https://doi.org/ 10.48047/AFJBS.6.Si2.2024.5945-5954

African Journal of Biological Sciences

Journalhomepage: http://www.afjbs.com

Research Paper Open Access

Clustering Breast Cancer Patients Using K-means: A Comprehensive Analysis of Socio-Economic, Genetic, and Clinical Variables for Enhanced Treatment Strategies

Mouas Fatima Ezahra[1], Rais Ghizlane[2], Moussaid Abdellatf[3], Fechtali Taoufiq[4]

[1] *Hassan II University Casablanca, Laboratory of Engineering Sciences and Biosciences, Morocco. Fatimazahramouas@gmail.com*

[2] *Medical Oncology Department, Faculty of Medicine and Pharmacy of Agadir, Ibn Zohr University, Souss Massa University Hospital, Agadir, Morocco. g.rais@uiz.ac.m*

[3] *National School of Science and System Analysis, Mohammed 5 university in Rabat, Morocco. Abdellatif.moussaid@gmail.com*

[4] *Hassan II University Casablanca, Laboratory of Engineering Sciences Laboratory of Engineering Sciences and Biosciences, Morocco. taoufiq.fechtali@fstm.ac.ma*

Abstract-This study analyzes 354 questionnaires from breast cancer patients at Hassan II Hospital in Agadir, Morocco, encompassing socio-economic, genetic, and clinical variables. Our approach employs the K-means clustering algorithm to partition patients into groups based on Euclidean distance, with the optimal number of clusters determined by the Elbow method. We visualize clustering results using Principal Axis Correlation (PAC) and identify key variables influencing cluster formation through a correlation matrix. The clusters are characterized by distinct patterns in age, menopause status, and parity, underscoring the significance of these factors in breast cancer treatment and patient grouping.

Keywords: Breast Cancer, Machine Learning, K-means, Unsupervised Learning.

1. Introduction

Breast cancer stands as a formidable global health challenge, ranking as the second leading cause of cancer-related mortality among women with a staggering 685,000 fatalities in 2020 [1]. This alarming statistic underscores the urgent necessity for innovative and optimized therapeutic approaches to confront the complexities of this disease. Breast cancer is intricately woven into the fabric of genetic mutations that disrupt normal cellular functioning, resulting in uncontrolled proliferation and infiltration into surrounding tissues [2]. To effectively address this challenge, a comprehensive understanding of the disease's dynamics and a nuanced exploration of therapeutic strategies are imperative.

Breast cancer's impact surpasses statistical figures, affecting diverse regions globally and underscoring the need for collaborative efforts to optimize treatment. Risk factors such as family history, obesity, alcohol consumption, and age at first menstruation provide a nuanced understanding of the disease [3]. This information serves as a valuable resource for tailoring personalized therapeutic strategies.

In this paper, we have collected a dataset composed of several factors related to patients from Hassan II Hospital in Morocco to perform clustering of these patients. The objective is to assist doctors in identifying the optimal grouping of cases to make informed decisions for each group.

In the remaining sections, we will provide a state-of-the-art review of works related to breast cancer, with a particular focus on clustering patients with breast cancer. Following this, we will discuss our approach, including the data used and the methodology employed. We will then present the obtained results and their discussion. Finally, we will conclude with our findings and recommendations for future research.

2. Related Works

Breast cancer treatment necessitates a multidisciplinary approach, encompassing a range of methods from conservative surgery to chemotherapy and immunotherapy, but Early detection still the crucial as it significantly increases patient survival rates and reduces both mortality and treatment costs [4]. Various studies have focused on employing machine learning methods for the early detection and classification of breast cancer, highlighting the potential of these technologies to improve diagnostic accuracy and treatment outcomes. In this way, Machine learning techniques have been extensively applied in the classification of mammography images, which are critical for breast cancer screening [5]. Methods such as Support Vector Machines (SVMs) [6], neural

networks [7], K-Nearest Neighbors (K-NN) [8], and others have shown promising results in this field. For example, Zheng et al. combined k-means clustering with SVM (K-SVM) to extract tumor features, achieving an impressive accuracy of 97.38% in breast cancer diagnosis [9]. This underscores the effectiveness of combining clustering algorithms with other machine learning techniques to enhance the classification of mammography images.

Also, Clustering algorithms, particularly k-means, are widely recognized in the field of data analysis due to their simplicity and effectiveness. K-means clustering is a popular choice for tasks such as image segmentation and pattern recognition, which are essential in the analysis of medical images [10]. In this way, Alam et al. developed an automated method for detecting clusters of microcalcifications in mammograms using morphological operations and contrast filters to reduce image noise [11]. Their approach achieved a classification accuracy of 94.48% for the Digital Database for Screening Mammography (DDSM) and 100% for the Mammographic Image Analysis Society (MIAS) database, demonstrating the potential of clustering methods in improving diagnostic accuracy. Also, several studies have evaluated different clustering methods for detecting microcalcifications in mammograms. For example, the study of [12] use a 2D median filter alongside clustering, achieving 90.0% sensitivity and 78.0% specificity on MIAS data. Another study [13] found that employing k-means with a 5×5 median filter yielded a sensitivity of 94.4% on MIAS data. And also, the study Darweesh et al. using hierarchical clustering to DDSM data resulted in lower performance, with only 38.8% accuracy and a 61.1% testing error. These findings highlight the variability in performance across different clustering techniques and datasets [14].

On the other hand, Various risk factors contribute to the development of breast cancer, including family history, obesity, alcohol consumption, and age at first menstruation [15]. Understanding these factors is critical for tailoring personalized therapeutic strategies. The integration of AI and machine learning in breast cancer research has the potential to significantly impact the field by enabling the analysis of large datasets to identify patterns and correlations that may not be immediately apparent through traditional methods. The application of AI, particularly clustering algorithms, in breast cancer research offers numerous benefits. By analyzing patient data, clustering algorithms can identify subgroups of patients with similar characteristics, which can inform more

personalized and effective treatment plans [16]. For instance, clustering patients based on genetic markers, tumor characteristics, and treatment responses can help in optimizing therapeutic strategies and improving patient outcomes.

3. Materials and Methods

Data

- ***Data Description***

This study relies on the analysis of 354 questionnaires from breast cancer patients collected at Hassan II hospital in Agadir, Morocco. The dataset encompasses socio-economic, genetic, and clinical variables, providing a comprehensive overview of factors potentially influencing breast cancer prognosis and treatment outcomes. The dataset contains 354 rows and 33 columns, representing various aspects of the patients' demographic, medical, and treatment information. The variables can be categorized into different types, including categorical, numerical, and binary variables (figure 1). Socio-economic variables include origin (rural or urban), level of study (ranging from illiterate to higher education), marital status (married, single, divorced, etc.), occupation, and financial situation (ranging from difficult to stable). Genetic and personal history variables include family history (yes or no), diabetes (yes or no), parity (number of children), menarche (age at first menstruation), menopause (yes or no), and oral contraceptives usage status. Clinical variables cover age, BMI (Body Mass Index), physical activity level (inactivity, active, etc.), tumor location (left breast or right breast), hormone receptor status (yes or no), HER2 status (yes or no), cancer histology (IDC, etc.), stage, SBR grade, tumor size range, anemia (yes or no), metastasis (yes or no), and extramural extension (yes or no). Psychological and lifestyle variables include psychological state (anxious, sad, etc.), coping mechanisms (yes or no), HT (hormone therapy), and cholesterol (yes or no). Treatment variables encompass various combinations of treatments like treatment with OP (operation), CTX (chemotherapy), HT (hormone therapy), RTH (radiation therapy), and TT (targeted therapy).

To better understand the composition of our dataset, we present a bar graph (figure 1) displaying the number of categorical and numerical variables information effectively.

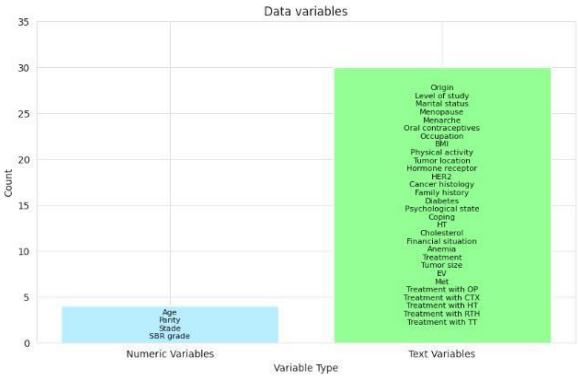

Figure 1. Data

- ***Data Preparation***

To prepare our data for analysis, we first verified the presence of missing values and anomalies. Given that our data was manually collected based on our questionnaire, we found no missing values or anomalies. Therefore, this step was not necessary. However, several variables in our dataset are categorical text variables. To handle these categorical variables effectively, we employed the Label Encoder [17]. This method maintains the integrity of our variables by mapping each category to a unique integer.

The Label Encoder is particularly useful because it preserves the categorical nature of the variables while converting them into a format suitable for machine learning algorithms. Each category is assigned a unique numerical label, facilitating the analysis without altering the inherent structure of the data.

Approach

Our approach consists of constructing several groups (clusters) based on the Euclidean distance between each row in our data. This is achieved using the k-means clustering algorithm, which aims to partition the data into k clusters, where each data point belongs to the cluster with the nearest centroid (mean of the points in the cluster) [18].

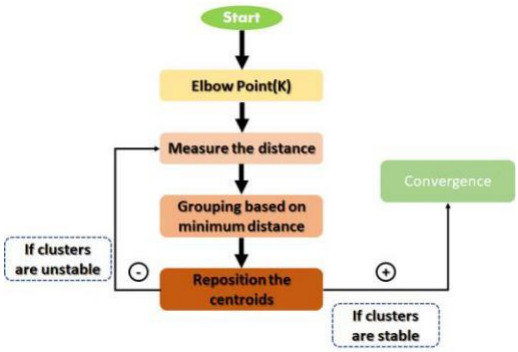

Figure 2. K-means Algorithm with Elbow Method

As mentioned in figure 2, the algorithm starts by Choosing the Number of Clusters (k) using the Elbow Method [19] that involves running k-means clustering on the dataset for a range of values of k (e.g., from 1 to 14) and computing the Within-Cluster Sum of Squares (WCSS) for each k. WCSS is the sum of squared distances between each point and the centroid of its assigned cluster. Mathematically, it is represented in equation 1:

$$\mathrm{WCSS} = \sum_{i=1}^{k} \sum_{\mathrm{xj} \in \mathrm{Ci}} \| \mathrm{xj} - \mu \mathrm{i} \|^2 \qquad (1)$$

where Ci is the i^{th} cluster, x_j is a data point in cluster Ci, and μ_i is the centroid of cluster Ci. As k increases, WCSS decreases. The goal is to find the "elbow point" where the rate of decrease sharply slows down, indicating the optimal number of clusters.

After that, each data point is assigned to the nearest centroid based on the Euclidean distance (formula 2).

$$\mathrm{d(x,y)} = \sqrt{\sum_i^n (x_i - y_i)^2} \qquad (2)$$

Once all points are assigned to clusters, the centroids are recalculated as the mean of

all points in the cluster. The new centroid μi of cluster Ci is computed as:

$$\mu_i = \frac{1}{|C_i|} \sum_{xj \in Ci} x_j \quad (3)$$

And finally, the process repeated until the centroids no longer change significantly, indicating convergence of the algorithm.

In our case, the elbow method (Figure 3) indicates that the optimal number of clusters to choose is 3. This is identified by observing the point where the Within-Cluster Sum of Squares (WCSS) starts to diminish at a slower rate, forming an "elbow" shape. Selecting 3 clusters balances the trade-off between having too many clusters, which might lead to overfitting, and too few clusters, which might not capture the inherent structure of the data effectively.

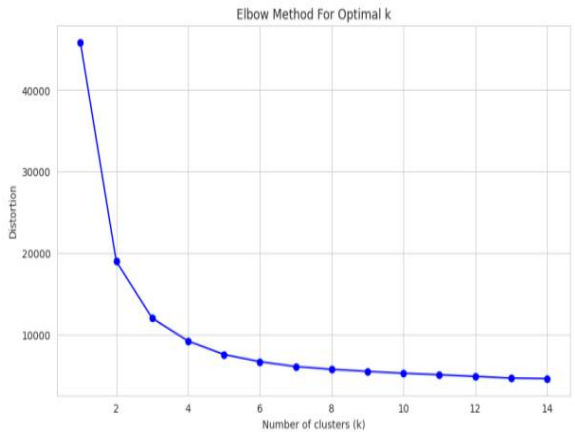

Figure 3. Elbow Results

After determining the optimal number of clusters, we proceed with the K-means algorithm using K=3 to partition the patients into three distinct clusters.

4. Results and Discussion

In order to visualize the results of K-means, we employed the Principal Axis Correlation (PAC) algorithm, which calculates the dependence between our variables and the newly derived cancer clusters. This method helps in understanding how the original variables influence the cluster formation. As shown in Figure 4, the three groups are well represented in the two components of PCA, demonstrating the robustness and reliability of our K-means clustering. This visualization confirms that the clusters are distinct and meaningful, indicating that the algorithm has effectively partitioned the data based on the underlying patterns.

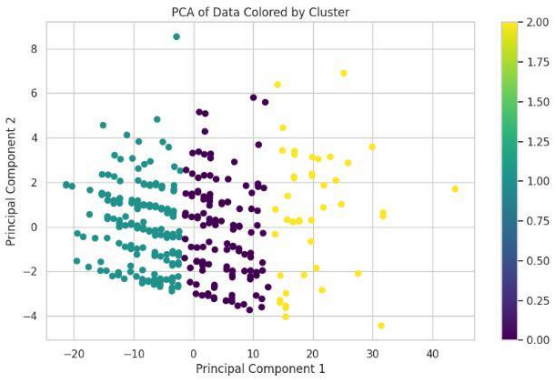

Figure 4. Clustering Results

Furthermore, we generated a correlation matrix between the original variables and the cluster assignments to identify which variables are most strongly associated with each cluster. By plotting this matrix as a heatmap (Figure 5), we were able to visually inspect the relationships and dependencies within the data.

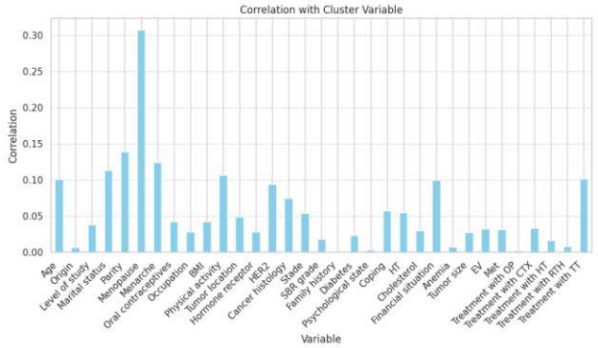

Figure 5. Correlation Matrix

As shown in Figure 5, all variables exhibit some degree of correlation with the clustering results, indicating that the clustering was influenced by a comprehensive set of variables. However, the most strongly correlated variables are Menopause, Age, Parity, and Physical Activity. These variables had the highest correlation coefficients, suggesting that they played a significant role in the formation of the clusters and are key factors in distinguishing between different patient groups.

To further analyze the results, we examined the distribution of these highly correlated variables by cluster using a group-by technique. This method allowed us to visualize how each category within these variables is distributed across the different clusters, providing deeper insights into the characteristics and differences between the patient groups.

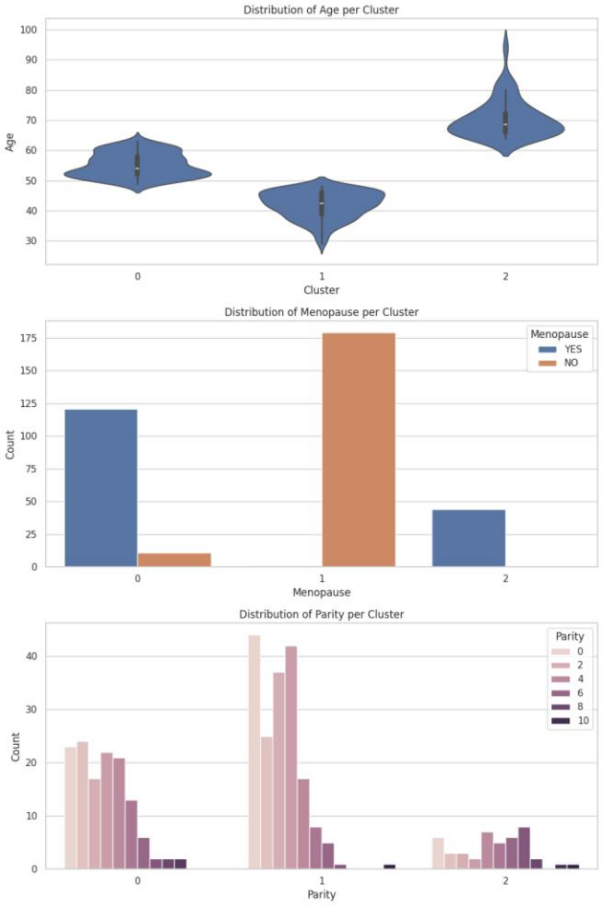

Figure 6. Distributions of the Most Correlated Variables

As shown in Figure 6, the distribution of age demonstrates that the K-means algorithm effectively segments the patients into three distinct age groups, with each cluster representing a specific age range. This segmentation validates the hypothesis that age is a crucial variable in breast cancer treatment, aligning with research that highlights age as a significant factor influencing treatment decisions and outcomes.

Similarly, Menopause status significantly distinguishes the clusters. Patients who have undergone menopause are not present in Cluster 1, while Cluster 2 predominantly

includes patients who have not experienced menopause. This suggests that menopause status plays a vital role in differentiating patient groups and underscores the need for tailored treatment approaches based on this variable. Some patients in Cluster 1 may still be grouped together based on other variables, indicating complex interactions between factors.

For the variable Parity, which represents the number of children a woman has, the clusters also show distinct patterns. All clusters include patients across the parity spectrum, but Cluster 1 has a higher concentration of patients with fewer children, while Cluster 2 includes more patients with four or more children. This distribution suggests that parity is a significant factor in understanding patient profiles and their corresponding treatment plans.

Overall, the group-by analysis of these key variables (Age, Menopause, and Parity) by cluster provides valuable insights into the distinct characteristics of each patient group, reinforcing the importance of these variables in the context of breast cancer treatment and cluster formation.

5. Conclusion

To conclude, our study demonstrates the effectiveness of the K-means algorithm in clustering breast cancer patients based on comprehensive socio-economic, genetic, and clinical data. The Elbow method indicated an optimal cluster number of three, and PAC analysis confirmed the robustness of these clusters. Key variables such as age, menopause status, and parity emerged as significant factors influencing cluster formation. This clustering approach can help in distinguishing patient groups, thereby facilitating more personalized and effective treatment plans. Also, in a hospital context, this method can help doctors quickly and reliably group patients, enhancing decision-making and treatment strategies for breast cancer management.

References

[1] A. H. Jafarian, M. Kooshkiforooshani, A. Rasoliostadi, and N. M. Roshan, Vascular mimicry expression in invasive ductal carcinoma; a new technique for prospect of aggressiveness, *Iran. J. Pathol.*, vol. 14, no. 3, p. 232, 2019.

[2] R. Ranjbarzadeh *et al.*, MRFE-CNN: multi-route feature extraction model for breast tumor segmentation in Mammograms using a convolutional neural network, *Ann. Oper. Res.*, vol. 328, no. 1, pp. 1021–1042, Sep. 2023, doi: 10.1007/s10479-022-04755-8.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, Cancer statistics, 2018, *CA.*

Cancer J. Clin., vol. 68, no. 1, pp. 7–30, Jan. 2018, 10.3322/caac.21442.

[4] M. Milosevic, D. Jankovic, A. Milenkovic, and D. Stojanov, Early diagnosis and detection of breast cancer, *Technol. Health Care*, vol. 26, no. 4, pp. 729–759, 2018.

[5] G. Meenalochini and S. Ramkumar, Survey of machine learning algorithms for breast cancer detection using mammogram images, *Mater. Today Proc.*, vol. 37, pp. 2738–2743, 2021.

[6] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai, SVM and SVM ensembles in breast cancer prediction, *PloS One*, vol. 12, no. 1, p. e0161501, 2017.

[7] F. F. Ting, Y. J. Tan, and K. S. Sim, Convolutional neural network improvement for breast cancer classification, *Expert Syst. Appl.*, vol. 120, pp. 103–115, 2019.

[8] S. F. Khorshid and A. M. Abdulazeez, ―Breast cancer diagnosis based on k-nearest neighbors: a review, *Pal Archs J. Archaeol. Egypt Egyptology*, vol. 18, no. 4, pp. 1927–1951, 2021.

[9] B. Zheng, S. W. Yoon, and S. S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, 2014.

[10] A. K. Dubey, U. Gupta, and S. Jain, Analysis of k-means clustering approach on the breast cancer Wisconsin dataset, *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 11, pp. 2033–2047, Nov. 2016, doi: 10.1007/s11548-016-1437-9.

[11] N. Alam, E. RE Denton, and R. Zwiggelaar, Classification of microcalcification clusters in digital mammograms using a stack generalization-based classifier, *J. Imaging*, vol. 5, no. 9, p. 76, 2019.

[12] E. Michael, H. Ma, H. Li, F. Kulwa, and J. Li, ―Breast Cancer Segmentation Methods: Current Status and Future Potentials, *BioMed Res. Int.*, vol. 2021, pp. 1–29, Jul. 2021, doi: 10.1155/2021/9962109.

[13] K. M. Prabusankarlal, P. Thirumoorthy, and R. Manavalan, Computer aided breast cancer diagnosis techniques in ultrasound: a survey, *J. Med. Imaging Health Inform.*, vol. 4, no. 3, pp. 331– 349, 2014.

[14] M. S. Darweesh *et al.*, Early breast cancer diagnostics based on hierarchical machine learning classification for mammography

images, *Cogent Eng.*, vol. 8, no. 1, p. 1968324, Jan. 2021, doi: 10.1080/23311916.2021.1968324.

[15] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review, *Cancers*, vol. 13, no. 17, p. 4287, 2021.

[16] W. L. Al-Yaseen, A. Jehad, Q. A. Abed, and A. K. Idrees, The Use of Modified K-Means Algorithm to Enhance the Performance of Support Vector Machine in Classifying Breast Cancer., *Int. J. Intell. Eng. Syst.*, vol. 14, no. 2, 2021, Accessed: Jun. 08, 2024. [Online]. Available: http://www.inass.org/2021/2021043017.pdf

[17] E. Poslavskaya and A. Korolev, Encoding categorical data: Is there yet anything ‗hotter ‘than one-hot encoding? arXiv, Dec. 28, 2023. Accessed: Jun. 08, 2024. [Online]. Available: http://arxiv.org/abs/2312.16930

[18] A.M. Ikotun, A.E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, *Inf. Sci.*, vol. 622, pp. 178–210, 2023.

[19] M. Cui, Introduction to the k-means clustering algorithm based on the elbow method, *Account. Audit. Finance*, vol. 1, no. 1, pp. 5–8, 2020.