali[2] and Mustafa Asaad Hasan[3]

[1]*College of Human Science, University of Thi-Qar, Thi-Qar, Iraq*
[2]*College of Veterinary Medicine, University of Al-Shatrah, Thi-Qar, Iraq*
[3]*University of Thi-Qar, Thi-Qar, Iraq*

Received 15 Aug. 2023, Revised 13 Feb. 2024, Accepted 19 Feb. 2024, Published 1 Mar. 2024

Abstract: Traditional metrics may not adequately assess performance in certain situations, whereas the Area Under Curve (AUC) offers a comprehensive perspective by considering both sensitivity and specificity. This method enhances interpretability, addresses limitations, and promotes the development of robust clustering algorithms. In unsupervised learning, utilizing AUC is a significant method for improving the precision and accuracy of machine learning models. Our work is inspired by several recent related works that implement approaches to manage the challenges of developing new metrics that can effectively assess and evaluate the performance of clustering algorithms. The research question relies on the concept of using an optimal metric for model evaluation of classification and clustering. Therefore, the paper investigates the use of the classification metric AUC for clustering validation purposes. The methodology we adopt is a hybrid clustering model because such a technique offers a robust model by combining the strengths of each model. The linkage approach directly impacts the clustering results, so we give significant attention to this feature in our implementation. Among the various linkage methods, we utilized single and average linkages. The Manhattan and Euclidean metrics are the distance measures used in this work. Thus, our contribution is to explore the benefit of using linkages and distance measurement in clustering with the help of the AUC metric. In addition, the entire proposed work and the contributions of this paper are evaluated and applied to the NSL-KDD dataset. Based on the proposed approach of using AUC with clustering, the Detection Rate (DR), False Alarm Rate (FAR), and other criteria are chosen to examine the model's results and capabilities.

Keywords: Hybrid, classification, Clustering, Evaluation Metrics, AUC, Linkages, Distance

1. INTRODUCTION

In the evaluation of clustering techniques, the roles of linkages and distance measurements are critically connected and share the same purpose, serving as the backbone for accurate data clustering. Based on the fact that distance measurements, such as Euclidean, Manhattan, and Cosine similarity, determine the similarity between items in a dataset [1], the choice of the distance measurement method directly affects the clustering outcome by influencing how the similarity between items is quantified. In hierarchical clustering, linkage methods include single, complete, and average methods; this process is further refined by determining how distances between clusters are estimated, affecting the ultimate results and accuracy of the clusters. The choice of distance metric affects clustering algorithms' precision, efficiency, and accuracy. The distance measurements have various performances by considering the data's specifications in their nature. Consequently, we are taking into account the clustering parameters that interchange with distance measurements. In this work, we consider the different distance measures and area under the curve (AUC) metrics on various clustering techniques to assess their performance on both artificial and real-life datasets. We show that the choice of a distance measure should be based on the dataset and clustering technique. Using AUC in clustering by recent works [2], [3], such as the area under the curve for clustering (AUCC) and the area under precision-recall curve, have been proposed as suitable clustering validation indexes (CVIs), contributing to developing a complete metric for supervised and unsupervised learning. This comprehensive approach is valuable for evaluating clustering algorithms as it is applied through the use of AUC classification metrics. In addition, leveraging both distance and linkage techniques alongside AUC is essential for enhancing the clustering performance and reliability. In addition, it underscores the complicated balance and importance of these elements in the clustering validation process. Specificity and sensitivity, which are provided by AUC, guarantee the potential of supplying comprehensive testing and evaluation for model performance. The proposed technique offers a complete solution for some limitations and shortcomings of the traditional metrics. Both Manhattan and Euclidean with single

E-mail address: ali.fattah@utq.edu.iq, waffaa@shu.edu.iq, mustafa.alkhafaji@utq.edu.iq

and average linkages have influenced this work to apply similarity and dissimilarity processes in more complicated datasets. The challenge in this study is that the clustering algorithms' nature is looking for data grouping based on shared features, whereas classification follows pre-classified datasets. Thus, the use of metrics interchangeable between them is the motivation of this paper. Common methods like k-means and hierarchical clustering, each being beneficial for specific purposes, have been the focus of much research [4]. This study innovates by merging these algorithms into a hybrid hierarchical clustering method, aimed at enhancing efficiency and accuracy, especially in anomaly detection scenarios [5]. Hybrid clustering models, known for their adaptability and robust handling of noise and outliers, offer an advanced approach to data analysis [6]. Yet, their complexity and the risk of overfitting necessitate thorough validation and understanding of dataset characteristics. Our research is particularly inspired by recent advancements in evaluating clustering outcomes. A novel approach, the AUCC, extends the AUC metric from supervised to unsupervised learning, providing a valuable measure for clustering performance [7]. With the help of AUC visualization capability, we can easily split the clustered groups and draw a receiver operating characteristic (ROC) curve for graphical distinction. At the same time, the interpretation of AUC values describes the model performance. In Figure 1-a, the process of merging clustering and classification metrics techniques is illustrated showing the conceptual idea of the proposed work. On the other side, Figure 1-b depicts the hybrid model of the k-means and agglomerative models and workflow of the data process. The implementation of the proposed work depends on the NSL-KDD dataset, which is an updated version of the KDD Cup 1999 dataset. The implementation of the chosen dataset is for intrusion detection as a best practice for clustering [8]. Comprising 42 features across various attack categories, this dataset provides a comprehensive platform for our study. Despite its advantages over its predecessor, the NSL-KDD dataset does present certain limitations, such as outdated information and incomplete attack type coverage [9][10]. Recognizing the inadequacy of traditional classification evaluation metrics like confusion matrices and log loss in clustering contexts, our study focuses on alternative evaluation methods [11]. Clustering algorithms, being unsupervised learning models, necessitate metrics that assess inherent patterns and similarities without a predefined target variable. We explore both internal metrics, such as Within Sum of Squares (WSS) and Silhouette Score, and external metrics, like the Adjusted Rand Index (ARI), to evaluate our hybrid clustering model [12]. Thus, this paper contributes to the field by developing and evaluating an anomaly prevention system using hybrid clustering methods on the NSL-KDD dataset. The efficiency and effectiveness of our model are assessed using AUC alongside other relevant metrics. The following section delves into related work, providing a thorough analysis of the current research landscape. Eventually, our paper focuses on three main research question and/or research gaps

- What are the metrics used in clustering and classification?
- limitations of multi-class and imbalanced datasets?
- AUC efficiency in classification and clustering?

The upcoming sections are as follows: The Literature Review comprehensively integrates and critically evaluates existing research related to the study's objectives, identifying gaps and the state of the art in the field. We aim to broaden the scholarly conversation, demonstrating how it addresses previously unexplored questions or builds upon existing knowledge. Next is the Methodology section, in which we clarify the adopted approach, including the dataset and algorithms used. This is followed by the Results and Discussion sections, where we showcase the proposed model's performance and discuss it in view of related work. Finally, we provide a proper conclusion and future work enhancement recommendations to overcome any limitations and challenges.

2. LITERATURE REVIEW

Data clustering has attracted the attention of numerous researchers. Several studies have been discussed by various authors in the literature [13]. The reviewed literature, published within the last eight years, appears in the most qualified journals. Our review process focuses on various aspects related to our main topic. Firstly, we investigate the use and fine-tuning of classification and clustering metrics for optimal performance evaluation. Secondly, we explore how AUC can be utilized for clustering validation, specifically. Thirdly, identify the challenge of using datasets that are primarily used for clustering purposes and applying them for AUC. Fourthly, we consider all the parameters and metrics that are involved in clustering validation and performance evaluation, such as linkages and distance measurements. Fifth, we also review research on hybrid models used in intrusion and anomaly detection. Thus, these criteria are the main subjects that are related to our proposed topic by which we try to find research gaps and motivation works. The review process we follow aims to identify a few key findings in each paper, such as the data used, models, metrics, results, challenges, resource requirements, time complexity, limitations, and future directions. Therefore, in this section, we analyze literature that focuses on the use of metrics in classification and clustering models.

The authors [14] utilize the Follower-Leading Clustering Algorithm (FLCA) for bibliometric analysis. FLCA is capable of clustering large datasets of bibliographic information to discern patterns and trends, which is crucial in fields such as medicine for synthesizing research findings. Work by [15] proposes a taxonomy for non-binary evaluation metrics, specifically focusing on anomaly scores. This taxonomy may provide insights into the differences and applications of these metrics in various scenarios. Also a review [16], focusing on hierarchical agglomerative clustering, k-means clustering, and mixture models, addresses crucial topics for cluster analysis practitioners. It likely delves into the statistical foundations, challenges, and recent developments in these clustering methods. [17] introduces the partitioning Davies-Bouldin index, a novel method for

evaluating clustering. It discusses the approach, its advantages, and potential applications across various clustering scenarios. A comprehensive overview by [18] of online clustering algorithms covers their evaluation methods and metrics. It also explores the applications of these algorithms and presents a benchmarking study, providing insights into their practical effectiveness and areas of application. [19] introduced a novel k-means algorithm optimized for climate data mining, accompanied by a new framework for evaluating clustering uncertainty. It highlights the algorithm's effectiveness in processing structured data and discusses its potential as a new standard in k-means clustering. [20] proposed a study to develop an overlapping community discovery approach based on the concept of local optimal expansion cohesion. In the experiments, the proposed algorithm produced significantly better results than the other algorithms, and the community structure following the division was much more acceptable. Also, [21] based on agglomerative algorithm researchers developed to identify potential sectors owned by a region by combining hybrid algorithm clustering and Location Quotient. As a result of this study, 279 PDRB/GDRP sectors were classified into two main groups: potential sectors of a region with 125 sectors, and the remaining 154 sectors were included in the lower sector for non-potential regions. Authors at [22] performed a comparison of OPTICS, DBSCAN, and the objective clustering technology, which is also presented in their paper. Based on the results of the simulation, the proposed technique appears highly effective.

Density-based algorithms were found to be more effective than agglomerative hierarchical algorithms in simulations [23]. By HCPE, agglomerative hierarchical clustering based on performance evaluation, the authors presented a method for reducing the number of models for high-order dynamical systems. In the study [24], researchers employed Support Vector Machines (SVM) to identify outliers in the KDD dataset. Utilizing the same dataset, authors in [25] developed IDS models using deep learning-based artificial neural networks, surpassing state-of-the-art detection accuracy and false alarm rate methods. Studies [26] and [27] proposed anomaly detection using decision trees and random forests (RF), demonstrating the effectiveness of a decision tree classifier for reliable intrusion detection on two datasets. Consequently, the True Positive Rate (TPR), False Positive Rate (FPR), and Detection Rate (DR) saw significant improvements using this method (FAR). The authors of [28] successfully identified four attacks in the KDD dataset with minimal false positives and negatives using a four-layered classification approach. Furthermore, they introduced a technique to simplify the method by reducing the number of features in the original dataset, potentially improving accuracy while decreasing complexity. However, they could have addressed any labeling errors that may have led to inaccurately categorized attacks. We employed various supervised, unsupervised, and outlier learning methods on the KDD dataset to address misclassified data, but our overall accuracy was lower than that in [29]. Classification and anomaly clustering methods based on machine learning approaches are widely applied to the KDD dataset, comprising four attacks with distinct traffic patterns. In [30], attack-type categorization using the KDD dataset was achieved with a low misclassification rate. Nonetheless, these models require adaptation for contemporary multi-cloud environments where threats are continually evolving and interrelated. Moreover, concerns have been raised about the KDDcup99 dataset's ability to accurately represent everyday network activity [31]. The KDDCUP '99 IDS dataset was used for data mining with support vector machines (SVM) to perform neural network-based categorization. In a 10-fold cross-validation experiment, the accuracy reached 90% for the training set but only 80% for the test set. Various clustering and classifier methods, including unsupervised cluster analysis approaches, have been described in intrusion detection system literature. However, the inaccurate clustering of specific datasets has hindered the effectiveness of attack detection systems. The paper [32] presents a new binary classification method that uses fewer attributes and emphasize the importance of evaluation metrics like MCC, ROC-AUC, and AUC-PR in assessing algorithm performance, especially with unbalanced data. It reveals that MCC outperforms both ROC-AUC and AUC-PR in imbalanced datasets and identifies random forest and gradient boosting as the top algorithms for bankruptcy prediction. However, the methodology only evaluates six machine learning models and three metrics, which may limit its generalizability and overlook other relevant metrics. In addition, it is specifically tailored for bankruptcy prediction, which may restrict its use in other areas, and doesn't account for the impact of varying hyper parameters or feature engineering on algorithm performance. The reference [33] used three datasets; 64 datasets from the Tabula Muris Compendium, 80 datasets from Julia Handl, and 2D synthetic data generated specifically for this study. This study introduces a novel validation metric of clustering. It combines three new indices AUIPRC, AUPRC, and SAUPRC beside other metrics such as; VRC, UCC, PBM, C/Sqrt(k), SWC, Dunn, DB, and C Index. Various clustering validation metrics were used in this study, including Precision, Recall curves, AUCC, Silhouette Width Criterion, Davies-Bouldin, C Index, Dunn, PBM, Calinski-Harabasz, Point Biserial, and Ratkowsky Lance. The main contribution of this study is the propos of new clustering metrics that deliver better results. The weakness of this work is that it does not explicitly showing the potential drawbacks and constraints of the new proposed metrics.

The paper [34] primarily focuses on evaluating risk prediction using AUC. The results promised and showing enhanced prediction accuracy for hierarchical data. They used linear mixed models to assess the prediction outcomes. The main contribution of this work is addressing the statistical challenge of estimating the AUC variance in the presence of hierarchical data dependency. However, a limitation of such work is that there is a challenge of longitudinal data structures in case of complex correlation. That because methodology's applicability in certain data scenarios which is affects model non-convergence. The authors [35] used

the Wisconsin Public Schools dataset. This dataset was best practice for social prediction challenges. The findings tell that it is crucial working towards resolving systemic inequities in order to improve learning outcomes. It is evident that addressing broader structural issues and inequalities will not lead for long-term success, as solely targeting individual students is insufficient. The AUC is the only sole metrics was used for real time evaluation. The DEWS tool has high rate of false positives for Latino and Black students, which can lead to stigmatization and ineffective interventions. This shows the need of developing more equitable and accurate tools that can effectively support struggling students, without presence biases.

To address these issues, we propose a solution to enhance the accuracy of the Intrusion Detection System (IDS) by developing hybrid models. These models blend conventional detection techniques with machine learning. We focus on creating a hybrid clustering model that seeks the best mix of algorithms and evaluation criteria. Additionally, we examine the effects of the hybrid approach on various parameters. Our paper introduces a comprehensive model to address a gap in existing literature. This model contrasts traditional hierarchical clustering with our novel hybrid approach. We utilize two distance measurements and assess them internally and externally using Area Under the Curve (AUC) metrics.

3. METHDOLOGY

The proposed methodology of this research work adopts the concept of applying the AUC metric for clustering evaluation purposes, involving several steps, each of which contributes to the effective evaluation of clustering models. While AUC is typically used for classification tasks, its application in clustering requires adapting the metric to assess the quality of the clusters effectively. The outline of this process is depicted in 1-a, and 1-b. Our methodology is the implementation of the suggested solution in this work. Therefore, in this section, we will explore the dataset used, the model used, how the metrics are applied, and what the other traditional clustering metrics are.

A. Dataset

The NSL-KDD dataset, an enhanced version of the KDD Cup 1999 dataset, is crucial for anomaly detection, particularly in clustering applications. It consists of 42 features per network connection, categorized into basic, content, and traffic features, which assist in the classification of connections into four attack types: DoS, U2R, R2L, Probing, and normal connections. The dataset is efficiently partitioned into a training set of approximately 125,000 examples and a testing set of nearly 22,500, providing a well-balanced mix of attack types. This feature makes it an ideal choice for the development and evaluation of anomaly detection models.

B. Clustering and AUC

There are a few steps to implement this core mission of the work. First, we applied data preparation techniques such as scaling, feature extraction, dimensionality reduction, and partitioning. Next, we applied the proposed hybrid model, as illustrated in Figure 1-a. Second, in label assignment during clustering, it is essential to note that data points do not come with labels. However, labels are necessary for AUC calculation [36]. Two approaches can be taken to assign labels: one is to use external labels, if available, to validate the clusters; the other approach is to assign labels based on the clustering result, treating each cluster as a separate class [30]. There are a few steps to implement this core mission of the work. First, we applied data preparation techniques such as scaling, feature extraction, dimensionality reduction, and partitioning. Next, we applied the proposed hybrid model, as illustrated in Figure 1-a. Second, it is essential to note that data points do not come with labels during label assignment during clustering. However, labels are necessary for AUC calculation [37]. Two approaches can be taken to assign labels: one is to use external labels, if available, to validate the clusters; the other approach is to assign labels based on the clustering result, treating each cluster as a separate class [38]. Then, measuring the similarity and dissimilarity is essential. Appropriate metrics should be chosen based on the data type and domain to measure these aspects. Pairwise measurements can be made to calculate the similarity/dissimilarity for each pair of data points. Factors like the threshold's value determination method and similarity measures would have an obvious effect on the results of AUC and ROC. Thus, clustering performance will consequently be affected due to these criteria of classification outcome [39]. The fifth step is quite important. In this step, we convert the clustering problem into a binary classification series. Then, we can use similarity/dissimilarity based on the chosen threshold value [40]. The most applicable step in our approach is to merge the classification by clustering data model. Therefore, clustering the data into their groups has been classified by the threshold's value. The next step is to construct the ROC curve after plotting the labeled data points [41]. Eventually, the AUC is computed after finding the FPR (false positive rate) and TPR (true positive rate). A higher AUC indicates a better distinction between similar and dissimilar data points [42]. Evaluating and adjusting the clustering algorithm, parameters, or data preprocessing steps based on the AUC score facilitates iterative model improvement. The paper's methodology focuses on three aspects, depicted by figure 1-b: first, the significant role of linkages in clustering. Linkage in clustering involves evaluating the similarities or differences between data points to form clusters. The main linkage types are single, complete, average, and centroid, each using distinct algorithms and distance measures like Euclidean or Manhattan. The choice of the linkage method is crucial as it shapes cluster structure and affects their size, distribution, and overall quality, impacting the accuracy and interpretability of clustering results. Secondly, as the paper focuses on developing an Intrusion Detection System using clustering and classification techniques, it examines two threshold levels on the NSL-KDD dataset for anomaly detection. The objective is to categorize intrusions in NSL-

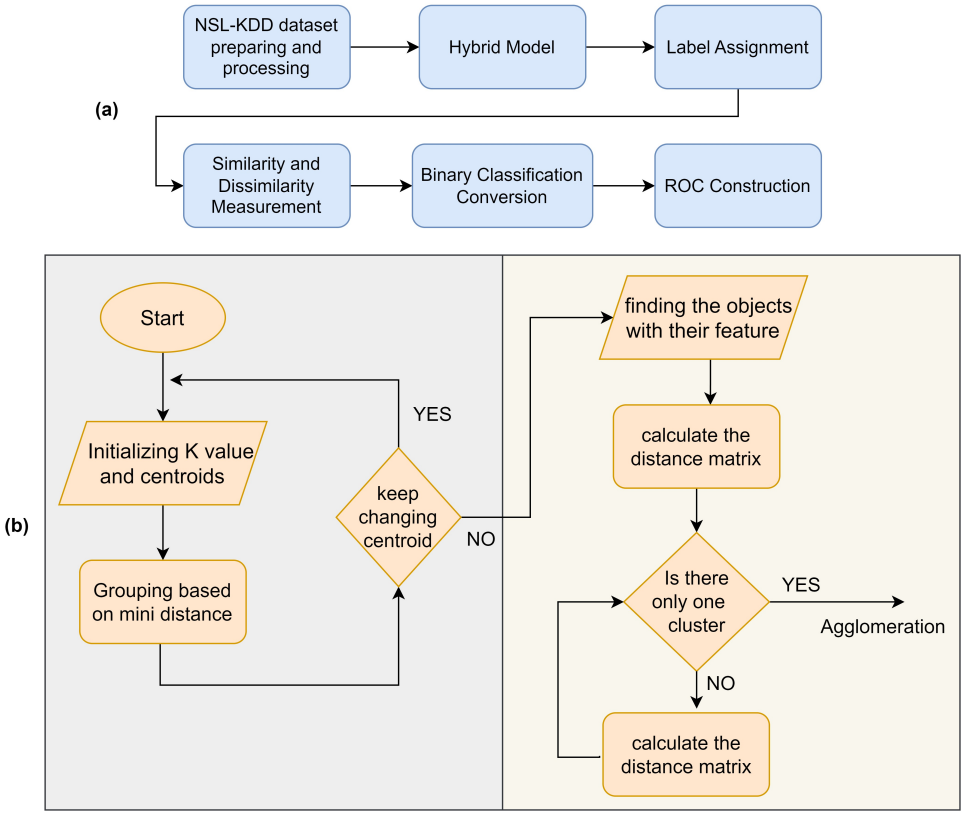

Figure 1. a- binary classification and clustering, b- hybrid model: k-means and agglomerative algorithms

KDD, split into training and testing sets for analysis. An innovative detection method combining these models is introduced and assessed with two thresholds. Evaluation metrics include detection rates, false alarm ratios, accuracy, F1 Score, and AUC. Thirdly, our methodology combines clustering (an unsupervised method grouping data by similarity) with classification (assigning predefined labels to new instances) to improve anomaly detection systems. This blend allows for detecting subtle anomalies and proactive system monitoring, leveraging both unsupervised and supervised learning strengths. We test this hybrid model using k-means and agglomerative algorithms on the D31 dataset, which comprises synthetic, multivariate data with 3100 samples across 31 categories, commonly used in clustering research [38]. The hybrid model is tested by k-means and agglomerative algorithms on the D31 dataset with 3100 synthetic samples over 31 categories.

C. Hybrid Model

Figure 1-b depicts a hybrid clustering model that combines k-means and agglomerative algorithms. The proposed hybrid model is illustrated by Figure 1-b which combines k-means and agglomerative algorithms. This process implemented by seven steps; initialize k and centroids, assign objects to clusters, update centroids (for each cluster, calculating new centroid as the mean of all objects in the cluster), create distance matrix for agglomerative clustering, merge clusters, update distance matrix, then last step is final agglomeration into a single cluster (if needed). It computes a distance matrix to determine cluster similarity and iteratively merges the closest clusters. This continues until a single cluster forms or the desired structure is reached, effectively leveraging both algorithms for improved clustering results. In the following section, we delve into the practical aspects of using AUC metrics to evaluate our hybrid clustering model. We'll examine the effectiveness of AUC in measuring performance, especially in identifying different cluster patterns. The methods for calculating AUC and its role in assessing clustering sensitivity and specificity will be highlighted. Furthermore, we'll discuss interpreting AUC values within the hybrid clustering model framework, which merges various clustering techniques.

4. RESULTS

In our study, the results aligned with the methodology section and demonstrated the validity of our experiments. The use of the D31 Dataset proved to be the best choice for developing a hybrid clustering technique. Figure 2 compares the accuracy of standard hierarchical and hybrid hierarchical methods in terms of the Rand index (an external criterion).

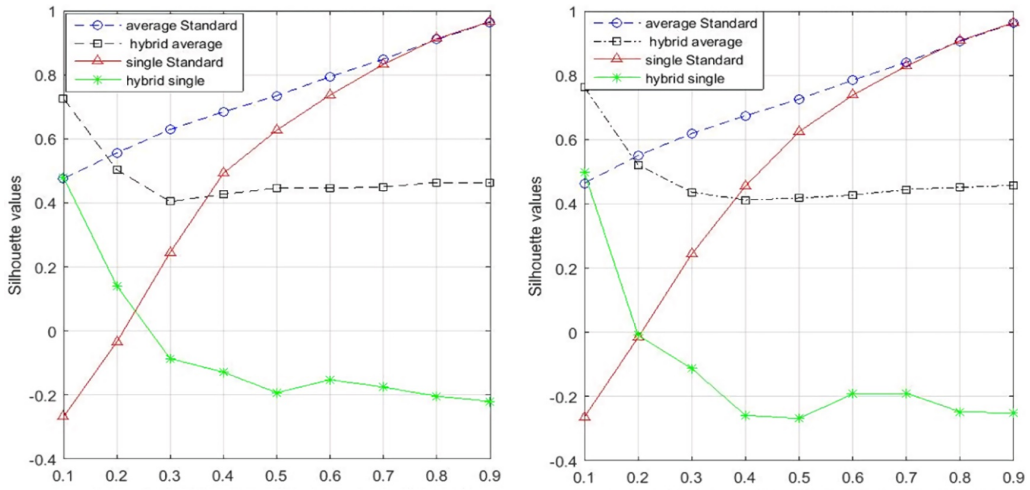

Figure 2. Hybrid and standard hierarchical comparison with Silhouette Index: Euclidean on the left and Manhattan on the right

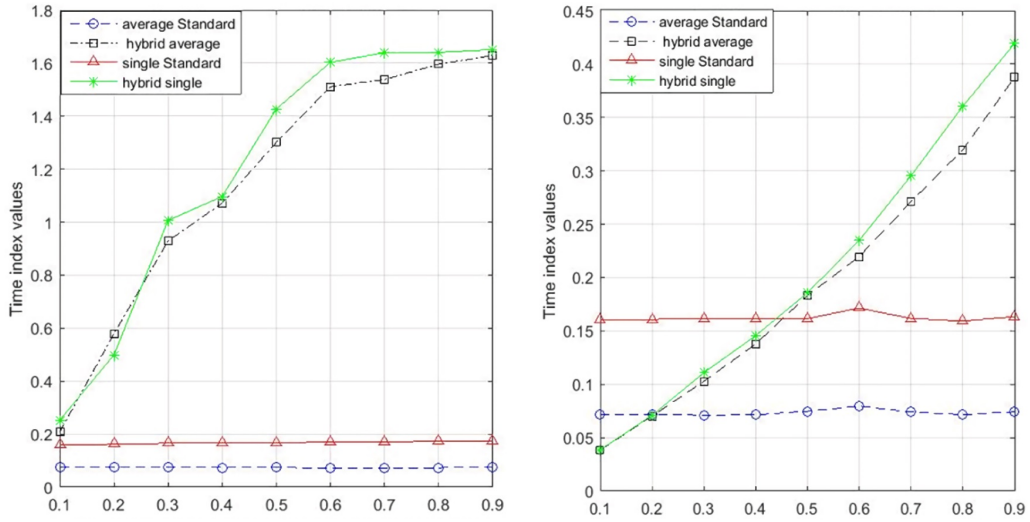

Figure 3. Hybrid and standard hierarchical comparison with time Index: Euclidean on the left and Manhattan on the right

We observed that the hybrid average method outperformed the standard method with both Euclidean and Manhattan distances. Furthermore, the hybrid single linkage displayed superior accuracy compared to the standard single hierarchical method when using Manhattan distance but fell short when using Euclidean distance. Figure 3 illustrates these comparisons on the D31 dataset by showing average and single linkages with Euclidean and Manhattan distances in terms of silhouette accuracy. Lastly, Figure 4 reveals that the standard single hierarchical outperforms hybrid hierarchical when cluster number ratios are above 0.4, except for when the number of clusters is equal to a ratio of 0.1. In this case, standard average hierarchical is superior to hybrid average hierarchical. As seen in this figure, the accuracy of hybrid average hierarchical is better than the accuracy of hybrid single hierarchical with two distances.

Figure 3 illustrates a comparison between the standard hierarchical method and the hybrid hierarchical method in terms of time. Our findings indicate that, in the hybrid hierarchical method (KH), as the number of ratios increases, so does the required computation time. However, the traditional hierarchical method remains unaffected by the ratios, maintaining a constant computation time since it utilizes the same 3100 data points and 35 clusters in each step. We also discovered that hybrid linkage with Euclidean distance takes less time than standard linkage with the same distance when the number of clusters is equal to ratios below 0.4. Conversely, with Manhattan distance, the standard method outperforms the hybrid method in terms of time efficiency for all cluster quantities. Data with lower ratios exhibit reduced computational costs. For D31 data, experiments reveal superior Rand index accuracy for the hybrid average method using Euclidean distance compared to the standard average method. In contrast, minimum values are found in average standard methods using Manhattan and single

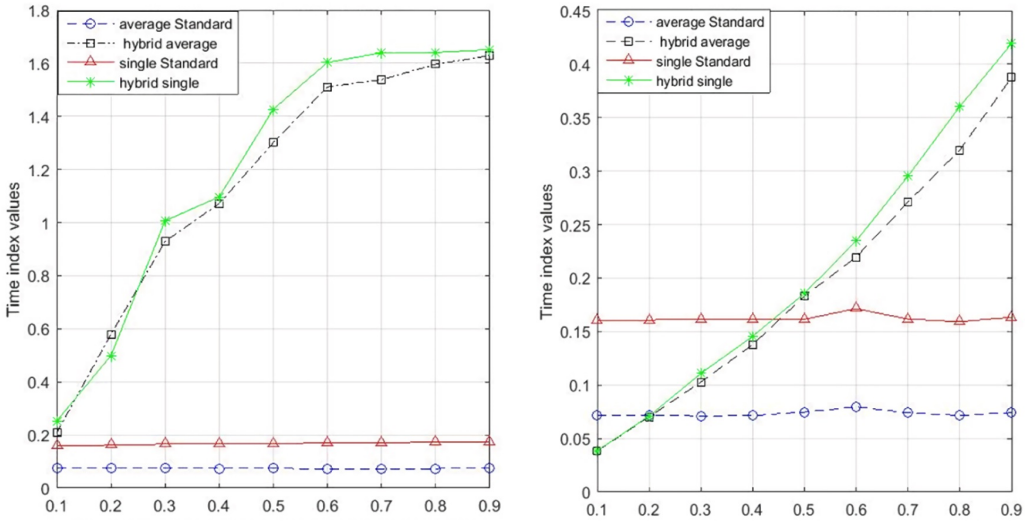

Figure 4. Hybrid and standard hierarchical comparison with time Index: Euclidean on the left and Manhattan on the right

TABLE I. Summary of the comparison results

Metric	Euclidean (0.0 – 0.45)		Manhattan (0.0 – 1.8)		Criteria
Time index	0.05 → 0.05	0.0	0.1 → 0.1	0.0	Average standard
	0.05 ↑ 0.39	0.1	0.25 ↑ 1.62	1.37	Hybrid average
	0.15 → 0.15	0.0	0.18 → 0.18	0.0	Single standard
	0.045 ↓ 0.42	0.375	0.25 ↑ 1.63	1.11	Hybrid single
	Euclidean (-0.4 – 0.1)		**Manhattan (-0.4 – 1.0)**		**Criteria**
Silhouette	0.48 ↑ 0.98	0.5	0.45 → 0.98	0.53	Average standard
	0.77 ↓ 0.46	0.31	0.79 → 0.43	0.36	Hybrid average
	-0.34 ↑ 0.98	0.64	-0.27 ↑ 0.98	1.25	Single standard
	0.45 ↓ -0.22	0.23	0.5 ↓ -0.26	0.24	Hybrid single
	Euclidean (0.96 – 0.995)		**Manhattan (0.965 – 1.0)**		**Criteria**
Rand index	0.973 ↓ 0.968	0.05	0.9739 ↓ 0.9677	0.006	Average standard
	0.9948 ↓ 0.974	0.02	0.9952 ↓ 0.9745	0.02	Hybrid average
	0.9772 ↓ 0.968	0.009	0.9765 ↓ 0.9677	0.008	Single standard
	0.9808 ↓ 0.9625	0.018	0.9823 ↓ 0.9767	0.005	Hybrid single

standard methods employing Euclidean distances within Rand indices. Surprisingly, identical values appear across both average and single standard methods regarding time indices. In summary, our comparison highlights four criteria—time, silhouette, and Rand indices—and assesses them based on Euclidean and Manhattan distances. Overall, Hybrid Average methods using Euclidean distance outperform standard average methods in terms of Rand index accuracy. Furthermore, Hybrid Average hierarchical methods excel over their Hybrid Single hierarchical counterparts across both distances when considering silhouette accuracy.

However, it's worth noting that KH is more efficient than average standard linkage when working with smaller cluster numbers or ratios. Table 1 summarizes such results. In both Hybrid Standard and Single Standard approaches, we can observe numerous small clusters and outliers. Larger clusters represent typical flow based on specific thresholds, while smaller ones may indicate potential anomalies. We generated ROC curves for both techniques using various threshold values. By arranging D samples in order with our Average Standard algorithm, we determined anomalous samples when dissimilarity values exceeded detection thresholds. We then established the ROC curve for the Average Standard algorithm.

A. AUC and clustering

Figures 5-9 display the ROC curves for Hybrid Standard, Single Standard, Hybrid Average, and Average Standard when applied to Test Probe, DoS, R2L, U2R, and mixed subsets. The Average Standard method yields the largest area under the curve for specific attack and mixed subset scenarios. It can identify numerous attacks while maintaining low false alarm rates (FARs). Average Standard outper-

TABLE II. DR and FAR of NSL-KDD data subsets

	Data Subsets	Average Standard	Hybrid Average	Single Standard	Hybrid Standard
Probe	Train	(0.95, 0.0680)	(0.95, 0.0635)	(0.99, 0.0810)	(0.97, 0.0910)
	Test	(0.99, 0.0660)	(0.93, 0.0795)	(0.85, 0.0285)	(1.00, 0.1095)
DoS	Train	(0.93, 0.0690)	(0.90, 0.0810)	(0.98, 0.0790)	(0.97, 0.0470)
	Test	(0.85, 0.0780)	(0.90, 0.0600)	(0.87, 0.0950)	(0.93, 0.2130)
R2L	Train	(0.68, 0.0960)	(0.18, 0.0980)	(0.38, 0.0760)	(0.38, 0.0725)
	Test	(0.52, 0.0895)	(0.30, 0.0950)	(0.40, 0.0335)	(0.42, 0.0635)
U2R	Train	(0.84, 0.0925)	(0.70, 0.0955)	(0.84, 0.0690)	(0.82, 0.0415)
	Test	(0.90, 0.0905)	(0.84, 0.0920)	(0.52, 0.0355)	(0.52, 0.0145)
Mix	Train	(0.91, 0.0900)	(0.87, 0.0950)	(1.00, 0.0900)	(0.93, 0.0800)
	Test	(0.89, 0.0800)	(0.84, 0.0980)	(0.80, 0.0840)	(0.77, 0.0500)

forms the other three methods in detecting low frequency attack types like U2R and R2L that resemble regular traffic patterns. Table 2 presents DR and FAR value pairs for all subsets. Generally, increasing DR leads to a higher FAR or vice versa. Table 2 represents four different models (Average Standard, Hybrid Average, Single Standard, Hybrid Standard) against five different subsets of data (Probe, DoS, R2L, U2R, Mix). The values represent model performance, presumably in terms of DR and (FAR). A brief interpretation of the provided data could be as follows: Probe subset: The Hybrid Standard model performed the best on the test set with DR 1.00 and FAR 0.1095. The Single Standard model had the highest DR on the training set, 0.99. DoS subset: The Hybrid Standard model had the highest DR on the test set, 0.93, while the Single Standard model had the highest DR on the training set, 0.98. R2L subset: Here, all models performed relatively poorly, especially on the training set. The best performance on the test set came from the Hybrid Standard model, with an DR of 0.42. U2R subset: The Average Standard model performed best on the test set, with an DR of 0.90. The Single Standard and Average Standard models performed equally well on the training set, with DR 0.84. Mix subset: The Average Standard model performed the best on the training set, 1.00. For the test set, the Average Standard model again performed the best, with an DR of 0.89. Overall, the Hybrid Standard model performed the best on the majority of the test sets, but not the training sets, suggesting it might have a good generalization capability.

The maximum value is (1.00, 0.1095) for the Single Standard in Test DoS. Hybrid Average tends to have higher accuracy in training but not always in testing. Single Standard often shows high test accuracy, indicating good generalization but with varying levels of consistency. Hybrid Standard seems to offer a balance between accuracy and consistency, especially in test scenarios. Also, examining the potential of overfitting in models with high training accuracy but lower test accuracy. Considering the balance between accuracy and consistency when choosing a model for a particular application.

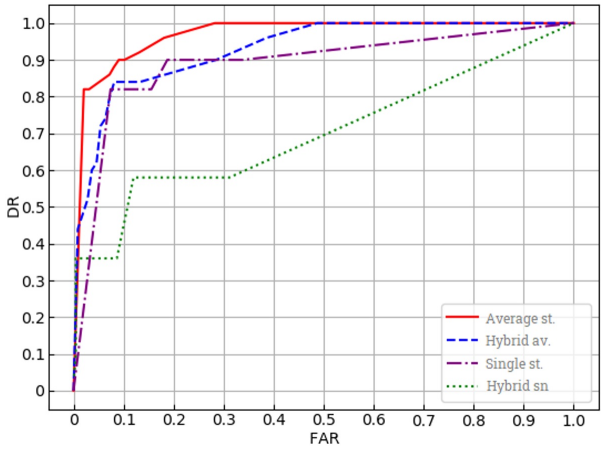

Figure 5. U2R subset

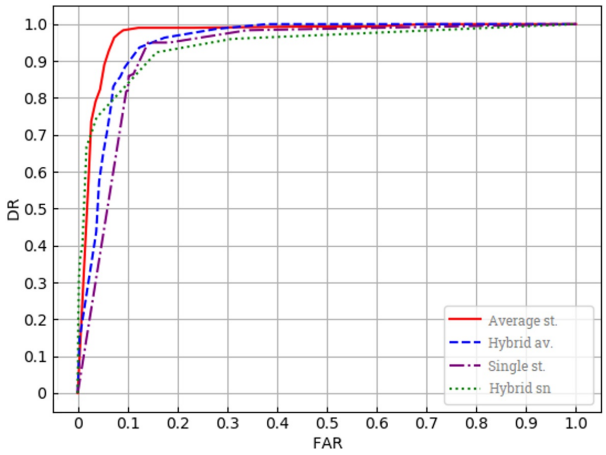

Figure 6. mixed subset

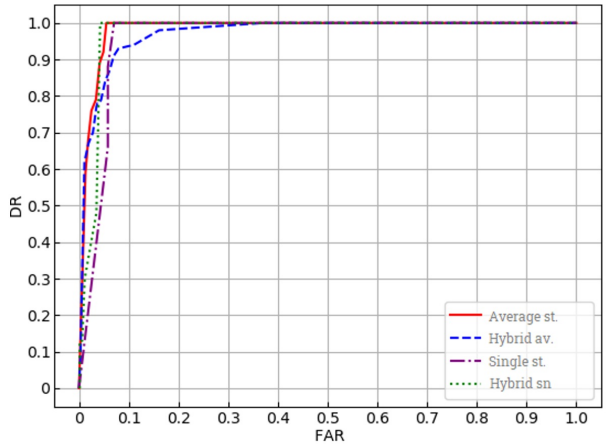

Figure 7. Probe subset

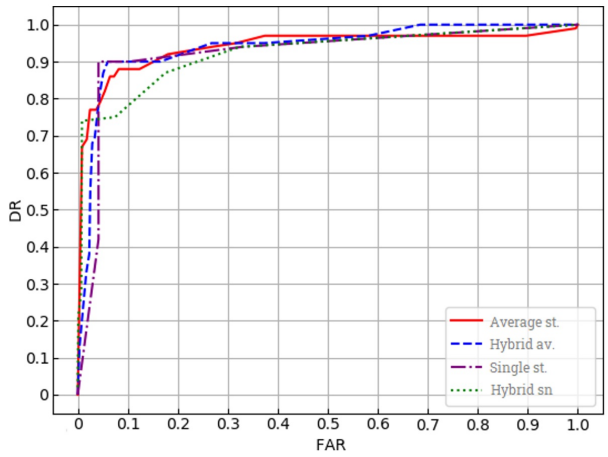

Figure 8. DoS subset

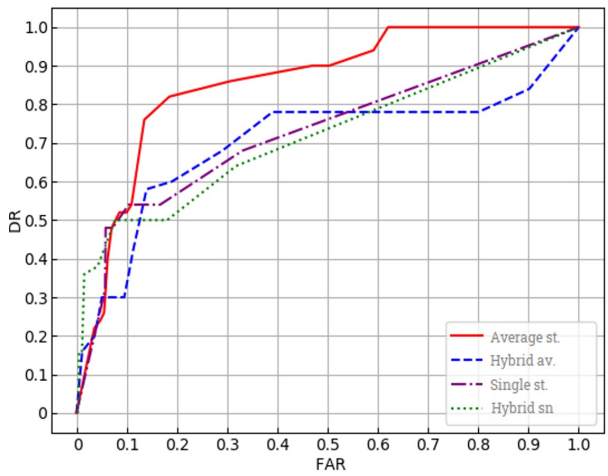

Figure 9. R2L subset

5. DISCUSSION

After implementing the methodology and getting results, in this section we analyze and interpret the results with discussion in regards into other related work. The figures 1 to 6 are evaluations of clustering algorithms, comparing 'average Standard,' 'hybrid average,' 'single Standard,' and 'hybrid single' over various metrics. The first two graphs show the Rand index values, which validate the similarity of different clusters. Because the variable on the x-axis increases, the Rand index generally decreases for the 'average Standard' and 'hybrid average,' indicating less similarity between clusters. The 'single Standard' and 'hybrid single' lines show more fluctuation. The next figures 3 and 4 illustrate silhouette values, a measure of how well each object lies within its cluster; higher values suggest better fit. Here, 'single Standard' performs poorly at low values of the variable but improves significantly as the variable increases, while 'average Standard' remains relatively stable. Last, the time index by figures 5 and 6 illustrate computational efficiency, with 'single Standard' showing low and stable times, suggesting faster performance, while 'hybrid single' times increase significantly with the variable, indicating slower performance. This could suggest that while 'single Standard' may be faster, it is less stable in terms of clustering quality, whereas 'hybrid' methods may offer a balance between clustering quality and computational efficiency. The results of figures 5 to 9 was applied on the NSL-KDD dataset. The average standard curve consistently performs well across all figures, maintaining a high DR at low levels of FAR. That means, the average standard method has a good balance of sensitivity and specificity. The hybrid average and hybrid standard curves show varying performance but generally follow the average standard curve closely, indicating that these hybrid methods are competitive. The single standard curve tends to lag behind the others, especially at lower FAR values. This might indicate that the single standard method has lower sensitivity or a higher rate of false negatives at certain thresholds. We interpret the results is that the average standard method offers the best performance in terms of both sensitivity and specificity, whereas the single standard method require adjustment or inherently be less capable in this context. The hybrid methods appear to be a compromise between the two, possibly combining elements of both to create a more balanced classifier. These interpretations would be more accurate with specific context on the data and the classification task these curves represent. Research work by [43] used the Benchmark, the Purely Spatial" dataset. he proposed metrics were tested in base of traditional scan statistics with spatial restrictions. he proposed metrics were tested in base of traditional scan statistics with spatial restrictions. This New evaluation metrics for spatial diseases clustering detection with highlighting the limitation of the traditional metrics such as precision and recall. The Main contribution is to introduce new metrics in the field. However, the limitation is that the benchmark dataset only has circle cluster shapes. Also, the need to present irregular shapes. The authors argue that there is a need

to consider the hyper parameters with the new metrics as they would significantly affect the results. We found there is no mention about computational resources, robustness and generalizability, and model complexity. [44] used three datasets: Yeast, Chronic disease, and Emotions. The MAPE and RMSE are the evaluation metrics used for clustering validation. K-fold cross-validation is the key of work used to assess the clustering algorithm. The clustering validation index is calculated by the clustering stability and the results outputs. This task is measured by comparing the probability of the training and test dataset of the same clustering by k-fold and RMSE. The proposed new method uses multi-label datasets. So, each data point in the clusters would have multiple labels at the same time. However, such a limitation is that there is no mention of outliers' impact on the clustering performance. Also, like others' work, no computational resources have been taken into account. The most important of this work is using the PCA, principle component analysis, and normalization techniques for dimensional reduction for best clustering performance quality. Also, this paper has not considered model complexity, robustness, or generalizability. By reference [45] utilized a real dataset called Yeast6. The contribution focuses on various distance measurements like Manhattan, Euclidean, CAN, COR, CHE, and BRY. The CWE-ENS supervised learning method was applied with clustering feature space. There is no novelty in such work. However, the authors try to show how the CWE-ENS can affect the clustering feature space. The limitation found is that each dataset needs particular analysis and custom parameters, so the results vary and depend on the dataset itself. The authors suggest Manhattan as a standard measurement. With an imbalanced dataset, the model accuracy was the best with Euclidean. Meanwhile, Manhattan has close but less results than Euclidean. In brief, the paper aims to test how supervised learning uses the predictive performance CWE-ENS method with clustering feature space. This work did not consider model complexity, robustness, or generalizability. The IBM Watson dataset as a real data from various telecom companies used by [46]. It is best for clustering because it has 26% customer churn rate. Different models have been used like; K-NN, random forest, and XGBoost. For churn prediction, the XGBoost was the best performance. All the other traditional metrics were used; accuracy, recall, precision, and F1 score. These is no any work related to the AUC. Even though there is no clear novelty of such work, the authors stated that the significance of work is delivered by making deep comparison of ML models to predict churn of customers. The data processing has an essential role for example, feature selection, filtering, and noise removal.

6. CONCLUSION

The hybrid clustering model would have extra consideration of accuracy, by incorporating with AUC metric. Applying specificity and sensitivity, that provided by AUC, with clustering model can help in offering robust evaluation approach. Results of applying different linkages and distance measurements have been applied and proven that hybrid standard model with average FAR reached 0.108 across all the test subsets. The value was the lowest among other results. The Data Subsets is divided into different subsets: Probe, DoS, R2L, U2R, and Mix. Each subset is further split into Train and Test groups. This suggests that the models are evaluated on different types of data. Also, there are four types: Average Standard, Hybrid Average, Single Standard, and Hybrid Standard. These could represent different modeling strategies or algorithms.

For future work, there couple of directions for future work. First Investigate the integration of advanced clustering algorithms (like DBSCAN, HDBSCAN, or Spectral Clustering) to see if they offer improvements over traditional methods. Also, Beyond AUC, consider evaluating clustering outcomes using other performance metrics like Precision-Recall AUC, F1 Score, or Silhouette Score for a more comprehensive analysis. Investigate the effectiveness of AUC in different clustering contexts and compare it with traditional clustering evaluation metrics. Discover the power of feature selection and engineering in enhancing clustering results and boosting their AUC evaluation. In respect to real-world applications and case studies: Apply clustering approach to real-world datasets in various domains (like healthcare, finance, or social media analytics) to validate its practical effectiveness. Conduct case studies focusing on specific challenges, such as imbalanced datasets or noisy data, and how your method addresses these issues.

REFERENCES

[1] S. Orozco-Arias, J. S. Piña, R. Tabares-Soto, L. F. Castillo-Ossa, R. Guyot, and G. Isaza, "Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements," 2020.

[2] M. Aamir and S. M. Ali Zaidi, "Clustering based semi-supervised machine learning for DDoS attack classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4, pp. 436–446, 2021.

[3] J. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Informative Evaluation Metrics for Highly Imbalanced Big Data Classification," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2022, pp. 1419–1426.

[4] M. Jain, G. Kaur, and V. Saxena, "A K-Means clustering and SVM based hybrid concept drift detection technique for network anomaly detection," *Expert Systems with Applications*, vol. 193, p. 116510, 2022.

[5] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," *Atmospheric Pollution Research*, vol. 11, no. 1, pp. 40–56, 2020.

[6] M. Torabi, S. Hashemi, M. R. Saybani, S. Shamshirband, and A. Mosavi, "A Hybrid clustering and classification technique for forecasting short-term energy consumption," *Environmental Progress & Sustainable Energy*, vol. 38, no. 1, pp. 66–76, jan 2019.

[7] P. A. Jaskowiak, I. G. Costa, and R. J. G. B. Campello, "The area under the ROC curve as a measure of clustering quality," *Data Mining and Knowledge Discovery*, vol. 36, no. 3, pp. 1219–1245, 2022.

[8] S. L. Supongmen Walling, "Performance Evaluation of Supervised Machine Learning Based Intrusion Detection with Univariate Feature Selection on NSL KDD Dataset," *National Institute of Technology Nagaland*, vol. PREPRINT V, 2023.

[9] R. Thomas and D. Pavithran, "A Survey of Intrusion Detection Models based on NSL-KDD Data Set," in *2018 Fifth HCT Information Technology Trends (ITT)*, 2018, pp. 286–291.

[10] K. Samunnisa, G. S. V. Kumar, and K. Madhavi, "Intrusion detection system in distributed cloud computing: Hybrid clustering and classification methods," *Measurement: Sensors*, vol. 25, p. 100612, 2023.

[11] M. Bhushan, J. Ángel Galindo Duarte, P. Samant, A. Kumar, and A. Negi, "Classifying and resolving software product line redundancies using an ontological first-order logic rule based method," *Expert Systems with Applications*, vol. 168, p. 114167, 2021.

[12] S. Manoharan, "Performance Analysis of Clustering Based Image Segmentation Techniques," *Journal of Innovative Image Processing*, vol. 2, no. 2, pp. 14–24, 2020.

[13] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104743, 2022.

[14] Mahnoor Chaudhry, Imran Shafi, Mahnoor Mahnoor, Debora Libertad Ramírez Vargas,, Ernesto Bautista Thompson and I. Ashraf, "A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective," *Symmetry*, vol. 15, no. 9, pp. 1–44, 2023.

[15] S. Sørbø and M. Ruocco, "Navigating the metric maze: a taxonomy of evaluation metrics for anomaly detection in time series," *Data Mining and Knowledge Discovery*, 2023.

[16] A. Jaeger and D. Banks, "Cluster analysis: A modern statistical review," *WIREs Computational Statistics*, vol. 15, no. 3, p. e1597, 2023.

[17] F. Ros, R. Riad, and S. Guillaume, "PDBI: A partitioning Davies-Bouldin index for clustering evaluation," *Neurocomputing*, vol. 528, pp. 178–199, apr 2023.

[18] J. Montiel, H.-A. Ngo, M.-H. Le-Nguyen, and A. Bifet, "Online Clustering: Algorithms, Evaluation, Metrics, Applications and Benchmarking," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 4808–4809.

[19] Q.-V. Doan, T. Amagasa, T.-H. Pham, T. Sato, F. Chen, and H. Kusaka, "Structural S k-means and clustering uncertainty evaluation framework (CUEF) for mining climate data," *Geoscientific Model Development*, vol. 16, no. 8, pp. 2215–2233, 2023.

[20] H. Liu, L. Fen, J. Jian, and L. Chen, "Overlapping Community Discovery Algorithm Based on Hierarchical Agglomerative Clustering," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 03, p. 1850008, jul 2017.

[21] T. A. Munandar, Azhari, A. Mushdholifah, and L. Arsyad, "Hierarchical Regional Disparities and Potential Sector Identification Using Modified Agglomerative Clustering," *IOP Conference Series: Materials Science and Engineering*, vol. 180, no. 1, p. 12074, 2017.

[22] S. Babichev, B. Durnyak, I. Pikh, and V. Senkivskyy, "An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms BT - Lecture Notes in Computational Intelligence and Decision Making," in *Advances in Intelligent Systems and Computing*, V. Lytvynenko, S. Babichev, W. Wójcik, O. Vynokurova, S. Vyshemyrskaya, and S. Radetskaya, Eds. Cham: Springer International Publishing, 2020, pp. 532–553.

[23] S. Al-Dabooni and D. Wunsch, "Model Order Reduction Based on Agglomerative Hierarchical Clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1881–1895, 2019.

[24] B. Mohammed and E. K. Gbashi, "Intrusion Detection System for NSL-KDD Dataset Based on Deep Learning and Recursive Feature Elimination," *Engineering and Technology Journal*, vol. 39, no. 7, pp. 1069–1079, 2021.

[25] A. Vinolia, N. Kanya, and V. N. Rajavarman, "Machine Learning and Deep Learning based Intrusion Detection in Cloud Environment: A Review," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2023, pp. 952–960.

[26] J. Liu, K. Xiao, L. Luo, Y. Li, and L. Chen, "An intrusion detection system integrating network-level intrusion detection and host-level intrusion detection," in *2020 IEEE 20th International Conference on Software Quality, Reliability and Security (QRS)*, 2020, pp. 122–129.

[27] R. Mogg, S. Y. Enoch, and D. S. Kim, "A Framework for Generating Evasion Attacks for Machine Learning Based Network Intrusion Detection Systems BT - Information Security Applications," in *Lecture Notes in Computer Science*, H. Kim, Ed. Cham: Springer International Publishing, 2021, pp. 51–63.

[28] A. R. Abdulla and N. G. M. Jameel, "A Review on IoT Intrusion Detection Systems Using Supervised Machine Learning: Techniques, Datasets, and Algorithms," *UHD Journal of Science and Technology*, vol. 7, no. 1 SE - Articles, pp. 53–65, mar 2023.

[29] Ahmed Alghazali and Zaid Hanoosh, "Using a Hybrid Algorithm with Intrusion Detection System based on Hierarchical Deep Learning for Smart Meter Communication Network," *webology*, vol. 2, no. 19, pp. 3850–3865, 2022.

[30] A. Kumar and T. K. Das, "Rule-based Intrusion Detection System using Logical Analysis of Data," in *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2023, pp. 129–135.

[31] M. Fernández-Sanjurjo, B. Bosquet, M. Mucientes, and V. M. Brea, "Real-time visual detection and tracking system for traffic monitoring," *Engineering Applications of Artificial Intelligence*, vol. 85, pp. 410–420, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197619301691

[32] P. A. Jaskowiak and I. G. Costa, "Clustering Validation with The Area Under Precision-Recall Curves," 2023.

[33] C. Bay, R. J. Glynn, J. M. Seddon, M.-L. T. Lee, and B. Rosner, "Evaluation of Risk Prediction with Hierarchical Data: Dependency Adjusted Confidence Intervals for the AUC," pp. 526–538, 2023.

[34] J. M. Corchado, G. Hernández, E. Herrera-Viedma, J. Parra-Dominguez, and M. E. Pérez-Pons, "Evaluation metrics and dimensional reduction for binary classification algorithms: a case study on bankruptcy prediction," *The Knowledge Engineering Review*, vol. 37, p. e1, 2022.

[35] K. Kwegyir-Aggrey, M. Gerchick, M. Mohan, A. Horowitz, and S. Venkatasubramanian, "The Misuse of AUC: What High Impact Risk Assessment Gets Wrong," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1570–1583.

[36] R. Duan, L. Gao, Y. Gao, Y. Hu, H. Xu, M. Huang, K. Song, H. Wang, Y. Dong, C. Jiang, C. Zhang, and S. Jia, "Evaluation and comparison of multi-omics data integration methods for cancer subtyping." *PLoS computational biology*, vol. 17, no. 8, p. e1009224, aug 2021.

[37] L. Qi, W. Wang, T. Wu, L. Zhu, L. He, and X. Wang, "Multi-Omics Data Fusion for Cancer Molecular Subtyping Using Sparse Canonical Correlation Analysis," *Frontiers in Genetics*, vol. 12, 2021.

[38] W. Huang, Y. Peng, Y. Ge, and W. Kong, "A new Kmeans clustering model and its generalization achieved by joint spectral embedding and rotation." *PeerJ. Computer science*, vol. 7, p. e450, 2021.

[39] H. Dunkel, H. Wehrmann, L. R. Jensen, A. W. Kuss, and S. Simm, "MncR: Late Integration Machine Learning Model for Classification of ncRNA Classes Using Sequence and Structural Encoding." *International journal of molecular sciences*, vol. 24, no. 10, may 2023.

[40] T. Osabe, K. Shimizu, and K. Kadota, "Differential expression analysis using a model-based gene clustering algorithm for RNA-seq data," *BMC Bioinformatics*, vol. 22, no. 1, p. 511, 2021.

[41] A. Dogan and D. Birant, "K-centroid link: a novel hierarchical clustering linkage method," *Applied Intelligence*, vol. 52, no. 5, pp. 5537–5560, 2022.

[42] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 6439–6475, 2023.

[43] R. C. Diniz, P. O. S. Vaz-de Melo, and R. Assunção, "Evaluating the Evaluation Metrics for Spatial Disease Cluster Detection Algorithms," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 401–404.

[44] A. N. Tarekegn, K. Michalak, and M. Giacobini, "Cross-Validation Approach to Evaluate Clustering Algorithms: An Experimental Study Using Multi-Label Datasets," *SN Computer Science*, vol. 1, no. 5, p. 263, 2020.

[45] G. Naidu, T. Zuva, and E. M. Sibanda, "A Review of Evaluation Metrics in Machine Learning Algorithms BT - Artificial Intelligence Application in Networks and Systems," R. Silhavy and P. Silhavy, Eds. Cham: Springer International Publishing, 2023, pp. 15–25.

[46] J. Klikowski and R. Burduk, "Distance Metrics in Clustering and Weighted Scoring Algorithm BT - Progress in Image Processing, Pattern Recognition and Communication Systems," M. Choraś, R. S. Choraś, M. Kurzyński, P. Trajdos, J. Pejaś, and T. Hyla, Eds. Cham: Springer International Publishing, 2022, pp. 23–33.

Ali Fattah Dakhil He has completed the degree in M.Sc., Computer Science from University of Salford by fall of 2013. His specialized work are data analysis, machine learning approaches, deep learning algorithms, computer vision, and clustering.

Waffaa M. Ali She received her degree in M.Sc., Computer Science from University Thi-Qar. She is instersted in machine learning, statistical approaches, deep learning algorithms, autonomous systems.

Mustafa Asaad Alkhafaji received the B.S degree in computer science from University of Thi-Qar in 2011 and the M.S. degree from University of Colorado Denver, USA, in 2017. He obtained PhD in information technology at University of Babylon. His interset is AI and automated projects.