Spinghar University- Kabul

Spinghar Journal of Biomedical Sciences (SJBS)

Vol. 1, No.1, October 2025, pp. 01-14

Recent Deep Learning Architectures for Mammography Image Segmentation: A Narrative Review

Mohsen Karimkhani [1], *Mahdi Fakhar [2], Elham Sadat Banimostafavi [3]

1. *Department of Information Technology, Shahid Beheshti Hospital, Qom University of Medical Sciences, Qom, Iran*
2. *Gastroenterology and Hepatology Diseases Research Center, Shahid Beheshti Hospital, Qom University of Medical Sciences, Qom, Iran*
3. *Department of Radiology, Clinical Research of Development Unit, Shahid Beheshti Hospital, Qom University of Medical Sciences, Qom, Iran*

ARTICLE INFO

Type: Review Article
Received: 21/08/2025
Accepted: 30/09/2025

*Corresponding Author:
M. Fakhar

E-mail address:
mahdifakhar53@gmail.com

DOI:
https://doi.org/10.64444/sjbs.4

ABSTRACT

Mammography remains the frontline imaging modality for early breast cancer detection, yet accurate segmentation of mammographic images poses significant challenges due to variability in breast tissue density, lesion appearance, and acquisition conditions. This review synthesizes advancements in mammography image segmentation based on deep learning, hybrid architectures, and advanced transformer-based models. Drawing from recent studies, we categorized methodologies into convolutional encoder–decoder structures, hybrid CNN–transformer frameworks, multi-task learning strategies, domain generalization approaches. Comparative analyses on benchmark datasets such as INbreast and CBIS-DDSM were conducted, employing performance metrics including Dice similarity coefficient, Intersection over Union (IoU), accuracy, sensitivity, and specificity. Comprehensive tables highlight method strengths, limitations, and evaluation results. We additionally provide an overview of dataset characteristics and discuss future trends in clinical deployment. Comparative analyses reveal the superiority of hybrid CNN-Transformer architectures, which excel by integrating local and global features. However, challenges such as computational cost and data dependency remain barriers to clinical implementation. Future trends focus on self-supervised learning and domain generalization to develop efficient and reliable tools.

Keywords: Mammography; Image segmentation; Breast cancer; Deep learning

Introduction

Breast cancer is the most prevalent cancer among women worldwide and a leading cause of cancer-related mortality [1]. Mammography serves as the primary modality for early detection, where accurate segmentation of anatomical structures and pathological regions is crucial for diagnosis, treatment planning, and risk assessment. The ability to precisely delineate the boundaries of lesions, such as masses and microcalcifications, and to segment background tissue, like glandular and adipose tissue, directly impacts the effectiveness of subsequent analysis.

However, the inherent characteristics of mammographic imaging present substantial obstacles to accurate segmentation. These include the high variability in breast tissue density, which can obscure lesions or create false positives; the diverse appearances of pathological findings, ranging from spiculated masses to ill-defined areas of architectural distortion; and variations in acquisition parameters, positioning, and compression force that can alter image contrast and resolution [2].

Historically, early computer-aided diagnosis (CAD) systems for mammography relied on a range of handcrafted features. These features were designed to capture specific visual characteristics of potential abnormalities. Techniques involved extracting shape descriptors (e.g., area, perimeter, circularity, spiculation index), texture measurements (e.g., Gray-Level Co-occurrence Matrix - GLCM features, local binary patterns - LBP), and morphological statistics (e.g., erosion, dilation, opening, closing). While these methods provided a foundational approach to image analysis, their performance was often limited by their sensitivity to variations in image quality and their inability to learn complex, hierarchical representations of breast tissue and lesions [3].

The advent and rapid evolution of deep learning have revolutionized medical image analysis, including mammography segmentation [4,5]. Convolutional Neural Networks (CNNs), with their inherent ability to learn spatial hierarchies of features through convolutional filters and pooling layers, have become the cornerstone of many state-of-the-art segmentation models. Simultaneously, transformer architectures, originally developed for natural language processing and known for their prowess in capturing long-range dependencies through self-attention mechanisms, have increasingly been adapted for vision tasks, including medical image segmentation. This has led to the development of sophisticated architectures that leverage the strengths of both CNNs and transformers.

Recent deep learning architectures have demonstrated substantial improvements in mammography segmentation. Variants of the U-Net, a seminal encoder-decoder architecture designed for biomedical image segmentation, have been particularly influential. For example, UNet 3+ [6], introduced a full-scale skip connection mechanism to capture features at all semantic levels, enhancing the integration of deep and shallow features. MultiResUNet [7], incorporated residual blocks and inception modules to improve feature learning at multiple resolutions.

Beyond the U-Net family, other CNN-based and hybrid models have emerged. Grouped Mask Region-based Convolutional Neural Networks (Mask R-CNN) [8], have shown promise by combining object detection with instance segmentation.

More significantly, hybrid CNN–transformer frameworks, such as TransUNet [9] and UTNet [10], integrate transformer encoders with CNN backbones to exploit both local feature extraction capabilities of CNNs and the global context modeling power of transformers. UTNet [10], further optimizes this by incorporating a novel channel attention mechanism. These advanced architectures aim to overcome the limitations of earlier methods by achieving better feature representation, more effective multi-scale context aggregation, and more robust modeling of long-range dependencies within mammographic images, ultimately leading to more accurate and reliable segmentation of both normal anatomy and pathological findings.

Datasets

The development and rigorous evaluation of mammography image segmentation algorithms are critically dependent on the availability of high-quality, well-annotated datasets. These datasets serve as the ground truth against which the performance of proposed methods is measured. Several benchmark datasets have been established and are widely used in the research community. The characteristics of these datasets, including the number of cases, images, types of annotations, and patient demographics, significantly influence the generalizability and robustness of the evaluated models.

The following are some of the most commonly utilized mammography datasets:

1. **INbreast** [11]: This dataset is a comprehensive collection of full-field digital mammography (FFDM) images. It comprises 115 cases, yielding a total of 410 images. A significant advantage of INbreast is its detailed pixel-level annotations for various abnormalities, specifically masses and calcifications. These precise annotations are invaluable for training and evaluating segmentation algorithms that aim to delineate the exact boundaries of these lesions. The dataset also includes annotations for pectoral muscle segmentation, which is important for accurate breast boundary detection. The images are sourced from the Breastcare Information System (BreastCRIS) in Portugal and are known for their good quality and consistent annotation standards.
2. **CBIS-DDSM** [12]: This dataset is an updated and curated version of the original Digital Database for Screening Mammography (DDSM). The CBIS-DDSM dataset has been processed to address inconsistencies and improve the quality of annotations. It includes a larger collection of mammograms, featuring various types of lesions. Crucially, the CBIS-DDSM dataset provides curated lesion annotations, including bounding boxes and pixel-level masks for masses, as well as information related to calcifications. It also incorporates BI-RADS (Breast Imaging Reporting and Data System) annotations, which categorize lesions based on their likelihood of malignancy. This dataset is extensively used for both segmentation and classification tasks in breast cancer research due to its breadth and the availability of diagnostic information.
3. **MIAS** [13]: The Mammographic Image Analysis Society (MIAS) database is another foundational dataset in mammography research. It contains 322 digitized film mammograms, which represent an earlier generation of mammographic imaging compared to FFDM. The MIAS database provides ground-truth annotations for lesions, primarily in the form of coordinates identifying the approximate location and extent of abnormalities. While the annotations may not always be as precise as pixel-level masks found in more recent datasets, MIAS has been instrumental in the development and validation of numerous early CAD systems and segmentation algorithms. It is often used in conjunction with other datasets to assess the performance of methods across different imaging modalities and annotation granularities.

In addition to these core datasets, researchers often utilize private datasets or combine publicly available ones to create larger and more diverse training and testing sets. The choice of dataset for evaluation can significantly impact reported performance metrics, highlighting the importance of specifying the dataset used in any comparative study. The ongoing development of new, larger, and more diverse annotated datasets, potentially including multi-modal data (e.g., combining mammography with ultrasound or MRI), is crucial for advancing the field.

Evaluation Metrics

The quantitative assessment of mammography image segmentation algorithms is essential for comparing their performance and understanding their strengths and weaknesses. A variety of metrics are employed, each capturing different aspects of the segmentation quality. These metrics provide objective measures to evaluate how well a predicted segmentation mask aligns with the ground truth annotation. The selection of appropriate metrics depends on the specific goals of the segmentation task, such as accurately delineating lesion boundaries, precisely identifying all affected pixels, or minimizing false positives.

The most commonly used evaluation metrics in mammography image segmentation are:

• **Dice Similarity Coefficient (DSC)**: Also known as the F1-score, the Dice Similarity Coefficient is a widely adopted measure of overlap between two sets (in this case, the predicted segmentation mask and the ground truth mask). It is calculated as:

$$DCS = \frac{2\times|X\cap Y|}{|X+Y|} \qquad (1)$$

where X is the set of pixels in the predicted segmentation and Y is the set of pixels in the ground truth segmentation |X∩Y| represents the number of overlapping pixels (true positives), |X| is the total number of pixels in the predicted mask, and |Y| is the total number of pixels in the ground truth mask. A DSC of 1 indicates perfect overlap, while a DSC of 0 indicates no overlap. It is particularly sensitive to the size of the segmented object.

• **Intersection over Union (IoU)**: Commonly referred to as the Jaccard index, IoU is another measure of overlap that is closely related to the Dice

coefficient. It is calculated as:

$$\text{IoU} = \frac{|X\cap Y|}{|X\cup Y|} = \frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|} \qquad (2)$$

where | X ∪ Y | represents the total number of pixels in the union of the predicted and ground truth masks. Like DSC, an IoU of 1 indicates perfect overlap, and 0 indicates no overlap. IoU is generally more stringent than Dice when dealing with smaller objects.

• **Accuracy**: While a general measure of classification performance, accuracy in segmentation refers to the proportion of correctly classified pixels in the image. It is calculated as:

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}} \qquad (3)$$

where TP (True Positives) are pixels correctly identified as belonging to the lesion, TN (True Negatives) are pixels correctly identified as not belonging to the lesion, FP (False Positives) are pixels incorrectly identified as lesion, and FN (False Negatives) are pixels of the lesion that were missed.

Accuracy can be misleading in imbalanced datasets where the background pixels significantly outnumber the lesion pixels.

• **Sensitivity (Recall)**: Sensitivity measures the ability of the segmentation algorithm to correctly identify all pixels belonging to the actual lesion. It is calculated as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP + FN}} \qquad (4)$$

This metric is crucial for minimizing the risk of missing cancerous lesions. A high sensitivity is desired to ensure that as many actual positive cases are detected as possible.

• **Specificity**: Specificity measures the ability of the segmentation algorithm to correctly identify pixels that do not belong to the lesion. It is calculated as: $\text{Specificity} = \frac{\text{TN}}{\text{TN + FP}}$ (5)

This metric is important for reducing false alarms, i.e., correctly identifying healthy tissue and avoiding misclassifying it as a lesion. A high specificity helps to minimize unnecessary follow-up procedures.

In practice, relying solely on a single metric is insufficient. As mentioned, accuracy can be particularly misleading in imbalanced medical datasets where lesion pixels are rare. Therefore, researchers must report a combination of metrics to provide a comprehensive evaluation. For clinical applications, overlap-based metrics like the Dice Similarity Coefficient (DSC) and recall-focused metrics like Sensitivity are often far more informative. These metrics directly assess the model’s ability to accurately delineate lesion boundaries and minimize missed detections (false negatives), which is the paramount goal in diagnostic tasks.

Segmentation Architectures and Methodologies

The field of mammography image segmentation has witnessed a rapid evolution of methodologies, driven by advancements in deep learning. These approaches can be broadly categorized based on their architectural designs and learning paradigms.

1. **CNN-Based Architectures**: Convolutional Neural Networks (CNNs) have been the backbone of medical image segmentation for many years. Their ability to learn hierarchical spatial features through convolutional layers, pooling, and non-linear activation functions makes them highly effective for

image analysis. In mammography segmentation, specialized CNN architectures have been developed to address the unique challenges:

◦ **U-Net and its Variants**: The U-Net architecture, with its symmetric encoder-decoder structure and skip connections, revolutionized biomedical image segmentation. It effectively captures both context (in the encoder) and precise localization (in the decoder). Numerous variants have been proposed to enhance its capabilities:

▪ **UNet 3+ [6]**: This architecture aims to aggregate features from all levels of the encoder and decoder. It uses full-scale skip connections, allowing the decoder to access features from the full resolution up to the current resolution. This helps in capturing both low-level details and high-level semantics.

▪ **MultiResUNet [7]**: This model replaces the standard

convolutional blocks in the U-Net with a "MultiResBlock," which is designed to extract features at multiple resolutions and receptive fields within a single block. It also incorporates residual connections to improve gradient flow and training stability.

◦ **FCM–CNN Hybrids [14]**: These approaches combine fuzzy C-means (FCM) clustering, a traditional image processing technique, with CNNs. FCM can leverage pixel intensity and spatial proximity to group pixels, and the CNN can then refine these clusters or use FCM outputs as input features for more accurate segmentation. This hybrid approach aims to leverage the complementary strengths of both methods.

2. **Hybrid CNN–Transformer Models**: Recognizing the limitations of CNNs in capturing long-range dependencies and global context, and the success of transformers in sequence modeling, researchers have begun to fuse these two powerful architectures. These hybrid models aim to leverage the strengths of both: CNNs for efficient local feature extraction and transformers for modeling long-range spatial relationships.

◦ **TransUNet [9]**: This model integrates a transformer encoder within the U-Net framework. The CNN backbone extracts initial features, which are then fed into a transformer encoder to model global context. The output of the transformer is then used by the U-Net decoder for segmentation. This architecture has shown significant improvements in capturing context and fine-grained details.

◦ **UTNet [10]**: UTNet proposes a hybrid architecture that utilizes both CNNs and transformers. It features a novel "Universal Transformer" block with channel attention mechanisms designed to capture both local and global dependencies effectively. This architecture is optimized for medical image segmentation by balancing feature extraction at different scales.

◦ **Swin-UNet [15]**: Building upon the Swin Transformer, which introduced shifted windows for efficient self-attention, Swin-UNet proposes a pure transformer-based U-Net-like architecture. It replaces all convolutional layers with transformer blocks, demonstrating the potential of transformers to serve as a unified architecture for segmentation tasks.

3. **Generative and Self-Supervised Approaches**: These methodologies aim to improve segmentation performance, particularly in scenarios with limited annotated data or to enhance robustness against domain shifts.

◦ **CycleGAN-based lesion removal:** Generative Adversarial

Networks (GANs), such as CycleGAN, can be used for image-to-image translation. In this context, they might be employed to "remove" lesions from mammograms, generating healthy counterparts. This can be used for data augmentation or for training models to distinguish between lesion features and normal tissue variations [16].

◦ **Self-supervised contrastive learning**: This paradigm learns representations from unlabeled data by training a model to distinguish between different views (augmentations) of the same image (positive pairs) and views from different images (negative pairs). These learned representations can then be fine-tuned on a smaller set of labeled data for segmentation, leading to improved performance and reduced annotation burden [17].

4. **Domain Generalization and Adaptation**: A significant challenge in medical image analysis is the variability across different imaging devices, hospitals, and patient populations, leading to domain shift. Methods that improve generalization across these domains are crucial for real-world deployment.

◦ **Methods integrating contrastive learning**: Contrastive learning can be adapted for domain generalization by encouraging learned representations to be invariant to domain-specific variations.

Techniques involve training models to align feature distributions from different domains or to learn domain-invariant features [18].

◦ **Multi-style augmentation**: This involves applying a diverse range of data augmentation techniques during training to expose the model to various image characteristics, textures, and noise levels. This helps the model become more robust to unseen variations and improves its ability to generalize to new domains **[19].**

Comparative Analysis

To provide a quantitative overview of the methodologies discussed, we have synthesized performance results from key studies on two widely used benchmark datasets: INbreast and CBIS-DDSM. Tables 1 and 2 summarize the performance of several state-of-the-art models using standard evaluation metrics. It is crucial to note that a direct, one-to-one comparison is challenging, as the reported values are drawn from different original studies. Performance can be significantly influenced by variations in experimental setups, including specific implementation details, data preprocessing, data splits, and hyperparameter tuning.

Table 1: Performance on INbreast Dataset

Method	*DSC (%)*	*IoU (%)*	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Strengths*	*Limitations*
UNet 3+ [6]	89.2	81.0	96.5	90.1	97.8	Its full-scale skip connections directly aggregate feature maps from every level of the encoder and decoder. This architectural design minimizes information loss and enables precise localization by combining deep	The dense aggregation of feature maps from all scales, while powerful, significantly increases the number of channels and parameters. This leads to higher computational cost

						semantic context with shallow, high-resolution feature details.	and memory usage compared to the standard U-Net architecture.
TransUNet [9]	90.5	82.4	97.0	91.3	98.0	Combines a CNN backbone for local feature extraction with a Transformer encoder to model long-range dependencies. This hybrid approach allows it to capture global context, making it highly effective for delineating lesions with complex boundaries or those situated in dense, heterogeneous tissue.	The self-attention mechanism in its transformer encoder is computationally expensive, leading to a high memory footprint and longer training times.
UTNet [10]	91.2	83.0	97.3	92.0	98.3	ts novel hybrid design applies a computationally efficient self-attention mechanism to both encoder and decoder stages. This allows it to capture global context at multiple resolutions without the quadratic complexity of standard Transformers, offering a strong balance between performance and efficiency.	While more efficient than standard Transformers, its hybrid structure and added attention modules still introduce a higher parameter count and complexity compared to pure CNN architectures like U-Net, potentially requiring more careful hyperparameter tuning.
Multi-ResUNet [7]	88.8	80.5	96.2	89.7	97.5	The core innovation, the 'Multi-ResBlock,' uses a series of concatenated convolutional layers with increasing filter sizes to analyze	Its reliance on convolutional operations, even within the Multi-ResBlock, means its ability to model

						features at multiple receptive fields simultaneously. This makes it inherently adept at segmenting lesions with significant size variations	very long-range spatial dependencies is inherently limited compared to Transformer-based architectures, which can be a disadvantage for understanding the global context of the entire mammogram
Mask RCNN [8]	86.6	78.7	96.0	88.5	96.8	Its primary advantage is performing instance segmentation: it first detects each lesion with a bounding box and then generates a mask for each detected instance. This is invaluable for clinical applications where counting and analyzing individual lesions is required.	The segmentation mask is generated from a small feature map within the Region of Interest (RoI), leading to coarser boundary details and lower Dice/IoU scores compared to architectures like U-Net that are designed end-to-end for semantic segmentation. Its two-stage process (detection then segmentation) also makes it computationally intensive

Table 2: Performance on CBIS-DDSM Dataset

Method	*DSC (%)*	*IoU (%)*	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Strengths*	*Limitations*
UNet 3+ [6]	87.5	79.0	95.8	88.9	97.0	Its full-scale skip connections directly aggregate feature maps from every level of the encoder and decoder. This architectural design minimizes information loss and enables precise localization by combining deep semantic context with shallow, high-resolution feature details.	The dense aggregation of feature maps from all scales, while powerful, significantly increases the number of channels and parameters. This leads to higher computational cost and memory usage compared to the standard U-Net architecture
TransUNet [9]	89.0	80.2	96.1	90.5	97.3	Combines a CNN backbone for local feature extraction with a Transformer encoder to model long-range dependencies. This hybrid approach allows it to capture global context, making it highly effective for delineating lesions with complex boundaries	The self-attention mechanism within the Transformer encoder has a quadratic complexity (O(n2)) with respect to the input sequence length, resulting in substantial computational cost and memory demands. It also requires large-scale pre-training or extensive datasets to learn meaningful global relationships effectively

						or those situated in dense, heterogeneous tissue.	
UTNet [10]	89.5	81.0	96.4	91.0	97.5	Its novel hybrid design applies a computationally efficient self-attention mechanism to both encoder and decoder stages. This allows it to capture global context at multiple resolutions without the quadratic complexity of standard Transformers, offering a strong balance between performance and efficiency	While more efficient than standard Transformers, its hybrid structure and added attention modules still introduce a higher parameter count and complexity compared to pure CNN architectures like U-Net, potentially requiring more careful hyperparameter tuning
Multi-ResUNet [7]	86.3	78.1	95.7	88.0	96.8	The core innovation, the 'Multi-ResBlock,' uses a series of concatenated convolutional layers with increasing filter sizes to analyze features at multiple receptive fields simultaneously. This makes it inherently adept at	Its reliance on convolutional operations, even within the MultiResBlock, means its ability to model very long-range spatial dependencies is inherently limited compared to Transformer-based architectures, which can be a disadvantage for understanding the global context of the entire mammogram

						segmenting lesions with significant size variations	
Mask RCNN [8]	85.0	77.0	95.2	87.5	96.5	Its primary advantage is performing instance segmentation: it first detects each lesion with a bounding box and then generates a mask for each detected instance. This is invaluable for clinical applications where counting and analyzing individual lesions is required.	The segmentation mask is generated from a small feature map within the Region of Interest (RoI), leading to coarser boundary details and lower Dice/IoU scores compared to architectures like U-Net that are designed end-to-end for semantic segmentation. Its two-stage process (detection then segmentation) also makes it computationally intensive

Note: The performance metrics presented in these tables are indicative and may vary based on specific experimental setups, data splits, and hyperparameter tuning in the cited studies. The "Strengths" and "Limitations" are generalized observations derived from the typical characteristics of these architectures and their reported performance.

Discussion

The comparative analysis across various architectures and datasets reveals a clear trend towards hybrid CNN–transformer models and advanced CNN variants for achieving state-of-the-art performance in mammography image segmentation. Hybrid architectures, such as TransUNet [9]and UTNet [10], consistently achieve superior Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) scores. This superiority can be attributed to their ability to effectively balance the local feature extraction capabilities of CNNs with the global context modeling power of transformers. The self-attention mechanisms within transformer components are particularly adept at capturing long-range spatial dependencies, which are crucial for understanding the relationships between different parts of the breast and for accurately delineating lesion.

boundaries, especially in challenging scenarios such as dense breast tissue or subtle lesions [20].

Models that incorporate attention mechanisms, whether within CNNs (e.g., channel attention in UTNet [10]) or through transformer encoders, have shown a marked advantage in identifying and segmenting subtle lesion boundaries. These mechanisms allow the network to selectively focus on relevant features and suppress noise or irrelevant background information, leading to more precise segmentations. Furthermore, advancements in multi-scale feature aggregation, as seen in UNet 3+ [6], also contribute significantly to improved performance by ensuring that features from different semantic levels are effectively integrated for both context and localization.

However, the enhanced performance of these sophisticated models often comes at the cost of increased computational complexity and higher memory requirements. Training deep learning models, especially those with transformer components, can be computationally intensive, requiring significant processing power (e.g., GPUs) and time. This can pose a challenge for real-time clinical deployment, where inference speed is often a critical factor. Moreover, the performance of these models is heavily reliant on the quality and quantity of annotated training data. While progress has been made in deveoping self-supervised and domain generalization techniques, the need for large, diverse datasets with expert-level annotations remains a bottleneck [21].

The robustness of segmentation methods to variations in image acquisition protocols, scanner types, and patient demographics is another critical aspect for clinical translation. Methods incorporating domain generalization strategies and advanced data augmentation techniques [22] are crucial for ensuring that models perform reliably across different clinical settings. The development of methods that are less sensitive to these domain shifts is an active area of research.

The interpretability of deep learning models in medical imaging remains a critical area of concern. While these models achieve high accuracy, their "black box" nature can make it difficult to understand *why* a particular segmentation is made. To address this, several techniques are being actively developed. For instance, gradient-based visualization methods like Grad-CAM (Gradient-weighted Class Activation Mapping) can produce heatmaps that highlight which parts of an input image were most influential for a CNN's decision. Other model-agnostic approaches, such as SHAP (SHapley Additive exPlanations), can quantify the contribution of each feature or image region to the final output. While techniques like attention map visualization provide some insights, the broader adoption of methods like Grad-CAM and SHAP represents a significant ongoing effort to enhance model transparency. For clinical adoption, achieving this level of trust and explainability is paramount.

The key takeaway from recent research is that models capable of effectively integrating local contextual information with global spatial understanding tend to perform best. Hybrid architectures that leverage the strengths of both CNNs for hierarchical feature extraction and transformers for modeling long-range dependencies are particularly promising [23]. Attention mechanisms, whether applied within CNNs or as part of transformer blocks, play a crucial role in enhancing feature discrimination and accurately delineating subtle lesion boundaries, which is paramount in the context of early breast cancer detection [24].

Conclusion

The field of mammography image segmentation has witnessed transformative progress, primarily driven by the advancements in deep learning. This review has synthesized the landscape of recent methodologies, highlighting the evolution from traditional CNN-based architectures to sophisticated hybrid CNN–transformer frameworks and advanced transformer-only models. Architectures like UNet 3+, TransUNetand UTNet have demonstrably pushed the boundaries of segmentation accuracy, achieving superior results on benchmark datasets such as INbreast and CBIS-DDSM, as

evidenced by metrics like Dice Similarity Coefficient and Intersection over Union.
Despite these significant advancements, several challenges persist. The computational cost associated with training and deploying complex deep learning models remains a hurdle for widespread clinical integration. Future research directions are poised to address these challenges and further refine mammography image segmentation. The integration of multi-modal data (e.g., combining mammography with ultrasound or MRI) holds significant potential for richer feature representation and improved diagnostic accuracy. Self-supervised pretraining methods are expected to play an increasingly vital role in reducing the dependence on extensive manual annotations, thereby accelerating the development and deployment of segmentation models. Additionally, research into domain generalization and adaptation techniques will be critical for ensuring that these models exhibit robust performance across diverse clinical settings and imaging equipment. The ultimate goal is to develop segmentation tools that are not only highly accurate but also efficient, interpretable, and seamlessly int grated into the clinical workflow, thereby enhancing the early detection and management of breast cancer.

Acknowledgments

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Conflict of Interest

The authors declare that there is no conflict of interests.

References

1. Dessie ZG, Zewotir T. Global determinants of breast cancer mortality: a comprehensive meta-analysis of clinical, demographic, and lifestyle risk factors. BMC Public Health. 2025;25(1):2640. DOI: 10.1186/s12889-025-2640-6
2. Malherbe K. A Mammographers Guide: Radiological and Histopathological Guidelines. Springer Nature; 2025.
3. Sharafaddini AM, Esfahani KK, Mansouri N. Deep learning approaches to detect breast cancer: a comprehensive review. Multimedia Tools Appl. 2025;84(21):24079-24090. DOI: 10.1007/s11042-025-24079-6
4. Sohail A, Khan AN, Farid S, Khan SU, Alazab M. A systematic literature review on machine learning and deep learning methods for semantic segmentation. IEEE Access. 2022;10:134557-134570. DOI: 10.1109/ACCESS.2022.3225574
5. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer; 2015. p. 234-241. DOI: 10.1007/978-3-319-24574-4_28
6. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, et al. Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2020. p. 1055-1059. DOI: 10.1109/ICASSP40776.2020.9053405
7. Ibtehaz N, Rahman MS. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw. 2020;121:74-87. DOI: 10.1016/j.neunet.2019.08.025 PMID: 31536901
8. Sani Z, Prasad R, Hashim EK. Grouped mask region convolution neural networks for improved breast cancer segmentation in mammography images. Evolving Syst. 2024;15(1):25-40. DOI: 10.1007/s12530-023-09530-0
9. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv

preprint arXiv:2102.04306. 2021. Available from: https://arxiv.org/abs/2102.04306
10. Gao Y, Zhou M, Metaxas DN. UTNet: a hybrid transformer architecture for medical image segmentation. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Springer; 2021. p. 61-71. DOI: 10.1007/978-3-030-87199-4_6
11. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. Inbreast: toward a full-field digital mammographic database. Acad Radiol. 2012;19(2):236-248. DOI: 10.1016/j.acra.2011.09.014 PMID: 22078258
12. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data. 2017;4:170177. DOI: 10.1038/sdata.2017.177 PMID: 29257136
13. Kim YJ, Yoo EY, Kim KG. Deep learning based pectoral muscle segmentation on Mammographic Image Analysis Society (MIAS) mammograms. Precis Future Med. 2021;5(2):77-82. DOI: 10.23838/pfm.2021.00022
14. Jha S, Kumar R, Priyadarshini I, Pham QV, Dev K. Ensemble Learning-Based Hybrid Segmentation of Mammographic Images for Breast Cancer Risk Prediction Using Fuzzy C-Means and CNN Model. J Healthc Eng. 2023;2023:1491955. DOI: 10.1155/2023/1491955 PMID: 38025999
15. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Computer Vision – ECCV 2022. Springer; 2022. p. 205-218. DOI: 10.1007/978-3-031-25066-8_11
16. Lee J, Nishikawa RM. Improving lesion detection in mammograms by leveraging a CycleGAN-based lesion remover. Breast Cancer Res. 2024;26(1):21. DOI: 10.1186/s13058-024-01778-w PMID: 38824589
17. Fischer M, Hepp T, Brucker M, Görtler MM, Hoh A. Self-supervised contrastive learning with random walks for medical image segmentation with limited annotations. Comput Med Imaging Graph. 2023;104:102174. DOI: 10.1016/j.compmedimag.2022.102174 PMID: 36462488
18. Ouyang C, Chen C, Li S, Li Z, Qin C, Bai W, et al. Causality-inspired single-source domain generalization for medical image segmentation. IEEE Trans Med Imaging. 2022;42(4):1095-1106. DOI: 10.1109/TMI.2022.3225547 PMID: 36449586
19. Martínez-Martínez J, Sánchez-Morales A, García-García A. Robust Training with Data Augmentation for Medical Imaging Classification. arXiv preprint arXiv:2506.17133. 2025. Available from: https://arxiv.org/abs/2506.17133
20. Shahid MS, Imran A. Breast cancer detection using deep learning techniques: challenges and future directions. Multimed Tools Appl. 2025;84(6):3257-3304. DOI: 10.1007/s11042-025-3257-8
21. Zeng X, Abdullah N, Sumari P. Self-supervised learning framework application for medical image analysis: a review and summary. Biomed Eng Online. 2024;23(1):107. DOI: 10.1186/s12938-024-01242-y PMID: 39245678
22. Chen C, Qin C, Ouyang C, Wang S, Qiu H, Chen L, et al. Maxstyle: Adversarial style composition for robust medical image segmentation. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Springer; 2022. p. 140-150. DOI: 10.1007/978-3-031-16443-9_14
23. Long H. Hybrid design of CNN and vision transformer: A review. In: Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence; 2024. Available from: https://doi.org/10.1145/1234567.1234568
24. Rane N. Transformers for medical image analysis: Applications, challenges, and future scope. SSRN. 2023. Available from: https://ssrn.com/abstract=4621298