JURNAL INFOTEL
VOL. 16, NO. 4, NOVEMBER 2024, PP. 837–854. DOI:10.20895/INFOTEL.V16I3.1142

RESEARCH ARTICLE

Optimizing Autism Spectrum Disorder Identification with Dimensionality Reduction Technique and K-Medoid

Galih Hendro Martono[1,*] and Neny Sulistianingsih[2]

[1,2]Department of Computer Science, Faculty of Engineering, Universitas Bumigora

*Corresponding email: galih.hendro@universitasbumigora.ac.id

Received: February 13, 2024; Revised: July 24, 2024; Accepted: December 03, 2024.

Abstract: This research addresses the challenges of diagnosing and treating Autism Spectrum Disorder (ASD) using dimensionality reduction techniques and machine learning approaches. Challenges in social interaction, communication, and repetitive behaviors characterize ASD. The dimension reduction used in this research aims to identify what features influence autism cases. Several dimension data reduction techniques used in this research include PCA, Isomap, t-SNE, LLE, and factor analysis, using metrics such as Purity, Silhouette Score, and the Fowlkes-Mallows index. The machine learning approach applied in this study is K-Medoid. By employing this method, our goal is to pinpoint the unique characteristics of autism that may facilitate the detection and diagnosis process. The data used in this research is a dataset collected for autism screening in adults. This dataset contains 20 features: ten behavioral features (AQ-10-Adult) and ten individual characteristics. The results indicate that Factor Analysis outperforms other methods based on Purity metrics. However, due to data structure issues, the t-SNE method cannot be evaluated using Purity metrics. PCA and LLE consistently provide stable Silhouette Scores across different k values. The Fowlkes-Mallows index results closely align, but t-SNE tends to yield lower values. The choice of algorithm requires careful consideration of preferred metrics and data characteristics. Factor analysis is adequate for Purity, while PCA and LLE consistently perform well. This research aims to improve the accuracy of ASD identification, thereby enhancing diagnostic and treatment precision.

Keywords: autism spectrum disorder, dimensionality reduction, K-Medoid clustering, machine learning, Purity metric

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

Autism Spectrum Disorder encompasses a range of mental diseases characterized by challenges in social interaction, communication, and repetitive behavior. According to [1], symptoms often manifest in early childhood, yet a formal diagnosis is commonly established later in life. The paper notes the prevalence of co-occurring disorders like epilepsy, depression, anxiety, and attention deficit disorders in children with ASD. Additionally, varying intellectual abilities are observed in individuals with ASD, spanning from severe conditions to higher functioning levels. Meanwhile, [2] defines ASD as a childhood-onset developmental disorder affecting the immature brain, leading to impaired social communication, interactions, and repetitive behaviors. Referencing the fifth edition of the DSM-5, the paper classifies ASD under neurodevelopmental disorders. Deficits in general mental abilities delineate intellectual disability in ASD, while communication disability involves impairments in language, speech, and communication skills [2].

Identifying Autism Spectrum Disorder (ASD) is essential to provide appropriate support and interventions tailored to the unique needs of individuals affected by the condition. Early and accurate identification allows for timely access to specialized education, therapeutic services, and assistive technology (AT), which can significantly enhance the quality of life and developmental outcomes for children with ASD [2] The positive effects of identification include improved communication skills, better social interactions, and increased participation in the community through the use of AT and tailored educational programs. On the other hand, the negative effects may involve the stigma associated with the diagnosis, potential discrimination, and the psychological impact on the child and their family. Despite these challenges, the benefits of early identification and intervention far outweigh the drawbacks, as they facilitate the integration of children with ASD into society and help them achieve their full potential.

[3] underscores the global significance of breast cancer as a health concern, leveraging machine learning algorithms such as k-nearest neighbor (KNN), Naïve Bayes, and support vector machine (SVM) to predict breast cancer recurrence [3]. The study highlights the superiority of the KNN algorithm, achieving 77.98% accuracy. In [4], autonomous learning style detection techniques are explored, with a novel algorithm utilizing prior knowledge and incorporating SVM, Naïve Bayes, and K-Nearest Neighbor (K-NN) classification algorithms. The study demonstrates Naïve Bayes as the most accurate at 91.48%, emphasizing the efficacy of prior knowledge compared to data-driven and literature-based approaches. Meanwhile, [5] delves into the analysis of electroencephalography (EEG) in autistic children, employing the Welch periodogram method, independent component analysis (ICA), and finite impulse response (FIR) filter for preprocessing. The study reveals differences in power spectral density values between normal and autistic EEG signals, particularly in the delta sub-band, providing potential insights for autism detection. Lastly, [6] addresses the critical issue of malnutrition in toddlers, employing various machine-learning methods. The random forest method is the most recommended, achieving high accuracy and contributing valuable insights for predicting toddlers' nutritional status early.

Research on Autism Spectrum Disorder (ASD) has employed various methodologies to address different aspects of the condition. In [2], individuals with ASD and non-ASD were identified through a machine learning approach, applying classification methods such as K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM) with linear basis functions—Decision Tree (DT) to

achieve optimal accuracy. Concurrently, [7] utilized eye-tracking visualization and machine learning to support individuals with ASD, emphasizing a diverse approach to ASD research. Additionally, [2, 4] contributed by developing technology to assist ASD students in their learning journey.

There is a difference in approach from current research, which emphasizes optimizing the grouping of ASD using dimension reduction techniques and the K-Medoid method. The concept in K-Medoid is the same as in K-Means. However, based on the concept, the partition method can still be carried out by minimizing the number of dissimilarities between each object and the reference point in question [8]. Although K-Medoid Clustering has advantages in cluster analysis [9], some limitations must be considered in the context of ASD grouping. These limitations primarily arise due to the complexity of ASD data requiring a specialized approach. The high variability in symptoms and characteristics of individuals with the Fowlkes-Mallows Index can make clustering difficult, even with effective clustering methods such as K-Medoid.

Additionally, managing high data dimensionality, as in the case of Fowlkes-Mallows Index data, is an additional challenge. Therefore, it is essential to identify the need for dimensionality reduction techniques to enable more efficient analysis and better interpretation. Although dimensionality reduction methods can help overcome data complexity, it is essential to remember that such methods must be carefully considered according to the specific characteristics of Fowlkes-Mallows Index data to produce more valid and meaningful results.

This research aims to construct a more effective identification of ASD by utilizing dimensionality reduction techniques and machine learning to improve accuracy in diagnosing and treating Fowlkes-Mallows Index. This research has significant relevance and utility in the context of Fowlkes-Mallows Index diagnosis and treatment, reflected in the need to identify ASD better to provide more specific and appropriate diagnostic and treatment approaches. In this context, this research can positively contribute to developing Fowlkes-Mallows Index identification methods, improving the quality of life of individuals with ASD through improvements in training and social interactions.

The paper begins with an introduction providing a comprehensive overview of ASD, highlighting its diagnosis and treatment challenges. The introduction emphasizes the need for effective identification and dimensionality reduction techniques to enhance accuracy in identifying ASD subtypes. The research method outlines the approach, applying K-Medoid clustering and various dimensionality reduction algorithms such as PCA, Isomap, t-SNE, LLE, and Factor Analysis. The result section details the evaluation process, presenting findings on Purity, Silhouette Score, and Fowlkes-Mallows Index for each algorithm and discussing their implications. The discussion interprets the results, highlighting the strengths and limitations of each method and offering insights into the choice of dimensionality reduction techniques based on evaluation metrics and data characteristics. Finally, the conclusion synthesizes the key findings, underscores the significance of the research in advancing ASD identification methods, and suggests future directions for refining clustering analyses in the context of ASD data.

2 Research Method

This section explains the approach used: data collection, dimensionality reduction, clustering with K-Medoid, and evaluation. These steps are crucial to ensure the accuracy and reliability of our research findings in exploring the effectiveness of dimensionality reduction techniques in the context of autism screening. These steps are illustrated in Figure 1.

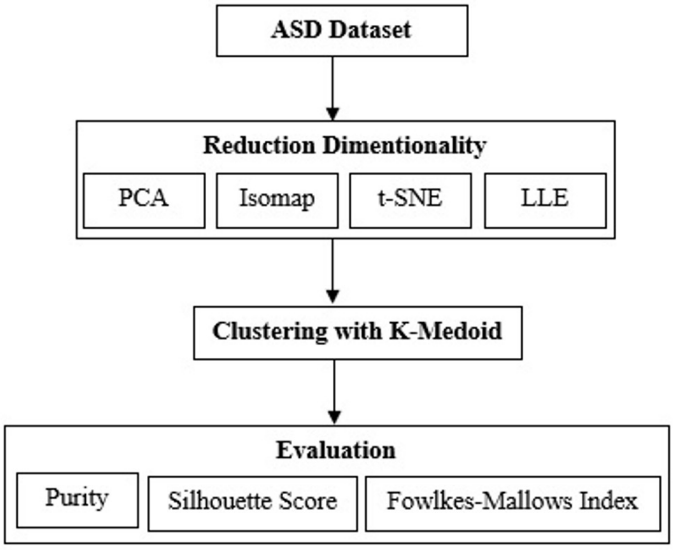


Figure 1: Research method.

2.1 Autism Spectrum Disorder's Dataset

The dataset used in this research is ASD data. This ASD dataset was obtained from www.kaggle.com from Dr. Fadi Tabtah. This dataset has various features that include answers to questions (A1 to A10), respondent age, score from Q-chat-10, gender, ethnicity, presence of jaundice at birth, family history of ASD, the person completing the test, reason for taking the screening, and class variable. These features provide insight into respondents' responses, personal characteristics, and family background that can be used for analysis or modeling in specific contexts, such as ASD screening. The amount of data in this dataset is 1054 data.

2.2 Reduction Dimensionality

The dimensionality reduction stage is carried out to reduce the number of variables or dimensions in the dataset but still retain the most relevant information so that several benefits, such as computational acceleration, increased model performance, and data simplification, can be generated. Several Dimensionality Reduction methods used in this

research include Principal Component Analysis (PCA), Isometric Mapping (ISOMAP), t-distributed Stochastic Neighbor Embedding (t-SNE), Locally Linear Embedding (LLE) and Factor Analysis (FA).

PCA is a multivariate analysis that reduces the complexity of datasets while preserving data covariance [10]. PCA looks for lines or vectors (principal components) where the data has maximum variance. The first principal component captures most of the variance, followed by the second principal component, and so on. Once the main components are identified, data can be projected into the subspace formed by these components. By removing components with low variance, PCA effectively reduces the dimensionality of the data without losing much information. Additionally, the eigenvalues and eigenvectors of the covariance matrix are used to determine the extent to which each principal component contributes to the total variance. The PCA algorithm is as follows [11].

PCA's Algorithm
Input: $X \in \mathbb{R}^{n \times d}$ **Output:** $Y \in \mathbb{R}^{n \times k}$
1. Create the covariance matrix $(X \cdot X^T)$. 2. Perform linear Eigen-decomposition on the covariance matrix $(X \cdot X^T)$ to derive Eigenvalues and their corresponding Eigenvectors. 3. Arrange the Eigenvalues in a descending order, sorting the corresponding Eigenvectors accordingly. 4. Formulate a matrix W with dimensions $d \times k$ by selecting the leading k Eigenvectors. 5. Transform the original matrix X using the matrix W to yield the updated subspace Y, expressed as $Y = X \cdot W$.

Isomap is a development of the classic Multidimensional Scaling (MDS) method, which uses geodesic distances [12]. Geodesic distance is defined as an edge that connects a path between one node and another node in a graph [13]. So, in the process, the following Equation 1 is used [12].

$$K = HD^G H \tag{1}$$

Where $D^{(G)}$ is geodesic distance, whereas H is a matrix which is the result of the dimension reduction carried out. Isomap focuses on maintaining the geodesic distance between all pairs of data points. Next, MDS is used to obtain a new matrix. The Isomap algorithm is as follows.

Isomap's Algoritm
Input: $X \in \mathbb{R}^{n \times d}$ **Output:** $Y \in \mathbb{R}^{n \times k}$
1. Develop the neighborhood graph of G. 2. Calculate the geodesic distance and build D^G. 3. Apply MDS to D^G to obtain new space Y.

Another method that can be used for dimension reduction is t-SNE. Like Isomap, t-SNE is an excellent non-linear method for visualizing data in high-dimensional space into low-dimensional space [12]. t-SNE initially utilizes Stochastic Neighbor Embedding (SNE)

on the dataset, transforming the Euclidean distances in high-dimensional space into conditional probabilities that indicate similarities between pairs of data points. The likeness between data point x_a and data point x_b is expressed through the conditional probability $P_{(a|b)}$, as defined in the Equation 2 [12].

$$P_{a|b} = \frac{\exp - \frac{\|x_b - x_a\|^2}{2\sigma^2}}{\sum_{a \neq k} - \frac{\|x_k - x_a\|^2}{2\sigma^2}} \tag{2}$$

Where $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$ are data points, and $\boldsymbol{\sigma}$ is the variance parameter. To measure the distance between the two data points, a Gaussian distribution is used with a given variance $\boldsymbol{\sigma^2}$.

Afterward, instead of using a Gaussian distribution, a 'Student t-distribution' with one degree of freedom is used, similar to the Cauchy distribution, to obtain the second set of probabilities $\boldsymbol{Q_{a|b}}$ in a low-dimensional space. If the high-dimensional data $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$ are correctly mapped to the low-dimensional data $\boldsymbol{y_a}$ and $\boldsymbol{y_b}$, then the similarity between $\boldsymbol{P_{a|b}}$ and $\boldsymbol{Q_{a|b}}$ becomes the same. Therefore, t-SNE minimizes the difference between these two probabilities from low to high dimensions. This difference is measured by optimizing the cost function (ϕ) of the Kullback-Leibler divergence sum, as shown below.

$$\phi = \sum_a \sum_b P_{a|b} \log \frac{P_{a|b}}{Q_{a|b}} \tag{3}$$

The t-SNE algorithm is as follow [13].

t-SNE's Algoritm
Input: $X \in \mathbb{R}^{n \times d}$ **Output:** $Y \in \mathbb{R}^{n \times k}$
1. Apply SNE to X to calculate the conditional probabilities $P_{a\|b}$ and $Q_{a\|b}$. 2. Map X to Y by minimizing the difference between $P_{a\|b}$ and $Q_{a\|b}$ based on the cost function ϕ.

LLE methods such as the Isomap and t-SNE methods are non-linear algorithms that try to reproduce local linear relationships between data neighbors. In dimensionality reduction, LLE tries to represent complex data in lower dimensions without changing the relative relationship between adjacent data. The LLE algorithm is as follows [13].

LLE's Algoritm
Input: $X \in \mathbb{R}^{n \times d}$ **Output:** $Y \in \mathbb{R}^{n \times k}$
1. Find c-nearest neighbors for each data point of X. 2. Calculate local weights W that linearly best rebuild data ($X^{'}$) from its neighbor. 3. Map $X^{'}$ to Y on k-dimensions using the same weights from step #2 by minimizing the cost.

Factor Analysis is a multivariate statistical method used to explore relationships between observed variables and identify latent factors that may influence variability in data.

The main goal of Factor Analysis is to identify the basic structure or factors that can explain the correlation pattern between variables. Besides that, Factor Analysis is used to find out the uniqueness among many attributes (variables) [14].

2.3 Clustering using K-Medoid

K-Medoid is a non-hierarchical clustering algorithm that operates on representative values known as medoids, serving as central points within each cluster. In contrast to K-Means, which utilizes averages, K-Medoid identifies central values that minimize the average dissimilarity to all objects within a cluster. The medoid is defined as the object whose average dissimilarity to others in the cluster is minimal, making it the most centrally located point in the dataset [15]. The method addresses a fundamental limitation of K-Means by relying on central values, called medoids, for each group.

While K-Medoid shares similarities with K-Means, a key distinction is how the cluster's center is determined. In K-Medoid, the center, or medoid, is selected based on the object in the middle of the cluster, rendering the method more robust against outliers than K-Means [9]. The K-Medoid algorithm is as follows.

K-Medoid's Algorithm
Input: X, k **Output:** μ, cluster assignment
1. Determine k. 2. Initialize medoids randomly. 3. Compute Euclidean Distances. 4. Update Medoids. 5. Repeat steps 3-4.

Euclidean Distances are used, mainly when calculating the distance between each data object and medoid when determining cluster membership. Equation 4 used to calculate the Euclidean distance [16].

$$d(x_i, \mu_j) = \sum_{i=1}^{n} (x_i - \mu_j)^2 \tag{4}$$

Where x_i is an object, μ_j is the value of medoid (center) to the j-th cluster.

2.4 Evaluation

The evaluation in this research was performed using multiple measures. These metrics include Purity, Silhouette Score, and Fowlkes-Mallows Index. Purity is used to determine the purity value of each cluster, representing the most appropriate member within the cluster. The purity value can be calculated using Equation 5 [16].

$$\mathrm{Purity}(y) = \frac{1}{N_y} \max(n_{xy}) \tag{5}$$

Where $\mathrm{Purity}(y)$ is the purity value of the y-th variable; N_y is the amount of data belonging to the y-th cluster; y is the cluster index.

The silhouette is a method of interpreting and validating consistency in data groups. The silhouette value measures how similar an object is to its group (cohesion) compared to other groups (separation). Silhouette values range from -1 to $+1$, where high values indicate that the object matches its group and poorly matches neighboring groups. If most objects have high values, the grouping configuration is considered appropriate. Conversely, the clustering configuration may have too many clusters if many objects have low or negative values. Silhouette Score can be calculated using Equation 6 [17].

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{6}$$

Where $s(i)$ is the silhouette value for the i object or data element, $a(i)$ is the average distance of the object i to all other objects that are in the same group (cohesion), and $b(i)$ is the average distance of object i to objects in other groups with the closest distance to object i (separation).

Besides Purity and Silhouette Score, the Fowlkes-Mallows Index can be used to assess the similarity between two clusterings resulting from a clustering algorithm or for gauging confusion matrices. This metric applies to various scenarios, including comparisons between hierarchical clusterings, a clustering outcome, and a benchmark. A higher Fowlkes–Mallows index value signifies a stronger resemblance between the clusters and the benchmark classifications, thus providing a quantitative measure of the degree of similarity in clustering outcomes. Fowlkes-Mallows Index can be calculated using Equation 7 [18].

$$\mathrm{FM} = \sqrt{\mathrm{PPV} \times \mathrm{TPR}} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}} \tag{7}$$

Where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. TPR is the true positive rate, also called sensitivity or recall, and PPV is the positive predictive rate, also known as precision.

3 Result

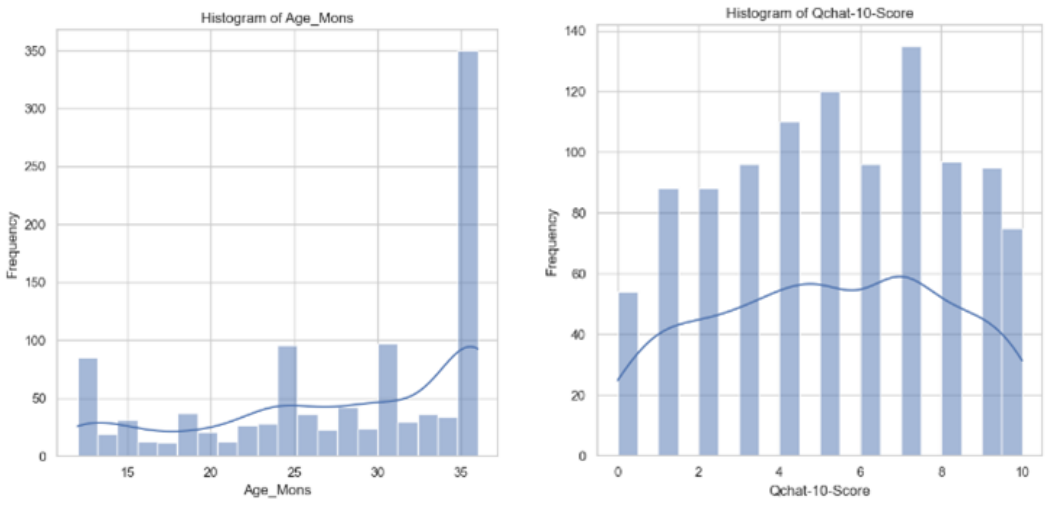
The tests carried out in this research used the ASD dataset. This dataset was previously analyzed to determine the correlation between each variable and obtain information about the variables used. Information about variables from the ASD dataset can be seen in Table 1

Furthermore, detailed information was gathered for the numeric variables (Age_Mons and Qchat-10-Score), highlighting the most frequently appearing values in these two variables. For the Age_Mons variable, the most common data points are above 35 months, indicating that most ASD cases recorded in this dataset occur in toddlers aged 35 months and older. In contrast, the Qchat-10-Score variable shows that the most frequent scores fall between 6 and 8, suggesting that the dataset typically records the presence of 6-8 symptoms in toddlers. 2(a) and 2(b) illustrate the frequency distribution for the variables Age_Mons and Qchat-10-Score, respectively.

Meanwhile, other categorical variables such as Sex, Ethnicity, Jaundice, and Family_mem_with_ASD were also analyzed. The results of the analysis related to these data show that for gender data, the toddlers recorded are more predominantly male than female as shown in 3(a). Furthermore, for Ethnicity data, the ethnicity of toddlers in the ASD

Table 1: Dataset ASD's summary

Variable	Definition	Type Data
A1-A10	Corresponding Q-chat-10-Toddler Features	Binary
Age_Mons	Age by months 5	Numeric
Qchat-10-Score	Score by Q-chat-10	Numeric
Sex	Sexuality	Categorical
Ethnicity	Ethnicity	Categorical
Jaundice	Whether the case was born with jaundice	Categorical
Family_mem_with_ASD	Whether any immediate family member has a PDD	Categorical
Who completed the test?	Parent, self, caregiver, medical staff, clinician, etc.	Categorical
Class/ASD Traits	ASD traits or No ASD traits (automatically assigned by the ASDTests app)	Categorical

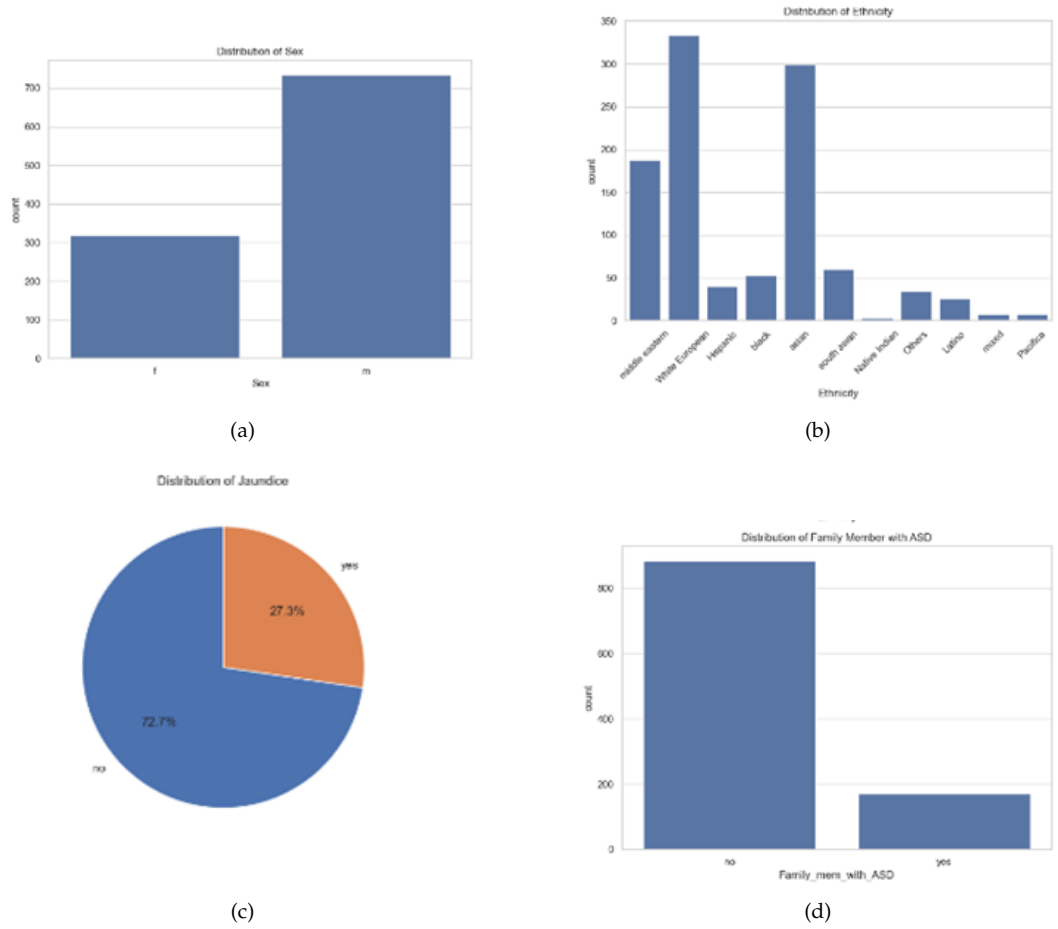


(a) (b)

Figure 2: Frequency of the data for the variables; 2(a) Age_Mons and 2(b) Qchat-10-Score.

dataset is dominated by White European ethnicity, followed by Asian and Middle Eastern, and then several other ethnicities such as South Asian, Black, Hispanic, and so on. 3(b) shows the distribution of ethnicity data. Meanwhile, for the jaundice variable, 72.7% of the data from toddlers were not accompanied by jaundice at birth. Then, for the variable Family_mem_with_ASD, which shows a history of ASD in the family, it is known that the data for toddlers who do not have a history of ASD in their family is more dominant than those who have a history of ASD. 3(c) and 3(d) show the data distribution for the variables Jaundice and Family_mem_with_ASD, respectively.

Furthermore, the correlation between variables was also analyzed using a heatmap. The heatmap reveals important relationships among the variables. The diagnostic assessments



(a) (b)

(c) (d)

Figure 3: Distribution of the data for the variables; 3(a) Sex, 3(b) Ethnicity, 3(c) Jaundice, and 3(d) Family_mem_with_ASD.

are generally positively correlated with each other, suggesting that children who score high on one assessment tend to score high on others. Age shows a weaker correlation with the diagnostic assessments and Qchat-10-Score, indicating that while age may influence the scores, it is not a strong determining factor. The Qchat-10-Score is more strongly correlated with individual diagnostic assessments, reflecting its comprehensive nature in capturing the various aspects of the assessments. This correlation matrix as shown in Figure 4 helps identify the key variables that are interrelated and can guide further analysis or interventions for children with ASD.

Next, with the ASD dataset, a dimension reduction process was carried out using several methods such as PCA, Isomap, t-SNE, LLE, and Factor Analysis. In each of these methods, the variables in the ASD dataset are reduced into two components that are represented by red and blue colors. The distribution of data in two-dimensional space resulting from PCA, Isomap, t-SNE, LLE, and Factor Analysis is shown in Figure 5, respectively. The

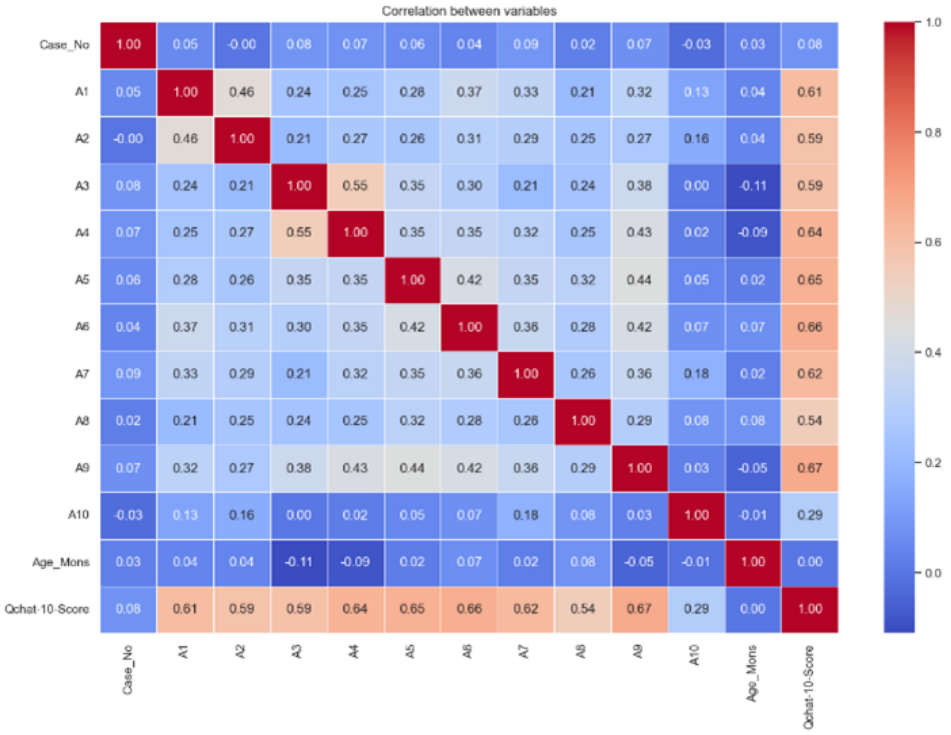


Figure 4: Correlation between variables.

frequency distribution of each ASD class in the dataset helps to see whether the distribution between ASD classes is balanced after the dimension reduction process is carried out using various methods.

Experiments were then conducted by testing the results of the dimension reduction method previously using K-Medoid. Implementing K-Medoid begins with determining the value of the k parameter (cluster) that will be used. In this study, the k values used ranged from 2 to 15 clusters. Evaluation is carried out using the Purity, Silhouette Score, and Fowlkes-Mallows Index.

In the evaluation with Purity, it was discovered that For k values ranging from 2 to 4, the Purity values remain constant at 0.691 for both PCA and Isomap, suggesting that the clusters formed by these methods exhibit similar levels of homogeneity. Besides that, as the number of clusters (k) increases beyond 4, there is a slight improvement in Purity for PCA, reaching 0.745 at k=5. Similarly, Isomap shows a consistent increase in Purity up to k=7, reaching 0.773.

Also, from evaluation using Purity, t-SNE demonstrates variable performance across different k values. While it maintains a relatively stable Purity for lower k values, it shows an increase at $k = 7$ (0.773) and remains constant for higher k values. Meanwhile, the Purity values for LLE remain constant at 0.691 across all k values, indicating consistent but relatively lower homogeneity compared to other methods. Factor Analysis shows a mixed performance, with increasing Purity up to $k = 11$ (0.706), after which it remains relatively stable for higher k values.

Table 2: Evaluation results with purity

k	PCA	Isomap	t-SNE	LLE	Factor Analysis
2	0.691	0.691	0.691	0.691	0.691
3	0.691	0.691	0.691	0.691	0.691
4	0.691	0.691	0.691	0.691	0.691
5	0.745	0.691	0.691	0.691	0.691
6	0.734	0.691	0.691	0.691	0.691
7	0.734	0.773	0.691	0.691	0.691
8	0.739	0.691	0.691	0.691	0.691
9	0.734	0.701	0.691	0.691	0.691
10	0.750	0.703	0.691	0.691	0.691
11	0.764	0.701	0.691	0.691	0.706
12	0.761	0.702	0.694	0.691	0.691
13	0.762	0.790	0.694	0.691	0.705
14	0.763	0.786	0.694	0.691	0.701
15	0.757	0.785	0.694	0.691	0.702

Table 3 presents the evaluation results using the Silhouette Score for different dimensionality reduction techniques (PCA, Isomap, t-SNE, LLE, and Factor Analysis) with varying numbers of clusters (k) ranging from 2 to 15. The Silhouette Score measures how well-defined the clusters are, with higher scores indicating better-defined clusters.

Observing the results, it's evident that the performance of each technique varies across different values of k. For PCA, the Silhouette Score starts relatively high at $k = 3$ but gradually decreases as k increases. Isomap shows a consistent performance with moderate scores across different k values. t-SNE initially performs well at $k = 2$ but struggles to

maintain high scores as k increases. LLE demonstrates a consistently high Silhouette Score, indicating robust cluster separations across different k values.

On the other hand, Factor Analysis shows lower scores than other techniques, suggesting less well-defined clusters. Table 4 presents evaluation results using the Fowlkes-Mallows Index for various dimension reduction techniques, including PCA, Isomap, t-SNE, LLE, and Factor Analysis, with the number of clusters (k) varying from 2 to 15. Fowlkes-Mallows Index measures the similarity between clustering results and ground truth.

First, we can see that the performance of dimensionality reduction techniques varies depending on the number of clusters used. At $k = 2$, t-SNE shows the highest Fowlkes-Mallows Index with a value of 0.540. At the same time, other techniques have values relatively close to each other, especially PCA, Isomap, and Factor Analysis, which have values of around 0.536. At $k = 3$, t-SNE again shows good performance, but as the number of clusters increases, other techniques such as PCA, Isomap, and Factor Analysis show better Fowlkes-Mallows Index.

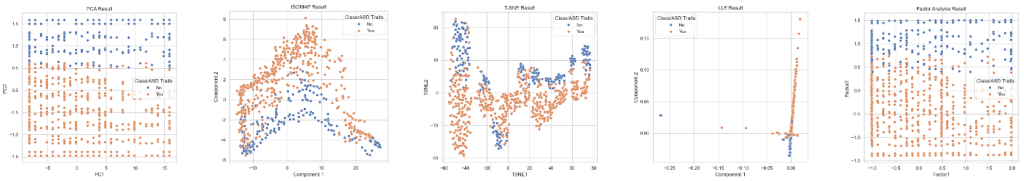
T-SNE generally provides stable and good clustering results for varying k, with the highest Fowlkes-Mallows Index at specific points. However, these results can depend significantly on the specific data used, and their interpretation should be done with caution. Besides that, Isomap and Factor Analysis also perform well in several scenarios. Visualization of the clustering results using the K-Medoids algorithm with PCA, Isomap, t-SNE, LLE, and Factor Analysis methods can be seen in Figure 6, respectively. Meanwhile, the distribution of the number of data points in each cluster $(k = 15)$ can be seen in Figure 7 for clustering results using the K-Medoids algorithm with PCA, Isomap, t-SNE, LLE, and Factor Analysis methods, respectively.

Table 3: Evaluation results with Silhouette Score

k	**PCA**	**Isomap**	**t-SNE**	**LLE**	**Factor Analysis**
2	0.624	0.594	0.499	0.773	0.366
3	0.628	0.537	0.479	0.595	0.409
4	0.608	0.597	0.516	0.582	0.413
5	0.488	0.544	0.533	0.570	0.377
6	0.511	0.528	0.524	0.562	0.410
7	0.490	0.531	0.515	0.554	0.386
8	0.479	0.468	0.513	0.550	0.420
9	0.468	0.445	0.516	0.544	0.414
10	0.458	0.433	0.512	0.550	0.400
11	0.474	0.415	0.512	0.544	0.417
12	0.465	0.418	0.518	0.547	0.412
13	0.438	0.515	0.514	0.545	0.399
14	0.452	0.473	0.512	0.543	0.400
15	0.462	0.498	0.513	0.537	0.386

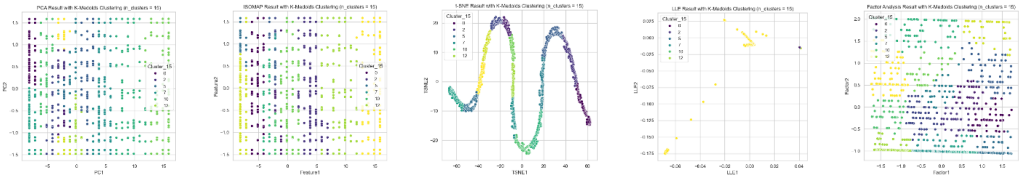
4 Discussion

The Analysis section of this study delves into the evaluation results of various dimension reduction algorithms, such as PCA, Isomap, t-SNE, LLE, and Factor Analysis, using two



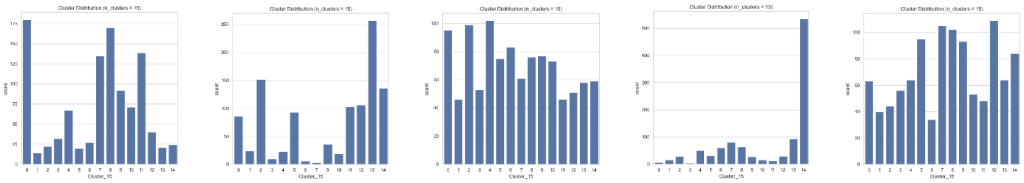
(a) (b) (c) (d) (e)

Figure 5: Distribution of data in two-dimensional space from; 5(a) PCA, 5(b) Isomap, 5(c) t-SNE, 5(d) LLE, and 5(e) Factor Analysis.



(a) (b) (c) (d) (e)

Figure 6: Visualization of the clustering results using the K-Medoids algorithm with ; 6(a) PCA, 6(b) Isomap, 6(c) t-SNE, 6(d) LLE, and 6(e) Factor Analysis.



(a) (b) (c) (d) (e)

Figure 7: Distribution of the number of data points in each cluster $(k = 15)$ with ; 7(a) PCA, 7(b) Isomap, 7(c) t-SNE, 7(d) LLE, and 7(e) Factor Analysis.

Table 4: Evaluation results with fowlkes-mallows index

k	PCA	Isomap	t-SNE	LLE	Factor Analysis
2	0.536	0.535	0.540	0.535	0.536
3	0.463	0.441	0.444	0.461	0.447
4	0.403	0.400	0.382	0.434	0.384
5	0.392	0.366	0.347	0.420	0.350
6	0.356	0.343	0.317	0.412	0.313
7	0.339	0.369	0.289	0.406	0.295
8	0.332	0.334	0.271	0.401	0.276
9	0.318	0.334	0.256	0.400	0.260
10	0.316	0.334	0.243	0.397	0.245
11	0.306	0.325	0.233	0.396	0.240
12	0.303	0.325	0.224	0.395	0.224
13	0.312	0.343	0.217	0.395	0.218
14	0.298	0.336	0.211	0.394	0.212
15	0.297	0.349	0.205	0.394	0.206

key metrics: Purity and Silhouette Score. Purity measures the consistency of labels within clusters, while the Silhouette Score assesses the quality of clustering based on the distance between clusters and the within-cluster cohesion. The findings indicate that most algorithms exhibit relatively consistent Purity values across different numbers of clusters (k values), with Factor Analysis showing higher Purity values at $k = 11$ and $k = 13$. However, the Silhouette Score metric reveals variability in algorithm performance, with t-SNE sometimes providing lower scores, indicating potential mismatches between the clustering and the underlying data structure. In contrast, PCA and LLE demonstrate more stable Silhouette Scores across different k values.

Several similarities and differences emerge when comparing these results with previous research, particularly in machine learning and classification algorithms. For instance, studies focusing on predicting breast cancer recurrence [3] and detecting autonomous learning styles [4] also leverage machine learning algorithms like k-nearest neighbor (kNN), Naïve Bayes, and SVM, albeit for different purposes. While the current study explores dimension-reduction techniques for grouping individuals with ASD, previous research primarily focuses on classification tasks within healthcare and education domains.

Moreover, machine learning in ASD research has seen diverse applications, including identifying individuals with ASD, supporting them through eye-tracking visualization, and developing technology to assist ASD students in their learning journey [2, 7]. These approaches highlight the multifaceted nature of ASD research and the importance of employing various methodologies to address different aspects of the condition.

However, the current study's emphasis on optimizing ASD grouping using dimension reduction techniques and the K-Medoid method represents a departure from previous approaches. While previous research has utilized machine learning for classification and support purposes, this study focuses on clustering ASD individuals based on symptomatology and characteristics. This shift in focus underscores the complexity of ASD data and the need for specialized approaches to group individuals with ASD effectively.

Furthermore, the limitations of K-Medoid clustering in the context of ASD grouping are addressed, particularly concerning the high variability in symptoms and characteristics among individuals with ASD. Despite the effectiveness of clustering methods like K-

Medoid, the complexity of ASD data presents challenges that may impede accurate grouping. The challenges highlight the importance of considering the unique characteristics of ASD data when applying clustering techniques and the need for further research to develop more specialized approaches tailored to the intricacies of ASD.

5 Conclusion

The evaluation results indicate that dimension reduction algorithms, including PCA, Isomap, t-SNE, LLE, and Factor Analysis, consistently produce Purity values across the tested k values. While high Purity signifies effective clustering, Factor Analysis stands out with superior Purity at $k = 11$ and $k = 13$, suggesting its potential advantage in maintaining cluster integrity. In terms of the Silhouette Score metric, variations in algorithm performance are observed, with t-SNE occasionally yielding lower scores, indicating potential challenges in representing data structure within clusters. In contrast, PCA and LLE exhibit more stable Silhouette Scores across the entire range of k values. Additionally, using the Fowlkes-Mallows Index, similar outcomes are observed among dimensionality reduction algorithms, with t-SNE consistently yielding lower values. Consequently, the choice of a dimensionality reduction algorithm should align with the preferred evaluation metric, considering the nuances revealed in the performance of each algorithm across different metrics and k values.

As a next step, this research could involve further exploration of the factors that influence the performance of dimensionality reduction algorithms on clustered data. A deeper analysis of data characteristics that may influence t-SNE performance can provide more detailed insights. Additionally, research could consider combining or adapting dimensionality reduction methods to improve clustering performance. These conclusions and findings provide a basis for further research that can provide deeper insight into the selection of dimensionality reduction algorithms in the context of clustering analysis.

Furthermore, the Fowlkes-Mallows Index measures the similarity between the resulting clustering and the ground truth. Higher values indicate better clustering. The results show that the dimensionality reduction algorithm provides similar results regarding the Fowlkes-Mallows Index, although there are minor variations in each k value. However, t-SNE tends to provide lower values than other methods.

The choice of dimensionality reduction algorithm should be carefully considered depending on the preferred evaluation metrics and data characteristics. Factor Analysis shows good performance on Purity, while PCA and LLE provide consistent results on the Silhouette Score and Fowlkes-Mallows Index. Careful consideration is required in choosing an algorithm that suits the purpose of clustering analysis and the data structure.

References

[1] A. Novianto and M. D. Anasanti, "Autism spectrum disorder (ASD) identification using feature-based machine learning classification model," *IJCCS*, vol. 17, p. 259, July 2023.

[2] Y. Purnama, F. A. Herman, J. Hartono, Neilsen, D. Suryani, and G. Sanjaya, "Educational software as assistive technologies for children with autism spectrum disorder," *Procedia Comput. Sci.*, vol. 179, pp. 6–16, 2021.

[3] I. K. A. Enriko, M. Melinda, A. C. Sulyani, and I. G. B. Astawa, "Breast cancer recurrence prediction system using k-nearest neighbor, naïve-bayes, and support vector machine algorithm," *J. Infotel*, vol. 13, pp. 185–188, Dec. 2021.

[4] M. S. Hasibuan and R. Z. A. Aziz, "Detection of learning styles with prior knowledge data using the SVM, K-NN and naïve bayes algorithms," *J. Infotel*, vol. 14, pp. 209–213, Aug. 2022.

[5] M. Melinda, I. K. A. Enriko, M. Furqan, M. Irhamsyah, Y. Yunidar, and N. Basir, "The effect of power spectral density on the electroencephalography of autistic children based on the welch periodogram method," *J. Infotel*, vol. 15, pp. 111–120, Feb. 2023.

[6] R. Gustriansyah, N. Suhandi, S. Puspasari, and A. Sanmorino, "Machine learning method to predict the toddlers' nutritional status," *J. Infotel*, vol. 16, Jan. 2024.

[7] F. Cilia, R. Carette, M. Elbattah, G. Dequen, J.-L. Guérin, J. Bosche, L. Vandromme, and B. Le Driant, "Computer-aided screening of autism spectrum disorder: Eye-tracking study using data visualization and deep learning," *JMIR Hum. Factors*, vol. 8, p. e27706, Oct. 2021.

[8] N. K. Kaur, U. Kaur, and D. Singh, "K-medoid clustering algorithm-a review," *Int. J. Comput. Appl. Technol*, vol. 1, no. 1, pp. 42–45, 2014.

[9] I. C. Dewi, B. Y. Gautama, and P. A. Mertasana, "Analysis of clustering for grouping of productive industry by k-medoid method," *International Journal of Engineering and Emerging Technology*, vol. 2, p. 26, Sept. 2017.

[10] E. Elhaik, "Principal component analyses (pca)-based findings in population genetic studies are highly biased and must be reevaluated. sci rep. 2022; 12 (1): 14683," tech. rep., Epub 2022/08/30. https://doi. org/10. 1038/s41598-022-14395-4 PMID: 36038559.

[11] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Comput. Sci. Rev.*, vol. 40, p. 100378, May 2021.

[12] B. Ghojogh, M. Crowley, F. Karray, and A. Ghodsi, "Multidimensional scaling, sammon mapping, and isomap," in *Elements of Dimensionality Reduction and Manifold Learning*, pp. 185–205, Springer, 2023.

[13] J. Bouttier, P. Di Francesco, and E. Guitter, "Geodesic distance in planar graphs," *Nuclear physics B*, vol. 663, no. 3, pp. 535–567, 2003.

[14] M. Usman, S. Ahmed, J. Ferzund, A. Mehmood, and A. Rehman, "Using PCA and factor analysis for dimensionality reduction of bio-informatics data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 5, 2017.

[15] E. Herman, K.-E. Zsido, and V. Fenyves, "Cluster analysis with K-Mean versus K-Medoid in financial performance evaluation," *Appl. Sci. (Basel)*, vol. 12, p. 7985, Aug. 2022.

[16] R. K. Dinata, S. Retno, and N. Hasdyna, "Minimization of the number of iterations in k-medoids clustering with purity algorithm," *Rev. D Intell. Artif.*, vol. 35, pp. 193–199, June 2021.

[17] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[18] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American statistical association*, vol. 78, no. 383, pp. 553–569, 1983.