

ORIGINAL ARTICLE

Enhancing Breast Cancer Segmentation in Mammography with UNet++ - Deep Learning Approach

Kimia Jalalian ^{1*} , Golnaz Hosseini ¹, Razieh Ghiasi ², Alireza Bosaghzadeh ³

¹ Department of Biomedical Engineering, Pooyesh Institute of Higher Education, Qom, Iran

² Department of Computer Engineering, Qom University, Qom, Iran

³ Department of Computer Engineering, Shahid Rajaei Teacher Training University, Tehran, Iran

*Corresponding Author: Kimia Jalalian

Received: 19 June 2024 / Accepted: 08 May 2025

Email: kimiajalalian33@gmail.com

Abstract

Purpose: Manually segmenting mammograms is time-consuming and subjective. Therefore, automatic segmentation of breast masses is necessary but poses significant challenges due to factors such as low signal-to-noise ratio, diverse mass shapes and sizes, varying contrast levels, and high false positive rates. To address these challenges, we have developed an automatic image segmentation method based on a comprehensive pre-processing pipeline.

Materials and Methods: Our proposed method consists of two phases: 1) the pre-processing phase, which includes denoising, contrast enhancement, image cropping, resizing, and augmentation of mammograms, and 2) the model design phase, where UNet++ is employed as an encoder-decoder-based network for segmenting breast masses. The encoder captures relevant information from various regions in the input image, while the decoder reconstructs the spatial location of the target region. We conducted extensive experiments on publicly available CBIS-DDSM and INbreast datasets to evaluate the performance of our proposed method. For a comprehensive assessment, we utilized evaluation metrics including Precision, True Positive Rate, Dice Score Coefficient, and Jaccard Index. Additionally, a confusion matrix was employed to evaluate segmentation accuracy, while violin plots depicted the distribution of results across different BI-RADS and ACR categories.

Results: Based on our findings, our proposed method demonstrates promising results with a precision rate of 92.33%, a True Positive Rate of 93.83%, a Dice Score Coefficient measuring 92.92%, and a Jaccard Index of 87.05% in the CBIS-DDSM dataset. Furthermore, to assess the generalizability of our proposed method, the INbreast dataset was used as an unseen test set. The results demonstrate a precision rate of 91.15%, a True positive rate of 91.15%, a Dice Score coefficient of 92.53%, and a Jaccard Index of 87.25%, indicating robust performance on data outside the training distribution.

Conclusion: The integration of UNet++ with a pre-processing pipeline in digital mammography has shown promising results in accurately segmenting breast masses. This method has the potential to significantly improve early breast cancer detection and reduce diagnostic errors in clinical practice while employing a relatively lightweight model.

Keywords: Breast Cancer; Semantic Segmentation; UNet++; Digital Mammography; Deep Learning.

1. Introduction

Breast cancer stands as a significant concern for women globally, with high rates of diagnosis and mortality [1]. Effective treatment and a reduction in mortality rates are associated with the early detection of breast cancer through the screening stage [2].

Mammography, MRI, CT, and PET are commonly used imaging techniques for breast cancer screening and diagnosis [3]. However, mammography is the primary screening method in clinical practice as it effectively detects suspicious lumps and abnormalities [4]. In mammography, the breast is compressed to reduce thickness, enhancing image clarity by minimizing tissue overlap and scatter radiation. Images are usually taken from multiple angles to ensure that all areas of the breast are examined. Mediolateral Oblique (MLO) and Cranio-Caudal (CC) are two common imaging views. Radiologists typically rely on both views to ensure accurate diagnosis and assessment. The MLO view captures the breast from the center of the chest outward, while the CC view is taken from above the breast. The CC view provides a comprehensive depiction of the entire breast. In contrast, the MLO view focuses on the upper-outer quadrant, offering the best visualization of the breast's lateral side, statistically the most common area for pathological changes.

Manual investigation of mammograms can lead to false diagnoses due to factors such as complex breast tissue composition, image noise, artifacts, variations in breast density, and the high volume of screening performed daily. Dense tissues can obscure tumors, particularly smaller ones, increasing the risk of false negatives in mammograms and reducing the early detection of cancer diagnosis. Recent research indicates that the sensitivity of mammography can decline to 62-68% in women with extremely dense breasts, considerably lower than in those with fatty breasts [5]. Therefore, there is a growing demand to automatically analyze mammograms, particularly for women with dense breast tissue, where manual interpretation often falls short [6]. It enhances the accuracy of mammogram examinations by precisely identifying the location and boundaries of tumors, thereby enabling effective treatment planning in clinical practice.

Therefore, Artificial Intelligence (AI)-based algorithms have been developed [7], [8], which demonstrated significant potential in accurately delineating suspicious tumor regions, providing valuable insights into diagnosis findings. One of the most effective AI techniques in this field is deep learning, a subset of machine learning that utilizes artificial neural networks with multiple hidden layers to learn complex patterns and representations from data. Deep learning algorithms, particularly those based on convolutional neural networks (CNNs), have achieved state-of-the-art performance in various image analysis tasks, including medical image segmentation [9]. Therefore, this has led to the development of computer-aided detection (CAD) systems that assist radiologists in interpreting mammograms [10], which effectively reduces false positives. Furthermore, these methods have the potential to significantly reduce patient examination time and alleviate the workload of radiologists [11].

In medical imaging, segmentation is used to identify and extract Regions of Interest (RoI) for further analysis. One prominent approach is the U-Net, which integrates high-level features from the decoder with low-level features from the encoder through skip connections [12]. Inspired by the success of U-Net, UNet++ was introduced to enhance performance by incorporating skip connections at different depths [13]. This model has demonstrated a high capacity for extracting deep and semantic features, leading to precise and detailed segmentation of masses. In recent years, many researchers have focused on breast tumor segmentation. This summary highlights some of the notable works in this domain.

Sun *et al.* [14] proposed an attention-guided dense-upsampling network (AUNet) for breast tumor segmentation in mammograms, achieving an average Dice Score Coefficient of 81.8% for CBIS-DDSM and 79.1% for INbreast. Rajalakshmi *et al.* [15] proposed the Deeply Supervised U-Net model (DS U-Net). Their method enhanced image contrast using the Contrast-Limited Adaptive Histogram Equalization (CLAHE) technique. It was evaluated on CBIS-DDSM and INbreast, which achieved a Dice Score Coefficient of 82.9% for CBIS-DDSM and 79% for INbreast.

Zeiser *et al.* [16] employed the U-Net model for mass segmentation in mammograms, achieving a

sensitivity of 92.32%, a specificity of 80.47%, and a Dice Score Coefficient of 79.39%. Baccouche *et al.* [17] introduced the Connected-UNet model, connecting two U-Nets with modified skip connections, which was evaluated on CBIS-DDSM and INbreast, and a private dataset. It has achieved Dice Score Coefficient of 89.52%, 95.28%, and 95.88% and Intersection over Union (IoU) of 80.02%, 91.03%, and 92.27%, respectively, on CBIS-DDSM, INbreast, and private datasets.

Min *et al.* [18] developed a Mask R-CNN model for mass segmentation. They applied Multi-scale Morphological Sifting (MMS) to convert grayscale mammograms into pseudo-color images. Then, they evaluated their model on the INbreast dataset, yielding an average Dice Score Coefficient of 0.88 for mass segmentation. Many existing approaches rely on increasingly complex and computationally intensive models to obtain accurate mass boundary information. In contrast, the primary goal of our study is to demonstrate that by leveraging advanced pre-processing techniques, it is possible to achieve high-quality segmentation of mammograms using a relatively lightweight segmentation model. The significant contributions of our study can be summarized as follows:

1. We underscore the importance of advanced pre-processing techniques in enhancing mammogram quality for segmentation.
2. We demonstrate that a lightweight segmentation model, when coupled with effective pre-

processing, can achieve competitive performance compared to more complex models.

3. We used image augmentation techniques to expand the dataset and mitigate overfitting.

4. We provide a comprehensive analysis comparing our model's performance against existing complex models on the CBIS-DDSM and INbreast datasets.

The subsequent sections of this paper are organized as follows: Section 2 introduces the proposed method for breast tumor segmentation. Section 3 outlines the experimental results, including detailed analysis and evaluation. Furthermore, section 4 presents a discussion of our method. Finally, Section 5 concludes this paper, outlining key insights and potential directions for future investigation.

2. Materials and Methods

In this section, we describe the datasets and our framework, which consists of two main phases, pre-processing and the mass segmentation model. Figure 1 shows the workflow of the proposed framework.

2.1. Dataset Description

We evaluated our proposed framework on the two publicly available datasets of INbreast [19] and CBIS-DDSM. The CBIS-DDSM (Curated Breast Imaging Subset of DDSM) dataset [20] is an enhanced version of the DDSM (Digital Database for Screening

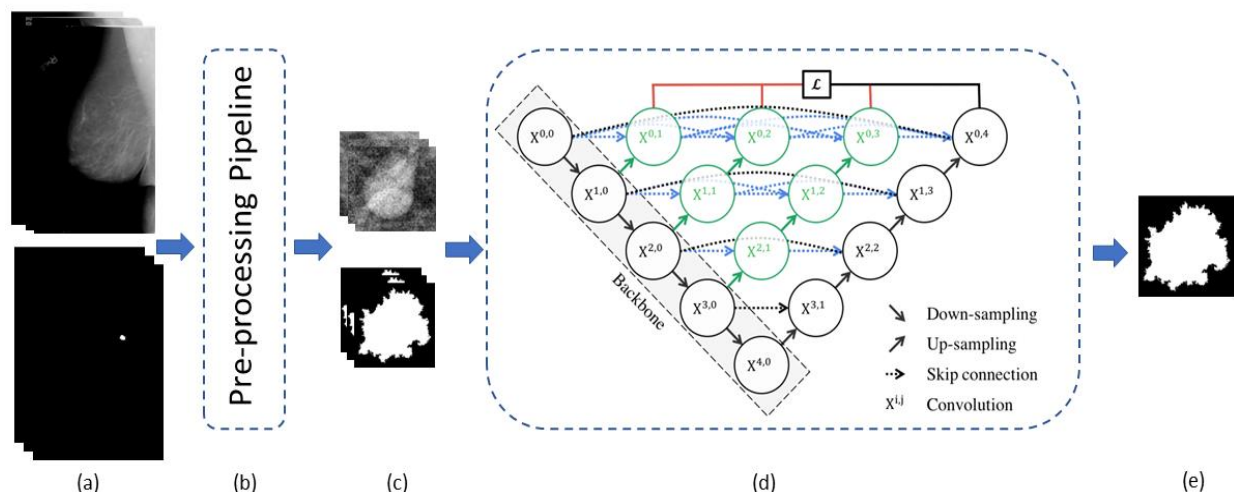


Figure 1. The overview of the proposed system. a) raw mammogram and ground truth, b) pre-processing pipeline, c) pre-processed mammogram and corresponding ground truth, d) UNet++ architecture 12, e) predicted mask

Mammography) dataset, including full mammograms, cropped images, and masks for masses and calcifications in DICOM format. In this paper, we utilized 1577 mammograms to implement our proposed framework, while calcifications were left for future investigation. The samples were divided into three subsets: 80% for training, 10% for validation, and 10% for testing, as outlined in Table 1. This dataset encompasses detailed metadata, including ACR (American College of Radiology) breast density, BI-RADS (Breast Imaging Reporting and Data System) categories, and pathology findings that facilitate validating the proposed algorithm. The BI-RADS is a standardized system for classifying report findings in breast imaging to enhance clinical management decisions. It assigns a category from 0 to 6 to describe the findings, where higher values indicate a greater probability of malignancy. Figure 2a shows the distribution of BI-RADS categories in the CBIS-DDSM dataset. Moreover, breast density is classified into four different groups according to the American College of Radiology (ACR) standards: entirely fatty (ACR A), scattered fibroglandular (ACR B), heterogeneously dense (ACR C), and extremely dense (ACR D). Figure 2b depicts the prevalence of ACR categories in the CBIS-DDSM dataset. As represented in Figure 3a and 3b, the CBIS-DDSM dataset is obtained from two different views: the Mediolateral Oblique (MLO) view and the Cranial-

Caudal (CC) view. The MLO view captures the breast from the center of the chest outward, while the CC view is taken from above the breast. The CC view provides a comprehensive depiction of the entire breast. In contrast, the MLO view focuses on the upper-outer quadrant, offering the best visualization of the breast's lateral side, statistically the most common area for pathological changes.

The INbreast dataset is a publicly available repository of Full-Field Digital Mammography (FFDM) images, collected from the Breast Centre at Centro Hospitalar de S. João in Porto, Portugal. This dataset comprises 410 full-resolution mammograms from 115 patients, including 90 cases from women with both breasts affected (four images per case) and 25 cases from mastectomy patients (two images per case). It covers four types of breast abnormalities, including mass, calcification, symmetry, and distortions. The images are provided in DICOM format, with corresponding ground truth annotation available in XML format. For this study, a subset of 107 mammogram-containing masses was selected to assess the proposed framework. The dataset includes both MLO and CC views, and samples of these views are represented in Figures 3c and 3d. Furthermore, the distribution of BI-RADS and ACR categories within the dataset is presented in Figure 4a and 4b, providing valuable insights into its clinical diversity.

Table 1. The number of samples before and after lesion-based RoI extraction and augmentation in the CBIS-DDSM dataset

Raw mammograms	ROIs	Training ROIs (80%)	Validation ROIs (10%)	Testing ROIs (10%)	Augmented Training ROIs
1577	1681	1360	152	169	4080

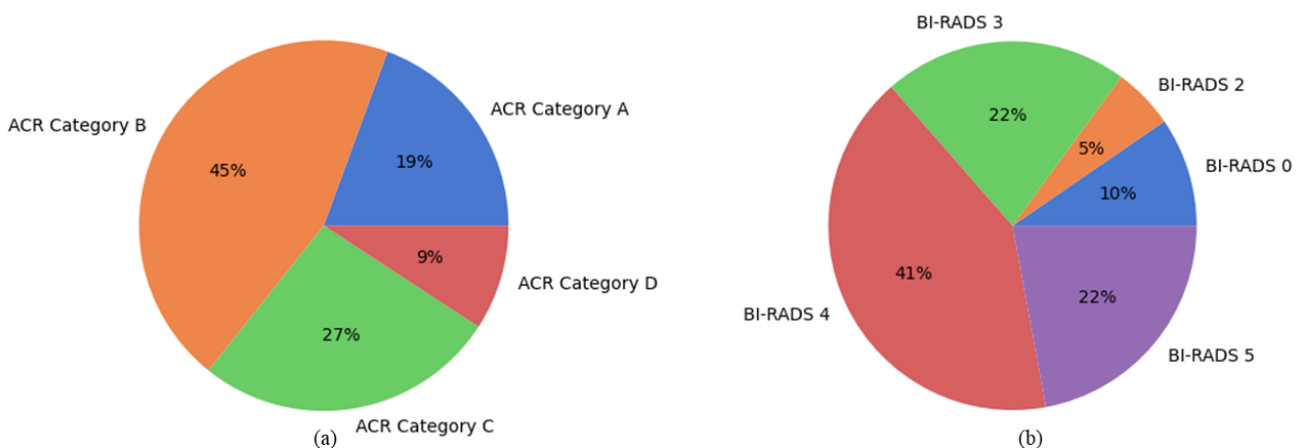


Figure 2. The distribution of a) ACR and b) BI-RADS categories in the CBIS-DDSM dataset

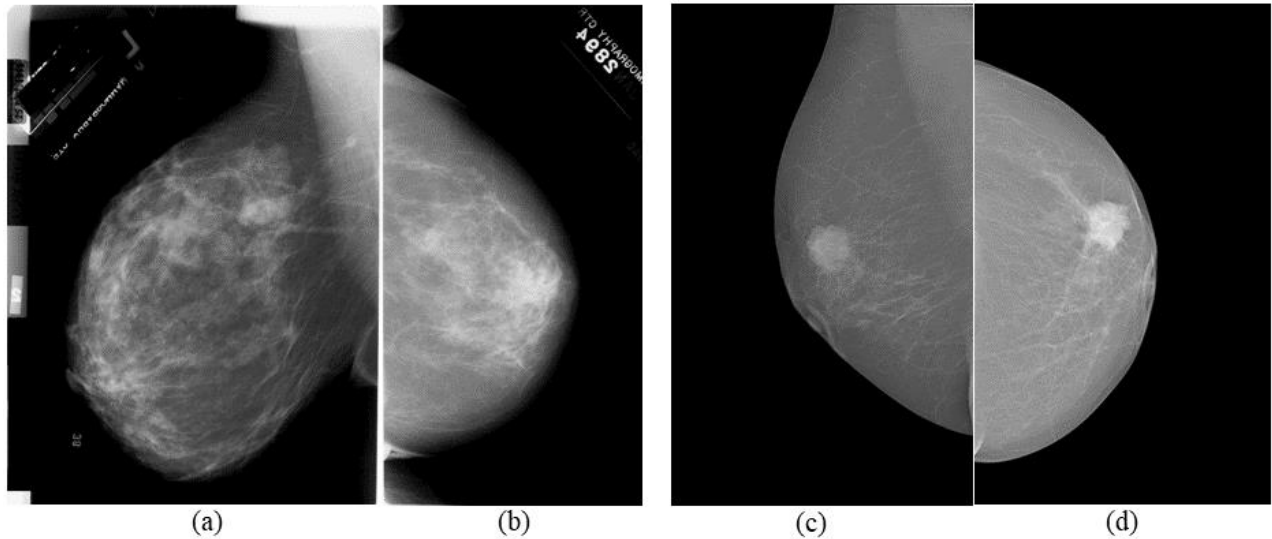


Figure 3. MLO and CC views of mammograms from the CBIS-DDSM dataset (a, b) and INbreast dataset (c, d)

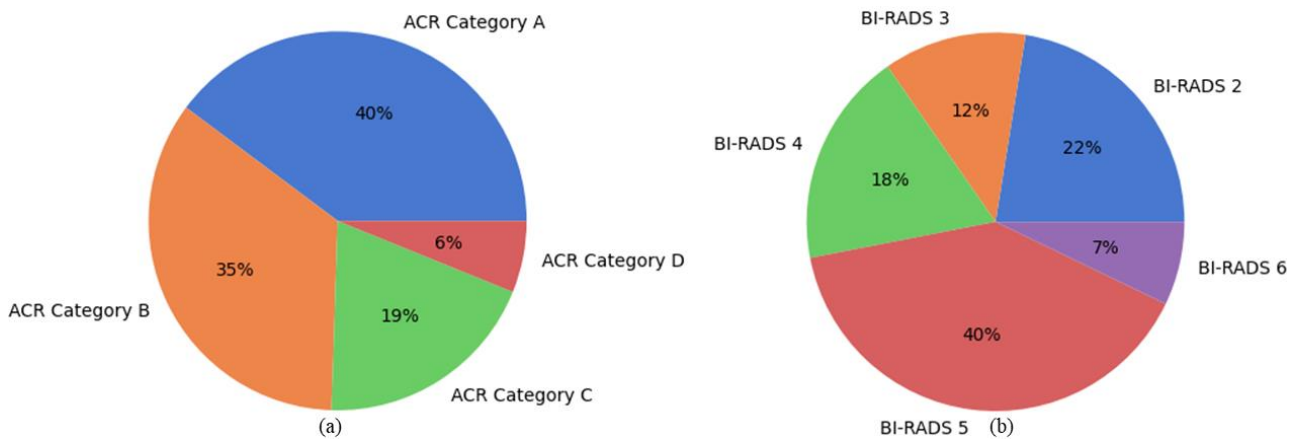


Figure 4. The distribution of a) ACR and b) BI-RADS categories in the INbreast dataset

2.2. Pre-processing Pipeline

The pre-processing phase plays a vital role in achieving accurate segmentation of mammography images, primarily due to challenges such as low signal-to-noise ratio (SNR) and low contrast [21]. To address this challenge, our proposed method incorporates a range of essential pre-processing steps to prepare the images for segmentation. The pre-processing steps employed in this study align with our previous paper [21], as depicted in Figure 5.

2.2.1. Denoising Techniques

Various types of noise, such as salt and pepper, speckle, Gaussian, and Poisson, can significantly degrade image quality, making it difficult to analyze

and interpret mammograms accurately [22]. To mitigate this issue, a range of denoising approaches has been developed. In this study, we employed the Wiener filter [23], [24], a spatial domain linear filter, to optimize image quality by reducing noise while preserving the essential image details and structures. Figure 5b shows the result of the Wiener filter on the mammogram. To evaluate the effectiveness of the Wiener filter, it was compared with other denoising methods, such as Gaussian denoising, median filtering, and the Non-Local Means (NLMeans) filter. The Gaussian denoising assumes the noise is Gaussian-distributed and applies simple filtering to reduce its effect on the image. The median filter replaces each pixel with the median value from its neighborhood, preserving edges better than mean filters. NLMeans [25] utilizes a non-local averaging

method that assesses the similarity between pixels over a larger area, effectively reducing noise while maintaining the structural integrity of the image. The comparative analysis of these denoising techniques reveals that Wiener excels in terms of performance metrics, including the Peak signal-to-noise ratio (PSNR), and Mean Square Error (MSE) [26]. Five random samples were selected, and denoising techniques were employed. As detailed in Table 2, the Wiener filter exhibits higher PSNR values and lower MSE compared to other approaches in all images, highlighting its effectiveness in removing noise while preserving detailed information in mammograms.

2.2.2. CLAHE Technique

Enhancing image quality is essential for revealing subtle features within images. The Contrast-Limited Adaptive Histogram Equalization (CLAHE) approach [27] is a powerful technique to improve image contrast, particularly for highlighting features such as tumors and other abnormalities [23], [24]. Unlike traditional histogram equalization, CLAHE operates on smaller image sections, referred to as tiles, enabling localized contrast enhancement. Bilinear interpolation

is employed to prevent the appearance of false boundaries between adjacent tiles. Figure 6 illustrates the effectiveness of the localized approach in CLAHE compared to histogram equalization, which processes the entire image and can lead to undesirable effects such as over-brightening and saturation. To further assess the efficacy of CLAHE in comparison to histogram equalization, the Contrast-to-Noise Ratio (CNR) is calculated, presenting the clarity of an image by comparing the contrast of the region of interest to the background noise. Five random samples were selected and enhanced using CLAHE and traditional histogram equalization. As presented in Table 3, the CNR of images enhanced by CLAHE was superior to images processed with traditional histogram equalization. These findings, along with Figure 5c, highlight the effectiveness of CLAHE in improving image quality for detecting breast abnormalities.

2.2.3. ROI Extraction

In this study, it is assumed that the positions of masses in the mammograms have been identified before the segmentation process. The primary objective is to refine the segmentation technique to

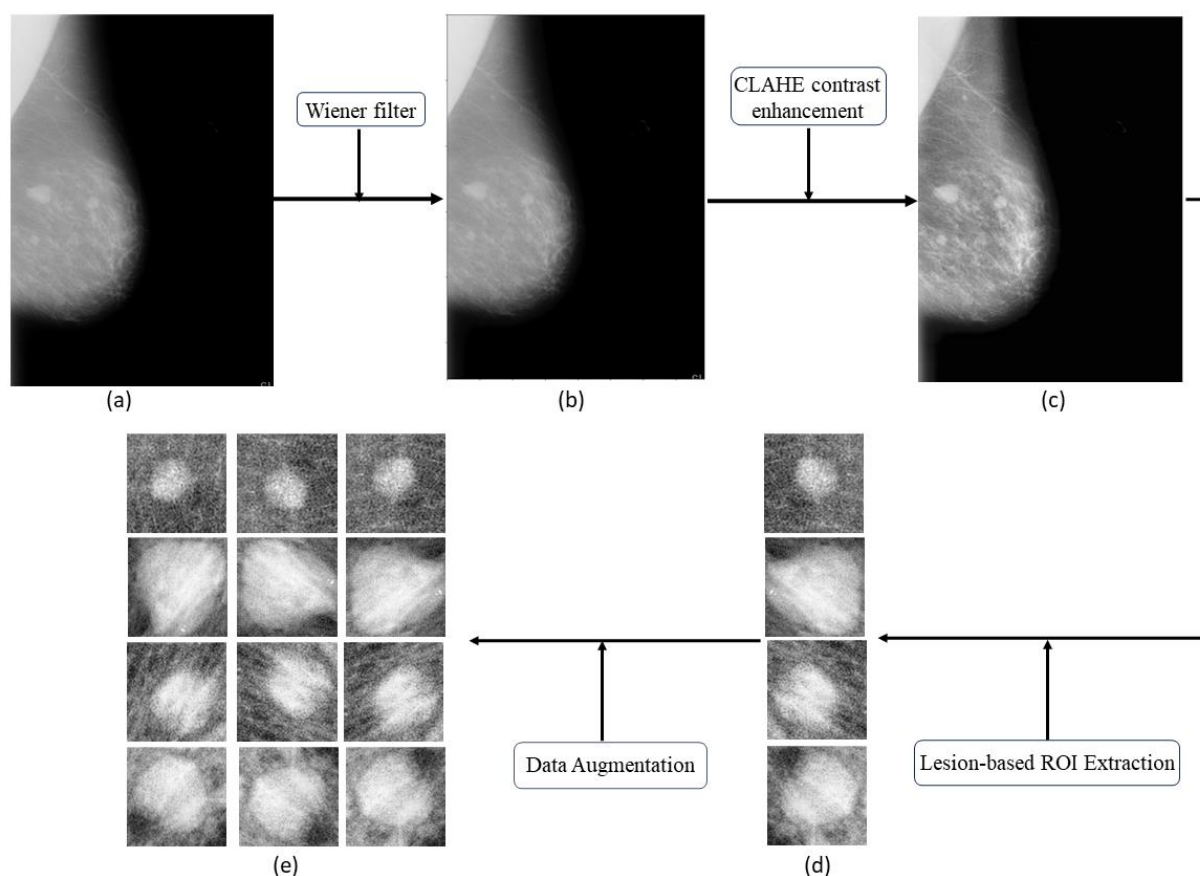


Figure 3. The pre-processing pipeline for mammogram inputs in the CBIS-DDSM dataset

Table 3. A comparison of various denoising methods by calculating PSNR and MSE using 5 mammograms in the CBIS-DDSM dataset

Image	PSNR				MSE			
Denoising Method	Gaussian	Median Filter	NL-Means	Wiener	Gaussian	Median Filter	NL-Means	Wiener
P_00209	36.11	44.00	42.84	44.82	15.9	2.58	3.37	2.14
P_01423	42.14	48.91	46.21	49.18	3.96	0.83	1.55	0.785
P_01151	37.05	43.98	43.43	45.01	12.80	2.59	2.94	2.04
P_00947	38.92	43.36	42.05	43.57	8.32	2.99	4.05	2.85
P_00160	35.84	41.26	40.33	41.55	16.93	4.86	6.02	4.54

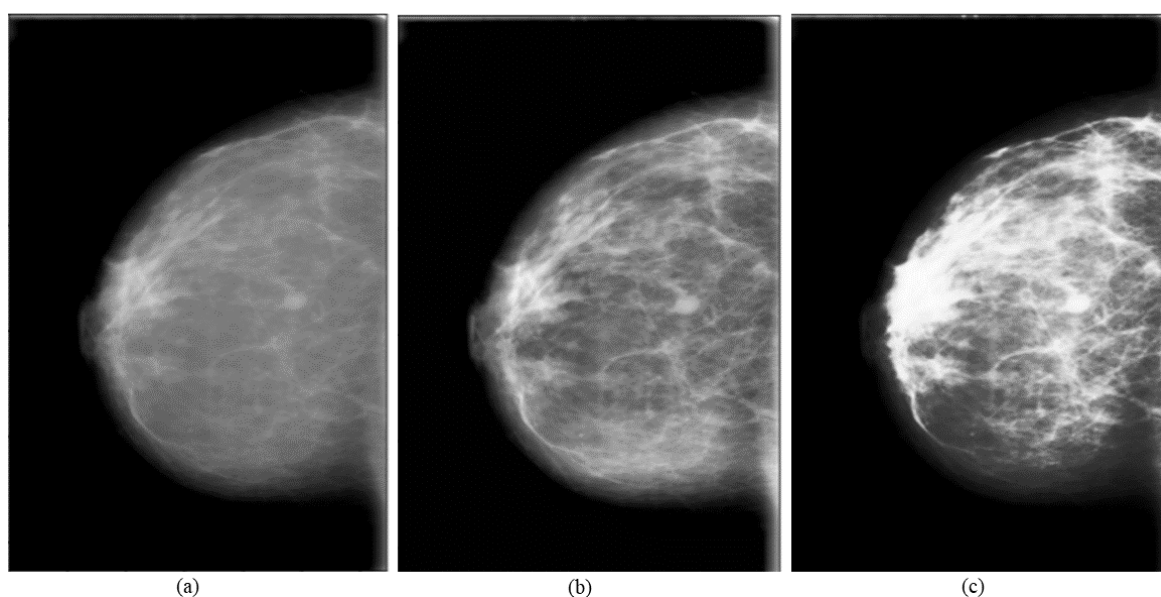
**Figure 4.** A comparison of contrast enhancement techniques on a sample from the CBIS-DDSM dataset (P_01423), a) input image, b) the input mammogram enhanced using CLAHE, c) Histogram Equalization**Table 2.** Performance comparison of two contrast enhancement techniques using the CNR metric in the CBIS-DDSM dataset

Image	Contrast to Noise Ratio (CNR)	
	CLAHE	Histogram Equalization
P_00209	1.04	0.96
P_01423	0.959	0.801
P_01151	0.826	0.695
P_00947	1.21	0.991
P_00160	1.07	0.90

isolate and analyze lesions accurately. Therefore, the entire mammogram is cropped around the lesion to ensure that subsequent analyses are focused solely on the region of interest (ROI). The number of ROI samples after this extraction step is presented in Table 1. In cases involving multiple tumors, this method is applied to each tumor individually, thereby increasing the number of samples. Figure 5d shows the cropped ROIs for each lesion.

2.2.4. Resize and Normalization

In the subsequent processing steps, the Regions of Interest (ROIs) undergo resizing and normalization to standardize their dimensions and pixel values for consistent analysis. ROIs are resized to a spatial resolution of 224×224 pixels. Min-max normalization is then applied to rescale the pixel intensities, standardizing the range of pixel values.

2.2.5. Data Augmentation

To improve the robustness and generalization ability of the model, data augmentation plays a crucial role, especially when working with limited or imbalanced datasets [28]. By generating additional data variations, the model becomes less prone to overfitting and can better generalize to unseen data. In this paper, we employed common data augmentation methods, including 90-degree image rotation and random flips, as shown in Figure 5e. These techniques preserve the original content while introducing diverse perspectives, enriching the training process. The number of samples after data augmentation is presented in Table 1. Notably, only the training set undergoes augmentation; the validation and test sets remain unchanged to ensure reliable evaluation, enhance model generalization, and prevent data leakage.

2.3. Semantic Segmentation

Semantic segmentation is a fundamental computer vision technique employed to partition an image into multiple meaningful segments, with each segment corresponding to a particular object or region in the image [29]. In this technique, every pixel in an image is classified into a specific category or class. The prediction is a pixel-level map where each pixel is labeled with one of the predefined classes, representing the objects or Regions of Interest (RoI). Semantic segmentation models are relatively lightweight compared to more complex approaches, such as panoptic and instance segmentation. These models focus on pixel-level classification without the added complexity of distinguishing individual object instances, which leads to fewer parameters and quicker inference times. For instance, UNet++ typically has around 9 million parameters, significantly lower than many instance or panoptic segmentation models that often exceed 30 million parameters (e.g., Mask R-CNN with ~44M parameters or Panoptic FPN with ~50M parameters). The reduced computational cost of semantic segmentation allows for practical deployment in clinical environments with limited hardware resources [30].

In this paper, we used the UNet++ model architecture [13] for breast mass segmentation. The UNet++ [13] network, an enhanced version of the

original U-Net [12], has consistently achieved superior results due to a novel skip connection and deep supervision. UNet++ employs a deeper network architecture and utilizes convolutional layers or additional blocks to enhance the model's learning capacity. Similar to U-Net, this model architecture consists of two main components: an encoder and a decoder, where the encoder and decoder sub-networks are connected through a series of nested, dense skip pathways, allowing for the efficient extraction of features from both high-level and low-level image details. The encoder stacks convolutional layers that consistently down-sample the image to extract information from it, while the decoder reconstructs the image features using the process of deconvolution. The redesigned skip pathways aim to reduce the semantic gap between the feature maps of the encoder and decoder subnetworks. Moreover, UNet++ incorporates mechanisms to capture multi-scale context information, enabling the model to understand objects and structures at different scales. To enhance the ability to handle objects of various sizes, dilated convolutions are used without increasing the number of parameters.

For additional exploration, we also employed the U-Net, Pyramid Scene Parsing Network (PSPNet) [31], and LinkNet [32] for breast mass segmentation. The PSPNet model incorporates a pyramid pooling module, aggregating contextual information from multiple regions of an image at different scales. This approach enhances the model's ability to discern complex structures and boundaries by capturing global and local dependencies. The LinkNet model, known for its computational efficiency, follows an encoder-decoder framework while preserving spatial information through skip connections. These connections enable the model to retain fine-grained details during the down-sampling process, thereby mitigating the loss of crucial anatomical features.

3. Results

In this section, we conduct extensive experiments to highlight the effectiveness of the suggested method in mass mammogram segmentation. First, we introduce essential evaluation criteria for assessing the semantic segmentation model's performance. The evaluation settings are explained in the following section. The

experimental results obtained from the proposed model and its analysis are described at the end of this section.

3.1. Evaluation Metrics

To evaluate the performance of the semantic segmentation model, we utilized various evaluation metrics, including the Jaccard Index, precision, True Positive Rate (TPR), Dice Score Coefficient [33], and Confusion Matrix. The Jaccard Index, also referred to as Intersection over Union (IoU), evaluates the similarity between the predicted mask and the ground truth. It computes the ratio of the intersection of the predicted and ground truth regions to the union of these regions as detailed as follows (Equation 1):

$$\text{Jaccard Index} = \frac{TP}{TP + FP + FN} \quad (1)$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. A Jaccard Index value of 1 indicates a perfect overlap between the predicted mask and ground truth, while 0 indicates no overlap.

The precision represents the model's capability to accurately identify positive pixels (correctly segmented pixels) among all pixels predicted as positive. A high precision value indicates effective minimization of false positives, accurately distinguishing positive from negative pixels (Equation 2).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

True Positive Rate (TPR), also known as sensitivity or recall, is defined as the ratio of true positives to the sum of true positives and false negatives, as depicted in Equation 3. It indicates the model's accuracy in identifying positive pixels.

$$TPR = \frac{TP}{TP + FN} \quad (3)$$

The Dice Score Coefficient (DSC), also known as the F1-score, is a harmonic mean of precision and recall. This metric proves especially advantageous in scenarios involving imbalanced positive and negative pixels within an image (Equation 4):

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (4)$$

In addition, the Confusion Matrix is a visual metric for evaluating the performance of semantic segmentation models. In this matrix, the rows depict the actual labels, whereas the columns represent the predicted labels. The values within each cell indicate the number of pixels that were correctly or incorrectly classified.

3.2. Experiment Configuration

In this section, we outline the experimental configuration used to evaluate the proposed method. The Adam optimizer and Dice loss function [34] were employed to enhance model training. Additionally, we tested five different batch sizes and learning rates to evaluate their effects on the loss function and the dice score during the training and validation stages. As summarized in Table 4, a batch size of 16 and a learning rate of $1e^{-3}$ yielded the highest dice scores and

Table 4. Tuning hyperparameters by changing different learning rates and batch size values using the CBIS-DDSM dataset

Hyperparameter	Value	Train Loss	Valid loss	Train DSC	Valid DSC
Learning Rate	1e-2	0.081	0.086	91.8%	91.54%
	1e-3	0.068	0.080	93.15%	92.07%
	1e-4	0.070	0.083	92.98%	91.85
	5e-2	0.093	0.101	90.65%	90.06%
	5e-3	0.456	0.462	40.54%	39.62%
Batch Size	4	0.072	0.081	92.78%	91.96%
	8	0.069	0.077	93.09%	92.34%
	16	0.065	0.079	93.50%	92.19%
	32	0.064	0.080	93.49%	92.08%
	64	0.064	0.081	93.64%	92.01%

lowest loss value across training and validation. Therefore, we selected these parameters for our final model configuration. Moreover, the model was trained for 100 epochs on a system equipped with an A100 GPU, utilizing the Python 3.7 and the PyTorch framework.

3.3. Results

3.3.1. Quantitative Evaluation of Models

In this section, we provide a comprehensive comparison between the UNet++ model and several baseline models for breast mass segmentation on the CBIS-DDSM dataset, using performance metrics such as TPR, Dice similarity coefficient (DSC), and the Jaccard index (Table 5). Furthermore, we assessed the efficacy of our method using the INbreast dataset as a testing cohort. It is important to note that the model was exclusively trained on the CBIS-DDSM dataset, and the INbreast dataset was reserved for evaluation purposes. The results are detailed in Table 6.

Table 5. The results of our proposed method on the INbreast dataset

Dataset	Precision	TPR	DSC	Jaccard Index
INbreast	92.83%	88.33%	89.92%	82.00%

3.3.2. Effect of Pre-processing on Model Performance

To investigate the impact of pre-processing techniques on model performance, we designed an experiment with four distinct groups: One group involved no Wiener and CLAHE techniques. Another group focused on using CLAHE, while a third group utilized only the Wiener filter. The final group incorporated both the CLAHE technique and the Wiener filter. The results of this experiment are

detailed in Table 7, highlighting how each pre-processing approach affects the model's performance.

3.3.3. Cross-Fold Validation

To assess the robustness and generalization capability of the proposed method, we performed cross-fold validation on the CBIS-DDSM dataset. The outcomes of this validation process, presented in Table 8, demonstrate the consistency of the model's performance across different folds.

3.3.4. Confusion Matrix Analysis

Figures 7 and 8 depict the normalized confusion matrices for the CBIS-DDSM and INbreast datasets. These matrices provide insights into the model's ability to correctly identify non-mass regions (True Negatives) and mass regions (True Positives) while maintaining low rates of False Positives and False Negatives.

3.3.5. Performance Based on Metadata Categories

We further evaluated the UNet++ model's performance by analyzing metadata categories, including BI-RADS assessment, and ACR density, which are crucial for clinical applications. To visualize the distribution of segmentation performance across these categories, we employed Violin Plots [35], combining features of box and density plots. In these plots, the X-axis represents clinical categories, including BI-RADS and ACR, while the Y-axis corresponds to the segmentation Dice Score Coefficient. For each violin, the dots along the line represent different samples in the testing set. A wider section of a violin plot indicates a higher concentration of data points, implying that more instances fall within that range of values. The median line marks the central tendency of the model's performance, while outliers

Table 6. The performance comparison between different model networks using the CBIS-DDSM dataset

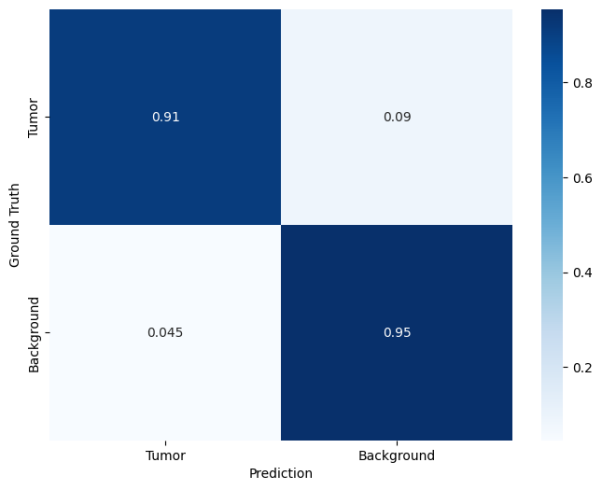
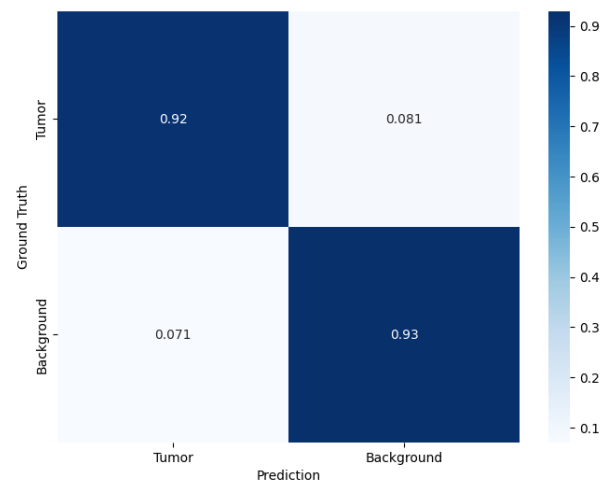
Model	Precision	TPR	Dice Score Coefficient (DSC)	Jaccard Index
UNet++	92.33% ± 0.5	93.83% ± 0.51	92.92% ± 0.03	87.05% ± 0.049
U-Net	92.09% ± 0.51	93.67% ± 0.15	92.71% ± 0.25	86.67% ± 0.45
PSPNet	92.29% ± 0.22	93.68% ± 0.24	92.83% ± 0.10	86.88% ± 0.17
LinkNet	92.29% ± 0.37	93.68% ± 0.49	92.83% ± 0.09	86.87% ± 0.15

Table 7. Results of 5-fold cross-validation in the CBIS-DDSM dataset

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Dice score	92.79%	94.89%	94.92%	94.44%	94.76%	94.36%
Jaccard index	86.63%	90.23%	90.27%	89.42%	90.03%	89.32%
precision	92.66%	94.18%	93.93%	93.39%	94.21%	93.67%
recall	92.91%	95.60%	95.93%	95.53%	95.31%	95.06%

Table 8. The comparison results between the input images with pre-processing and without pre-processing using the CBIS-DDSM dataset

Category	Precision	Recall	DSC	Jaccard Index
w/o CLAHE and Wiener	91.71% $\pm$ 0.29	93.46% $\pm$ 0.25	92.4% $\pm$ 0.05	86.23% $\pm$ 0.09
Wiener	91.86% $\pm$ 0.36	93.62% $\pm$ 0.38	92.58% $\pm$ 0.1	86.49% $\pm$ 0.17
CLAHE	92.11% $\pm$ 0.36	93.09% $\pm$ 0.46	92.43% $\pm$ 0.05	86.27% $\pm$ 0.08
Full pre-processing	92.33% $\pm$ 0.5	93.83% $\pm$ 0.51	92.92% $\pm$ 0.03	87.05% $\pm$ 0.04

**Figure 5.** Confusion matrix of the proposed framework on the INbreast dataset**Figure 6.** Confusion matrix of the proposed framework on the test set of the CBIS-DDSM dataset

appear as dots outside the main distribution. Figures 9 and 10 illustrate the Dice score coefficient results across different BI-RADS and ACR categories in CBIS-DDSM and INbreast datasets, respectively.

3.3.6. Qualitative and Visual Analysis of Segmentation Metrics

Figure 11 showcases qualitative examples of segmentation masks generated by the UNet++ model, compared with baseline models, such as U-Net, LinkNet, and PSPNet, using the CBIS-DDSM dataset.

Moreover, Figure 12 demonstrates the prediction results on the INbreast dataset, which was used as an unseen dataset. This visual comparison highlights the differences in segmentation quality among the models. To present a more effective evaluation, the corresponding DSC value is included for each

prediction. Four samples with the lowest performance were selected to further evaluate the limitations of our proposed method, as depicted in Figure 13.

3.3.7. Convergence of the Training Process

The convergence of a deep learning model is typically determined by monitoring changes in loss and accuracy throughout the training process. Figures 14 and 15 show the curves of the training and validation loss and accuracy of the UNet++ as the number of iterations increases during the training process.

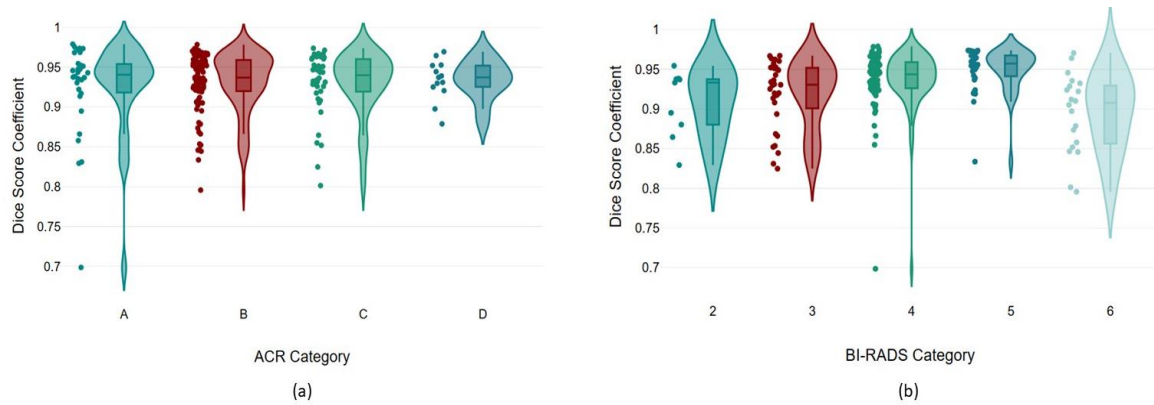


Figure 9. Result analysis in the CBIS-DDSM dataset based on (a) ACR, and (b) BI-RADS

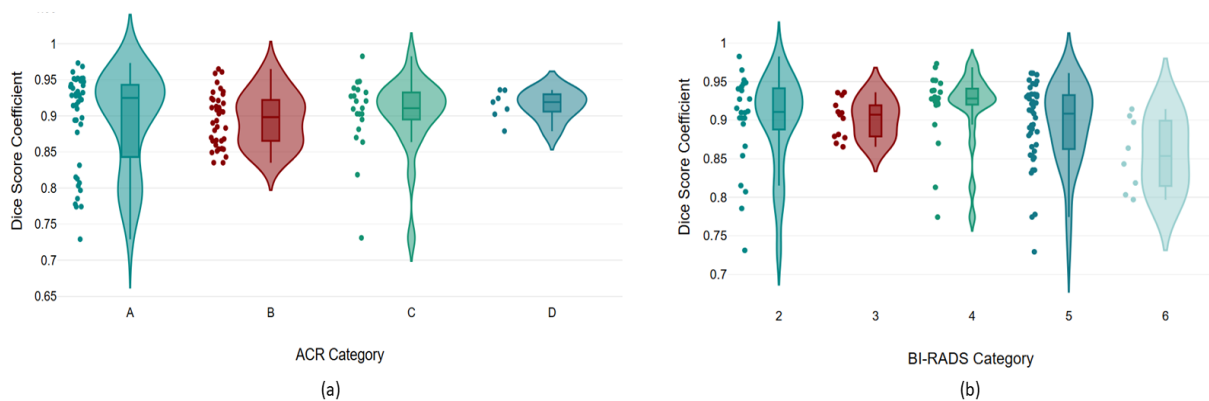


Figure 7. Result analysis in the INbreast dataset based on (a) ACR, and (b) BI-RADS

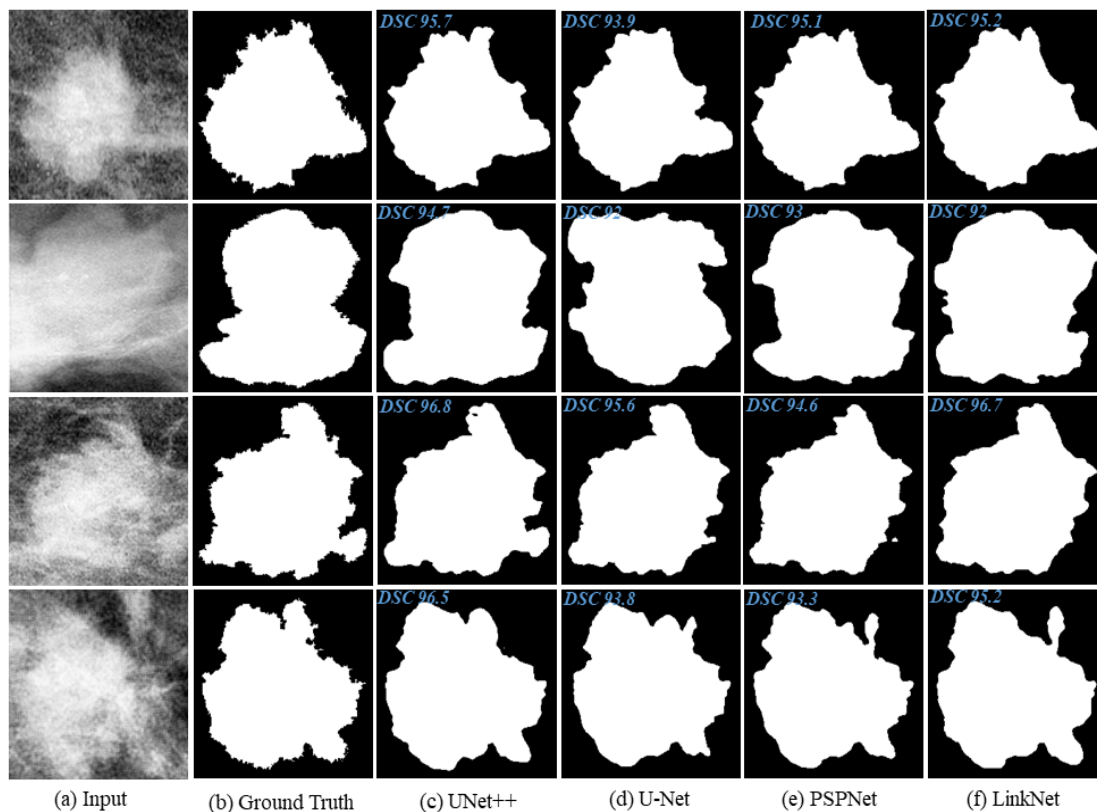


Figure 11. Qualitative predictions of 4 mammogram samples from the CBIS-DDSM dataset. a) pre-processed mammography images, b) ground truths, the outputs of c) UNet++, d) U-Net, e) PSPNet, f) LinkNet

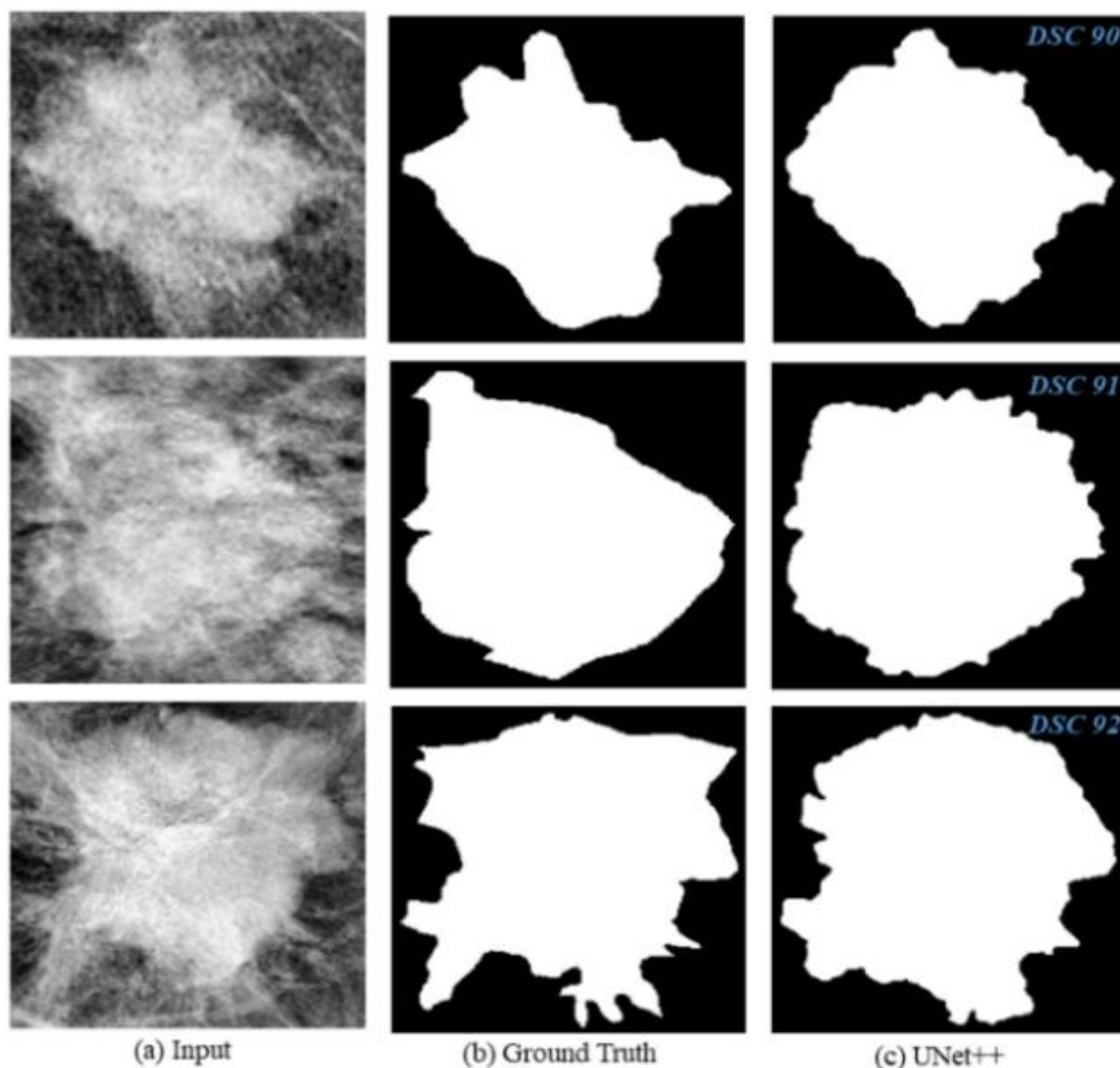


Figure 8. Qualitative predictions of 3 mammogram samples from the INbreast dataset. a) pre-processed mammography images, b) ground truths, c) the outputs of UNet++

4. Discussion

The comparative analysis presented in [Table 5](#) highlights the superior performance of the UNet++ model compared to baseline models, such as U-Net, LinkNet, and PSPNet, in breast mass segmentation tasks on the CBIS-DDSM dataset. This superiority is evident across all evaluation metrics, including precision, TPR, DSC, and Jaccard index. Furthermore, testing on the INbreast dataset shows that the model maintains consistent performance despite being trained on a different dataset, as outlined in [Table 6](#). This result highlights the model's generalization capability, an essential factor for real-world clinical applications where models must handle diverse

imaging conditions.

The experimental results summarized in [Table 7](#) demonstrate the impact of various pre-processing techniques on the performance of the UNet++ model. The combination of CLAHE and Wiener filtering yielded the highest improvement in segmentation metrics, while the absence of pre-processing resulted in comparatively lower performance. To ensure the reliability and generalization capability of our framework, we utilized cross-fold validation across datasets. The results in [Table 8](#) indicate a consistent performance of the UNet++ model across all folds. The normalized confusion matrices illustrated in

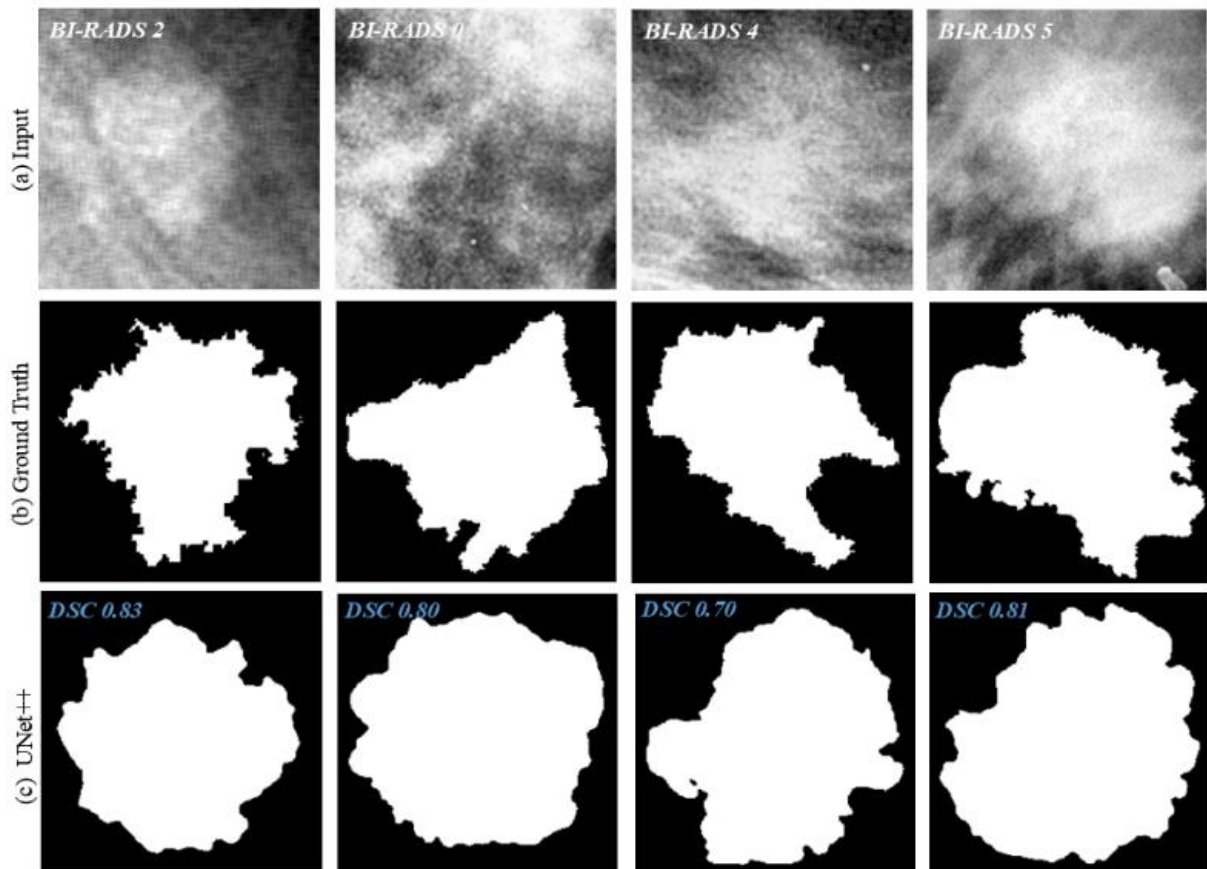


Figure 10. Qualitative predictions of the 4 samples with the lowest performance from the CBIS-DDSM dataset. a) Pre-processed mammography images, b) ground truths, c) the outputs of the UNet++

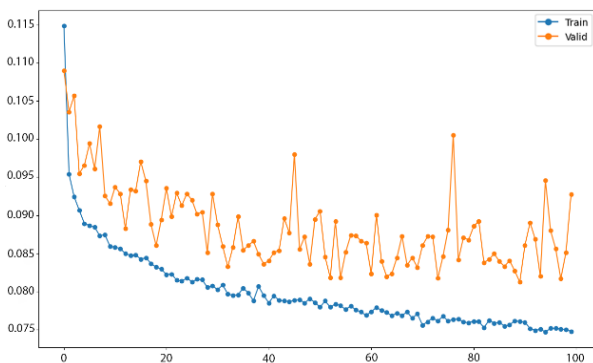


Figure 9. The training and validation loss curves of our proposed method.

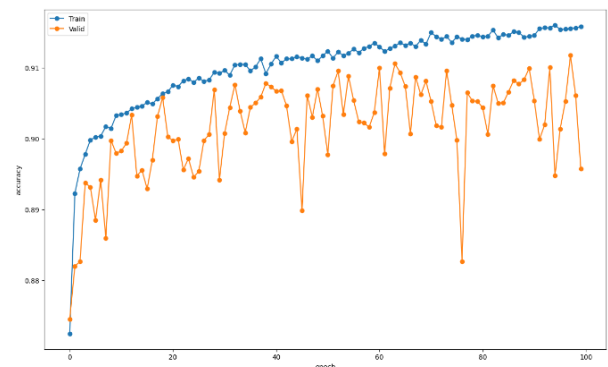


Figure 15. The training and validation accuracy curves of our proposed method

Figures 7 and 8 demonstrate that the model is reliable and effectively distinguishes between mass and non-mass regions. This reliable performance is vital for clinical applications to minimize diagnostic errors, improve patient outcomes, and enhance trust in AI-assisted diagnostic tools. Our analysis, depicted in Figure 9b and Figure 10b, demonstrates that the proposed method maintains acceptable DSC values

across different BI-RADS categories. Accurate segmentation in higher BI-RADS categories, which are associated with a greater likelihood of malignancy, can aid radiologists in making informed decisions regarding biopsy recommendations and treatment planning. Additionally, breast density is a significant factor influencing the risk of breast cancer and the sensitivity of mammographic detection. Figure 9a and

Figure 10a show that ACR category 2 demonstrates the highest variability in segmentation accuracy across both datasets, whereas ACR categories 3 and 4 exhibit more stable and reliable performance. Although CBIS-DDSM and INbreast show similar performance trends, slight differences are observed due to dataset-specific characteristics.

Figure 11 presents qualitative comparisons of segmentation outputs from UNet++ and other baseline models. The visual analysis reveals that UNet++ produces segmentation masks that closely match the ground truth, achieving a high DSC value. Figure 13 presents four representative samples exhibiting low segmentation performance (DSC 0.70-0.83), as identified by the violin plot in Figure 9. The first row (P_106_LEFT_MLO) corresponds to a benign lesion (BI-RADS 2) with a circumscribed margin and clear boundaries, as confirmed by pathological assessment [36]. However, the ground truth annotation inaccurately depicts the lesion with a spiculated and irregular shape. The second row (P_1834_Right_CC) involves an oval mass with an obscured margin and an incomplete pathological examination (BI-RADS 0), suggesting diagnostic uncertainty. The images in the third (P_432_LEFT_MLO) and fourth (P_1394_LEFT_MLO) rows, categorized as BI-RADS 4 and 5, respectively, represent a spiculated mass with irregular margins. Spiculated lesions present significant challenges in mammographic

analysis due to their complex morphology and interaction with surrounding breast tissue [37].

The training and validation loss and accuracy curves shown in Figures 14 and 15 indicate that the UNet++ model is learning effectively from the data, achieving a balance between bias and variance. The absence of overfitting implies that the model captures the underlying data distribution well, generalizing effectively to unseen data.

To further validate the effectiveness of our method, Table 9 compares the proposed method against several state-of-the-art segmentation methods reported in recent literature. The results demonstrate that our method consistently outperforms existing methods across all key evaluation metrics, including DSC, precision, and the Jaccard index.

5. Conclusion

In this study, we have introduced a breast mass segmentation method that utilizes an optimized pre-processing pipeline and the UNet++ deep learning method. Although more complex models may achieve marginally better results, our method offers a balance between high segmentation accuracy and lower computational cost. This balance makes it more suitable for real-world clinical applications, where efficiency and scalability are essential. Extensive experiments on the CBIS-DDSM and INbreast

Table 9. Performance comparison of the proposed system and state-of-the-art methods on the CBIS-DDSM dataset

Reference	Pre-processing	Model	DSC	Jaccard Index	Recall/Sensitivity
Sun <i>et al.</i> [14]	Removing artifacts, normalization	Attention U-Net	81.8%	-	-
Rajalakshmi <i>et al.</i> [15]	CLAHE	Deeply supervised U-Net (DS U-Net)	82.9%	-	-
Baccouche <i>et al.</i> [17]	ROI extraction, denoising, CLAHE, normalization, augmentation	Connected-UNet	89.52%	80.02%	-
Tsochatzidis <i>et al.</i> [38]	ROI extraction, augmentation	UNet+	72.2%	56.5%	-
Su <i>et al.</i> [39]	Removing artifacts, CLAHE	YOLO-LOGO	74.5%	64%	-
El-Banby [40]	Removing artifacts, CLAHE, normalization, and augmentation	U-Net	87.98%	-	90.58%
Ours	Wiener denoising, CLAHE, removing artifacts, normalization, ROI extraction, augmentation	UNet++	92.92% ± 0.03	87.05% ± 0.04	93.83% ± 0.51

datasets demonstrated that our proposed method consistently outperforms traditional segmentation models, including U-Net, LinkNet, and PSPNet, across evaluation metrics such as precision, TPR, DSC, and Jaccard index. Notably, our findings highlight the model's generalization ability, as it maintains high performance on an independent dataset despite being trained on a different one. This robustness is essential for ensuring reliability in diverse clinical environments where imaging conditions may vary significantly. Mammograms are known for their low contrast and low SNR ratio, making them challenging to analyze. To overcome these challenges, we have applied a combination of CLAHE and Wiener filtering, which enhances the quality of mammogram images. These advancements have the potential to greatly improve the accuracy and efficiency of breast cancer screening and diagnosis, ultimately leading to better patient outcomes. Our comparative analysis confirmed that pre-processing plays a crucial role in enhancing model performance.

Notably, the highest segmentation accuracy was achieved when both contrast enhancement and noise reduction techniques were applied, highlighting the effectiveness of combining these methods. The qualitative comparisons with baseline models revealed that UNet++ produces segmentation masks that are more closely aligned with ground truth annotations, further reinforcing its effectiveness. By analyzing segmentation performance across different BI-RADS and ACR categories, our method maintains acceptable DSC values across varying risk levels and breast density categories. It can aid radiologists in making informed decisions regarding biopsy recommendations and treatment planning. Despite the promising results of this study, several limitations need to be addressed in future research. One key limitation is the lack of diversity and clinical balance in the used datasets. Increasing dataset diversity, especially in terms of BI-RADS categories and ACR classifications, will enable a more comprehensive evaluation of the model's performance across different clinical scenarios. To enhance the model's generalizability, future studies should incorporate larger and more diverse datasets, particularly those including real-world hospital data and expert annotations from radiologists.

Future research should also focus on extending the segmentation capabilities of UNet++ by integrating additional modules for lesion classification, risk assessment, pathology information, patient history, and decision support. Furthermore, improving model adaptability to various imaging conditions and optimizing computational efficiency will be essential for broader clinical adoption, particularly in resource-limited and remote healthcare settings.

Acknowledgment of AI Assistance

To ensure clarity and coherence, AI-based tools were utilized to enhance the grammar and structure of this manuscript. However, all scientific analyses, interpretations, and conclusions were solely developed by the authors.

References

- 1- Rebecca L Siegel, AN Giaquinto, and Ahmedin Jemal, "Cancer statistics, 2024. CA: a cancer journal for clinicians." (2024).
- 2- Stephen W Duffy *et al.*, "Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women." *Cancer*, Vol. 126 (No. 13), pp. 2971-79, (2020).
- 3- James V Fiorica, "Breast cancer screening, mammography, and other modalities." *Clinical obstetrics and gynecology*, Vol. 59 (No. 4), pp. 688-709, (2016).
- 4- Kwan Hoong Ng and Malai Muttarak, "Advances in mammography have improved early detection of breast cancer." *Journal-Hong Kong College of Radiologists*, Vol. 6pp. 126-31, (2003).
- 5- Ritse M Mann *et al.*, "Breast cancer screening in women with extremely dense breasts recommendations of the European Society of Breast Imaging (EUSOBI)." *European radiology*, Vol. 32 (No. 6), pp. 4036-45, (2022).
- 6- Leila Abdelrahman, Manal Al Ghamdi, Fernando Collado-Mesa, and Mohamed Abdel-Mottaleb, "Convolutional neural networks for breast cancer detection in mammography: A survey." *Computers in biology and medicine*, Vol. 131p. 104248, (2021).
- 7- Eman Justaniah, Areej Alhothali, and Ghadah Aldabbagh, "Mammogram segmentation techniques: A review." *International Journal of Advanced Computer Science and Applications*, Vol. 12 (No. 5), (2021).
- 8- Wajeeha Ansar and Basit Raza, "Breast cancer segmentation in mammogram using artificial intelligence and image processing: A systematic review." *Current Chinese Science*, Vol. 3 (No. 1), pp. 3-22, (2023).

- 9- Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi, "Medical image segmentation using deep learning: A survey." *IET image processing*, Vol. 16 (No. 5), pp. 1243-67, (2022).
- 10- Epimack Michael, He Ma, Hong Li, Frank Kulwa, and Jing Li, "Breast cancer segmentation methods: current status and future potentials." *BioMed research international*, Vol. 2021 (No. 1), p. 9962109, (2021).
- 11- Solene Bechelli, "Computer-aided cancer diagnosis via machine learning and deep learning: a comparative review." *arXiv preprint arXiv:2210.11943*, (2022).
- 12- Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation." in *Medical image computing and computer-assisted intervention-MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, (2015): Springer, pp. 234-41.
- 13- Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "Unet++: A nested u-net architecture for medical image segmentation." in *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*, (2018): Springer, pp. 3-11.
- 14- Hui Sun *et al.*, "AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms." *Physics in Medicine & Biology*, Vol. 65 (No. 5), p. 055005, (2020).
- 15- N Ravitha Rajalakshmi, R Vidhyapriya, N Elango, and Nikhil Ramesh, "Deeply supervised u-net for mass segmentation in digital mammograms." *International Journal of Imaging Systems and Technology*, Vol. 31 (No. 1), pp. 59-71, (2021).
- 16- Felipe André Zeiser *et al.*, "Segmentation of masses on mammograms using data augmentation and deep learning." *Journal of digital imaging*, Vol. 33pp. 858-68, (2020).
- 17- Asma Baccouche, Begonya Garcia-Zapirain, Cristian Castillo Olea, and Adel S Elmaghraby, "Connected-UNets: a deep learning architecture for breast mass segmentation." *NPJ Breast Cancer*, Vol. 7 (No. 1), p. 151, (2021).
- 18- Hang Min *et al.*, "Fully automatic computer-aided mass detection and segmentation via pseudo-color mammograms and mask r-cnn." in *2020 IEEE 17th International symposium on biomedical imaging (ISBI)*, (2020): IEEE, pp. 1111-15.
- 19- Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso, "Inbreast: toward a full-field digital mammographic database." *Academic radiology*, Vol. 19 (No. 2), pp. 236-48, (2012).
- 20- Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research." *Scientific data*, Vol. 4 (No. 1), pp. 1-9, (2017).
- 21- D Narain Ponraj, M Evangelin Jenifer, P Poongodi, and J Samuel Manoharan, "A survey on the preprocessing techniques of mammogram for the detection of breast cancer." *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 2 (No. 12), pp. 656-64, (2011).
- 22- Ardra Mariya Joseph, M Grace John, and Anto Sahaya Dhas, "Mammogram image denoising filters: A comparative study." in *2017 Conference on emerging devices and smart systems (ICEDSS)*, (2017): IEEE, pp. 184-89.
- 23- Indra Kanta Maitra, Sanjay Nag, and Samir Kumar Bandyopadhyay, "Technique for preprocessing of digital mammogram." *Computer methods and programs in biomedicine*, Vol. 107 (No. 2), pp. 175-88, (2012).
- 24- R Ramani, N Suthanthira Vanitha, and S Valarmathy, "The pre-processing techniques for breast cancer detection in mammography images." *International Journal of Image, Graphics and Signal Processing*, Vol. 5 (No. 5), p. 47, (2013).
- 25- Antoni Buades, Bartomeu Coll, and J-M Morel, "A non-local algorithm for image denoising." in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, (2005), Vol. 2: Ieee, pp. 60-65.
- 26- G Arvind Kumar and Ashok Kusagur, "Evaluation of image denoising techniques a performance perspective." in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, (2016): IEEE, pp. 1836-39.
- 27- Etta D Pisano *et al.*, "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms." *Journal of digital imaging*, Vol. 11pp. 193-200, (1998).
- 28- Connor Shorten and Taghi M Khoshgoftaar, "A survey on image data augmentation for deep learning." *Journal of big data*, Vol. 6 (No. 1), pp. 1-48, (2019).
- 29- Alberto Garcia-Garcia, S Orts-Escolano, S Oprea, V Villena-Martinez, and J Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation. arXiv 2017." *arXiv preprint arXiv:1704.06857*, (2020).
- 30- Imran Qureshi *et al.*, "Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends." *Information Fusion*, Vol. 90pp. 316-52, (2023).
- 31- Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network." in *Proceedings of the IEEE conference on*

- computer vision and pattern recognition*, (2017), pp. 2881-90.
- 32- Abhishek Chaurasia and Eugenio Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation." in *2017 IEEE visual communications and image processing (VCIP)*, (2017): IEEE, pp. 1-4.
 - 33- Dominik Müller, Iñaki Soto-Rey, and Frank Kramer, "Towards a guideline for evaluation metrics in medical image segmentation." *BMC Research Notes*, Vol. 15 (No. 1), p. 210, (2022).
 - 34- Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations." in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, (2017): Springer, pp. 240-48.
 - 35- Antonio Galli, Stefano Marrone, Gabriele Piantadosi, Mario Sansone, and Carlo Sansone, "A pipelined tracer-aware approach for lesion segmentation in breast DCE-MRI." *Journal of imaging*, Vol. 7 (No. 12), p. 276, (2021).
 - 36- Helene Berment, V Becette, M Mohallem, F Ferreira, and P Chérel, "Masses in mammography: What are the underlying anatomopathological lesions?" *Diagnostic and interventional imaging*, Vol. 95 (No. 2), pp. 124-33, (2014).
 - 37- Song Liu, Xiao-Dong Wu, Wen-Jian Xu, Qing Lin, Xue-Jun Liu, and Ying Li, "Is there a correlation between the presence of a spiculated mass on mammogram and luminal a subtype breast cancer?" *Korean journal of radiology*, Vol. 17 (No. 6), pp. 846-52, (2016).
 - 38- Lazaros Tsochatzidis, Panagiota Koutla, Lena Costaridou, and Ioannis Pratikakis, "Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses." *Computer methods and programs in biomedicine*, Vol. 200p. 105913, (2021).
 - 39- Yongye Su, Qian Liu, Wentao Xie, and Pingzhao Hu, "YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms." *Computer methods and programs in biomedicine*, Vol. 221p. 106903, (2022).
 - 40- Ghada M El-Banby, Nourhan S Salem, Eman A Tafwek, and Essam N Abd El-Azziz, "Automated abnormalities detection in mammography using deep learning." *Complex & Intelligent Systems*, Vol. 10 (No. 5), pp. 7279-95, (2024).