Breast Cancer Prediction Using Ultrasound Images

Kaustubh Raykar*, Najeeb Fariduddin Saiyed†, Nachiket Ropia‡
*Symbiosis Institute of Technology, Pune kaustubh.raykar.btech2021@sitpune.edu.in
†Symbiosis Institute of Technology, Pune, India najeeb.saiyed.btech2021@sitpune.edu.in
‡Symbiosis Institute of Technology, Pune, India nachiket.ropia.btech2021@sitpune.edu.in

***Abstract*—Breast cancer is a prevalent disease that affects mostly women, and early diagnosis will expedite the treatment of this ailment. Recently, machine learning (ML) techniques have been employed in biomedical and informatics to help fight breast cancer. Extracting information from data to support the clinical diagnosis of breast cancer is a tedious and time-consuming task. The use of machine learning and feature extraction techniques has significantly changed the whole process of a breast cancer diagnosis. This research work proposed a machine learning model for the classification of breast cancer. To achieve this, a support vector machine (SVM) was employed for the classification, and Principal copmponent analysis (PCA) and linear discriminant analysis (LDA) was employed for feature extraction. We measured our model's feature extraction performance in principal component analysis (PCA) and Linear Discriminant Analysis (LDA) for classification. A comparative analysis of the proposed model was performed to show the effectiveness of the feature extraction, and we computed missing values based on the classifier's precision, recall and f1 score.**

Keywords: Breast cancer diagnosis; Principal component analysis; Linear discriminant analysis; Support vector machine; classification

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. Literature Survey

TABLE I

Paper Title	Authors	Year	Methodology	Results
Early diagnose breast cancer with PCA-LDA based FER and neuro-fuzzy classification system	R. Preetha S. Vinila Jinny	2020	PCA LDA	ANNFIS classifier-98.6%
Machine Learning Classification Techniques for Breast Cancer Diagnosis	David A. Omondiagbe Shanmugam Veeramani Amandeep S. Sidhu	2019	LDA SVM	SVM-98.82%
The role of Linear Discriminant Analysis for Accurate Breast Cancer prediction	Onyinyechi Jessica Egwom Oko Michael Ogar	2021	PCA LDA SVM	PCA–SVM-98.5% LDA–SVM-99.2%
Label-free diagnosis of breast cancer based on serum protein purification assisted surface-enhanced Raman spectroscopy	Yamin Lin Jiamin Gao	2021	PCA LDA	PLS-SVM - 94.67%
An Example of Performance Comparison of Supervised Machine Learning Algorithms Before and After PCA and LDA Application: Breast Cancer Detection	Seda Kaya Mete Yağanoğlu	2020	PCA LDA KNN Random forest Decision tree Naive Bayes algorithm, SVM Logistic regression	Logistic Regression-96.49%
Evaluation of features and classifiers for classification of early-stage breast cancer	RC Conceicao M O'halloran M Glavi E Jones	2020	PCA LDA SVM MVA	SVM-89%
Machine learning as an indicator for breast cancer prediction	Tahsin Mohammed Shadman Fahim Shahriar Akash Mayaz Ahmed	2021	Decision Tree K-Neighbors LDA Logistic Regression Naïve Bayes SVM	LDA-SVM-97.77%, PCA-SVM- 94.78%
Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis	Sara Ibrahim Saima Nazir Sergio A Velastin	2021	Decision Tree LDA QDA KNN, Naïve Bayes Classifier PNN SVM AdaBoost Fuzzy Sugeno (FSC)	k-NN classifier-98.69%
Dimensionality reduction-based breast cancer classification using machine learning	Kuhu Gupta Rekh Ram Janghel	2019	PCA LDA Fuzzy logic Neural networks	PCA -96.58 Backpropagation Neural Network

IV. 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.

V. 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.

VI. 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.

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 is an essential step in data analysis, which involves preparing and transforming the data to improve its accuracy and quality for further analysis. Exploratory data analysis (EDA) is done to gain insights into the image data, such as the distribution of the features, correlation between features, and identifying any outliers.

To reduce dimensionality, a comparative study between Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) is performed. This involves applying both techniques to the selected features and comparing their performance in terms of accuracy, precision, recall, and F1-score.

After selecting the features, predictive models are developed using Support Vector Machine (SVM) with the reduced dimensions from PCA and LDA. Performance evaluation is done to evaluate the developed models using metrics such as accuracy, precision, recall, and F1-score.

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.

VII. Methodology

A. PCA

Principal component analysis is a statistical technique used to reduce the dimensionality of high-dimensional datasets by identifying the principal components that explain the most variance in the data. For our application, PCA works by extracting features from the ultrasound images (i.e., texture features, shape features, and intensity features) And these are reduced to lower-dimensional subspaces and by machine learning algorithms it is used to build predictive models for our breast cancer diagnosis.

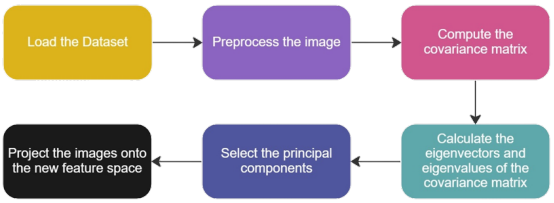

Fig. 1: Working of PCA

B. LDA

Linear discriminant analysis (LDA) is a supervised technique used to classify data into different categories. By transforming the original feature space into a lower-dimensional space. It can be used to project new data points into this lower-dimensional space for classification. It can be used to identify features that are more relevant for distinguishing between benign and malignant tumors. By using LDA to identify these features, we can build predictive models that can accurately classify new ultrasound images as either benign or malignant.

C. SVM

Support Vector Machine (SVM) is a machine learning algorithm used for classification analysis. It is used to classify images by finding the hyperplane that maximally separates the two classes in the feature space. By training SVM on labeled data , it can learn to identify patterns and features that distinguish between cancerous from non-cancerous tumors. It can handle large datasets with high dimensionality and is robust to noise and outliers. It is also shown to outperform other machine learning algorithms in terms of accuracy and efficiency making it a good choice in the field of breast cancer diagnosis.

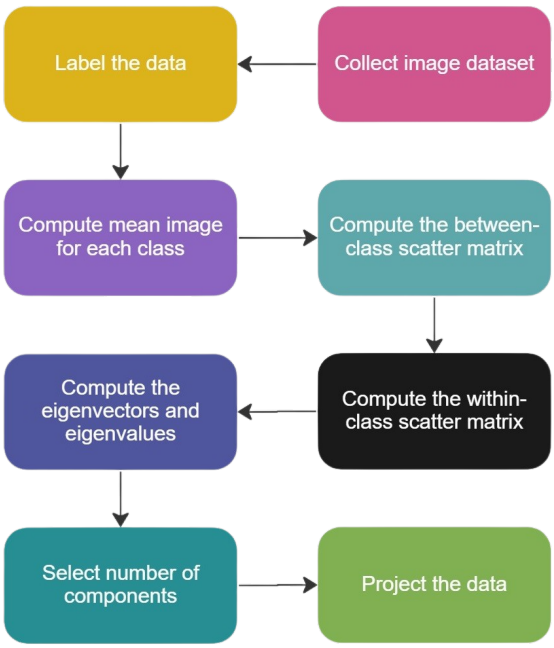

Fig. 2: Working of LDA

VIII. Results and Analysis

A. Comparing PCA and LDA

1) Using Graphs: PCA

After applying PCA to the ultrasound images, a scatter plot is plotted between first principal component on x-axis and second principal component on y-axis. In the resulting plot, the purple dots represent the benign cluster, the green dots represent the malignant clusters, and the yellow dots represent the clusters for normal.

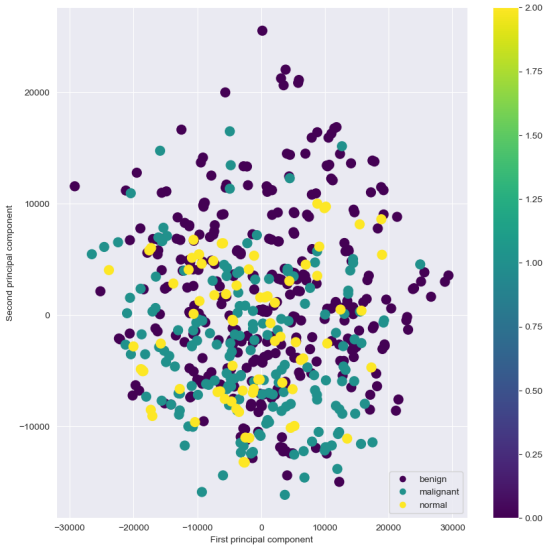

Fig. 3: PCA Scatter plot

The conclusions that can be observed from the figure ** are - the circular clustering suggests that the variance in the data is

evenly distributed across the principal components and there is no significant patterns. The reason for evenly distributed may be because the the images in the dataset contains similar features, making it difficult to distinguish between them.

Overall, applying PCA to the breast cancer dataset doesn't really help us to capture the intrinsic structure of our data and make any meaningful observation on our dataset.

LDA

The scatter plot in Figure 12, shows the plot between the first discriminant component (on x-axis) and the second discriminant component (on y-axis). The purple dots represent the normal class, the green dots represent the malignant class, and the yellow dots represent the other class.

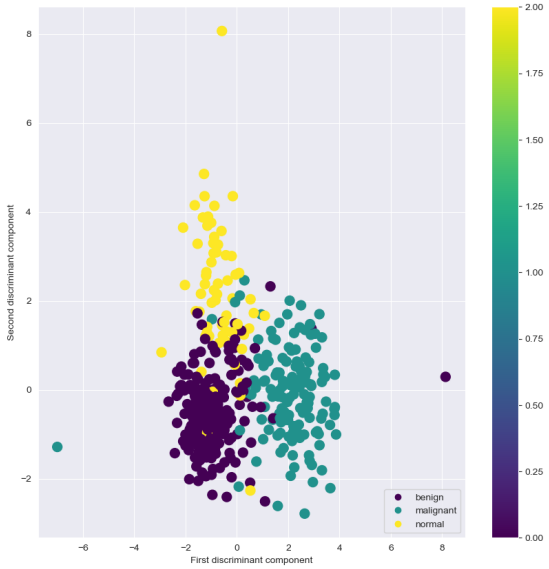

Fig. 4: LDA Scatter plot

From the plot we can observe a clear separation between each class, these three clusters serve as good boundaries between the classes. Outliers are present and it has handled it pretty well which can be seen in cases for the benign, malignant, and normal clusters.

Both PCA and LDA can be useful. PCA can help to extract relevant features from the ultrasound images and reduce the number of dimensions in the data, making it easier to analyze and visualize. However, PCA may not be the best choice for classification tasks, as it does not consider the class labels. On the other hand, LDA can help to extract discriminant features that can improve the accuracy of breast cancer prediction from ultrasound images. By maximizing the separation between the classes, LDA can help to identify the features that are most relevant for distinguishing between benign and malignant tumors.

In summary, PCA and LDA are both useful dimensionality reduction techniques, but they serve different purposes. Here LDA seems to be doing better work at classifying the data into its three classes for analysis from the scatter plots.

2) By performance: To evaluate and compare the performance of the SVM classifier on the outputs of LDA and PCA, we adopted three performance metrics: precision, recall, and F1 score.

Precision: the fraction of relevant instances among the retrieved instances. A high precision means that the classifier has a low false positive rate, i.e., it does not classify non-relevant instances as relevant.

Recall: the fraction of relevant instances that are retrieved. A high recall means that the classifier has a low false negative rate, i.e., it does not miss relevant instances.

F1 score: a weighted average of precision and recall, which gives more weight to the lower value. It is a measure of the classifier's overall accuracy.

We observed that the SVM classifier gave the following results. On training without PCA and LDA:
Precision: 68.10%, **Recall:** 68.58%, **F1 score:** 68.04%

We applied PCA and LDA to the dataset to reduce dimensionality. We then presented the classification results to determine which of the two-dimensionality reduction techniques is more effective for breast cancer diagnosis.

To apply PCA we needed to find the right number of components that would explain at least more than 90% of the variance in the dataset. So we plotted a graph between several components and explained the variance ratio.

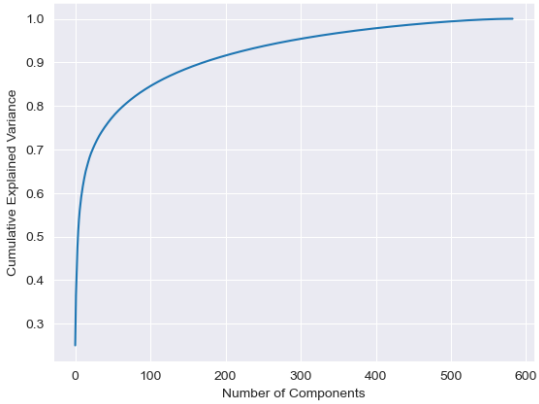

Fig. 5: Plot of Cumulative Explained Variance Ratio

from Figure, we observed that the number of components that are explaining 90% of the variance is achieved when components are 172.

Now for these 172 components, we get the evaluation metrics as follows on applying SVM;
Precision: 71.27%, **Recall:** 66.02%, **F1 score:** 59.95%

Similarly, for LDA we can only plot for n_components =2 because the number of classes in our dataset is 3 and the maximum number of components we could use in LDA would be 2 (i.e., n_components $\leq$ min(n_features, n_classes - 1) =

min(n_features, 3-1) = min(n_features,2)). On the other hand, PCA is an unsupervised learning method that does not require class labels and does not have the same restriction as LDA.

And this problem of having more classes than features is not specific to images, but it can occur in any dataset where the number of classes is larger than the number of features. However, in the case of image data, it is common to have a large number of features (i.e., pixels) which can make the dimensionality of the data very high.

So on displaying the evaluation metrics on LDA of n_components = 2 is:
Precision: 69.65%, **Recall:** 39.10%, **F1 score:** 42.05%

On comparing the results of the SVM classifier. We see that PCA dimensionality reduction shows slightly improved precision, but lower recall and F1 score. This suggests that PCA may have compressed some important information that was useful for the classifier.

The SVM classifier with LDA dimensionality reduction has a higher precision than the baseline, but a much lower recall and F1 score. This suggests that LDA may have to overfit the training data, resulting in poor generalization to new data.

In general, dimensionality reduction techniques like PCA and LDA can be useful for reducing the computational complexity of a classifier and improving its performance, especially when dealing with high-dimensional data like images. However, we can see that ultrasound images it doesn't classify the images well and during the dimensionality reduction process, it has led to a loss of information and a decrease in performance as the transformed features do not capture the relevant information for the classification task.

B. Comparing with PCA and without PCA

To compare if PCA affects our dataset, we applied PCA and displayed image compression (which aims to preserve important features and reduces image quality) and also displayed SNR (Signal-to-Noise Ratio) which is a measure of the quality of an image after compression. A high SNR indicates that the compressed image closely resembles the original image and has good quality, while a low SNR indicates that the compressed image is significantly different from the original and may have poor quality.

On applying PCA we get Image compression greater than 2 and SNR values more than 40. Image compression and SNR value for one of the images using PCA is displayed as shown in Figure 13.

Compression ratio: 2.68 SNR: 48.91 dB

For this specific ultrasound image, we get a compression ratio of 2.68 indicating that the size of the original grayscale image has been reduced by a factor of 2.68 after applying PCA compression. In other words, the compressed image requires only 1/2.68 or approximately 37.31% of the storage space required by the original image while retaining most of its important features.

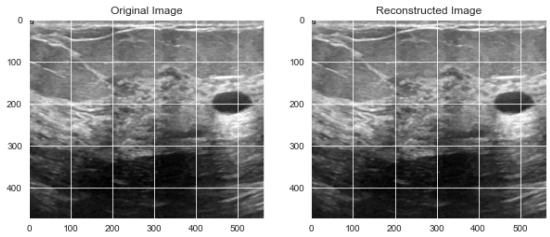

Fig. 6: Original Image vs PCA - Reconstructed Image

The SNR of 48.91 dB is a measure of the quality of the compressed data compared to the original data. A higher SNR indicates that the compressed data is closer in quality to the original data. In your case, the SNR of 48.91 dB indicates that the compressed data has a high level of fidelity to the original data, and that the compression process has not introduced a significant amount of noise or distortion.

Overall, these results suggest that the PCA algorithm has been successful in compressing the ultrasound image dataset while preserving the important information, and that the compressed data is of high quality compared to the original data.

Now on running this on SVM we see that

C. Comparing LDA and without LDA

On applying LDA to the dataset we see that image compression stays as 1.00 meaning no image compression has occurred and SNR as 0.00 dB, indicating that there was no signal-to-noise ratio improvement.

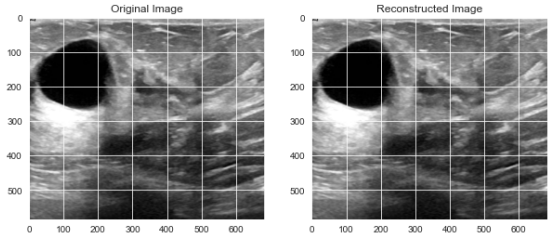

Fig. 7: Original Image vs LDA - Reconstructed Image

Compression ratio: 1.00 SNR: 0.00 dB

Therefore, LDA is a supervised machine learning algorithm used for classification and dimensionality reduction, and it is not used to calculate image compression or signal-to-noise ratio.The main objective of LDA is to find a linear combination of the predictor variables that best separates the different groups, while minimizing the within-group variance and maximizing the between-group variance. In other words, LDA seeks to find the projection of the data that maximizes the separability of the classes while preserving as much information as possible. An image compression and SNR value for one of the image using lda is displayed as shown in Figure 14 .

D. Performance Evaluation

In this study, we applied unsupervised machine learning algorithms to classify breast ultrasound images into three categories: normal, benign, and malignant. We used the dataset of breast ultrasound images from Baheya Hospital for early detection and treatment of Women's Cancer, Cairo, Egypt, which consisted of 780 images from 600 female patients.

We evaluated the performance of the algorithms using precision, recall, and F1-score metrics. The results of our applied algorithms are shown in Figure 8.

The best performing algorithm in terms of F1-score is the unsupervised machine learning algorithm without PCA-LDA, with a score of 68.04%. This algorithm also achieved the highest precision and recall scores at 68.10% and 68.58%, respectively. The unsupervised machine learning algorithm with PCA performed slightly worse with an F1-score of 59.95%. The unsupervised machine learning algorithm with LDA performed the worst, with an F1-score of 42.05%.

Overall, our results suggest that unsupervised machine learning algorithms can be effective for classifying breast ultrasound images and potentially reducing the need for manual interpretation. However, further research is needed to improve the performance of these algorithms and to evaluate their effectiveness in real-world clinical settings.

SVM	Precision	Recall	F1-Score
Without PCA-LDA	68.10%	68.58%	68.04%
With PCA	71.27%	66.02%	59.95%
With LDA	69.65%	39.10%	42.05%

Fig. 8: Performance comparison of SVM using dimensionality reduction techniques

IX. Conclusion

Comparison of the performance of LDA and PCA for dimensionality reduction of breast cancer diagnosis has been presented. Results obtained show that PCA and LDA may not be that effective for grey scaled ultrasound images.

It is important to note that breast cancer is a complex disease that requires a multidisciplinary approach for effective management. Regular self exams, clinical breast exams, and mammograms can help detect breast cancer at an early stage, when it is most treatable. The development of an individualized treatment plan, taking into account the specific characteristics of each case, is crucial for improving patient outcomes and survival.

X. Summary

Both PCA and LDA can be used as effective dimensionality reduction techniques for ultrasound images of breast cancer. LDA outperformed PCA in terms of classifying the three classes, indicating that LDA is a better technique for classification problems where the goal is to maximize the separation between classes. The use of SVM as a classification algorithm for both PCA and LDA reduced the dimensionality of the data. The results of the comparative study indicate that LDA combined with SVM is a promising approach for classifying ultrasound images of breast cancer. The findings of this study can be used to aid clinicians in the accurate diagnosis of breast cancer and to guide future research on the development of more effective classification algorithms for ultrasound images.

Overall, the comparative study between PCA and LDA on the breast cancer dataset of ultrasound images using SVM for evaluation provides valuable insights into the performance of these techniques and their potential applications in the field of medical imaging.

References

1. WHO. 2022 Cancer. Available online: https:www.who.int/news-rooms/factsheet/details/cancer (accessed on 2 May 2022).
2. Labrèche, F.; Goldberg, M.S.; Hashim, D.; Weiderpass, E. Breast cancer. In Occupational Cancers; Springer: Berlin/Heidelberg,
Germany, 2020; pp. 417–438. [CrossRef] 3. Kumar, V.; Misha, B.K.; Mazzara, M.; Thanh, D.N.; Verma, A. Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications. In Advances in Data Science and Management; Springer: Berlin/Heidelberg, Germany, 2019; pp. 435–442.
4. Meera, C.; Nalini, D. Breast cancer prediction system using data mining methods. Int. J. Pure Appl. Math. 2018, 119, 10901–10911.
5. Rathi, M.; Pareek, V. Hybrid approach to predict breast cancer using machine learning techniques. Int. J. Comput. Sci. Eng. 2016, 5, 125–136.
6. Way, G.P.; Sanchez-Vega, F.; La, K.; Armenia, J.; Chatila, W.K.; Luna, A.; Greene, C.S. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Rep. 2018, 23, 172–180.e3. [CrossRef] [PubMed]
7. Rajbharath, R.; Sankari, I.; Scholar, P. Predicting breast cancer using random forest and logistic regression. Int. J. Eng. Sci. Comput. 2017, 7, 10708–10813.
8. Luque, C.; Luna, J.M.; Luque, M.; Ventura, S. An advance review on text mining in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2019, 9, 1302. [CrossRef]
9. Hassan, M.; Hamada, M. Genetic algorithm for improving prediction accuracy of multi-criteria recommender systems. Int. J.

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.