Texture-based feature selection for Breast Ultrasound Classification Using Sum and Difference Histograms

Selección de características basada en textura para la clasificación de imágenes de ultrasonidos mamario mediante histogramas de suma y diferencia

Julio César Medina Flores[1], Juan José Cárdenas Cornejo[2], Miguel Ángel Armenta Loredo[1], Leonardo Martínez Jiménez[1], Dora Luz Almanza Ojeda[2]

[1] Multidisciplinary Studies Department, Engineering Division, Campus Irapuato-Salamanca, University of Guanajuato, Yuriria, Guanajuato, México. jc.medinaflores@ugto.mx, ma.armentaloredo@ugto.mx, leonardomj@ugto.mx,
[2] Engineering Division, Campus Irapuato-Salamanca, University of Guanajuato, Salamanca, Guanajuato, México. jj.cardenasc@ugto.mx, dora.almanza@ugto.mx

Abstract

Automatic analysis of lesions in breast ultrasound images is a valuable support tool for the early diagnosis of breast cancer; however, speckle noise, low contrast, and variability in tissue appearance pose challenges for classification systems. This work proposes a methodology for classifying breast ultrasound images using Machine Learning algorithms, based on a feature descriptor derived from the Sum and Difference of Histograms (SDH). We conducted the experimental evaluation using the publicly available Breast Ultrasound Images (BUSI) dataset, which comprises 780 images distributed among 437 benign, 210 malignant, and 133 normal cases. To determine the most suitable configuration, we evaluated different preprocessing techniques, multiple window sizes, and 24 Machine Learning (ML) algorithms, and automatically selected the combination that achieved the best performance. The results show that the proposed descriptor adequately discriminates among breast lesion categories; the best configuration used a Boosted Trees classifier with a 21×21 window, achieving an F1-score of 90.35% and an accuracy of 86.15%. These results indicate that the proposed methodology offers a lower-computational-complexity alternative to Deep Learning (DL)-based approaches while maintaining competitive performance for developing computer-aided diagnosis systems.

Keywords: Machine Learning (ML); Breast Ultrasound Images (BUSI); Sum and Difference of Histograms (SDH); classification.

Introduction

Breast cancer is among the diseases with the highest incidence and is a leading cause of mortality among women worldwide, representing a major public health concern. It is estimated that approximately 2.3 million new cases of breast cancer and nearly 670,000 deaths associated with this disease were recorded in 2022 (Kim *et al*., 2025). According to the World Health Organization, early detection remains one of the most important factors for increasing survival and improving treatment effectiveness (WHO, 2026).

Among imaging modalities used to detect and characterize breast lesions, ultrasound has gained considerable relevance because it is non-invasive, low-cost, and does not involve ionizing radiation. It is particularly useful as a complementary method to mammography, especially in women with dense breast tissue, in whom mammographic sensitivity may be reduced. Recent studies have shown that incorporating supplemental ultrasound into screening programs significantly increases the cancer detection rate in this group of patients (Heywang-Köbrunner *et al*., 2026). Nevertheless, interpreting ultrasound images remains challenging due to speckle noise, low contrast, variations in tissue intensity, and the high similarity between healthy and pathological structures. These factors hinder delineation of regions of interest and affect diagnostic accuracy (Li *et al*., 2022).

To overcome these limitations, the development of methodologies based on the quantitative analysis of medical images has intensified in recent years. In this context, the incorporation of radiomics has demonstrated considerable potential for extracting descriptive texture features and improving the classification of breast lesions (Gu & Jiang, 2022). Likewise, hybrid models that integrate radiomic features with clinical

information have achieved AUC values above 0.92 in differentiating BI-RADS 4 lesions (Zhang *et al*., 2025), demonstrating progress toward increasingly objective and reliable diagnostic systems.

In parallel, Computer-Aided Diagnosis (CAD) systems have incorporated digital image processing and artificial intelligence techniques to support the detection and classification of breast lesions. Recent studies have documented progress in this field (Andreadis *et al*., 2025), while specific approaches, such as BREAST-CAD, have demonstrated the potential of machine learning algorithms to achieve competitive results in classification tasks (Masoud *et al*., 2025). Furthermore, deep learning techniques have accounted for a substantial portion of research in this area, achieving outstanding performance across different medical imaging datasets (Sabry *et al*., 2026).

The publicly available Breast Ultrasound Images (BUSI) dataset, developed by (Al-Dhabyani *et al*., 2019), has become a primary reference dataset for evaluating breast lesion classification methods. Its public availability, together with the inclusion of benign, malignant, and normal images acquired under real clinical conditions, has provided a framework for comparing algorithms based on both Machine Learning (ML) and Deep Learning (DL) (Al-Dhabyani *et al*., 2019). Although the BUSI dataset has been used to evaluate classification algorithms, recent literature shows a clear trend toward the development of DL-based architectures. These studies include convolutional networks, hybrid models, and attention mechanisms, whereas studies based exclusively on texture descriptors and traditional ML algorithms are considerably less frequent (Kamsari *et al*., 2025) (Akash Chauhan & Indrajeet Kumar, 2025) (Ogut *et al*., 2025).

This scenario highlights the remarkable progress of deep learning architectures; however, it also underscores the need to explore alternatives that maintain competitive performance without requiring extensive computational resources. Consequently, there remains an opportunity to develop feature descriptors that efficiently represent the textural information in ultrasound images, leveraging the benefits of traditional ML algorithms without necessarily relying on deep architectures.

From this perspective, the present study proposes and evaluates a feature descriptor based on the sum and difference of histograms for representing breast ultrasound images. The descriptor is evaluated using the BUSI dataset across multiple ML algorithms, with various preprocessing strategies, multiple window sizes, and different classifiers, to analyze the influence of these factors on its discriminative capability and to identify the configuration that yields the best performance for the automatic classification of breast lesions. In this way, the study seeks to contribute to the development of CAD systems with lower computational complexity, providing an efficient alternative to support the early detection of breast cancer.

Proposed Methodology

Figure 1 presents a general diagram of the proposed methodology. The procedure is structured into five main stages: image preprocessing, feature extraction using a descriptor based on the sum and difference of histograms (SDH), dimensionality reduction, feature vector construction, and classification using different Machine Learning algorithms. To analyze the influence of preprocessing techniques on the descriptor's performance, three strategies were independently evaluated: gamma correction, histogram adjustment, and Contrast Limited Adaptive Histogram Equalization (CLAHE). For each strategy, the complete methodological pipeline was executed independently, allowing its effect on the classification process to be compared and the configuration providing the best performance to be identified.

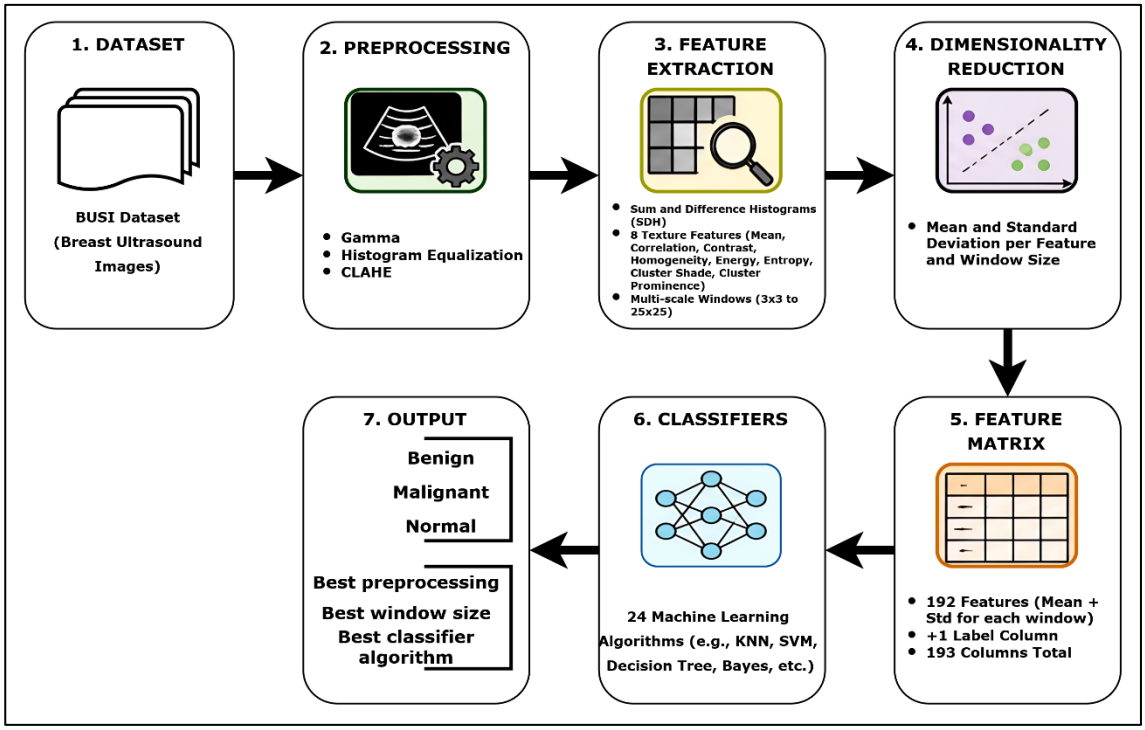
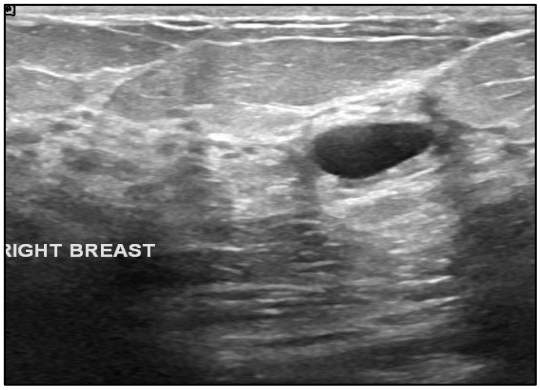


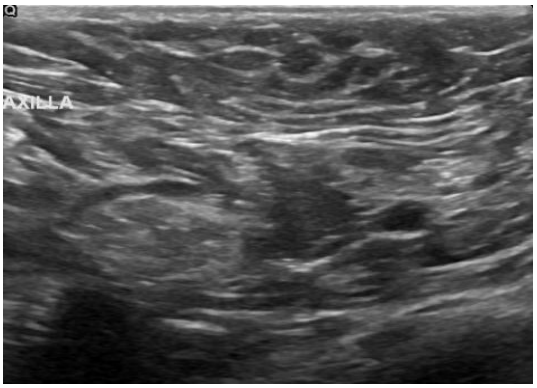
Figure 1. Block diagram of the methodology.
Source: Authors' own elaboration.

DATASET

The experimental evaluation of the proposed methodology was conducted using the publicly available Breast Ultrasound Images (BUSI) dataset, developed by (Al-Dhabyani *et al*. 2020). This dataset was acquired through a collaboration between the Faculty of Computer and Informatics at Benha University (Egypt) and Baheya Hospital for Early Detection and Treatment of Women's Cancer to provide a reference dataset for developing and evaluating computer-aided diagnosis systems based on breast ultrasound images.

The dataset consists of 780 ultrasound images from 600 patients, distributed across three diagnostic categories: benign, malignant, and normal, as shown in Table 1. All images were acquired under real clinical conditions and are provided with their corresponding diagnostic labels, enabling their use for supervised classification tasks. Figure 2 presents a representative example of each category in the dataset. Furthermore, an imbalance in the number of images per class is evident and should be taken into consideration when evaluating the performance of the classification models.





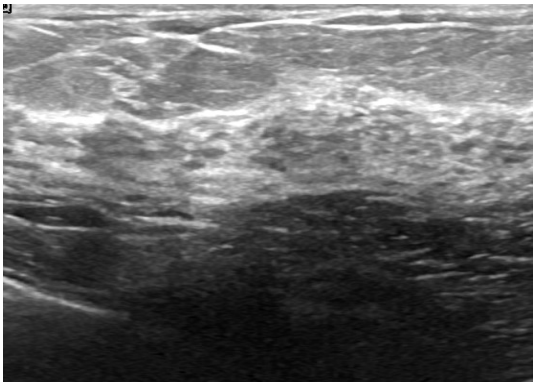


Figure 2. Example images from the BUSI database, from left to right and top to bottom: benign, malignant, and normal. Source: https://www.kaggle.com/datasets/sabahesaraki/breast-ultrasound-images-dataset

***Table 1**. Components of the BUSI database.*

Pathology	Number of images
Benign	437
Malignant	210
Normal	133

Due to its public availability, the diversity of clinical cases, and its widespread use in recent research, the BUSI dataset has become a reference for evaluating and comparing methodologies based on both Machine Learning (ML) and Deep Learning (DL) (Dan *et al*., 2024). In the present study, this dataset was used to evaluate the discriminative capability of the proposed feature descriptor across different classification algorithms.

Preprocessing

The images in the BUSI dataset underwent preprocessing to improve their visual quality and facilitate the extraction of discriminative features. Because ultrasound images commonly exhibit speckle noise, low contrast, and variations in intensity distribution, factors that may affect the representation of textural information (Li *et al*., 2022), three image enhancement methods were evaluated: gamma correction, histogram adjustment, and Contrast Limited Adaptive Histogram Equalization (CLAHE).

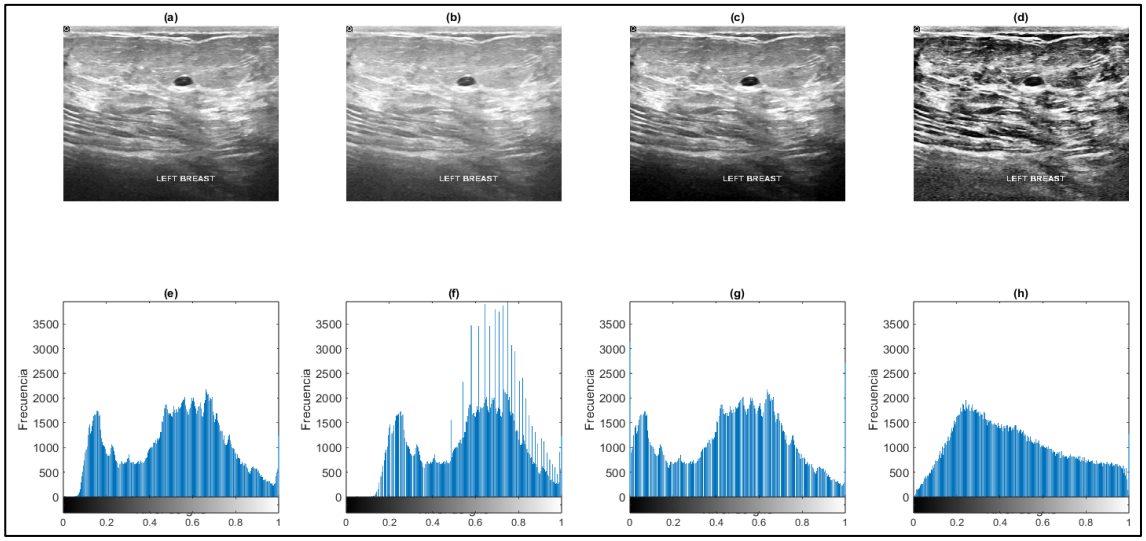


Figure 3. illustrates the preprocessing techniques used on an image from the BUSI dataset. The top row displays: (a) the original image, (b) gamma correction, (c) histogram adjustment, and (d) CLAHE. The bottom row shows the corresponding histograms: (e) the original, (f) gamma correction, (g) histogram adjustment, and (h) CLAHE.
Source: Authors' own elaboration.

Gamma correction was applied with a value of $\gamma = 0.75$ to modify the relationship between intensity levels and enhance structures of interest without significantly altering the information contained in the image. Histogram adjustment redistributed gray levels to increase global contrast, whereas CLAHE applied contrast-limited adaptive histogram equalization, improving the visibility of local details while reducing excessive noise amplification.

To assess how preprocessing affects the proposed descriptor's performance, each method was tested independently. The entire methodological pipeline was run separately for each of the four strategies, from feature extraction to classification. This approach enabled a comparison of the effects of each preprocessing strategy, window size, and the classifier that produced the best classification results.

Feature extraction

After preprocessing, the images underwent feature extraction using a descriptor based on Sum and Difference Histograms (SDH) (Unser, 1986). This descriptor represents the spatial distribution of intensity levels in the image by analyzing relationships among neighboring pixels, providing relevant information about the texture of breast tissues.

To capture the anisotropy and orientation of tissue structures, sum and difference matrices are computed using spatial displacements along four principal directions: horizontal (0°), vertical (90°), and diagonal (45° and 135°). The resulting matrices from the four directions are then averaged element-wise to obtain a rotation-invariant, omnidirectional representation. This procedure also ensures that the resulting feature vector is rotation-invariant (Haralick *et al*., 1973), which is particularly relevant in medical ultrasound, where the transducer angle may vary during clinical image acquisition. Furthermore, averaging consolidates global textural information, reduces the dimensionality of the feature space, and mitigates the risk of computational overfitting (Unser, 1986), while compensating for spurious variations introduced by acoustic speckle patterns.

Subsequently, to analyze texture at different spatial scales, feature extraction was performed using a sliding-window scheme. Based on the image's median width and height, we defined an analysis region and applied a fixed 5-pixel displacement between consecutive windows. We used square windows with odd dimensions ranging from 3×3 to 25×25, increasing the window size in increments of two (3×3, 5×5, 7×7, ..., 25×25). Odd-sized windows allow a central pixel to be defined, facilitating analysis of the neighborhood surrounding each image element (Ghalati *et al*., 2021).

For each generated sub-window, SDH matrices were constructed, from which eight second-order statistical features widely used in texture analysis were computed: Mean, Correlation, Contrast, Homogeneity, Energy, Entropy, Cluster Shade, and Cluster Prominence. These features capture different properties of the spatial distribution of intensity levels and provide a quantitative representation of the textural behavior of breast tissue.

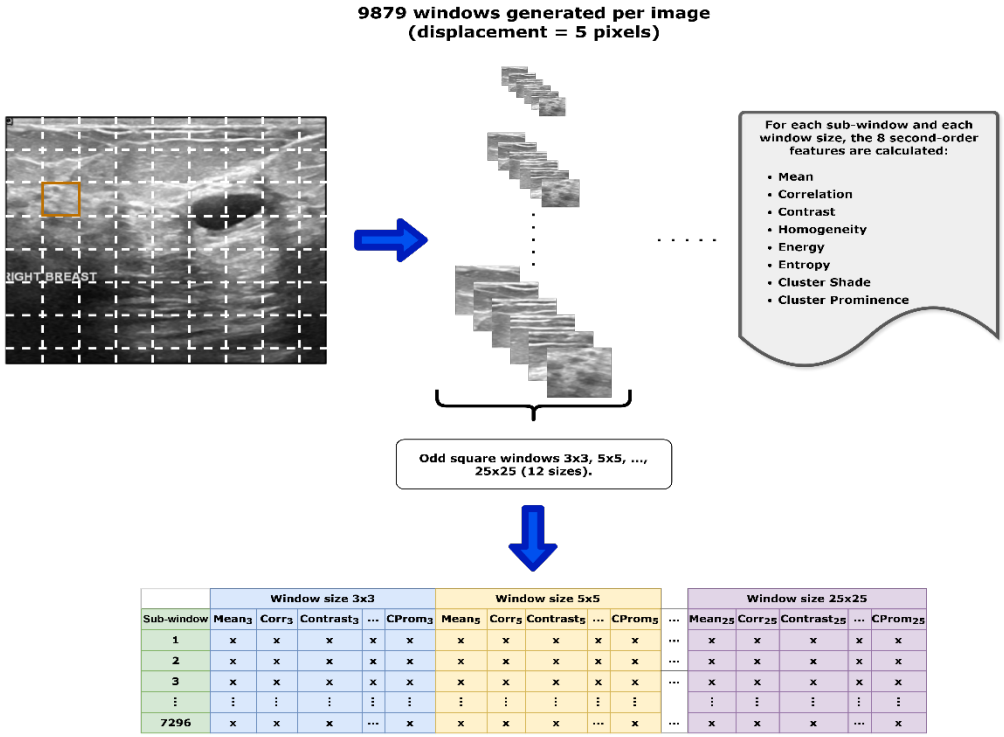


Figure 4. Feature extraction process.
Source: Authors' own elaboration.

The equations for each feature are presented below, along with their interpretation in the context of image processing. In these expressions, the variables $\boldsymbol{i}$ and $\boldsymbol{j}$ represent the possible values of the sum and difference histograms of gray levels, while $\boldsymbol{P_s(i)}$ and $\boldsymbol{P_d(j)}$ denote their corresponding normalized histograms:

Mean: Represents the average intensity value within the analyzed region, estimating the overall brightness or predominant gray level of the tissue.

$$\mu = \frac{1}{2}\sum_i i \cdot P_s(i) \tag{1}$$

Correlation: Quantifies the linear dependency between neighboring pixels. High values indicate organized textural patterns and structures with strong spatial continuity, whereas low values reflect more disorganized distributions.

$$Correlation = \frac{1}{2}\left[\sum_i (i-2\mu)^2 P_s(i) - \sum_j j^2 P_d(j)\right] \tag{2}$$

Contrast: Measures local intensity variations between adjacent pixels. High values indicate abrupt changes in texture, which are often associated with well-defined edges and heterogeneous structures.

$$Contrast = \sum_{j} j^2 P_d(j) \tag{3}$$

Homogeneity: Measures the degree of uniformity in the intensity distribution. High values correspond to regions with gradual changes between neighboring pixels, whereas low values indicate greater structural heterogeneity.

$$Homogeneidad = \sum_{j} \frac{1}{1+j^2} P_d(j) \tag{4}$$

Energy: reflects how ordered or uniform the texture is. Higher values suggest repetitive patterns and consistent distributions, while lower values imply more complex structures.

$$Energy = \sum_{i} P_s(i)^2 + \sum_{j} P_d(j)^2 \tag{5}$$

Entropy: Quantifies the randomness in the intensity distribution. High values indicate complex, heterogeneous textures, whereas low values correspond to more uniform regions.

$$Entropy = -\sum_{i} P_s(i) \log_2[P_s(i)] - \sum_{j} P_d(j) \log_2[P_d(i)] \tag{6}$$

Cluster Shade: Measures the degree of asymmetry in the intensity distribution relative to its mean. High absolute values indicate skewed distributions and greater texture irregularity.

$$Cluster\ Shade = \sum_{i} (i - 2\mu)^3 P_s(i) \tag{7}$$

Cluster Prominence: Evaluates the degree of concentration of the distribution around its mean and is sensitive to pronounced peaks and distributions with high kurtosis.

$$Cluster\ Prominence = \sum_{i} (i - 2\mu)^4 P_s(i) \tag{8}$$

Using these features together allows for creating descriptors that can differentiate between normal tissue, benign lesions, and malignant lesions. This information forms the foundation for the spatial feature aggregation stage, where the descriptor's behavior is summarized to produce the feature vector used in machine learning algorithms.

Dimensionality reduction (Global spatial feature aggregation)

Once the textural features were obtained for each window size, a compact descriptor for each complete image was needed. To this end, we first analyzed the dimensions of all images in the BUSI dataset, revealing variations in both width and height. To establish a uniform extraction procedure, we defined an analysis region using the median of the minimum and maximum dimensions across the dataset, with a fixed displacement of 5 pixels between consecutive sliding-window positions. Given the maximum window size (25 × 25 pixels) and the specified displacement, the total number of horizontal and vertical positions at which the SDH descriptor could be evaluated was determined. The number of windows generated in each direction was computed using Equations (9) and (10), and the total number of sub-windows evaluated per image was obtained using Equation (11).

$$W = Width = \left(\frac{N_W - W_{max}}{S}\right) + 1 \tag{9}$$

$$H = Height = \left(\frac{N_H - W_{max}}{S}\right) + 1 \tag{10}$$

$$Subwindows = W * H \quad (11)$$

Where: N_W = image width, N_H = image height, W_{max} = window size, and S = stride.

For each generated sub-window, the eight statistical features described in the previous section were computed across the 12 window sizes evaluated, yielding 96 features per position (8 features × 12 window sizes). Consequently, each image produced a large volume of information, equivalent to thousands of spatially distributed feature vectors, substantially increasing the computational cost during the classification stage.

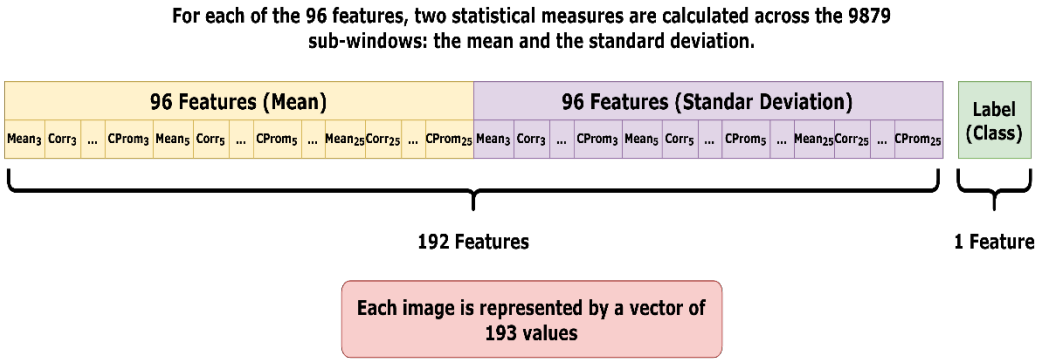


Figure 5. Final vector for each image.
Source: Authors' own elaboration.

To reduce the volume of information generated, we applied a global spatial aggregation process. For each of the 96 features, the arithmetic mean and standard deviation were computed across all sub-windows of the image. Thus, each image was represented by a 192-dimensional feature vector (96 means and 96 standard deviations), preserving both the average behavior and the spatial variability of the SDH descriptor across the entire image.

Characteristics matrix

After the global spatial aggregation stage, each image in the dataset was represented by a 192-dimensional feature vector consisting of the means and standard deviations calculated from the 8 textural features across the 12 window sizes evaluated. We then appended the corresponding diagnostic class label (benign, malignant, or normal) to each vector, resulting in a final matrix with 193 columns. In this representation, each row corresponds to an image from the BUSI dataset, and each column represents a descriptive feature; the last column contains the class label. This organization provides a homogeneous structure for all samples, facilitating their processing with supervised learning algorithms.

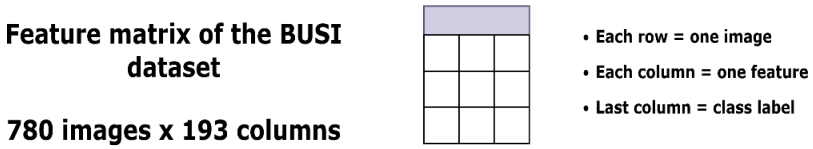


Figure 6. Final Matrix.
Source: Authors' own elaboration.

The feature matrix serves as the input to the classification stage, enabling evaluation of the proposed descriptor's performance under different preprocessing configurations, window sizes, and ML algorithms.

Classifiers

The resulting feature matrix served as input to the supervised classification process. To evaluate the discriminative capability of the proposed descriptor, 24 machine learning algorithms from different classification families were analyzed, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Random Forests, and Ensemble Learning techniques. Each algorithm was evaluated using the same training and test splits across the methodological configurations, considering the three preprocessing techniques and the twelve window sizes used for feature extraction. This strategy enabled consistent comparisons across models and allowed analysis of each configuration's influence on classification performance.

Because the BUSI dataset is imbalanced, we primarily evaluated classifier performance using the F1-score, which better reflects the balance between precision and recall than accuracy in imbalanced classification problems. Additionally, we calculated other performance metrics and present them in the Results section. Finally, the proposed methodology enabled the automatic identification of the combination of preprocessing technique, window size, and ML algorithm that achieved the best performance in breast lesion classification, thereby providing the optimal configuration for the developed automated classifier.

Results

The experimental results evaluating the proposed multiclass classification methodology on the BUSI dataset are presented. To ensure the methodological validity and generalization capability of the evaluated models, the experiments were conducted using two distinct configurations of the dataset. Initially, the dataset published by Al-Dhabyani *et al*. contained 780 images classified as normal (133 images), benign (437 images), and malignant (210 images). However, subsequent research (Pawłowska *et al*., 2023) revealed critical inconsistencies in this distribution. Specifically, researchers identified 155 duplicate images, representing 19% of the total dataset, along with significant discrepancies in the corresponding reference binary lesion masks. Likewise, eight identical lesion images were incorrectly labeled across multiple classes, and 70 images corresponded to anatomical structures outside the breast.

This data overlap poses a high risk of data leakage, potentially causing the algorithm to memorize patterns instead of learning true textural features, resulting in unnaturally high classification accuracy. To ensure realistic diagnostic performance, we first conducted experiments on the original dataset with leakage to create a baseline comparable to earlier studies. Next, we performed a second set of experiments on the cleaned, leakage-free dataset, which was limited to 4450 images after removing duplicates, auxiliary images, and inconsistencies. Under this dual validation scheme, the evaluation focused on analyzing the influence of three key factors on the performance of the proposed methodology: (1) the image processing method, (2) the sliding-window size used for feature extraction through the SDH, and (3) the machine learning algorithms employed. To determine which enhancement strategy best supports feature extraction, we compared the overall performance of the four evaluated approaches: original images, gamma correction, histogram adjustment, and CLAHE.

Because ultrasound images vary in spatial dimensions due to differences in imaging equipment, acquisition settings, and scanning depth, and to establish a robust quantitative procedure, we performed a dimensional analysis of the two evaluated datasets: the dataset with information leakage and the cleaned dataset. Table 2 presents the dimensional statistics for both datasets.

Table 2. *Statistical analysis of dimensions in the original BUSI dataset images, comparing cases with and without data leakage*.

Dataset	Number of images	Class distribution	Height (H, pixels)	Width (W, pixels)	Resolution Used
Dataset with data leakage	780	Benign: 437 Malign: 210 Normal: 133	Minimum: 310 Maximum: 719 Mean: 504 Median: 474	Minimum: 190 Maximum: 1048 Mean: 616 Median: 564	[474 x 564]
Dataset without data leakage	450	Benign: 222 Malign: 164 Normal: 64	Minimum: 512 Maximum: 512 Mean: 512 Median: 512	Minimum: 512 Maximum: 512 Mean: 512 Median: 512	[512 x 512]

Processing images with variable dimensions poses several challenges: (1) when images have different sizes, the relative scale of the noise and the filter vary across samples; and (2) SDH descriptor extraction evaluates 12 odd-sized window configurations. If image dimensions vary, each image would generate a different number of local subwindows, preventing the construction of regular tabular arrays. To address this challenge, we resized the images in each dataset to their respective reference median resolution using bicubic interpolation. By fixing the resolution at the dataset median, we minimize scaling-induced geometric distortion while constraining the analysis space so the preprocessing algorithms and multiscale scanning operate under identical spatial conditions.

Substituting the corresponding numerical values into Equations (9) to (11) with $W_{max} = 25$ px y $S = 5$ px:

- Total number of subwindows evaluated in the dataset with data leakage (780 images): **7,581,600**.
- Total number of subwindows evaluated in the dataset without data leakage (450 images): **4,321,800**.

For each local subwindow at each of the 12 window sizes, eight texture descriptors are computed, resulting in a specialized feature matrix per image. This process generates approximately 10,000 local feature vectors per image. Due to this volume of data, the Global Spatial Aggregation step calculates the mean and standard deviation across subwindows to summarize the information. This reduces the 4.32 million subwindows from the cleaned dataset into consistent, compact vectors with exactly 192 features per image. These vectors are then assembled into a global tabular matrix of 450 images—360 for training (80%) and 90 for testing (20%)—ready for use with 25 machine learning classifiers.

Table 3 presents the best results obtained on the dataset with data leakage. This scenario serves as a baseline for comparison, in which the distribution of texture features was altered, thereby influencing classifier performance.

***Table 3**. Comparison of the performance of each preprocessing technique applied to the BUSI dataset with data leakage.*

Processing Technique	Window Size	Classifier	Macro F1-Score	Accuracy
No preprocessing	25x25	Boosted Trees	89.27%	89.10%
Gamma Correction	23x23	Boosted Trees	89.64%	87.18%
Histogram adjustment	7x7	Quadratic SVM	75.81%	79.49%
CLAHE	23x23	Boosted Trees	88.41%	87.18%

To counteract this bias and assess the descriptor's true generalization capability, Table 4 presents diagnostic performance on the cleaned dataset, which includes only clinically valid breast ultrasound images.

Table 4. *Comparison of the performance of each preprocessing technique applied to the BUSI dataset without data leakage*

Preprocessing Technique	Window Size	Classifier	Macro F1-Score	Accuracy
No preprocessing	19x19	Boosted Trees	86.89%	82.22%
Gamma correction	23x23	Boosted Trees	85.65%	78.89%
Histogram adjustment	11x11	Boosted Trees	70.11%	70.00%
CLAHE	21x21	Boosted Trees	90.35%	86.15%

To comprehensively illustrate the impact of the proposed parametric variability, namely, the interaction between the 12 window sizes and the 24 classifiers, as well as the tabulated results, heatmaps were generated to show both the accuracy and the Macro F1-Score for each experimental configuration.

Figure 7 shows the performance of the SDH technique for extracting features from the original images without any preprocessing. This analysis was conducted to observe how data leakage creates regions of artificially high performance that diminish when the methodology is evaluated on the cleaned dataset. In other words, without data leakage, baseline performance decreases and depends strongly on specific window sizes.

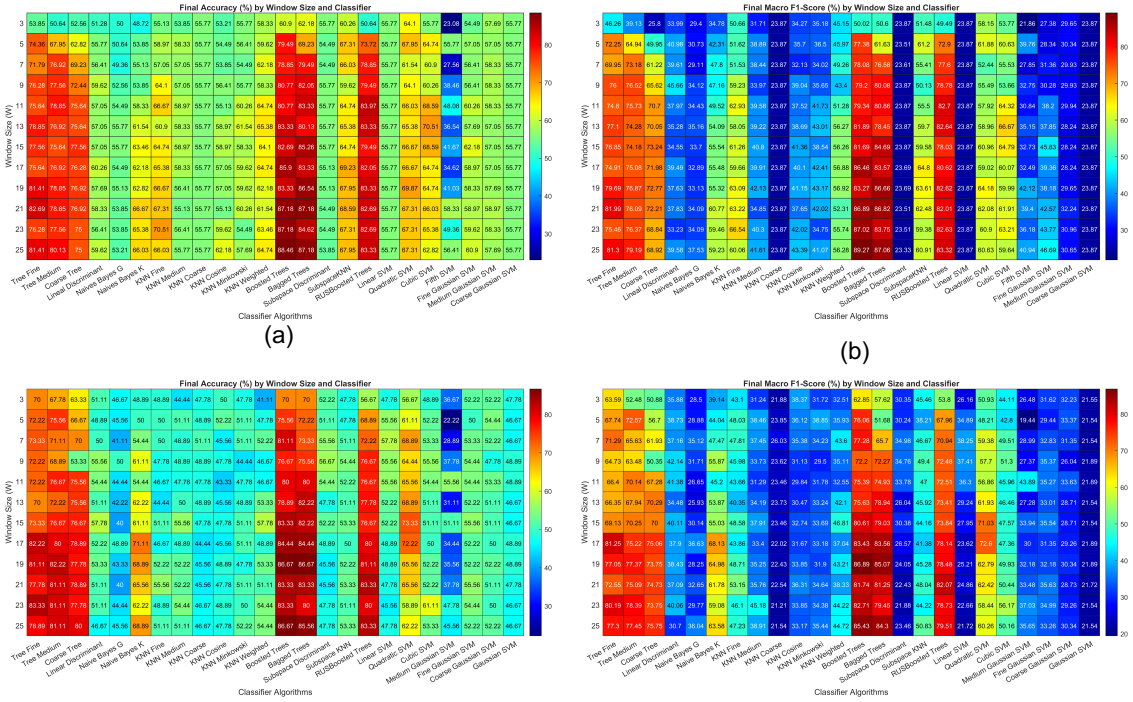


Figure 7. Heatmaps of the performance obtained from original images without preprocessing. With data leakage: (a) accuracy, (b) Macro F1-Score. Without data leakage: (c) accuracy, (d) Macro F1-Score.
Source: Authors' own elaboration.

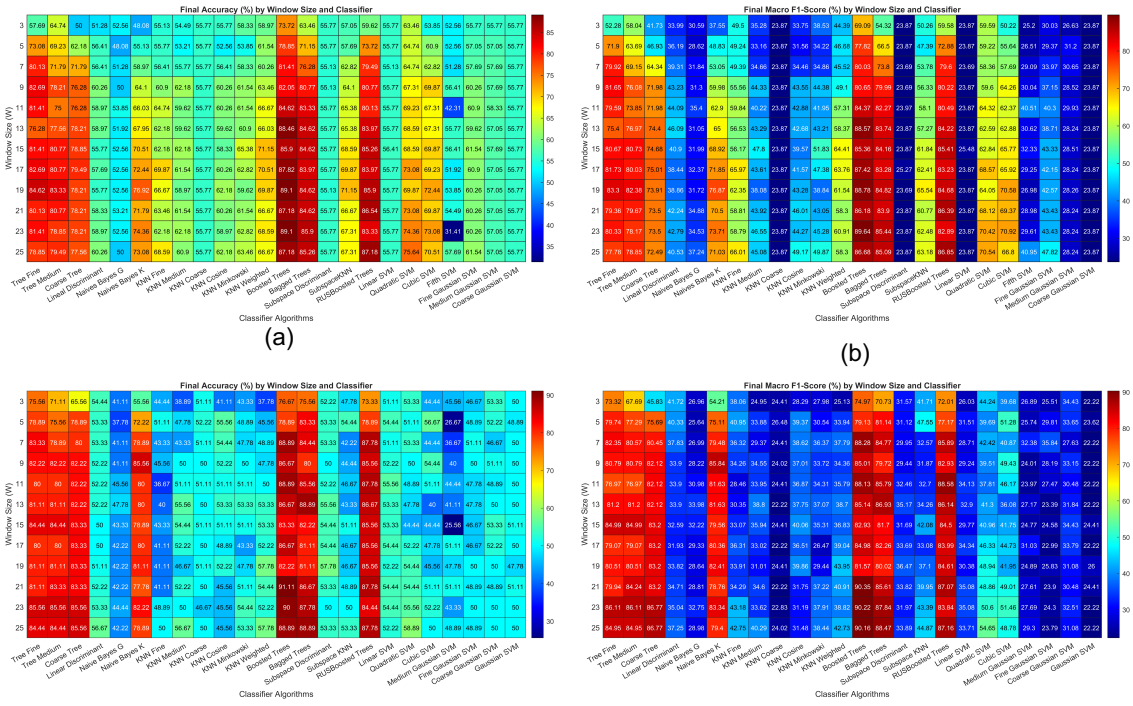


Figure 8. Heatmaps of the performance obtained from preprocessed images. With data leakage: (a) accuracy, (b) Macro F1-Score. Without data leakage: (c) accuracy, (d) Macro F1-Score.
Source: Authors' own elaboration.

Subsequently, Figure 8 shows the impact of applying the optimal preprocessing technique in each scenario. In the presence of data leakage, gamma correction was the best preprocessing method in terms of its overall impact on classification performance. Conversely, when evaluated using the cleaned dataset, CLAHE achieved the best overall performance.

To compare the proposed methodology with the state of the art, Table 5 presents results from our approach alongside those reported in other studies that classify using the same BUSI dataset.

Table 5. Performance comparison between the proposed methodology and state-of-the-art methods.

Study	Methodology	Dataset Used	Accuracy	Macro F1-score
Jayakrishnan & Katal (2026)	Radiomics + Bagged Tree	Original BUSI (with leakage)	98.96%	Not reported
Mishra *et al*. (2021)	Radiomics + ML	Original BUSI (with leakage)	97.40%	94.00%
Chauhan & Kumar (2025)	VGG16 + SVM	Original BUSI (with leakage)	91.02%	Not reported
Ogut *et al*. (2025)	Spatial-Geometric + CNN	Original BUSI (with leakage)	87.90%	83.30%
Ru, Zhu & Shi (2024)	CNN + GNN + Federated Learning	BUSI, BUS y OASBUD	**87.60%**	83.30%

Our Proposal (Baseline)	SDH + Gamma + Boosted Trees	Original BUSI (with leakage)	87.18%	89.64%
Our Proposal (Strict)	SDH + CLAHE + Boosted Trees	Original BUSI (without leakage)	86.15%	90.35%

As shown in Table 5, several studies implement approaches based on convolutional neural networks and attention mechanisms, achieving accuracy and F1-score values ranging from 91% to 98%. However, when these studies use the original BUSI dataset, the reported data leakage may compromise the validity of these results. Conversely, among studies reporting comparable metrics, such as the *Spatial-Geometric + CNN* approach with an F1-score of 83.30%, our proposed methodology demonstrates superior performance. In a clean, strict, data-leakage-free setting, our model (SDH + CLAHE + Boosted Trees) achieves a competitive F1-score of 90.35%.

This result supports and validates the proposed methodology against the prevailing trend of using deep learning architectures. Although such models are powerful, they operate as “black boxes”, require substantial computational resources, and are highly susceptible to memorizing dataset-specific parameters when the data are not rigorously curated. In contrast, the main disadvantage of our methodology compared with deep models is its inherent sensitivity to empirical parameterization: the system requires manual definition and iterative evaluation of the window size to extract optimal texture information, lacking the algorithmic capability of Deep Learning to automate hierarchical, multiscale learning of image features.

Analysis and Discussion

The results obtained allow the behavior of the Sum and Difference of Histograms (SDH)-based descriptor to be analyzed from two complementary perspectives: its performance on the original BUSI dataset, used as a baseline for comparison with previous studies, and its performance on a cleaned version of the dataset, in which the effects associated with image duplication and data leakage are reduced. This distinction is particularly relevant because subsequent investigations following the publication of the BUSI dataset have identified inconsistencies in its composition. Pawłowska *et al*. (2023) documented 235 duplicated images, including cases in which equivalent lesions appear in different diagnostic categories, images corresponding to axillary regions, and discrepancies in the reference masks.

In the data leakage scenario, the configuration using SDH, gamma correction, and Boosted Trees achieved a Macro F1-Score of 89.64%, whereas the best result without preprocessing was 89.27%. These results indicate that the descriptor has considerable discriminative capability even without learned representations from deep neural networks. However, these values should be interpreted only as a baseline, since duplicate imagers may cause highly similar samples to appear in both the training and test sets. Consequently, a direct comparison of these values with results reported using the original BUSI dataset may overestimate the generalization capability of the models. The comparison also shows that high accuracy reported on BUSI does not necessarily imply better generalization. The systematic review by Dan *et al*. (2024) indicates that although Deep Learning has shown promising results in breast ultrasound, considerable variability persists across studies, and there remains insufficient evidence to conclude that these systems consistently outperform specialists in real-world clinical scenarios. The authors also emphasize the need to standardize evaluation protocols and to jointly consider metrics such as sensitivity, specificity, AUC, and accuracy.

From this perspective, the Macro F1-Score of 90.35% achieved with the SDH + CLAHE + Boosted Trees configuration is particularly relevant. Although it does not exceed the accuracy values reported by deep learning architectures evaluated on the original BUSI dataset, it delivers competitive performance in a more stringent evaluation setting and, importantly, relies on an explicit feature representation and a traditional classifier. This considerably reduces architectural complexity compared with deep neural network solutions while providing a feature representation that is amenable to analysis and interpretation.

Another advantage of the proposed methodology is the controlled dimensionality of its representation. The procedure transforms local information obtained through SDH into a global 192-dimensional vector by combining means and standard deviations. This approach preserves information about the average behavior

and spatial variability of the texture without directly introducing thousands of local feature vectors into the classification process. In contrast, Deep Learning architectures automatically learn hierarchical representations from the data, but their interpretation is generally less straightforward, and their computational requirements may be higher. Nevertheless, the proposed methodology has an important limitation: the window size is a critical hyperparameter that must be determined experimentally. The best performance depends on the interaction among the window size, preprocessing technique, and classification algorithm. Therefore, although the SDH descriptor reduces the complexity of the representation process, it still requires manual parameterization, unlike deep architectures that can learn multiscale representations during training.

Finally, the results suggest that the main contribution of the proposed methodology does not necessarily lie in surpassing the maximum accuracy reported for BUSI, but rather in demonstrating that a relatively compact texture descriptor, combined with traditional machine learning algorithms, can achieve competitive performance when dataset quality is properly controlled. Therefore, comparisons should follow equivalent protocols and avoid interpreting results across different datasets, data partitions, tasks, and evaluation metrics as directly comparable.

Conclusions

This study proposed and comprehensively evaluated a methodology for classifying lesions in breast ultrasound images, based on texture feature extraction using the Sum and Difference of Histogram (SDH) and machine learning algorithms. The study revealed a critical bias in the current literature arising from data leakage in the BUSI dataset and established a series of validation experiments using a rigorously cleaned dataset. Under these conditions, the optimal model, consisting of CLAHE preprocessing, a 21 x 21 pixel window size, and the Boosted Trees classifier, achieved a Macro F1-Score of 90.35%, demonstrating notable robustness to class imbalance.

It is concluded that the multidirectional SDH descriptor, when combined with adaptive equalization techniques and tree-based ensembles, is a highly competitive, interpretable, and computationally efficient alternative to Deep Learning models. As future work, we plan to develop adaptive windowing mechanisms or automatic multiresolution analysis and to validate this architecture on other independent datasets, enabling efficient implementation in computer-aided systems for clinical ultrasonography.

References

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

Andreadis, T., Gasteratos, A., Seimenis, I., & Koulouriotis, D. (2025). Application of CAD systems in breast cancer diagnosis using machine learning techniques: A systematic review. *Bioengineering, 12*(7), 456. https://doi.org/10.3390/bioengineering12111160

Chauhan, A., & Kumar, I. (2025). Deep feature extraction and optimized VGG16-SVM architecture for breast cancer characterization. *Discover Computing, 28*, 208. https://doi.org/10.1007/s10791-025-09736-6

Dan, Q., Xu, Z., Burrows, H., Bissram, J., Stringer, J. S. A., & Li, Y. (2024). Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: a systematic review. *npj Precision Oncology, 8*, 21. https://doi.org/10.1038/s41698-024-00514-z

Gao, S., Liu, J., Li, L., Yang, D., Miao, Y., Zhang, X., Han, Q., Shi, Y., Wu, J., & Zhang, K. (2026). Application of deep learning technology in breast cancer: A systematic review of segmentation, detection, and classification approaches. *Biomedical Engineering Online, 25*(19). https://doi.org/10.1186/s12938-025-01502-5

Ghalati, M. K., Nunes, A., Ferreira, H., Serranho, P., & Bernardes, R. (2022). Texture Analysis and Its Applications in Biomedical Imaging: A Survey. *IEEE Reviews in Biomedical Engineering, 15*, 222-246. https://doi.org/10.1109/RBME.2021.3115703

Gu, J., & Jiang, T. (2022). Ultrasound radiomics in personalized breast management: Current status and future prospects. *Front. Oncol*. 12:963612. https://doi.org/10.3389/fonc.2022.963612

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-3*(6), 610-621. https://doi.org/10.1109/TSMC.1973.4309314

Heywang-Köbrunner, S. H., Elsner, S. A., Haußmann, E., Hacker, A., Grieger, P., Hadwiger, M., Hertlein, M., & Katalinic, A. (2026). Supplemental breast ultrasound in mammography screening for women with critically dense breasts. *Cancers, 18*(10), 1631. https://doi.org/10.3390/cancers18101631

Jayakrishnan, K., & Katal, N. (2026). Radiomics-integrated machine learning framework for quantitative breast cancer diagnosis. *Frontiers in Artificial Intelligence, 9*, 1752413. https://doi.org/10.3389/frai.2026.1752413

Kamsari, M., Sadeghi, S., de Oliveira, G. G., Alves, A. M., Sarshar, N. T., Anari, S., & Ranjbarzadeh, R. (2025). The role of data augmentation and attention mechanisms in UNet and ConvNeXt architectures for optimizing breast tumor segmentation. *Scientific Reports, 15*, 45268. https://doi.org/10.1038/s41598-025-29841-2

Kim, J., Harper, A., McCormack, V., Sung, H., Houssami, N., Morgan, E., Mutebi, M., Garvey, G., Soerjomataram, I., & Fidler-Benaoudia, M. M. (2025). Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nat Med **31**,* 1154 – 1162 (2025). https://doi.org/10.1038/s41591-025-03502-3

Li, X., Wang, Y., Zhao, Y., y Wei, Y. (2022). Fast Speckle Noise Suppression Algorithm in Breast Ultrasound Image Using Three-Dimensional Deep Learning. *Frontiers in Physiology, 13*, 880966. https://doi.org/10.3389/fphys.2022.880966

Masoud, R. M., Bakir, R. M. A., Saraya, M. S., & Ayyad, S. M. (2025). BREAST-CAD: A computer-aided diagnosis system for breast cancer detection using machine learning. *Technologies, 13*(7), 234-245. https://doi.org/10.3390/technologies13070268

Mishra, A. K., Roy, P., Bandyopadhyay, S., & Das, S. K. (2021). Breast ultrasound tumor classification: A machine learning-radiomics-based approach. *Expert Systems, 38*(7), e12713. https://doi.org/10.1111/exsy.12713

Ogut, Z., Karaduman, M., & Yildirim, M. (2025). Clinically focused computer-aided diagnosis for breast cancer using SE and CBAM with multi-head attention. *Tomography, 11*(12), 138. https://doi.org/10.3390/tomography11120138

Pawłowska, A., Karwat, P., & Żołek, N. (2023). Letter to the Editor. Re: "[Dataset of breast ultrasound images by W. Al-Dhabyani, M. Gomaa, H. Khaled & A. Fahmy, Data in Brief, 2020, 28, 104863]". *Data in Brief, 48*, 109247. https://doi.org/10.1016/j.dib.2023.109247

Ru, J., Zhu, Z., & Shi, J. (2024). Spatial and geometric learning for classification of breast tumors from multi-center ultrasound images: A hybrid learning approach. *BMC Medical Imaging, 24*, 133. https://doi.org/10.1186/s12880-024-01307-3

Sabry, M., Balaha, H. M., Ali, K. M., Mahmoud, A., Gondim, D., Ghazal, M., Soliman, T. H. A., & El-Baz, A. (2026). AI-Driven Breast Cancer Diagnosis: A Systematic Review of Imaging Modalities, Deep Learning, and Explainability. *Cancers, 18(8), 1305*. https://doi.org/10.3390/cancers18081305

Unser, M. (1986). Sum and Difference Histograms for Texture Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8*(1), 118-125. https://doi.org/10.1109/TPAMI.1986.4767760

World Health Organization (WHO). (2026). Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Zhang, Q., Gao, J., Agyekum, E.A., Zhu, L., Jiang, C., Du, S., & Yin, L. (2025). Combined clinical-ultrasound radiomics model for BI-RADS 4 masses. *American Journal of Translational Research, 17*(8), 6370-6380. https://doi.org/10.62347/SBKU2090