ISSN:2302-8149
e-ISSN:2540-9719

K-Means and Fuzzy C-Means Cluster Food Nutrients for Innovative Diabetes Risk Assessment

[1]Irma Darmayanti*, [2]Dinar Mustofa, [3]Nurul Hidayati, [4]Inka Saputri
[1,3,4]Department of Informatics Engineering, Amikom Purwokerto University, Banyumas, Indonesia
[2] Departement of Informatics, Amikom Purwokerto University, Banyumas, Indonesia
[1,2,3,4]Jl. Let. Jend. Pol. Soemarto, Purwanegara, Purwokerto Utara, Banyumas, Jawa Tengah.
*e-mail: irmada@amikompurwokerto.ac.id

(***received***: 30 August 2024, ***revised***: 12 September 2024, ***accepted***: 15 September 2024)

Abstract

Packaged food and beverages often pose a risk of increasing diabetes when consumed regularly. This study aims to classify these products based on their nutritional content listed on the labels, with a focus on identifying diabetes risk. The methods employed include K-Means and Fuzzy C-Means, K-Means is used to determine initial center of cluster, while Fuzzy C-Means enhances the clustering by assigning probabilistic memberships to each data point. These methods are applied to products sold in stores in Banyumas Regency, Central Java, Indonesia. This research is the first to combine these two methods in the context of product clustering based on nutritional labels. The results indicate that packaged food and beverage products can be classified into high-risk and low-risk clusters for diabetes. Consequently, this study provides important guidance for consumers in choosing healthier.

Keywords: Diabetes Risk, Packaging Labels, Product Classification, K-Means, Fuzzy C-Means

1 Introduction

Diabetes is among the most prevalent metabolic diseases globally, with a prevalence that continues to increase in many countries [1]. Type 2 diabetes is the most widespread form of the disease, is frequently linked to lifestyle risk factors like poor eating habits and insufficient physical activity [2], [3]. A diet high in fat, sugar, and natrium/sodium has been shown to contribute to the development of insulin resistance, which is one of the primary causes of type 2 diabetes [4]. Nutrition labels on food packaging provide important information about the content of fat, sugar, protein, and sodium, which can be used to assess the risk of diabetes in consumers [5] . Clustering methods such as K-Means [6] and Fuzzy C-Means [7], [8] can be used to group food products based on their nutritional content, allowing for the products identification that pose a high or low risk for diabetes in patients [9], [10].

Packaging for food and beverages frequently acts as a major source of foods high in calories, fat, and sugar factors that can elevate the risk of developing diabetes. Research has consistently shown a strong correlation between the regular intake of processed foods and sugary drinks and the growing incidence of type 2 diabetes in many populations. As a result, it is crucial for consumers to be aware of the nutritional content of the products they consume regularly, with nutrition labels playing a key role in helping assess potential health risks, including the risk of diabetes [11].

Clustering techniques based on machine learning are commonly applied across different fields to categorize data with similar traits. When it comes to nutrition, these methods can reveal patterns in nutritional content that might not be immediately obvious to the average consumer. A good example is K-Means, a popular clustering algorithm that groups data by similarity, such as the nutritional information found on food labels [12]. A more sophisticated approach, such as Fuzzy C-Means, offers greater flexibility in clustering data by allowing a single product to belong to multiple groups with varying probabilities. This means that instead of being confined to just one category, a product can share characteristics with several clusters [13]

Applying K-Means and Fuzzy C-Means techniques to cluster food products based on their nutritional content offers a more precise way to identify items that may pose a higher risk for diabetes. By analyzing data from food products available in Banyumas Regency, Central Java, Indonesia, this study introduces a novel method for assessing diabetes risk through nutrition labels. The integration of

ISSN:2302-8149
e-ISSN:2540-9719

both techniques enables a more thorough examination of consumption habits and related health risks, while also helping consumers make more informed decisions when selecting healthier products [14].

2 Literature Review

Recent research on the use of K-Means related to diabetes in Indonesia was conducted by researchers from Universitas Jenderal Achmad Yani [15]. The research applied the algorithm of K-Means Clustering to classify risk of diabetes mellitus (DM). An evaluation at k=2 revealed that the clusters had mixed data, featuring a Silhouette Coefficient = 0.5716 and a Davies-Bouldin Index = 0.672. The attributes used were VLDL, AGE, HDL, CLASS, HbA1c, BMI, LDL, Gender, Urea, Chol, Cr, and TG. The results of the study showed that scatter plot visualizations revealed a relatively even distribution of data within each cluster, indicating that each cluster had nearly homogeneous characteristics for each value of k.

Another study focuses on elderly patients with diabetes by analyzing their blood test results [16]. The K-Means method will be used to identify relevant clusters, specifically two clusters for elderly individuals with high and low levels of diabetes. The results of the K-Means Clustering, supported by a DBI value of -0.597, produced six clusters, with cluster0 being the most optimal, containing 57 elderly individuals with the highest levels of diabetes.

A study closely related to this research is the one conducted by Samudra University [17]. The study categorizes food and beverage products based on their risk level for diabetes, according to the nutrition labels on these products. The method of clustering employed is K-Means algorithm. The dataset consists of 38 food and beverage products. The results of the study indicate that 20 products are classified as high risk for diabetes, while 18 products are classified as low risk.

The Fuzzy method is often combined with other clustering techniques related to diabetes, such as KNN[16],[17]. The Fuzzy model that is closest to this research is the Fuzzy C-Means model [20], [21]. This research combine K-Means [17] and Fuzzy C-Means [20], [21] methods to cluster packaged food and beverage products available in stores in Banyumas Regency, Central Java, Indonesia, based on the nutritional content listed on the product labels. This approach, combining such methods for analyzing nutritional labels, has not been previously undertaken in research.

This literature review concludes that the K-Means method has been extensively used in diabetes-related research in Indonesia, particularly in classifying diabetes risk based on medical attributes and data from elderly patients. Earlier studies have applied the K-Means algorithm to effectively identify groups with high and low diabetes risk through data clustering. Moreover, similar research has grouped food and beverage products by diabetes risk using nutritional labels, though these studies primarily relied on the K-Means method alone. The Fuzzy approach, especially Fuzzy C-Means, is frequently combined with other techniques like KNN. This research presents a novel method by integrating K-Means and Fuzzy C-Means to cluster packaged food and beverage products in Banyumas based on their nutritional information—an approach not previously explored in other studies.

3 Research Method

This study employs a quantitative method with an experimental research design to evaluate the effectiveness of the K-Means and Fuzzy C-Means methods in clustering food and beverage products based on fat, protein, sugar, and sodium content for diabetes risk assessment. The dataset comprises 399 products, with nutritional labels recorded from various supermarkets in the Banyumas region, Central Java. Analysis is conducted using Python programming, utilizing the Pandas library for data processing, Numpy for numerical operations, Matplotlib for data visualization, and scikit-learn along with scikit-fuzzy for implementing the method of K-Means and Fuzzy C-Means. Listed below are the top 3 and bottom 3 data points from the dataset used.

Table 1. The Dataset

No	Products	Fat	Protein	Sugar	Natrium
1	Sprite	1	0	54	0.11
2	Coca Cola	0	0	39	0.025
3	Fanta Strawberry	0	0	11	0.025
…	…	…	…	…	…
397	Monde Cream Crackers	5	4	0	0.19
398	Monde Butter Cookies	8	2	6	0.015
399	Murgerbon Strawberry	10	3	5	0.035

The steps of the experiment using the K-Means method can be seen in the following figure.

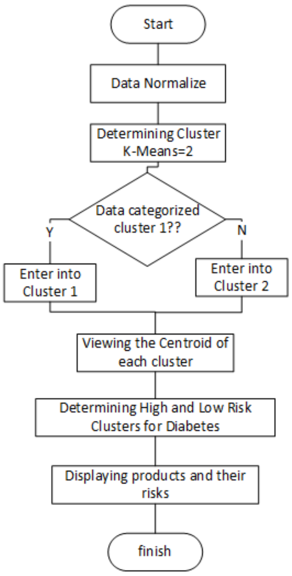

Start
Data Normalize
Determining Cluster K-Means=2
Data categorized cluster 1??
Y
N
Enter into Cluster 1
Enter into Cluster 2
Viewing the Centroid of each cluster
Determining High and Low Risk Clusters for Diabetes
Displaying products and their risks
finish

Figure 1. K-Means experiment steps

The explanation of Figure 1 above is as follows:

1. **Data Normalization:** This aims to normalize numerical data (Fat, Protein, Sugar, Sodium) so that they have a mean of zero and a standard deviation of one. Normalization is crucial in many machine learning algorithms to ensure that all features have the same scale, which can enhance the model's performance.
2. **Determining K-Means Clusters = 2:** This step applies K-Means clustering on the normalized nutritional data to divide it into two clusters. The clustering results are added as a new column in the original DataFrame for further analysis.
3. **Viewing the Centroid of Each Cluster:** This aims to retrieve the positions of the centroids from the K-Means clustering results, revert these centroid values to the original scale, and display the centroid values in the form of a DataFrame. This step allows us to see the original values of the features 'Fat,' 'Protein,' 'Sugar,' and 'Sodium' for each cluster identified by the K-Means algorithm.

4. **Determining High-Risk and Low-Risk Clusters for Diabetes:** This step identifies which cluster is at high risk and which is at low risk for diabetes patients.
5. **Displaying Products and Their Risks:** This step displays the products in the dataset and the results of their grouping, indicating whether they pose a high or low risk to diabetes patients.

The steps for data clustering using Fuzzy C-Means are as follows:

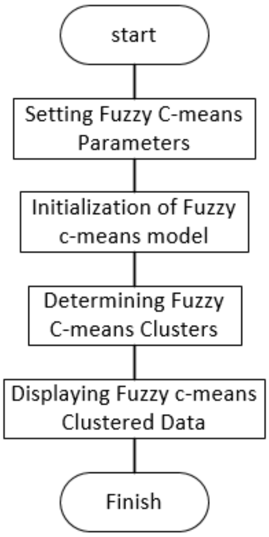

start
Setting Fuzzy C-means Parameters
Initialization of Fuzzy c-means model
Determining Fuzzy C-means Clusters
Displaying Fuzzy c-means Clustered Data
Finish

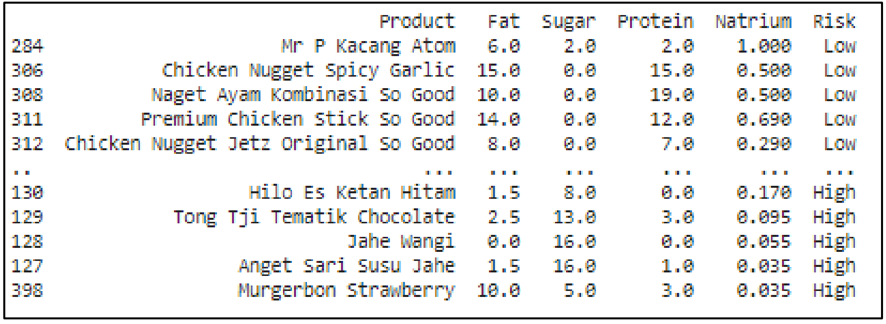
Figure 2. Fuzzy C-Means experiment steps

The explanation of Figure 2 above is as follows:

1. **Setting Fuzzy Parameters:** The fuzzy parameter values are set as follows: number of clusters = 2, maximum iterations = 1000, error threshold = 0.005, and random_state = 42.
2. **Initializing the Fuzzy C-Means Model:** This step involves determining the initial values for the Fuzzy C-Means model.
3. **Determining Fuzzy C-Means Clusters:** This step assigns data points to the previously established clusters.
4. **Displaying Fuzzy C-Means Clustered Data:** This step aims to display the products that have been clustered.

4 Results and Analysis

The clustering results using K-means can be seen in the following figure



```
                              Product   Fat  Sugar  Protein  Natrium  Risk
284                    Mr P Kacang Atom   6.0    2.0      2.0    1.000   Low
306          Chicken Nugget Spicy Garlic  15.0    0.0     15.0    0.500   Low
308          Naget Ayam Kombinasi So Good  10.0    0.0     19.0    0.500   Low
311         Premium Chicken Stick So Good  14.0    0.0     12.0    0.690   Low
312  Chicken Nugget Jetz Original So Good   8.0    0.0      7.0    0.290   Low
..                                    ...   ...    ...      ...      ...   ...
130                  Hilo Es Ketan Hitam   1.5    8.0      0.0    0.170  High
129           Tong Tji Tematik Chocolate   2.5   13.0      3.0    0.095  High
128                           Jahe Wangi   0.0   16.0      0.0    0.055  High
127                 Anget Sari Susu Jahe   1.5   16.0      1.0    0.035  High
398                Murgerbon Strawberry  10.0    5.0      3.0    0.035  High
```

Figure 3. K-Means Clustering Result

Figure 3 above shows the clustering results of products using the K-Means method. It is evident that some packaged beverages are at high risk of triggering diabetes. The scatter plot results from using K-Means are displayed in the following figure.

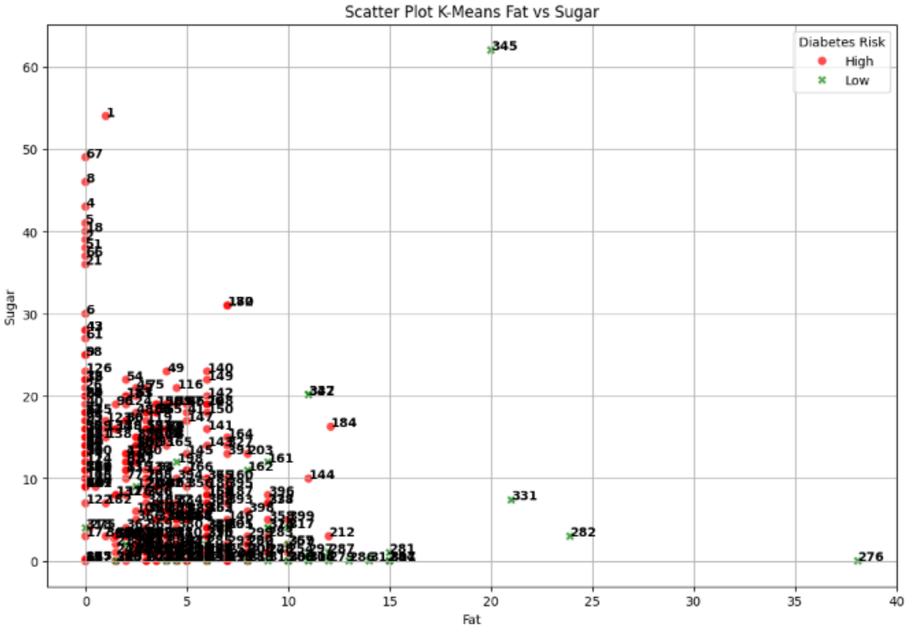
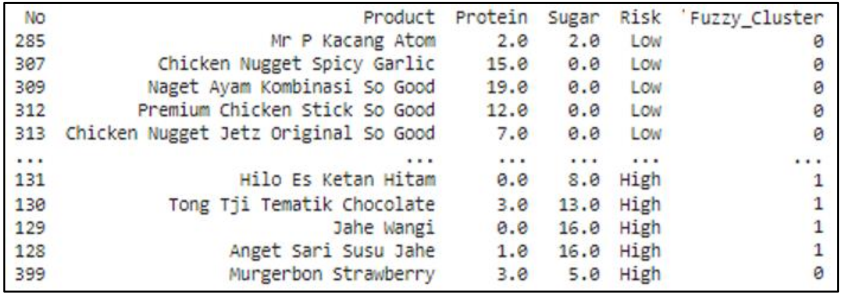


Figure 4. K-Means Clustering Scatter Plot Results

Figure 4 is a scatter plot showing the clustering results using the K-Means algorithm, which illustrates the relationship between fat content and sugar content in products labeled with diabetes risk. The plot reveals that many red dots are concentrated at lower values of fat and sugar content, indicating that many high-risk diabetes products do not have very high fat and sugar content. Green dots are more variably distributed, with some having higher fat and sugar content, suggesting that low-risk diabetes products can have a wider range of fat and sugar content.

In this plot, products with low fat and sugar content tend to be grouped in the high diabetes risk cluster. Products with high fat and sugar content tend to be in the low diabetes risk cluster, which may require further interpretation related to other factors influencing diabetes risk. There are some outliers, such as products numbered 345, 331, 282, and 276, which have very high fat or sugar content but are categorized as low diabetes risk. Products numbered 189 and 184 are also noteworthy as they have very high sugar content but fall into different categories.

The clustering results using Fuzzy C-Means can be seen in the following figure.



```
No                            Product  Protein  Sugar  Risk  'Fuzzy_Cluster
285                      Mr P Kacang Atom      2.0    2.0   Low              0
307          Chicken Nugget Spicy Garlic     15.0    0.0   Low              0
309           Naget Ayam Kombinasi So Good     19.0    0.0   Low              0
312         Premium Chicken Stick So Good     12.0    0.0   Low              0
313  Chicken Nugget Jetz Original So Good      7.0    0.0   Low              0
...                                   ...      ...    ...   ...            ...
131                  Hilo Es Ketan Hitam      0.0    8.0  High              1
130            Tong Tji Tematik Chocolate      3.0   13.0  High              1
129                          Jahe Wangi      0.0   16.0  High              1
128                  Anget Sari Susu Jahe      1.0   16.0  High              1
399                 Murgerbon Strawberry      3.0    5.0  High              0
```

Figure 5. Clustering Results Using Fuzzy C-Means

Figure 5 above shows the clustering of products using Fuzzy C-Means. The 'Risk' column represents the clustering using K-Means, where low-risk products are in cluster 0 and high-risk products are in cluster 1. It can be seen that low-risk K-Means falls into cluster 0 of Fuzzy C-Means, while high-risk falls into cluster 1. There are also high-risk products that fall into cluster.

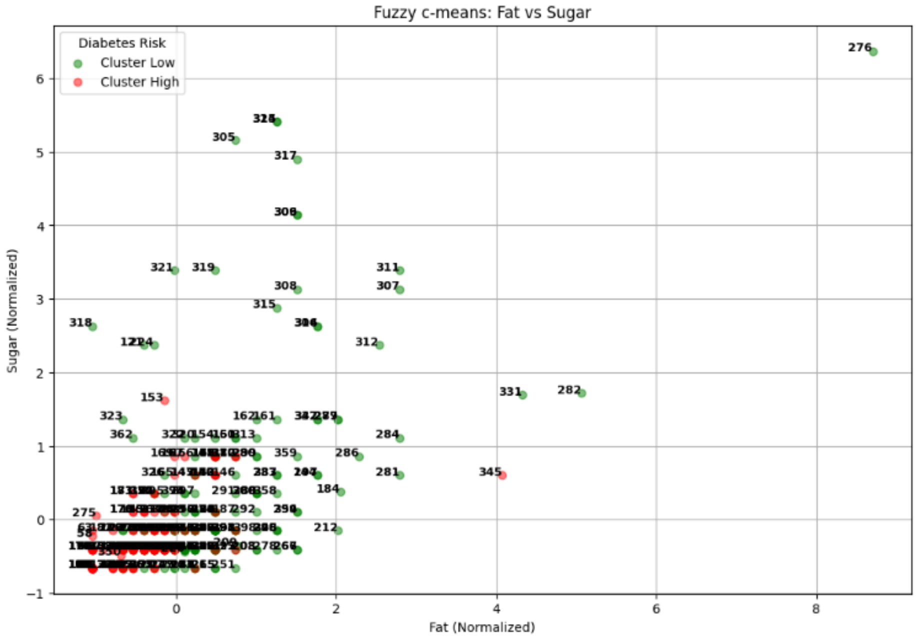


Figure 6. Fuzzy C-Means Clustering Scatter Plot Results

Figure 6 is a scatter plot that illustrates the clustering results using the algorithm of Fuzzy C-Means, which illustrates the relationship between fat content and sugar content in products labeled with diabetes risk. The plot depicts the data distribution, where many green dots are concentrated around lower values of fat and sugar content, indicating that many low diabetes risk products have relatively low fat and sugar content. Some green dots are more variably distributed, with some products having higher fat and sugar content. Red dots, indicating high diabetes risk, are scattered around products with varying fat and sugar content but are more concentrated in the lower range.

In this plot, products with low fat and sugar content tend to be grouped in the low diabetes risk cluster. Some products with higher fat content and lower sugar content are in the low diabetes risk cluster, which may require further analysis to understand the factors influencing this classification. There are some outliers, such as product number 276, which has very high fat and sugar content but is still in the low diabetes risk cluster. Product number 345 is also noteworthy as it has high fat content but falls into the high diabetes risk cluster.

5 Conclusion

Both clustering methods, K-Means and Fuzzy C-Means, are effective in identifying food and beverage products with high and low risk of diabetes. The results of this study can provide important guidance for consumers in selecting healthier products based on their nutritional content. There are some outliers indicating that products with high fat and sugar content can fall into the low-risk cluster, which requires further analysis to understand the factors influencing this classification. This study makes a significant contribution by combining the two clustering methods for the analysis of product nutrition labels, which has not been done before, and can serve as a basis for further research in assessing diabetes risk based on product nutritional content.

Acknowledgement

This research was funded by the Directorate General of Higher Education, Research, and Technology. The Ministry of Education and Culture through the Penelitian Dosen Pemula (PDP) scheme for the 2024 fiscal year.

References

[1] R. Birjais, A. K. Mourya, R. Chauhan, and H. Kaur, “Prediction and diagnosis of future diabetes risk: a machine learning approach,” *SN Appl Sci*, vol. 1, no. 9, Sep. 2019, doi: 10.1007/s42452-019-1117-9.

[2] L. Wang, X. Wang, A. Chen, X. Jin, and H. Che, “Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model,” *Healthcare (Switzerland)*, vol. 8, no. 3, 2020, doi: 10.3390/healthcare8030247.

[3] O. : Marendra, S. Kartolo, and A. H. Santoso, “Hubungan Frekuensi Konsumsi, Asupan Energi, Lemak, Gula, dan Garam dalam Fast Food dengan Kejadian Obesitas pada Siswa/I SMP X Yogyakarta,” *EBERS PAPYRUS*, vol. 28, no. 1, p. 38, 2022.

[4] U. e. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, “An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study,” *Sensors*, vol. 22, no. 14, Jul. 2022, doi: 10.3390/s22145247.

[5] L. Paper *et al.*, “Cross-sectional comparisons of dietary indexes underlying nutrition labels: nutri-score, Canadian ‘high in’ labels and Diabetes Canada Clinical Practices (DCCP),” *Eur J Nutr*, vol. 62, no. 1, pp. 261–274, Feb. 2023, doi: 10.1007/s00394-022-02978-w.

[6] D. Westari, “Performa Comparison of the K-Means Method for Classification in Diabetes Patients Using Two Normalization Methods,” *International Journal of Multidisciplinary Research and Analysis*, vol. 04, no. 01, Jan. 2021, doi: 10.47191/ijmra/v4-i1-03.

[7] Simeftiany Indrilemta Lomo, Endang Darmawan, and Sugiyarto, “Cluster analysis of type II Diabetes Mellitus Patients with the Fuzzy C-means method,” *Annals of Mathematical Modeling*, vol. 3, no. 1, pp. 24–31, Jun. 2023, doi: 10.33292/amm.v3i1.28.

[8] S. Surono and E. Darmawan, “The Risk Cluster in Type 2 Diabetes Mellitus Based on Risk Parameters Using Fuzzy C-Means Algorithm,” 2023. [Online]. Available: *https://doi.org/11.26554/sti.2223.8.1.17-24*

[9] H. Naz, T. Saba, F. S. Alamri, A. S. Almasoud, and A. Rehman, “An Improved Robust Fuzzy Local Information K-Means Clustering Algorithm for Diabetic Retinopathy Detection,” *IEEE Access*, vol. 12, pp. 78611–78623, 2024, doi: 10.1109/ACCESS.2024.3392032.

[10] S. Kusumadewi, L. Rosita, and E. G. Wahyuni, “Performance of Fuzzy C-Means (FCM) and Fuzzy Subtractive Clustering (FSC) on Medical Data Imputation,” *ComTech: Computer, Mathematics and Engineering Applications*, vol. 15, no. 1, pp. 29–40, May 2024, doi: 10.21512/comtech.v15i1.11002.

[11] Z. Chen *et al.*, “Ultra-Processed Food Consumption and Risk of Type 2 Diabetes: Three Large Prospective U.S. Cohort Studies,” *Diabetes Care*, vol. 46, no. 7, pp. 1335–1344, Jul. 2023, doi: 10.2337/dc22-1993.

[12] M. Mehedi Hassan, S. Mollick, and F. Yasmin, “An unsupervised cluster-based feature grouping model for early diabetes detection,” *Healthcare Analytics*, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100112.

[13] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, “Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis,” Jan. 01, 2021, *KeAi Communications Co.* doi: 10.1016/j.ceh.2020.11.001.

ISSN:2302-8149
e-ISSN:2540-9719

[14] K. El Moutaouakil, A. Yahyaouy, S. Chellak, and H. Baizri, “An Optimized Gradient Dynamic-Neuro-Weighted-Fuzzy Clustering Method: Application in the Nutrition Field,” *International Journal of Fuzzy Systems*, vol. 24, no. 8, pp. 3731–3744, Nov. 2022, doi: 10.1007/s40815-022-01358-0.

[15] R. Gestavito, A. Id Hadiana, F. Rakhmat Umbara, and U. Jenderal Achmad Yani Jl Terusan Jenderal Sudirman, “Pengelompokan Tingkat Risiko Penyakit Diabetes Melitus Menggunakan Algoritma K-Means Clustering,” *Jurnal Masyarakat Informatika Unjani*, vol. 8, no. 1, pp. 16–35, 2024.

[16] I. Tri Gustiane and T. Suprapti, “Clustering Hasil Cek Darah Diabetes Lansia Menggunakan Metode K-Means di Posbindu Kp. Lebakjero Desa Ciherang,” *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 2, 2024.

[17] N. Husna, F. Hanum, and M. F. Azrial, “Pengelompokkan Produk Kemasan yang Harus Dihindari Penderita Diabetes Menggunakan Algoritma K-Means Clustering,” *InfoTekJar (Jurnal Nasional Informatika dan Teknologi Jaringan)*, vol. 4, no. 1, pp. 167–174, Sep. 2019, doi: 10.30743/infotekjar.v4i1.1484.

[18] A. A. Jasim, L. R. Hazim, H. Mohammedqasim, R. Mohammedqasem, O. Ata, and O. H. Salman, “e-Diagnostic system for diabetes disease prediction on an IoMT environment-based hyper AdaBoost machine learning model,” *Journal of Supercomputing*, Jul. 2024, doi: 10.1007/s11227-024-06082-0.

[19] S. Dörterler, H. Dumlu, D. Özdemir, and H. Temurtaş, “Hybridization of Meta-heuristic Algorithms with K-Means for Clustering Analysis: Case of Medical Datasets,” *Gazi Journal of Engineering Sciences*, vol. 10, no. 1, pp. 1–11, Apr. 2024, doi: 10.30855/gmbd.0705n01.

[20] S. J. Setu, F. Tabassum, S. Jahan, and Md. I. Islam, “Detection of Diabetes using Combined ML Algorithm,” *International Journal of Intelligent Systems and Applications*, vol. 16, no. 1, pp. 11–23, Feb. 2024, doi: 10.5815/ijisa.2024.01.02.

[21] O. Virgolici, B. Virgolici, and " Carol, “Diabetes Prediction Using Machine Learning Techniques: A Brief Overview Diabetes & its Complications,” vol. 8, p. 2024, 2024.