Journal of Current Science and Technology, July - September 2025 Vol. 15 No. 3, Article 125

ISSN 2630-0656 (Online)

Cite this article: Saiyed, N., & Kantipudi, P. (2025). Advancing breast cancer detection: A comparison of PCA and LDA methods in analyzing ultrasound imagery. *Journal of Current Science and Technology*, *15*(3), Article 125. https://doi.org/10.59796/jcst.V15N3.2025.125

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Advancing Breast Cancer Detection: A Comparison of PCA and LDA Methods in Analyzing Ultrasound Imagery

Najeeb Saiyed[1], and MVV Prasad Kantipudi[2,*]

[1]Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India
[2]Department of Electronics and Telecommunication, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India

[*]Corresponding Author; E-mail: prasad.kantipudi@sitpune.edu.in

Received 24 March 2025; Revised 6 May 2025; Accepted 6 May 2025; Published online 15 June 2025

Abstract

Early and accurate detection of breast cancer via ultrasound imaging is essential, yet the high dimensionality of raw ultrasound features can hinder classifier performance and increase computational burden. Comparison between Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for feature reduction in a breast-cancer ultrasound diagnostic pipeline, alongside t-SNE for exploratory visualization. The research utilized 1,200 breast ultrasound images with 400 benign, 400 malignant, and 400 normal images obtained from Baheya Hospital (Cairo, Egypt). Minority classes were balanced using data augmentation techniques like rotation and flipping. PCA reduced the data to 172 components, preserving 90% of data variance, while LDA used two components. t-SNE generated a two-dimensional visual representation. Classifiers, including Support Vector Machine (SVM), Random Forest (RF), and XGBoost, were trained on: (a) the full feature set, (b) PCA-reduced data, and (c) LDA-reduced data. Evaluation metrics included precision, recall, and F1-score. Compression ratio and signal-to-noise ratio (SNR) measured image compression via PCA.Without reduction, XGBoost achieved the highest F1-score (76.97%), precision (77.40%), and recall (76.55%). PCA yielded a modest precision gain (XGBoost: 78.65%) but reduced recall and net F1-score (76.37%). LDA significantly degraded performance (XGBoost F1: 63.99%; RF F1: 60.13%; SVM F1: 42.05%). PCA compression reduced image size by 2.68x with an SNR of 48.91 dB, while LDA offered no compression benefit. t-SNE visualization revealed clear non-linear class clusters, underscoring the dataset's intrinsic complexity. For ultrasound-based breast cancer diagnosis, preserving full high-dimensional features and using a powerful non-linear model (e.g., XGBoost) yields optimal accuracy. PCA is best reserved for storage or runtime efficiency, LDA for scenarios with very low dimensional constraints, and t-SNE for exploratory data analysis. This comparative study highlights that dimensionality reduction may harm performance in complex imaging data and recommends context-specific use of PCA and LDA to avoid loss of critical diagnostic information.

Keywords: *breast cancer classification; principal component analysis; linear discriminant analysis; feature extraction*

1. Introduction

Breast cancer is one of the deadliest diseases worldwide, claiming millions of lives every year (NDTV World, 2024). Timely treatment relies heavily on early detection. Among the diagnostic tools, ultrasonic imaging has become a key modality, utilizing sophisticated image processing techniques to enhance diagnostic accuracy. Integrating intelligent models with imaging technologies has proven highly effective in improving the prediction and classification of breast cancer using clinical imaging data by extracting pertinent features for analysis.

The integration of medical imaging with intelligent models has gained significant attention in recent years, promising advancements in accuracy, efficiency, and reliability of disease prediction. These models support healthcare professionals by complementing traditional diagnostic approaches. Specifically, in breast cancer diagnosis, intelligent models help address the limitations of manual interpretation, which can be error-prone and time-consuming.

Ultrasound is a cost-effective and non-invasive imaging modality widely used in breast cancer diagnosis, playing a key role in detecting suspicious lesions. However, interpretation is manual, subjective, and dependent on radiologist expertise, increasing variability and the risk of missed diagnoses. Intelligent models trained on large, annotated datasets can identify complex patterns, subtle anomalies, and feature correlations that traditional analysis may overlook. This capability reduces diagnostic errors, accelerates evaluations, and improves clinical workflows. Recent advancements in intelligent imaging systems and ensemble learning methods have shown remarkable success in breast cancer detection and classification. These models extract high-dimensional features from ultrasound images to differentiate between benign and malignant tissues accurately (Maiprasert, & Kitbumrungrat, 2023; Pechprasarn et al., 2023).

Although mammography remains the clinical gold standard, ultrasound imaging is increasingly favored for its cost-effectiveness and radiation-free nature, particularly for women with dense breast tissue or in low-resource settings. However, modern ultrasound systems produce hundreds of high-resolution images per exam, resulting in extremely high-dimensional feature spaces that can challenge classifier performance. While dimensionality-reduction techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are widely adopted in clinical AI pipelines (Zhao et al., 2020), their true impact on diagnostic accuracy, especially for subtle tissue differentiation, remains underexplored. This study positions dimensionality reduction at the center of the diagnostic pipeline and empirically evaluates how PCA and LDA affect the performance of machine learning models for breast cancer classification using ultrasound imagery.

1.1 Literature Review

Research on breast cancer imaging has been increasingly focused on two related challenges: how compression can effectively reduce the large and highly correlated feature spaces captured by medical images, and how well the compressed representation can support accurate classification and it turns out PCA and LDA are dominating the dimension reduction field whereas nonlinear methods like t-distributed Stochastic Neighbor Embedding (t-SNE) are better used on exploratory visualization aids, yet feature trimming can erase subtle features like textural cues that help to distinguish between benign, malignant and normal breast tissue.

The research by Li et al., (2023) demonstrated that linear Support Vector Machine (SVM) achieved superior performance than LDA and multiple mathematical models for classifying 926 breast nodules through analysis of all Breast Imaging Reporting and Data System (BI-RADS) descriptors. At the same time, PCA served as an exploratory tool to visualize class overlap (Li et al., 2023). The combination of 107 radiomic features and peripheral-immune markers by Zhao et al., (2023) yielded a logistic model with an area under the curve (AUC) of 0.91, which declined when low-variance filters were applied (Zhao et al., 2023).

Multiple multicenter cohorts replicated this phenomenon throughout 2024. Support-vector machines achieved the best AUC score of 0.92 when used with all features for pre-operative axillary-burden prediction whereas an LDA baseline experienced almost 9 percentage points of reduced performance (Yao et al., 2024a). The same research team used original intratumoral and peritumoral signatures as inputs in a separate neoadjuvant chemotherapy study where gradient-boosted trees achieved a macro-AUC higher than 0.80 across three molecular subtypes (Yao et al., 2024b). The research of Wu et al., (2024) showed that Extreme Gradient Boosting (XGBoost) with Boruta-selected features working on five unaltered features, produced better results than nine alternative pipelines which contained either PCA or LDA.

Region-based studies continually demonstrate that the use of dimensionality reduction depends on data type: in the case of mammography, Wongnil et al., (2024) retained full resolution pixels and interfaced an AlexNet best region-CNN with an SVM, exceeding a YOLOv4 single-pass detector on 3,265 images. Conversely, in a tabular clinical situation, Panyamit et al., (2022), compressed 12 heart-failure tabular variables into three principal components to match the accuracy of the full-feature set at 96 %, thus demonstrating that PCA that is selected properly can reduce most of the variables when spatial information is irrelevant.

By contrast, LDA now serves mostly as a transparent baseline. Recent HER-2 studies strongly supported the point: For instance, both Xie et al., (2025) and Luo et al., (2025) each utilized thousands of unreduced radiomic and deep features, resulting in external test AUCs above 0.86 with augmented forests or transformer backbone feature extraction, while omitting LDA from the final comparison. Even for datatypes comparable to histopathology, where LDA might be more suitable (e.g., tumor-infiltrating lymphocytes), Liu et al., (2025) simply kept the five most important variables But, again, the model had first searched all 1,712 features and descriptors before using LDA.

PCA serves as an explicit data reduction technique in these applications, not for predictive analysis purposes (Fernandes et al., 2024). A 12-kilobyte nomogram developed by Zhang et al., (2024) used six rotated principal components derived from ten sonographic descriptors to distinguish mucinous carcinoma from fibro-adenoma with 92 % accuracy, although this did not exceed baseline performance. The combination of DenseNet embedding compression to 256 PCs followed by LASSO reduced GPU memory requirements by 60% without compromising the Ki-67 prediction AUC at 0.84 (Cen et al.,2025). The research confirms that using PCA as a compression method provides operational benefits for storage and runtime efficiency but does not enhance predictive discrimination power.

Across diagnostic imaging domains, dimensionality-reduction techniques are applied with distinct purposes: Principal Component Analysis (PCA) is typically used for data compression and runtime efficiency, t-Distributed Stochastic Neighbor Embedding (t-SNE) serves primarily as an exploratory visualization tool to highlight latent data structure, and Linear Discriminant Analysis (LDA) is generally retained as a transparent, interpretable baseline, but seldom used for complex, high-dimensional, nonlinear datasets like ultrasound. In most cases, researchers either adopt these methods based on convention or apply them separately in different pipelines. Despite their prevalence, no prior study has performed a direct, head-to-head empirical comparison of PCA and LDA under identical experimental conditions using the same dataset and classifier architectures within a clinically relevant context like breast ultrasound imaging. This absence leaves a methodological gap in understanding how classical reduction techniques perform relative to each other, and whether they are truly suitable in preserving diagnostic signals embedded in medical imaging data. Therefore, this study seeks to systematically evaluate PCA, LDA, and t-SNE within a unified experimental framework, employing multiple classifiers and consistent datasets, to inform evidence-based decisions on dimensionality-reduction strategies for breast cancer detection using ultrasound images.

2. Objectives

This study aims to evaluate the impact of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) on the diagnostic performance of machine learning models for breast cancer classification using ultrasound imagery, and to compare their effectiveness across multiple classifiers with support from t-SNE visualization.

3 Methods

3.1 Principal Component Analysis (PCA)

PCA is a technique used to reduce the dimensionality of high-dimensional datasets by identifying and prioritizing the principal components that account for the greatest variance within the data. In this application, PCA is used to extract critical features from ultra-sound images, including texture, shape, and intensity characteristics. These extracted features are subsequently projected onto lower-dimensional subspaces, which serve to optimize data representation. Leveraging these reduced-dimension features, machine learning algorithms are then implemented to develop predictive models, facilitating the accurate diagnosis of breast cancer as shown in Figure 1.

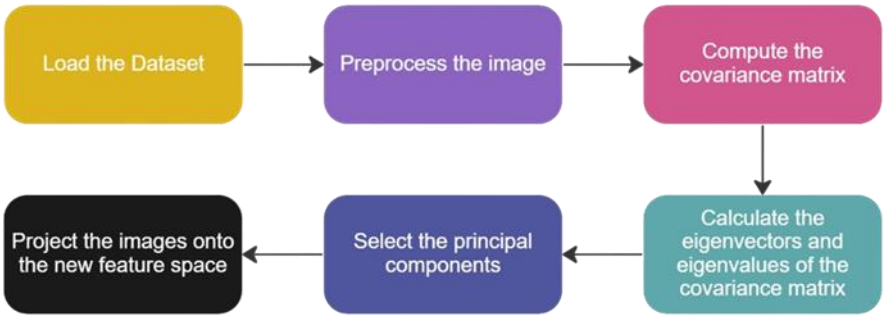

Figure 1 Overview of PCA feature extraction pipeline applied to breast ultrasound images

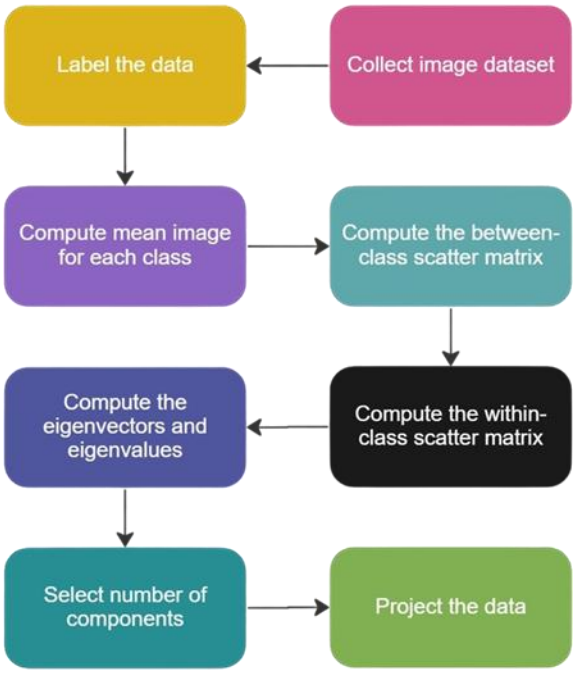

Figure 2 Workflow of LDA-based supervised dimensionality reduction for tumor classification

3.2 Linear Discriminant Analysis (LDA)

LDA is a supervised dimensionality reduction technique commonly used to classify data into distinct categories. It works by transforming the original feature space into a lower-dimensional space while preserving class-discriminatory information. Projecting new data points into this reduced space allows LDA facilitates efficient classification tasks. In our research, LDA is applied to identify features that are most relevant for distinguishing between benign and malignant tumors. By leveraging these discriminative features, predictive models can be constructed to accurately classify ultrasound images, enabling the differentiation of benign from malignant cases with enhanced precision as shown in Figure 2.

3.3 t-distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE transforms high-dimensional feature vectors into 2D or 3D maps, placing points nearer to one another if they are more correlated. The algorithm converts distance data between points into probability distributions before it places points through repeated iterations to match low-dimensional distributions with high-dimensional ones. The t-SNE algorithm provides outstanding results for exploratory visualizations because it highlights local structures, showing benign and malignant scans alongside normal scans as distinct clusters on the 2D or 3D maps. We include t-SNE as a diagnostic tool instead of a classification tool to validate the separability of features within the data.

3.4 Support Vector Machine (SVM)

SVM is a classification algorithm used in image analysis. The concept of SVM involves finding the optimal hyperplane that best separates two classes in the feature space. SVM trains the algorithm to recognize patterns and features effective in distinguishing cancerous tumors from non-cancerous ones, by learning on labeled datasets. Importantly, SVM also can deal with large high dimensional datasets while remaining robust to noise and outliers. In addition, SVM has long been found to be superior to the other machine learning algorithms in terms of classification accuracy and computational efficiency and thus is an obvious choice for breast cancer diagnosis and other similar applications (Tambe et al., 2025).

3.5 Random Forest

Random Forest is an ensemble method that learns many decision trees which are trained on bootstrap sample of data. Individual trees usually overfit but averaging votes from each tree dramatically reduces variance, making the model both robust to noise and resistant to overfitting. It is used in this paper because of its ability to capture complex, non-linear patterns between pixel-level textures, and to handle high-dimensional feature sets.

3.6 Extreme Gradient Boosting (XGBoost)

XGBoost, which stands for Extreme Gradient Boosted decision trees, is a technique that builds many smaller trees sequentially where each one ensures it learns from the mistakes of the previous trees. This boosting strategy captures subtle, non-linear interactions within ultrasound images. We chose XGBoost because it performs well with high-dimensional data.

3.7 About the Dataset

In this paper, we work with a well-rounded dataset collected from breast ultrasound images, which is considered one of the major resources for advanced research in breast cancer diagnosis. The dataset was gathered using state-of-the-art imaging technologies from Baheya Hospital, a specialized breast cancer hospital in Cairo, Egypt (Al-Dhabyani et al., 2020). As a result, it comprises high-quality data with clinically significant material, enabling the development and analysis of various machine learning models for detecting and categorizing cancer cases more accurately using ultrasound images. The dataset includes three classes: benign, malignant, and normal. Each class had 400 images while training to ensure class balance, and minority class where data augmented using rotation and flipping techniques.

Since early detection is critical in breast cancer, this dataset given here serves as a very good base for the development of an automated diagnostic system. This dataset enables machine learning models to extract subtle patterns and features from ultrasound images with minimal constraints from human absence or variability due to human error and inter-observer differences. The diverse and clinically representative samples in the dataset make it an ideal source of comprehensive training and testing of models aimed at accurate and generalizable diagnostics.

4 Results and Analysis

4.1 Comparing PCA and LDA

4.1.1 Using Graphs

PCA was applied to reduce dimensional space on the dataset then a scatter plot of the data was generated with the first principal component plotted on the x-axis and the second on the y-axis. The resulting scatter plot from Figure 3 shows different clusters, purple dots represent the benign cases, green dots represent the malignant cases and yellow dots represent the normal cases. This visualization helps to show the separability of the clusters and offers insights into the structure of the data when using PCA.

Observations from Figure 3 suggest that the circular clustering pattern implies even dispersion of variance across principal components, without any significant patterns in the data. This even distribution may result from the images having similar features, making it difficult for the algorithm to differentiate between classes. Hence, using PCA on the breast cancer dataset does not appear effective in capturing the inherent data structure or uncovering significant variables.

While the scatter plot presented in Figure 4 depicts the relationship between the first discriminant component (x-axis) and the second discriminant component (y-axis), as obtained through LDA. This visualization demonstrates the improved separability achieved through LDA, suggesting its potential for identifying class-discriminatory features in the dataset.

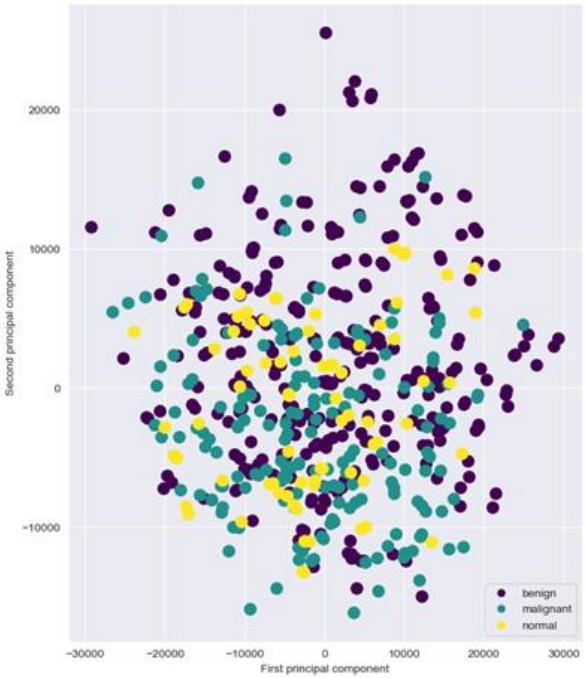

Figure 3 PCA scatter plot showing data distribution across the first two principal components for benign, malignant, and normal breast tissue

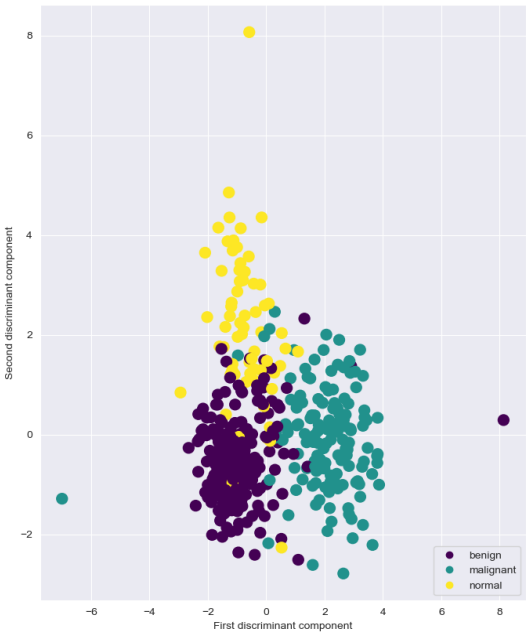

Figure 4 LDA projection revealing class separation across two discriminant components

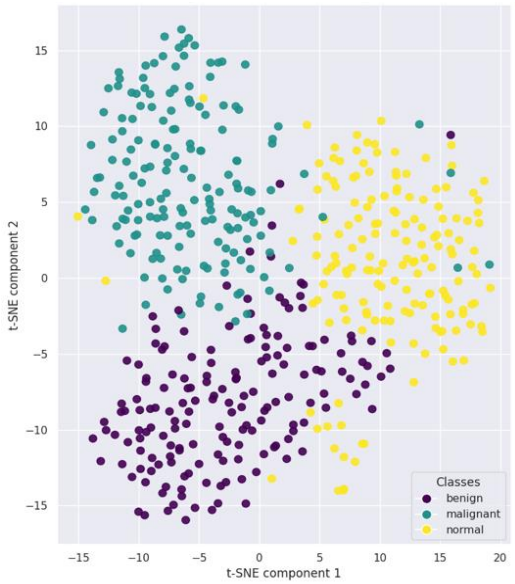

Figure 5 t-SNE 2D visualization of ultrasound feature embeddings, illustrating clear cluster separability among benign, malignant, and normal classes

Figure 5 illustrates a two-dimensional t-SNE projection of the three classes. Cluster of benign concentrate in the lower-left quadrant, malignant class on the upper-left, and normal class forms a well-separated cloud to the right. Only a thin band of mixed colors appear along the cluster borders, indicating limited class overlap and few malignant classes and one benign class in normal class cluster. This clear spatial segregation confirms that the extracted features capture meaningful anatomical differences and validates t-SNE's value as an exploratory sanity check before formal modelling.

Overall, Figures 3, 4, and 5 show that PCA's top-variance axes compress the ultrasound data efficiently but leave benign, malignant and normal scans heavily intermixed, so variance alone is not diagnostic. LDA, on the other hand, realigns the space and starts to peel the groups apart but still benign and malignant overlap shows that a purely linear split is insufficient. t-SNE finally exposes the full, curved structure of the data, producing three well-defined clusters with minimal overlap and indicating that the remaining separability is non-linear. Hence, t-SNE enables better visualization, especially for radiologists who can map and identify out-of-distribution cases for further review. Even data curators can quickly detect mislabeled images and remove noise before training.

4.1.2 By Performance

To compare and evaluate the performance of the SVM classifier on the PCA and LDA outputs, we define three key performance metrics: precision, recall, and F1 score. Precision is defined as the fraction of relevant instances among the retrieved instances. A high precision score indicates a low false positive rate, meaning the classifier rarely labels irrelevant instances as relevant. Recall is the fraction of relevant instances that are successfully retrieved. A high recall shows that the classifier has a small false negative rate (few missed instances). F1 score is a weighted harmonic mean of precision, giving more weight to the lower of the two values.It is a good overall measure for assessing the classifier to which it applies.

We first trained the SVM classifier without using PCA or LDA, and the results are shown in Table 1. We then applied PCA and LDA to the dataset to reduce its dimensionality and evaluate their effect on classification performance. For PCA analysis, the optimal number of components was determined to explain at least 90% of the dataset's variance. A graph was plotted showing how the number of components relates to the explained variance ratio (solid line). The smallest number of components was then selected to retain most of the dataset's information while reducing dimensionality. The classifier was trained on the PCA-reduced dataset using SVM and compared with the LDA-based results.

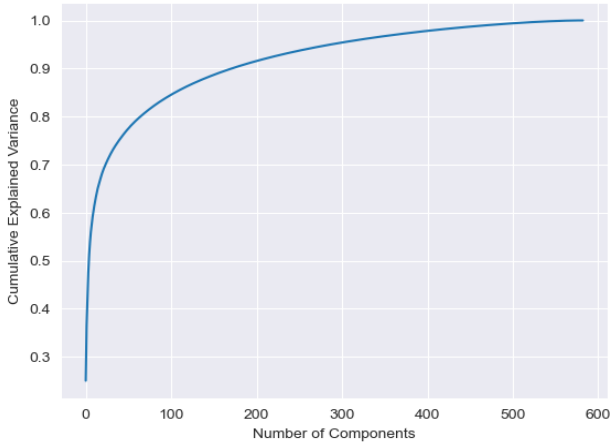

Figure 6 Cumulative explained variance curve used to select the number of PCA components retaining 90% of data variance

Table 1 Performance Comparison of SVM Classifier with and Without Dimensionality Reduction

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%

Figure 6 shows that 172 components are needed to explain 90% of the variance in the dataset. The evaluation metrics corresponding to these 172 components are presented in Table 1. Similarly, when using LDA, only two components can be plotted, as the dataset contains three classes. The maximum number of LDA components is limited to two, based on the formula $n \leq \min(n\_features, n\_classes - 1) = \min(n\_features, 2)$. In contrast, PCA is an unsupervised method that does not require class labels and is not subject to this restriction.

The issue of having more classes than features is not unique to image datasets; it can occur in any dataset where the number of classes exceeds the number of features. However, in the case of image data, it is common to encounter very high-dimensional inputs, such as pixel-level features, which increases the relevance of dimensionality reduction techniques. Table 1 presents the evaluation metrics for LDA using two components.

Comparing the SVM classifier results using PCA and LDA for dimensionality reduction reveals several notable performance trends. With PCA, dimensionality reduction led to a slight improvement in precision but a noticeable decline in recall and F1 score. This suggests that although PCA effectively removed redundant features and improved precision by reducing false positives, it may have also discarded crucial information needed to distinguish relevant instances. As a result, the classifier struggled to identify all true positives, lowering recall and the overall F1 score.

The SVM classifier with LDA-based dimensionality reduction achieved higher precision than the baseline model. However, both recall and F1 score were significantly lower. This indicates that while LDA successfully extracted discriminative features that reduced false positives, it may have overfit the training data. Overfitting typically leads to poor generalization, as the model becomes too closely tailored to the training data, reducing predictive performance on unseen instances.

From Table 1, SVM without PCA or LDA demonstrates the highest F1 score of 68.04%. As the best-performing algorithm in terms of F1 score, it also achieved its peak precision of 68.10% and the highest recall of 68.58%. SVM using PCA achieved an F1 score of 59.95%, yielding slightly inferior results compared to the baseline. While it improved precision marginally, the trade-off in recall reduced its overall performance. SVM with LDA achieved an F1 score of 42.05%, establishing itself as the worst-performing algorithm among those studied. This significant drop in performance suggests that LDA may have discarded essential information or overfit the data, hindering its classification capability.

Table 2 Random Forest Classifier Metrics Across Full, PCA-Reduced, and LDA-Reduced Feature Sets

Random Forest	Precision	Recall	F1-Score
Without PCA-LDA	73.20%	**74.05%**	**73.62%**
With PCA	**75.10%**	71.88%	73.45%
With LDA	70.56%	52.49%	60.13%

Table 3 XGBoost Classifier Performance on Full Feature Set Versus PCA and LDA Dimensionality Reduction

XGBoost	Precision	Recall	F1-Score
Without PCA-LDA	77.40%	**76.55%**	**76.97%**
With PCA	**78.65%**	74.20%	76.37%
With LDA	71.12%	58.46%	63.99%

From Table 2, we see that Random Forest performs with the best balance of precision, recall, and F1 score when all original features are used. This suggests that the full feature set provides Random Forest with the necessary information to classify instances effectively. Using PCA increases precision by approximately 2%, but there is a slight drop in recall, and the F1 score remains nearly unchanged. This indicates that while PCA may filter out some noise, it also removes subtle features that contribute to recall performance. In contrast, applying LDA further degrades performance. Recall drops to 52%, and the F1 score falls to 60%, indicating that LDA removes textural details essential for Random Forest's effectiveness. This suggests that LDA oversimplifies the feature space, leading to a loss of key information needed for accurate classification.

The XGBoost model using the unreduced feature set delivered the highest F1 score of 76.97% in Table 3, indicating that gradient-boosted trees perform best when the full dataset is retained, even with its inherent nonlinearity. While PCA slightly improves precision, it reduces recall and maintains an equivalent F1 score. This suggests that PCA is more beneficial for accelerating training rather than enhancing accuracy. Its dimensionality reduction does not preserve all the information critical for optimal model performance.

In contrast, LDA severely degrades performance. The high-order texture interactions that XGBoost relies on are largely eliminated in LDA's bottleneck processing. This results in an almost 18-percentage-point drop in recall and a corresponding decline in F1 score to the mid-60s. Deploying boosted-tree ensembles for breast cancer ultrasound classification requires a broad feature space. Minimizing the use of aggressive linear reducers like LDA is essential to preserve the complexity that these models are designed to exploit.

4.2 Comparing with PCA and without PCA

To evaluate the effectiveness of PCA on our dataset, we applied PCA and assessed its impact on image compression. This process aims to preserve important features while reducing image quality. We then measured the Signal-to-Noise Ratio (SNR), which indicates the quality of the image after compression. A high SNR signifies that the compressed image closely resembles the original and maintains good quality. In contrast, a low SNR suggests that the compressed image differs significantly from the original and may exhibit poor quality.

From Figure 7, the ultrasound image achieves 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 approximately 37.31% of the original storage space while retaining most of its important features. Additionally, the image yields a Signal-to-Noise Ratio (SNR) of 48.91 dB. A higher SNR indicates that the compressed data is more similar in quality to the original. In this case, an SNR of 48.91 dB reflects a high level of fidelity, suggesting that the compression process preserved image quality effectively and introduced minimal noise or distortion.

4.3 Comparing LDA and without LDA

LDA is not an image compression method. When applied, we observe that the compression ratio remains at 1, meaning no compression has occurred. Similarly, the SNR is 0 dB, indicating no improvement in signal-to-noise ratio. These results can be seen in Figure 8. LDA is a supervised machine learning algorithm used for classification and dimensionality reduction. It aims to find a projection of the data that maximizes class separability while preserving as much relevant information as possible.

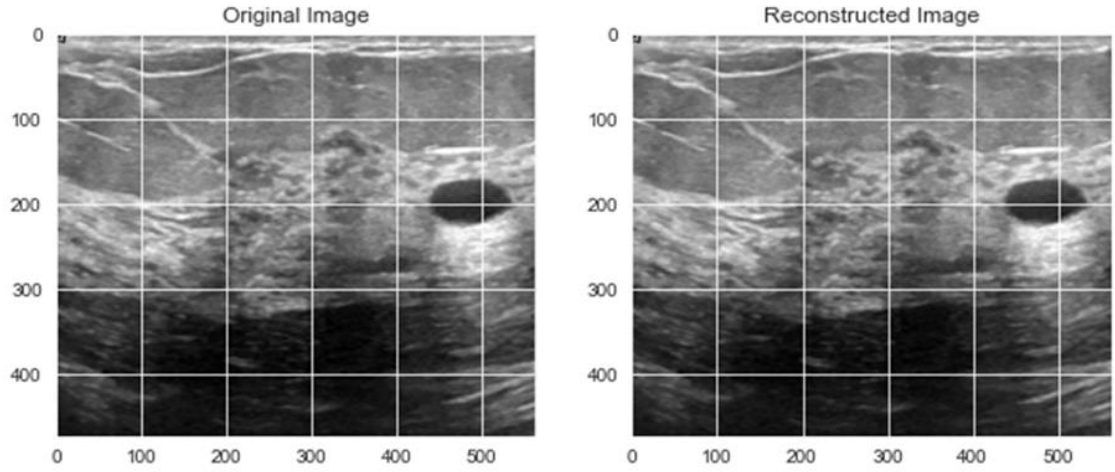

Figure 7 Comparison between original ultrasound image and PCA-reconstructed image after dimensionality reduction (SNR = 48.91 dB; compression ratio = 2.68×)

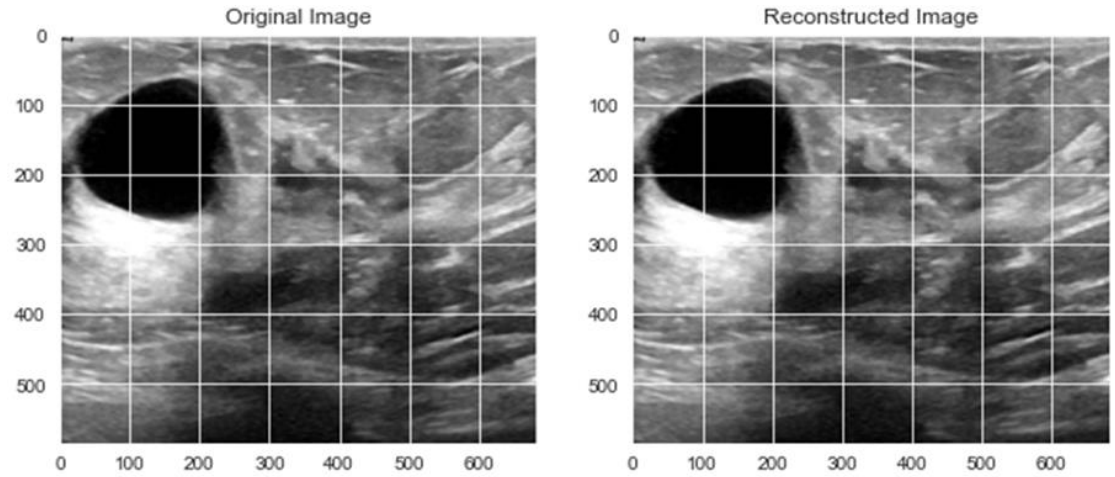

Figure 8 Original image vs. LDA-processed reconstruction; no compression or visual fidelity preserved due to class-dependent projection

5. Conclusion

This study compared three dimensionality reduction techniques-PCA, LDA, and t-SNE-paired with SVM, Random Forest, and XGBoost classifiers on a balanced set of 1,200 breast ultrasound images. PCA reduced image size by 37% and slightly improved precision, while LDA enhanced visual class separation in 2D but eliminated essential high-order features. t-SNE revealed that fundamental class differentiation follows a non-linear pattern. The highest F1 scores were achieved with the full feature set: XGBoost ≈ 77%, Random Forest ≈ 74%, and SVM ≈ 68%, indicating that neither PCA nor LDA provided performance benefits and in some cases, led to degradation. Automated diagnosis of breast cancer via ultrasound should prioritize using the complete high-dimensional feature set, with XGBoost as the preferred model. PCA may be reserved for storage or runtime efficiency, LDA for low-feature or rare scenarios, and t-SNE for exploration analysis rather than as a preprocessing step.

In general, dimensionality reduction techniques like PCA and LDA are valuable tools for reducing the computational complexity of machine learning models, especially when handling high-dimensional datasets such as medical imaging. These methods aim to simplify the feature space while retaining the most critical information, and they may improve model performance under certain conditions. However, for ultrasound image data, the application of PCA and LDA proved less effective. The dimensionality reduction process removed key information needed for accurate

classification, resulting in a loss of performance. This suggests that the feature transformations generated by PCA and LDA did not preserve the intrinsic data structure necessary for effective breast cancer diagnosis. A key finding was that retaining all original features led to better F1 scores than those obtained after dimensionality reduction. This highlights the sensitivity of classification models to even minor changes in the feature space. While dimensionality reduction can reduce computational load and enhance efficiency, it must also preserve the discriminative features required for identifying malignancies. Overall, this study emphasizes the need to align preprocessing strategies with the complexity of the data and the demands of clinical accuracy.

6. References

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, *28*, Article 104863. https://doi.org/10.1016/j.dib.2019.104863

Cen, Q., Wang, M., Zhou, S., Yang, H., & Wang, Y. (2025). Multi-center study: ultrasound-based deep learning features for predicting Ki-67 expression in breast cancer. *Scientific Reports*, *15*(1), Article 10279. https://doi.org/10.1038/s41598-025-94741-4

Fernandes, V., Carvalho, G., Pereira, V., & Bernardino, J. (2024). Analyzing data reduction techniques: an experimental perspective. *Applied Sciences*, *14*(8), Article 3436. https://doi.org/10.3390/app14083436

Li, L., Deng, H., Ye, X., Li, Y., & Wang, J. (2023). Comparison of the diagnostic efficacy of mathematical models in distinguishing ultrasound imaging of breast nodules. *Scientific Reports*, *13*(1), Article 16047. https://doi.org/10.1038/s41598-023-42937-x

Liu, B., Gu, X., Xie, D., Zhao, B., Han, D., Zhang, Y., ... & Fang, J. (2025). An Ultrasound-based Machine Learning Model for Predicting Tumor-Infiltrating Lymphocytes in Breast Cancer. *Technology in Cancer Research & Treatment*, *24*, Article 15330338251334453. https://doi.org/10.1177/15330338251334453

Luo, S., Chen, X., Yao, M., Ying, Y., Huang, Z., Zhou, X., ... & Huang, C. (2025). Intratumoral and peritumoral ultrasound-based radiomics for preoperative prediction of HER2-low breast cancer: a multicenter retrospective study. *Insights into Imaging*, *16*(1), Article 53. https://doi.org/10.1186/s13244-025-01934-6

Maiprasert, D., & Kitbumrungrat, K. (2023). Multinomial logistic regression analysis of breast cancer. *Journal of Current Science and Technology*, *2*(1), 23–31. Retrieved from https://ph04.tci-thaijo.org/index.php/JCST/article/view/595

NDTV World. (2024). *Breast cancer to cause a million deaths a year by 2040: Report*. Retrieved from https://www.ndtv.com/world-news/breast-cancer-to-cause-a-million-deaths-a-year-by-2040-report-5451027

Panyamit, T., Sukvivatn, P., Chanma, P., Kim, Y., Premratanachai, P., & Pechprasarn, S. (2022). Identification of factors in the survival rate of heart failure patients using machine learning models and principal component analysis. *Journal of Current Science and Technology*, *12*(2), 336–348. https://doi.org/10.14456/jcst.2022.26

Pechprasarn, S., Wattanapermpool, O., Warunlawan, M., Homsud, P., & Akarajarasroj, T. (2023). Identification of important factors in the diagnosis of breast cancer cells using machine learning models and principal component analysis. *Journal of Current Science and Technology*, *13*(3), 642–656. https://doi.org/10.59796/jcst.V13N3.2023.700

Tambe, S. N., Potharaju, S., Amiripalli, S. S., Tirandasu, R. K., & Jaidhan, B. J. (2025). Interdisciplinary research for predictive maintenance of MRI machines using machine learning. *Journal of Current Science and Technology*, *15*(1), Article 78. https://doi.org/10.59796/jcst.V15N1.2025.78

Wongnil, J., Krisanachinda, A., & Lipikorn, R. (2024). Breast cancer characterization using region-based convolutional neural network with screening and diagnostic mammogram. *Journal of Associated Medical Sciences*, *57*(3), 8–17. https://he01.tci-thaijo.org/index.php/bulletinAMS/article/view/269765

Wu, J., Ge, L., Guo, Y., Xu, D., & Wang, Z. (2024). Utilizing multiclassifier radiomics analysis of ultrasound to predict high axillary lymph-node tumour burden in node-positive breast-cancer patients: A multicentre study. *Annals of Medicine*, *56*(1), Article 2395061. https://doi.org/10.1080/07853890.2024.2395061

Xie, H., Tan, T., Li, Q., & Li, T. (2025). Revolutionizing HER-2 assessment: multidimensional radiomics in breast cancer

diagnosis. *BMC Cancer*, *25*, Article 265. https://doi.org/10.1186/s12885-025-13549-7

Yao, J., Jia, X., Zhou, W., Zhu, Y., Chen, X., & Zhan, W. (2024a). Predicting axillary response to neoadjuvant chemotherapy using peritumoral and intratumoral ultrasound radiomics in breast-cancer subtypes. *iScience*, 27(9), Article 110716. https://doi.org/10.1016/j.isci.2024.110716

Yao, J., Zhou, W., Xu, S., Jia, X., & Zhan, W. (2024b). Machine-learning-based breast-tumor ultrasound radiomics for pre-operative prediction of axillary sentinel lymph-node metastasis burden in early-stage invasive breast cancer. *Ultrasound in Medicine & Biology*, *50*(2), 229-236. https://doi.org/10.1016/j.ultrasmedbio.2023.10.004

Zhang, L., Wang, L., Liang, R., He, X., & Jiang, J. (2024). An effective ultrasound features-based diagnostic model via principal component analysis facilitated differentiating subtypes of mucinous breast cancer from fibroadenomas. *Clinical Breast Cancer*, 24(7), e583–e592.e3. https://doi.org/10.1016/j.clbc.2024.05.007

Zhao, X., Guo, J., Nie, F., Chen, L., Li, Z., & Zhang, H. (2020). Joint principal component and discriminant analysis for dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(2), 433-444. https://doi.org/10.1109/TNNLS.2019.2904701

Zhao, M., Zheng, Y., Chu, J., Liu, Z., & Dong, F. (2023). Ultrasound-based radiomics combined with immune status to predict sentinel lymph-node metastasis in primary breast cancer. *Scientific Reports*, *13*, Article 16918. https://doi.org/10.1038/s41598-023-44156-w