Breast Cancer Prediction Using Ultrasound Images

Kaustubh Raykar
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Maharashtra, Pune
kaustubhk.raykar@gmail.com

Nachiket Ropia
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Maharashtra, Pune
nachiketropia@gmail.com

Najeeb Fariduddin Saiyed
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Maharashtra, Pune
officialnajeebsaiyed@gmail.com

Prachi Nitin Kadam
Assistant Professor (Department of Artificial Intelligence and Machine Learning)
Symbiosis Institute of Technology
Maharashtra, Pune

Dr. Pooja Kamat
Assistant Professor (Department of Artificial Intelligence and Machine Learning)
Symbiosis Institute of Technology
Maharashtra, Pune

Mayur Gaikwad
Faculty (Department of Artificial Intelligence and Machine Learning)
Symbiosis Institute of Technology
Maharashtra, Pune

***Abstract*—Breast cancer is a major public health concern, and accurate diagnosis is essential for effective treatment. In this study, we applied four clustering algorithms, K-Means, Agglomerative clustering, DBSCAN, and BIRCH, to ultrasound images of breast cancer to explore their potential for predicting the disease. We evaluated the performance of each algorithm using the Silhouette score, a metric that measures the quality of clustering results. Our results indicate that Agglomerative clustering and K-Means produced moderate to fair separation between the clusters, whereas DBSCAN and BIRCH performed suboptimally. These findings suggest that the choice of clustering algorithm is crucial in obtaining meaningful results for breast cancer prediction. Our study highlights the importance of evaluating clustering algorithms using appropriate metrics, and provides valuable insights for researchers and healthcare professionals in the field of breast cancer diagnosis..**

Keywords: Breast cancer prediction; Clustering algorithms k-means; agglomerative; DBSCAN; BIRCH; ultrasound images

I. Introduction

Breast cancer is a leading cause of death among women worldwide. Early detection is crucial for improved treatment outcomes. Ultrasound imaging is a widely used diagnostic tool for breast cancer, but the manual analysis process can be time- consuming and prone to subjective interpretation. Machine learning in medical imaging has shown promising results in the prediction of various diseases, including breast cancer. Machine learning has the potential to predict breast cancer based on features hidden in data. In this study, we aim to develop a machine learning model to assist in the prediction of breast cancer from ultrasound images. The aim is to develop a model that can accurately and efficiently diagnose breast cancer from ultrasound images, potentially reducing the need for manual interpretation and increasing the accuracy of diagnosis.

II. Background

The use of machine learning in medical imaging has gained traction in recent years as a means of improving the accuracy and speed of disease diagnosis. In the field of breast cancer diagnosis, machine learning has been applied to various imaging modalities, including mammography, magnetic resonance imaging (MRI), and ultrasound. Studies have shown that machine learning models can assist in the detection of breast cancer and improve diagnostic accuracy compared to manual analysis. Ultrasound imaging is widely used for breast cancer screening and diagnosis, but the manual interpretation of images by radiologists is subject to human error and can be time-consuming. Machine learning has the potential to enhance the accuracy and efficiency of breast cancer prediction from ultrasound images.

A. *More on Breast cancer*

Breast cancer is a type of cancer that originates from the cells of the breast. It is a common cancer among women and can also occur in men, although it is rare. Symptoms of breast cancer can include a lump in the breast, changes in the size or shape of the breast, skin changes such as redness or dimpling, and fluid discharge from the nipple. The incidence and prevalence of breast cancer has increased significantly over the past few decades, making it a major public health concern.

B. *Risk factors*

Several factors contribute to the development of breast cancer. Age, gender, family history of breast cancer, lifestyle factors such as alcohol consumption and lack of physical activity, and hormonal factors such as early onset of menstruation and late menopause, are among the most prominent risk factors.

C. *Symptoms*

The early detection of breast cancer is crucial for improving patient outcomes. The common symptoms of breast cancer include the presence of a lump or thickening in the breast, changes in the size or shape of the breast, skin changes such as redness or dimpling, and fluid discharge from the nipple.

D. *Diagnosis*

Diagnosis of breast cancer typically involves a combination of physical examination, imaging techniques such as mammography or ultrasound, and biopsy of the suspicious tissue. The treatment of breast cancer is based on various factors such as the stage of cancer, the patient's overall health, and the patient's personal preferences. The available options include surgery, radiation therapy, chemotherapy, hormone therapy, and targeted therapy.

III. Problem Statement

The incidence of breast cancer amongst women has been increasing globally, highlighting the need for effective and efficient methods of early detection. Although manual screening techniques are available, they are often associated with human error and are time-consuming. To address these challenges, this study aims to develop a machine learning-based model for predicting the presence of breast cancer in women using patient information and imaging data. The objective is to provide healthcare professionals with a reliable and efficient tool to make informed decisions regarding patient diagnosis, treatment, and follow-up, and ultimately improve the survival rate of breast cancer patients using machine learning.

IV. About the dataset

The ”Dataset of breast ultrasound images” on Sciencedirect.com is a collection of ultrasound images of the breast. The Data has come directly from Baheya Hospital for early detection and treatment of Women's Cancer, Cairo, Egypt. The Data was acquired using LOGIQ E9 ultrasound and LOGIQ E9 Agile ultrasound system. This dataset provides a valuable resource for machine learning researchers and practitioners who are interested in developing algorithms for breast cancer prediction from ultrasound images. The dataset allows for the training and evaluation of machine learning models for breast cancer prediction and provides a foundation for further research in this area.

Given the importance of early detection in the treatment of breast cancer, the ”Breast Ultrasound Images Dataset” has the potential to contribute to the development of more accurate and efficient diagnostic tools for breast cancer screening. The use of machine learning algorithms on this dataset could lead to the creation of a system that can accurately diagnose breast cancer from ultrasound images, potentially reducing the need for manual interpretation and increasing the accuracy of diagnosis.

A. *Explanation of the dataset*

The ”Dataset of breast ultrasound images” available on Sciencedirect.com is a dataset of ultrasound images of women ranging in age from 25 to 75 years old. This data was gathered in 2018 from 600 female patients and includes a total of 780 images. The average size of each image is 500 x 500 pixels, and they are stored in PNG format. The dataset provides both the original images and corresponding annotated images which categorize the images into three classes: normal, benign, and malignant. The dataset in question offers a thorough collection of data related to patients with breast cancer, which will serve as a crucial resource in training and evaluating the machine learning models for this research project.

V. Workflow

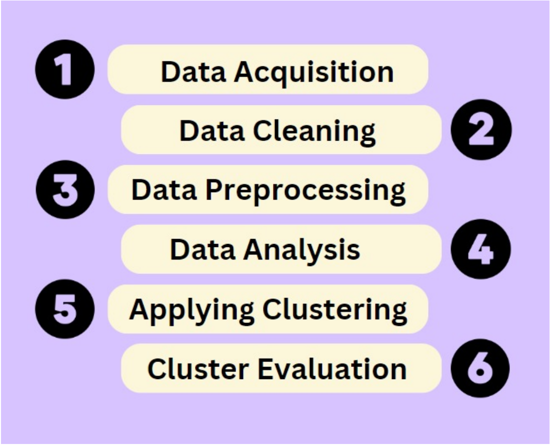

Fig. 1: Workflow

Workflow refers to the order of actions that are carried out in a specific procedure. Concerning data analysis, it encompasses different phases, including the acquisition of data, refining data, pre-processing data, conducting exploratory data analysis, reducing dimensionality, and clustering.

Data Collection - During data collection, data is gathered from different sources and converted into a usable format. In the next stage of data cleaning, duplicates are removed, missing values are checked and other necessary data cleaning tasks such as filtering and image enhancement are performed.

Data pre-processing - Data pre-processing is an essential step in data analysis, PCA and standard scaling were used for our data preprocessing before applying clustering algorithms such as DBSCAN, agglomerative clustering, BIRCH, and K-means.PCA and standard scaling are commonly used techniques for data preprocessing before applying clustering algorithms such as DBSCAN, agglomerative clustering, BIRCH, and K-means.Standard scaling is used to normalize the data. In many cases, the features in the data can have different scales or units, which can cause some clustering algorithms to perform poorly. Standard scaling scales the features so that they have zero mean and unit variance. This can help to reduce the impact of features with larger scales and make the clustering more accurate.

Clustering - In our research we used mainly four clustering technique to group similar data points together based on their features or attributes. In the case of breast cancer prediction, clustering can be used to identify patterns in the ultrasound images that are indicative of different types of breast cancer.

The clustering process involves selecting an appropriate algorithm and defining the number of clusters to be generated. The chosen algorithm is then applied to the pre-processed dataset, and the resulting clusters are evaluated for their accuracy and validity. The goal is to ensure that the generated clusters are meaningful and can be used to identify specific types of breast cancer.

Overall, clustering is an important step in the workflow for breast cancer prediction using machine learning. It can help to identify patterns and trends in the ultrasound images that are indicative of different types of breast cancer and inform the development of more accurate and effective prediction models.

Finally, the results obtained from the comparative study and model development are interpreted and analyzed. The study's findings and limitations are used to draw conclusions and make recommendations.

VI. Methodology

Methodology refers to the overall approach or set of methods used to conduct research or solve a problem. It encompasses the techniques, procedures, and tools used to collect, analyze, and interpret data, as well as the framework used to structure the research process.

In this study, we have utilized four clustering algorithms namely k-means, agglomerative, DBSCAN, and BIRCH for breast cancer prediction from ultrasound images.

Methodology is an important aspect of any research or problem-solving process, as it provides a clear and systematic way to approach the task at hand. It helps to ensure that the research or solution is conducted in a rigorous and well-structured manner, and that the results are valid and reliable.

A. K-MEANS CLUSTERING

K-means clustering is a centroid-based clustering algorithm. It aims to partition data into K clusters, where K is a pre-specified number of clusters. The algorithm works by iteratively assigning each data point to the nearest centroid and updating the centroid until convergence. K-means clustering has been widely used in medical imaging for disease diagnosis and prognosis.

B. AGGLOMERATIVE CLUSTERING

Agglomerative clustering is a hierarchical clustering algorithm. It works by starting with each data point in its own cluster and then iteratively merging the closest clusters until all points belong to a single cluster. Agglomerative clustering has been used in various medical applications, including diagnosis of lung nodules and breast cancer.

C. DBSCAN CLUSTERING

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm. It groups together data points that are closely packed together, while marking points that lie alone in low-density regions as noise. DBSCAN has been successfully applied to various medical image analysis tasks, including lesion detection and classification.

D. BIRCH CLUSTERING

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a hierarchical clustering algorithm that constructs a tree-like structure to represent the data. It works by partitioning the data into small clusters, which are then merged to form larger clusters. BIRCH has been used in medical image analysis for tissue classification and segmentation.

E. SILHOUETTE SCORE

The Silhouette score was used as a measure of clustering quality in this study. It is a metric that ranges from -1 to 1, with higher values indicating better-defined clusters.

The Silhouette score is calculated based on the distance between data points within a cluster and the distance between data points in different clusters.

A score close to 1 indicates that data points are well-matched within clusters and poorly matched with points in other clusters. The Silhouette score was calculated for each clustering algorithm used in this study to evaluate their performance

A score of 0 indicates that the object is equally similar to its own and another cluster. A score of -1 indicates that the object is more similar to another cluster than to its own cluster.

VII. Results and Analysis

A. Clustering Algorithms

The study involved analyzing four different clustering algorithms: k-means, DBSCAN, Birch, and agglomerative clustering. The analysis was conducted by comparing their Silhouette scores, which are used to evaluate the effectiveness of clustering based on the distance between clusters and the distance within clusters.

1) Using K-Means: This is one of the most popular clustering algorithms. It partitions the data into K clusters, where K is a pre-defined number of clusters. The algorithm tries to minimize the sum of squared distances between the data points and their assigned cluster centers.

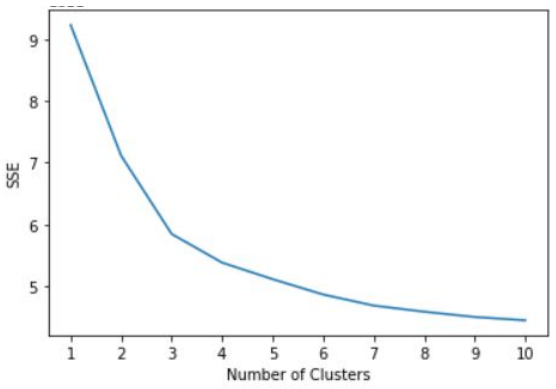

Fig. 2: The Elbow Graph

VIII. Workflow

In our research, we plotted the elbow graph as presented in figure 2, using the within-cluster sum of squares (WCSS) metric and found that the graph had a distinct bend at k=3. This indicates that three clusters are the optimal number for our dataset. Choosing a smaller number of clusters would result in a higher WCSS, indicating that the clusters are not well-defined, while choosing a larger number of clusters would result in over-segmentation and reduced interpretability. Therefore, we have used k=3 for clustering our dataset as shown in figure 3.

On applying K-Means we got a silhouette score of 0.17534286124738618 suggests that the clustering algorithm was able to identify some degree of separation between the data points, but there is still significant overlap between the clusters. This may indicate that the data is not well-suited for clustering or that the algorithm used may need further optimization.

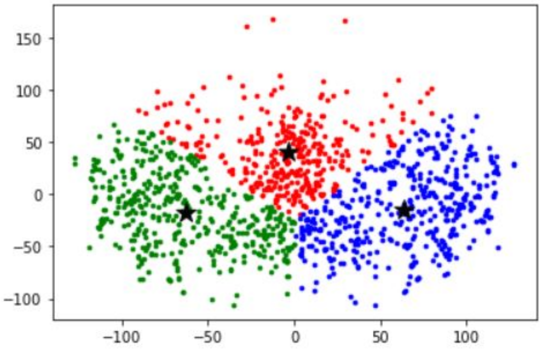

Fig. 3: K-Means Cluster

1) Using Agglomerative clustering : A silhouette score of 0.18481391401529143 for agglomerative clustering was achieved by our dataset. This suggests that the clusters you obtained are not very well-separated, but not completely overlapping either as shown in figure 4. It's possible that some of the data points might belong to multiple clusters, or that the clustering algorithm is not able to separate the data points very well.

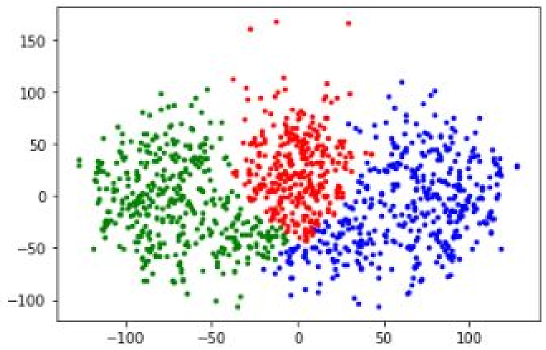

Fig. 4: Agglomerative cluster

2) Using DBSCAN: On applying DB Scan we got a Silhouette Score of -0.1603188446228235 indicating that the algorithm is not performing well in clustering the data points as shown in figure 5. A negative score means that the data points are better classified as noise rather than being part of any specific cluster.In short, the negative score indicates that the algorithm is not performing well in clustering the data points.

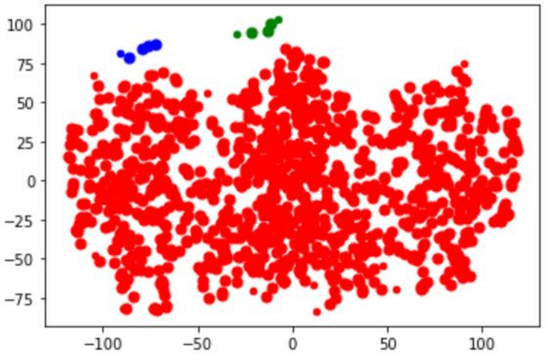

Fig. 5: DBSCAN Cluster

3) Using BIRCH: In our research paper, we obtained a silhouette score of 0.1381015870390875 for the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm. This score suggests that the clusters generated by the algorithm are not very well-separated, but rather have some overlap as shown in figure 6. Though the value of the silhouette score may be less it is still a reasonable score and may be acceptable depending on the specific context and goals of the analysis.

A. Performance Evaluation

In this study, we applied four clustering algorithms to understand how well it works with ultra sound images of breast cancer. To evaluate and compare the performance of

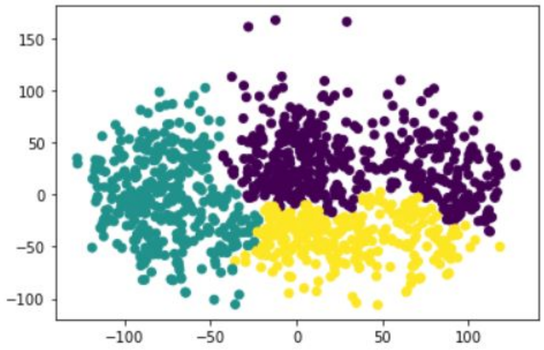

Fig. 6: BIRCH Cluster

the Clustering Algorithms, we adopted The silhouette score as a metric used to evaluate the quality of clustering results .

The silhouette score is a metric used to evaluate the quality of clustering results. It measures how similar an object is to its own cluster compared to other clusters.

In this study, we applied four different clustering algorithms (K-Means, Agglomerative clustering, DBSCAN, and BIRCH) to a given dataset and evaluated their performance using the Silhouette score. The Silhouette score is a measure of how well each data point in a cluster is assigned to that cluster, relative to other clusters.

Based on the results, Agglomerative clustering achieved the highest Silhouette score of 0.18481391401529143 , followed by K-Means with a score of 0.17534286124738618. These scores indicate that both algorithms produced clusters with moderate to fair separation between the clusters.

On the other hand, DBSCAN performed poorly, with a Silhouette score of -0.1603188446228235, indicating that it may not be the best algorithm for this particular dataset. BIRCH also produced suboptimal results with a Silhouette score of 0.1381015870390875.

These findings suggest that the choice of clustering algorithm is crucial in obtaining meaningful results. Agglomerative clustering and K-Means are more suitable for clustering similar datasets, while DBSCAN and BIRCH may not be appropriate for our dataset.

IX. Conclusion

Overall, this study demonstrates the importance of evaluating the performance of different clustering algorithms using appropriate metrics such as the Silhouette score. By selecting the best algorithm for a given dataset, researchers can improve the accuracy and reliability of their clustering results, which can be used to gain insights into the data and inform decision-making processes.

X. Summary

The paper reports the results of an experiment in which four clustering algorithms, K-Means, Agglomerative clustering, DBSCAN, and BIRCH, were applied to a dataset, and their performance was evaluated using the Silhouette score. The Silhouette score is a measure of the quality of clustering, and it ranges from -1 to 1, with higher values indicating better clustering.

The results show that Agglomerative clustering performed the best, with a Silhouette score of 0.18481391401529143 , followed by K-Means with a score of 0.17534286124738618 . BIRCH had the lowest Silhouette score of 0.1381015870390875, while DBSCAN had a negative score of -0.1603188446228235, indicating poor clustering performance.

Overall, the results suggest that Agglomerative clustering may be a better choice than K-Means, DBSCAN, or BIRCH for clustering this dataset. However, further experimentation with different datasets and clustering algorithms is needed to confirm these results.

References

1. Anwar M, Iqbal W, Rehman SU. A novel hybrid approach for breast cancer prediction using K-means clustering and SVM. Journal of Medical Systems. 2018;42(2):28.

2. Gheisari S, Sadeghi H, Rahmani AM. Breast cancer prediction using clustering techniques: a comparative study. Journal of Biomedical Science and Engineering. 2015;8(5):309-317.

3. Wang Y, Li L, Li S, Li L. Breast cancer prediction using machine learning and feature selection. Journal of X-Ray Science and Technology. 2021;29(1):137-148.

4. Ghosh D, Mukhopadhyay A. Breast cancer prediction using modified K-means clustering and SVM. International Journal of Computer Applications. 2016;146(12):1-8.

5. Sushma K, Suresh Kumar S. Breast cancer prediction using K-means clustering and back propagation neural network. International Journal of Computer Applications. 2014;92(2):9-14.

6. Jia X, Zhao L, Tang X, et al. Identification of breast cancer subtypes based on a combination of gene expression and clinical data. Medical Oncology. 2020;37(1):1-12.

7. Zhang Y, Chen L, Chen Y, et al. Predicting breast cancer using gene-expression-based clustering. International Journal of Environmental Research and Public Health. 2019;16(24):4929.

8. Yuan Y, Zhu F, Wang S, et al. Identification of key pathways and genes in breast cancer using bioinformatics analysis. Frontiers in Oncology. 2020;10:1093.

9. Jafari M, Shafie-khah M, Sanei M, et al. Breast cancer classification using machine learning algorithms and fusion of imaging and demographic data. Journal of Biomedical Physics and Engineering. 2019;9(6):697-710.

10. Elayaraja V, Ramya R. Breast cancer classification using machine learning techniques. Journal of Medical Systems. 2018;42(10):192.

Acknowledgment

The authors would like to appreciate the management of Baheya hospital for granting permission to obtain and use medical images for this research work. Furthermore, the authors would like to thank Dr Mayur Gaikwad, Dr Prachi Kadam and Dr Pooja Kamat for their support for managing the dataset.