JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 1, Desember 2025, hlm. 61 – 68 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i1.4265
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

COMPARISON OF CLUSTERING MODELS FOR GROUPING LIFESTYLE PATTERNS AND OBESITY FACTORS

Khalid Al Mas Ud[1*], Fathoni[1], Hafiz Muhammad Kurniawan[2]
[1]Information System, Sriwijaya University
[2]Master of Science in Data Science, INTI International University
email: *09031182227027@student.unsri.ac.id

Abstract: Obesity is an escalating global health concern, with unhealthy lifestyle patterns contributing significantly to its development. This study aims to evaluate and compare three clustering techniques for categorizing lifestyle patterns and obesity-related factors: K-Means, Agglomerative Clustering, and Gaussian Mixture Model (GMM). The data used in this study is sourced from the Food Nutrition dataset, which includes variables such as dietary habits, physical activity, and socio-economic status. The three clustering methods were assessed using evaluation metrics such as Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). The findings revealed that K-Means exhibited the best performance in terms of cluster separation with a Silhouette Score of 0.5559, while GMM showed better flexibility in handling more complex data. Although Agglomerative Clustering produced acceptable results, it had a higher overlap between clusters compared to the other methods. This study offers valuable insights into selecting the most appropriate clustering technique based on the data characteristics.

Keywords: agglomerative; clustering; GMM; k-means; lifestyle patterns; obesity

Abstrak: Obesitas menjadi masalah kesehatan yang semakin meningkat di seluruh dunia, dengan pola hidup yang tidak sehat berperan besar dalam perkembangannya. Penelitian ini bertujuan untuk membandingkan tiga metode clustering dalam mengelompokkan pola gaya hidup dan faktor yang memengaruhi obesitas, yaitu K-Means, Agglomerative Clustering, dan Gaussian Mixture Model (GMM). Data yang digunakan diperoleh dari dataset Food Nutrition yang mencakup informasi terkait pola makan, aktivitas fisik, serta faktor sosial-ekonomi. Ketiga metode tersebut diuji dengan menggunakan beberapa metrik evaluasi, seperti Silhouette Score, Davies-Bouldin Index (DBI), dan Calinski-Harabasz Index (CHI). Hasil penelitian menunjukkan bahwa K-Means memiliki kinerja terbaik dalam hal pemisahan klaster, dengan nilai Silhouette Score sebesar 0.5559, sementara GMM lebih fleksibel dalam menangani data yang lebih kompleks. Meskipun Agglomerative Clustering memberikan hasil yang dapat diterima, tumpang tindih antar klaster lebih besar dibandingkan dengan kedua metode lainnya. Penelitian ini memberikan pemahaman yang lebih baik mengenai pemilihan metode clustering yang tepat berdasarkan karakteristik data yang digunakan.

Kata kunci: agglomerative; clustering; GMM; k-means; obesitas; pola gaya hidup

INTRODUCTION

Obesity is a growing health problem in various countries, including Indonesia [1]. causes of obesity relate to unhealthy eating patterns, lack of physical activity, and socio-economic factors [2]. These factors are further exacerbated by irregular lifestyles, which can affect long-term health [3]. Therefore, it is important to analyze lifestyle patterns associated with obesity, as this can provide a clearer picture of its contributing factors.

A commonly used method in data analysis is clustering, which is used to group data based on the similarity of existing features [4]. Clustering techniques enable researchers to uncover hidden patterns within large and complex datasets, such as those found in lifestyles and eating habits related to obesity [5]. By applying clustering techniques, important patterns can be identified that provide insights into the underlying causes of obesity, which can then serve as a foundation for more targeted health interventions [6]. Moreover, previous research has indicated that analyzing lifestyle patterns can play a key role in the creation of more effective public health programs [7].

This study focuses on Comparing Clustering Models for Grouping Lifestyle Patterns and Obesity Factors using three of the most commonly used methods: KMeans, Agglomerative Clustering, and Gaussian Mixture Model (GMM). Each of these methods has its own advantages and disadvantages in clustering data with specific characteristics [8]. For instance, KMeans is often used for its simplicity in implementation and computational efficiency, while Agglomerative Cluste ring is more suitable for analyses with complex hierarchical structures [9]. GMM, on the other hand, provides a more flexible probabilistic approach and can handle more varied and unstructured data [4].

Previously, various studies have been conducted using clustering methods to identify lifestyle patterns related to obesity. For example, KMeans is often chosen for its simplicity in implemen tation and computational efficiency [7]. Meanwhile, Agglomerative Clustering is more suitable for analyses with hierar chical structures, and GMM provides a more flexible probabilistic approach [3]. These methods have also been used to analyze various other health phenomena related to eating habits and physical activity [10].]. However, even though many studies have used these methods, few have compared their performance in the context of grouping lifestyles associated with obesity [4].

The objective of this study is to assess and compare the effectiveness of three clustering models using evaluation metrics such as the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). Through this comparison, the goal is to identify the most suitable model for classifying data related to lifestyle patterns and obesity, as well as to offer recommendations for the best approach for future analysis [6]. This research will also consider the role of socio-economic factors in determining obesity patterns among different populations, which has been a major topic in previous obesity-related studies [11].

METHOD

This study uses a quantitative approach with clustering analysis methods to classify lifestyle patterns related to obesity. The data used in this research is obtained from the Food Nutrition Dataset available on Kaggle [12]. The analysis is carried out using three popular clustering methods, namely K-means, Agglomerative Clustering, and Gaussian Mixture Model (GMM) [13]. This study aims to evaluate the performance of each method in grouping data associated with obesity and lifestyle patterns.

The data used in this study is taken from the Food Nutrition Dataset available on the Kaggle platform. This

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 1, Desember 2025, hlm. 61 – 68 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i1.4265
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

dataset includes various variables related to eating patterns and physical activity, which are important in identifying factors associated with obesity [14]. The varia bles measured in this dataset include eating habits, physical activity, and socio-economic factors. This dataset also provides additional information related to age, gender, and anthropometric measu rements (e.g., weight and height), which can be used to calculate body mass index (BMI) and analyze the relationship betw een lifestyle and obesity [15].

K-means Method

K-means is a widely used clustering algorithm that groups data into K clusters based on their distance from the cluster centroids. The process for applying K-means is as follows:

1. Decide on the number of clusters (K) to be formed.
2. Randomly select K points to serve as the initial centroids.
3. Assign each data point to the closest centroid by calculating the Euclidean distance between the data point and the centroid.
4. Update the centroid by computing the average of all data points within the cluster.
5. Continue repeating steps 3 and 4 until the centroids stabilize and no longer change.

The formula for calculating the Euclidean distance between two points x and y is [16]:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Description:
x and y are the two data points whose distance will be calculated;
n is the number of features or dimensions in the data;
x_i and y_i is the value of the k-th features of data point x and y;$d(x,y)$ is Eucli dean distance between the two data points x and y, which is used to deter mine the proximity of data in feature space.

Where x_i and y_i represent the values of the p-th feature at data points x and y. The goal of the K-means objective function is to reduce the total squared distance between the data points and their corresponding cluster centroids [16]:

$$J = \sum_{1=1}^{K} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2 \quad (2)$$

Where μ_i is the centroid of cluster C_j.

Description:
J is the value of the K-means objective function to be minimized;
K is the number of clusters selected.
C_i is the-i cluster;
$x_j \epsilon C_i$ indicates that data belongs to cluster C_i;
μ_i is the center of the-i cluster, calculated as the average of all data in that cluster;
$\| x_j - \mu_i \|^2$ is the squared distance betwee data $x_j$and the cluster center μ_i.

Agglomerative Clustering Method

Agglomerative Clustering is a hierarchical clustering technique that progressively combines data points into clusters. Initially, each data point is treated as its own cluster, and then the most similar clusters are merged together. In each iteration, the distance between two clusters is computed using a distance measure, such as single linkage, complete linkage, or average linkage. The formula for calculating the distance between two clusters C_i and C_j in the single linkage method is [17]:

$$d(C_i, C_j) = \min_{x \in C_i y \in C_j} \|x - y\| \quad (3)$$

Description:
$d(C_i,C_j)$ is the distance between two clusters C_i and C_j;

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 1, Desember 2025, hlm. 61 – 68 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i1.4265
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

$x \in C_i$ indicates that is an element x in cluster C_i, and $y \in C_j$ indicates that is an element in cluster C_j;
$\|x - y\|$ is the distance between two data points x and y that are in clusters $C_i$and C_j.

Where x is an element in cluster C_i and C_j, and $\| x - y \|$ is the Euclidean distance between those elements.

Gaussian Mixture Model (GMM) Method

The Gaussian Mixture Model (GMM) is a probabilistic approach that assumes the data is derived from a mixture of multiple normal (Gaussian) distributions. GMM is employed to identify subgroups within the data and to cluster the data points based on the likelihood of their association with each normal distribution. It represents the data as a combination of several normal distributions, characterized by mean μ and covariance Σ parameters. The probability density function for GMM is [18]:

$$f(x) = \sum_{i=1}^{K} \pi_i \mathcal{N}\big((x|\mu_i, \Sigma_i)\big) \quad (4)$$

Description:
$f(x)$ is the probability that the data x comes from a mixture of normal distributions;
K is the number of Gaussian components in the mixture model;
π_i is the weight for the th Gaussian component the-i, indicating the proportion of data that corresponds to that component (with $\sum_{i=1}^{K} \pi_i = 1$);
$\mathcal{N}\big((x|\mu_i, \Sigma_i)\big)$ is the normal Gaussian distribution function with mean μ_i and covariance matrix $\Sigma_i$for the-i component;
χ is the data for which the distribution will be predicted.

Where π_i is the weight of the i, and $\mathcal{N}(x \mid \mu_i, \Sigma_i)$ is the normal distribution with mean $\mu_i$and covariance Σ_i.

Model Evaluation Metrics

Once the clustering is completed, the effectiveness of each clustering model is assessed using various metrics, including the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). These metrics help to measure how accurately each method (K-means, Agglomerative Clustering, GMM) organizes the obesity data according to the variables in the Food Nutrition Dataset.

RESULT AND DISCUSSION

In this research, three distinct clustering techniques K-means, Agglomerative Clustering, and Gaussian Mixture Model (GMM) were utilized to classify lifestyle patterns and factors associated with obesity. Each method was evaluated using data that had been preprocessed. The outcomes from each method were assessed with the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI) metrics. The subsequent section presents the results and discussion based on the application of these three methods.

K-means Clustering Result

In this study, K-means was used to cluster data related to lifestyle patterns and obesity. The number of clusters used was selected using the Elbow Method to determine the optimal number of clusters. The clustering results showed that K-means successfully produced three clearly separated clusters. The Silhouette Score obtained was 0.5559, indicating that K-means provided good cluster separation and significant distinction

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 1, Desember 2025, hlm. 61 – 68 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i1.4265
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

between the clusters formed. The clusters generated by K-means can be seen in Figure 1.

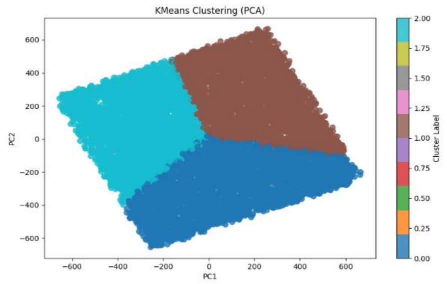

Figure 1. Visualization of the clustering results with K-means for three clusters.

which illustrates the distribution of data into three clearly separated clusters [19].

The evaluation of these results is in line with previous studies that have shown K-means to be effective for clustering data with clear and simple structures [16]. In terms of processing efficiency, K-means is faster than the other methods, making it a suitable choice for large datasets with a known number of clusters [12].

Agglomerative Clustering Result

Agglomerative Clustering produced three clusters similar to K-means, but the way the clusters were merged is different. This merging process is depicted in the dendrogram in Figure 2, which shows how the clusters were hierarchically combined. Agglomerative Clustering tends to be more sensitive to relationships between data points, but the Davies-Bouldin Index (DBI) obtained shows that Agglomerative Clustering has a higher score compared to K-means, indicating greater overlap between clusters. This suggests that while Agglomerative Clustering is effective for hierarchical grouping, it is more prone to overlapping clusters compared to K-means [20].

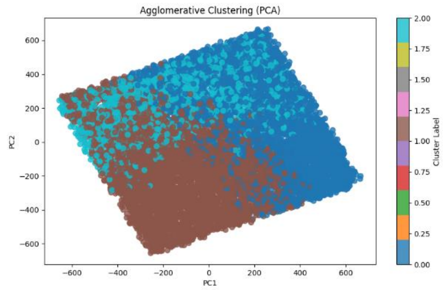

Figure 2. Dendrogram of cluster merging in Agglomerative Clustering.

Gaussian Mixture Model (GMM) Result

The Gaussian Mixture Model (GMM) method was applied by selecting three Gaussian components using the Bayesian Information Criterion (BIC) to determine the optimal number of components. GMM, as a probabilistic method, is more flexible in clustering data that are not distributed spherically. The GMM results showed a Log-Likelihood of 3500, which is lower compared to K-means and Agglomerative, but suggests that this model is better at handling complex data distributions [21].

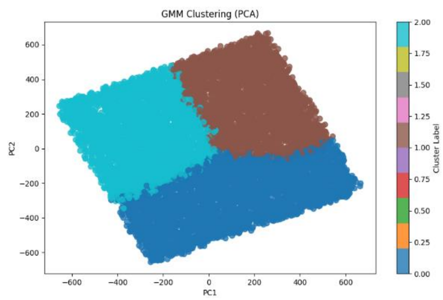

Figure 3. Visualization of clustering results using Gaussian Mixture Model (GMM) on the obesity data.

These results are consistent with previous research showing that GMM excels at handling data with non-spherical distributions and provides smoother and more flexible cluster separation [8]. Even though the Log-

Likelihood is lower, GMM offers advantages in dealing with more complex data [5].

Evaluation of Clustering Results

Table 1 shows the evaluation results of the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI) metrics for each clustering method applied.

Table 1. Clustering Result Evaluation

Method	evaluation of results metrics		
	Silhouette Score	DBI	CHI
KMeans	0.5559	0.5476	19724.1532
Agglomerative	0.5355	0.5111	18197.7873
GMM	0.5595	0.5381	19511.1707

Based on the evaluation results, K-means demonstrated better perform ance in terms of cluster separation and clustering quality according to the Silhouette Score and DBI. Although GMM performed better at handling more complex distributions, K-means was more efficient in processing time for data with a clear cluster structure [22]. Agglomerative Clustering gave acce ptable results but showed higher overlap between clusters compared to K-means and GMM, especially in terms of more overlapping cluster separation, as reflected in its higher DBI score.

CONCLUSION

his study aims to compare three clustering methods, namely K-means, Agglomerative Clustering, and Gaussian Mixture Model (GMM), in grouping lifestyle patterns and obesity factors. Based on evaluation results using the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI) metrics, the following conclusions can be drawn:

K-means produces fairly good cluster separation with a Silhouette Score of 0.5559, although it is less effective in handling data with complex distributions or data that are not spherically structured. This method offers efficient processing time and is suitable for datasets with clearly structured clusters. This research provides important insights into selecting the appropriate clustering method, depending on the characteristics of the data used. GMM is more recommended for data with complex distributions, while K-means can be considered for applications requiring efficiency and speed, particularly for data with clearer and simpler structures [6].

Agglomerative Clustering shows lower performance with a Silhouette Score of 0.5355. Nevertheless, this method has the advantage of hierarchical analysis, but it is unable to produce optimal cluster separation compared to K-means and GMM [9].

Gaussian Mixture Model (GMM) shows the best results with a Silhouette Score of 0.5595. GMM is more flexible in grouping data with complex and non-spherical distributions. This makes it the most superior choice in terms of the quality of the resulting clusters, although it requires longer processing time compared to K-means [23].

Overall, GMM is a superior method for handling more complex data with non-spherical distributions. K-means, though more time-efficient, remains a good choice for simpler and more structured datasets. On the other hand, Agglomerative Clustering, despite being useful for hierarchy-based analysis,

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 1, Desember 2025, hlm. 61 – 68 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i1.4265
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

has limitations in cluster separation compared to the other two methods.

Given these results, it is suggested that future studies investigate alternative clustering techniques, like DBSCAN or Fuzzy C-means, which are capable of managing data with irregular distributions and do not necessitate specifying the number of clusters in advance. Additionally, applying dimensionality reduction methods such as PCA or t-SNE is recommended to enhance the efficiency of processing data with high dimensions.

In addition, to enhance the validity of the results, subsequent experiments can involve larger and more diverse datasets, as well as consider variations in data collection. Further research can also explore the use of ensemble methods that combine various clustering algorithms to produce more robust results and reduce dependence on a single method.

BIBLIOGRAPHY

[1] M. Brauer *et al.*, "Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *Lancet*, vol. 403, no. 10440, pp. 2162–2203, May 2024.

[2] Y. Hu, Y. Zhang, J. Zhong, Y. Wang, E. Zhou, and F. Hong, "Association between obesity phenotypes and dietary patterns: A two-step cluster analysis based on the China multi-ethnic cohort study," *Prev. Med. (Baltim).*, vol. 187, p. 108100, 2024.

[3] Q. Wang, M. Yang, K. Chen, F. Zheng, Z. Zhang, and W. Niu, "Clustering unhealthy lifestyle factors in Chinese children and adolescents with overweight or obesity," *BMC Pediatr.*, vol. 25, no. 1, p. 226, 2025.

[4] J. F. López-Gil, J. Brazo-Sayavera, A. García-Hermoso, E. M. de Camargo, and J. L. Yuste Lucas, "Clustering Patterns of Physical Fitness, Physical Activity, Sedentary, and Dietary Behavior among School Children," *Child. Obes.*, vol. 16, no. 8, pp. 564–570, Oct. 2020.

[5] J. Kim *et al.*, "Physical Activity Pattern of Adults With Metabolic Syndrome Risk Factors: Time-Series Cluster Analysis," *JMIR Mhealth Uhealth*, vol. 11, p. e50663, 2023.

[6] G. S. Mohamed Khamis, N. S. Alqahtani, S. Munadi Alanazi, M. M. Alruwaili, M. S. Alenazi, and M. A. Alrawaili, "Using Fuzzy C-Means clustering and PCA in public health: A machine learning approach to combat CVD and obesity," *Informatics Med. Unlocked*, vol. 57, p. 101666, 2025.

[7] K. Ahmad, S. A. Keramat, G. M. Ormsby, E. Kabir, and R. Khanam, "Clustering of lifestyle and health behaviours in Australian adolescents and associations with obesity, self-rated health and quality of life," *BMC Public Health*, vol. 23, no. 1, p. 847, 2023.

[8] A. J. Grant, D. Gill, P. D. W. Kirk, and S. Burgess, "Noise-augmented directional clustering of genetic association data identifies distinct mechanisms underlying obesity," *PLoS Genet.*, vol. 18, no. 1, pp. 1–24, 2022.

[9] R. González-Martos *et al.*, "Unsupervised clustering of biochemical markers reveals health profiles associated with function and survival in active aging," *Sci. Rep.*, vol. 15, no. 1, p.

DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i1.4265
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

30546, 2025.

[10] Y. Wasnyo *et al.*, "Clustering of diet and physical activity behaviours in adolescents across home and school area-level deprivation in Cameroon, South Africa, and Jamaica.," *BMC Public Health*, vol. 24, no. 1, p. 3234, Nov. 2024.

[11] R. Thirumalaiselvi and D. Gomathi, "Healthy eating behaviors of girl children using clustering techniques: A questionnaire study," *i-manager's J. Comput. Sci.*, vol. 10, no. 1, p. 1, 2022.

[12] A. Wosiak, M. Krzywicka, and K. Żykwińska, "Assessing the Impact of Physical Activity on Dementia Progression Using Clustering and the MRI-Based Kullback–Leibler Divergence," *Appl. Sci.*, vol. 15, no. 2, p. 652, Jan. 2025.

[13] I. R. Paucar, C. Yactayo-Arias, and L. Andrade-Arenas, "Predictive Models in Mental Health Based on Unsupervised Data Clustering," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 9, 2025.

[14] M. A. Mizani *et al.*, "Identifying subtypes of type 2 diabetes mellitus with machine learning: development, internal validation, prognostic validation and medication burden in linked electronic health records in 420 448 individuals," *BMJ Open Diabetes Res. Care*, vol. 12, no. 3, p. e004191, Jun. 2024.

[15] G.-E. Yie *et al.*, "Plasma metabolite based clustering of breast cancer survivors and identification of dietary and health related characteristics: an application of unsupervised machine learning," *Nutr. Res. Pract.*, vol. 19, no. 2, p. 273, 2025.

[16] D. Geovani, Z. Umari, and S. Ramadini, "Cluster Analysis of Obesity Risk Levels Using K-Means And DBScan Methods," *Comput. Eng. Appl. J.*, vol. 13, no. 3, pp. 10–24, Oct. 2024.

[17] E. Setiawati, U. D. Fernanda, S. Agesti, M. Iqbal, and M. O. A. Herjho, "Implementation of K-Means, K-Medoid and DBSCAN Algorithms In Obesity Data Clustering," *IJATIS Indones. J. Appl. Technol. Innov. Sci.*, vol. 1, no. 1, pp. 23–29, Jan. 2024.

[18] D. E. Coral *et al.*, "Subclassification of obesity for precision prediction of cardiometabolic diseases," *Nat. Med.*, vol. 31, no. 2, pp. 534–543, 2025.

[19] M. M. Mottalib, J. C. Jones-Smith, B. Sheridan, and R. Beheshti, "Subtyping Patients With Chronic Disease Using Longitudinal BMI Patterns," *IEEE J. Biomed. Heal. Informatics*, vol. 27, no. 4, pp. 2083–2093, 2023.

[20] Z. Zhou *et al.*, "Volumetric visceral fat machine learning phenotype on CT for differential diagnosis of inflammatory bowel disease," *Eur. Radiol.*, vol. 33, no. 3, pp. 1862–1872, 2023.

[21] M. Mehedi Hassan, S. Mollick, and F. Yasmin, "An unsupervised cluster-based feature grouping model for early diabetes detection," *Healthc. Anal.*, vol. 2, p. 100112, 2022.

[22] Z. Lin *et al.*, "Machine Learning to Identify Metabolic Subtypes of Obesity: A Multi-Center Study," *Front. Endocrinol. (Lausanne).*, vol. 12, Jul. 2021.

[23] M. Rivera-Ochoa *et al.*, "Clustering Health Behaviors in Mexican Adolescents: The HELENA-MEX Study," *Res. Q. Exerc. Sport*, vol. 95, no. 1, pp. 281–288, Jan. 2024.