2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871091

Comparative Analysis of Current Deep Learning Networks for Breast Lesion Segmentation in Ultrasound Images

Margarida R. Ferreira, Helena R. Torres, Bruno Oliveira, João Gomes-Fonseca, Pedro Morais, Paulo Novais, and João L. Vilaça

***Abstract*— Automatic lesion segmentation in breast ultrasound (BUS) images aids in the diagnosis of breast cancer, the most common type of cancer in women. Accurate lesion segmentation in ultrasound images is a challenging task due to speckle noise, artifacts, shadows, and lesion variability in size and shape. Recently, convolutional neural networks have demonstrated impressive results in medical image segmentation tasks. However, the lack of public benchmarks and a standardized evaluation method hampers the networks' performance comparison. This work presents a benchmark of seven state-of-the-art methods for the automatic breast lesion segmentation task. The methods were evaluated on a multi-center BUS dataset composed of three public datasets. Specifically, the U-Net, Dynamic U-Net, Semantic Segmentation Deep Residual Network with Variational Autoencoder (SegResNetVAE), U-Net Transformers, Residual Feedback Network, Multiscale Dual Attention-Based Network, and Global Guidance Network (GG-Net) architectures were evaluated. The training was performed with a combination of the cross-entropy and Dice loss functions and the overall performance of the networks was assessed using the Dice coefficient, Jaccard index, accuracy, recall, specificity, and precision. Despite all networks having obtained Dice scores superior to 75%, the GG-Net and SegResNetVAE architectures outperform the remaining methods, achieving 82.56% and 81.90%, respectively.**

***Clinical Relevance*— The results of this study allowed to prove the potential of deep neural networks to be used in clinical practice for breast lesion segmentation, also suggesting the best model choices.**

I. Introduction

Breast cancer is the most prevalent cancer among women and one of the leading causes of death worldwide [1]. In 2020, there were nearly 2.3 million new diagnosed cases and 685,000 deaths from breast cancer [1]. Therefore, its early detection is crucial for successful treatment, thus reducing the mortality rate. Due to its non-invasive, non-ionizing radiation, inexpensive, and real-time nature, ultrasound is one of the most frequent and effective approaches in breast cancer diagnosis [2]. Indeed, ultrasound is one of the primary diagnostic imaging modalities in several clinical applications [3], [4]. However, breast cancer diagnosis using ultrasound is challenging due to the speckle noise, low signal-to-noise ratio, shadows that make the boundaries of tumors ambiguous, as well as the highly variable tumor sizes and shapes [2]. Furthermore, in traditional clinical practice, the contours of the lesion in breast ultrasound (BUS) images are usually delineated manually by radiologists, which is time-consuming and dependent on the observers [2]. Thus, several approaches for automatic lesion segmentation have been proposed [5].

In early research, the main segmentation methods included threshold-based, region-growth-based, and active contour-based methods [6]. These approaches often need extra manual intervention or a parameter selection process, which still leads to inter and intra-observer variations [6]. Recently, deep learning algorithms, particularly convolutional neural networks (CNNs), have superseded the image processing field, showing promising results for BUS segmentation [5]. In 2015, Ronneberger *et al*. proposed the U-Net, a pixel-to-pixel, end-to-end CNN for medical image segmentation [7]. In recent years, many U-Net-like architectures have been proposed, such as Residual U-Net [8], Attention U-Net [9], Dense U-Net [10], U-Net++ [11], and Dynamic U-Net (DynUNet) [12]. Moreover, other encoder-decoder architectures were also explored, such as Semantic Segmentation Deep Residual Network with Variational Autoencoder (SegResNetVAE) [13]. In [14], a transformer was used as the encoder of a network termed U-Net Transformers (UNETR). Specifically for the breast lesion segmentation task, Wang *et al.* proposed a novel residual feedback network (RF-Net) that improves performance by increasing the confidence of inconclusive pixels [5]. Iqbal *et al.* presented a multiscale dual attention-based network (MDA-Net) for lesion segmentation in BUS images [15]. Similarly, Xue *et al.* suggested a deep CNN, the Global Guidance Network (GG-Net), operated with a global guidance block and lesion boundary detection modules to optimize the BUS lesion segmentation [16].

Although many BUS segmentation methods have been investigated, most were evaluated using relatively small private datasets with different quantitative metrics, which

This work was funded by the projects "NORTE-01-0145-FEDER-000045" and "NORTE-01-0145-FEDER-000059", supported by the Northern Portugal Regional Operational Programme (NORTE 2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (FEDER). It was also funded by national funds, through the FCT (Fundação para a Ciência e a Tecnologia) and FCT/MCTES in the scope of the project UIDB/05549/2020, UIDP/05549/2020 and LASI-LA/P/0104/2020.

The authors also acknowledge FCT, Portugal and the European Social Fund, European Union, for funding support through the "Programa Operacional Capital Humano" (POCH) in the scope of the PhD grants SFRH/BD/136721/2018 (B. Oliveira) and SFRH/BD/136670 (H. R. Torres).

M. R. Ferreira is with 2Ai – School of Technology, IPCA, Barcelos, Portugal and with Algoritmi Center, School of Engineering, University of Minho, Guimarães, Portugal (email: amrferreira@ipca.pt).

H. R. Torres and B. Oliveira are with 2Ai – School of Technology, IPCA, Barcelos, Portugal, with Algoritmi Center, School of Engineering, University of Minho, Guimarães, Portugal, with Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal, and with ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal (email: htorres@ipca.pt, boliveira@ipca.pt).

J. Gomes-Fonseca, P. Morais, and J. L. Vilaça are with 2Ai – School of Technology, IPCA, Barcelos, Portugal (email: pmorais@ipca.pt, jlfonseca@ipca.pt, jvilaca@ipca.pt).

P. Novais is with Algoritmi Center, School of Engineering, University of Minho, Guimarães, Portugal (email: pjon@di.uminho.pt).

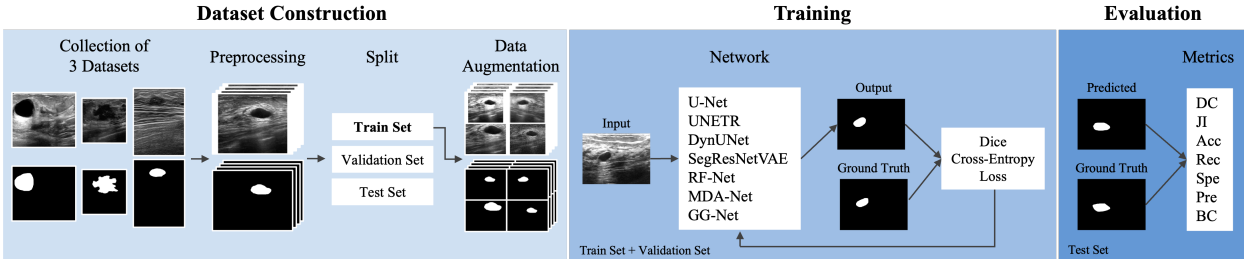

Figure 1. Graphical summary of the comparative study.

hampers the comparison of their performance. Thus, there is a need for a benchmark to compare state-of-the-art methods using a large and common dataset. In this study, we present a benchmark for breast lesion segmentation in 2D ultrasound images. Seven recent deep learning architectures were evaluated on a wide database composed of three publicly available datasets. By conducting comparative experiments, the strengths of the models can be analyzed and the networks that yield better results can be identified.

II. Methodology

A. Deep Learning Networks

Seven neural networks were selected to study the performance of deep-learning approaches for breast lesion segmentation. Fig. 1 shows a graphical summary of the comparative study. Specifically, the following methods were evaluated: U-Net [7], DynUNet [12], SegResNetVAE [13], UNETR [14], RF-Net [5], MDA-Net [15], and GG-Net [16]. U-Net and the models derived from it were selected to identify whether the recently developed models for BUS segmentation, namely RF-Net, MDA-Net, and GG-Net, offer advantages over general image segmentation strategies. UNETR was also studied to quantify the added value of a transformer network, a novel architecture that has shown promising results in other image segmentation tasks [14].

U-Net is one of the most used approaches in biomedical image segmentation. It has an encoder responsible for learning global contextual representations and a decoder dedicated to pixel/voxel-wise semantic prediction. Moreover, skip connections combine the encoder's and decoder's outputs at multiple resolutions, allowing the recovery of lost spatial information [7]. DynUNet is a dynamic implementation of U-Net, which automatically generates the decoder part to any given encoder. This model is more flexible since residual connection and deep supervision are supported [17]. SegResNetVAE is a residual U-Net with autoencoder regularization. Unlike U-Net, it has a new architecture for encoder blocks and a variational autoencoder branch in the decoder, which reconstructs the input and has a regularization effect in the presence of limited data [13]. The UNETR architecture is a Vision Transformer [18] generalization with multi-head self-attention to effectively capture the global multiscale information [14]. RF-Net, based on encoder-decoder architecture, learns residual representations of hardly-predicted pixels and feeds them into encoder blocks to increase the confidence of the hardly-predicted pixels [5]. MDA-Net contains a multiscale fusion block to overcome fixed receptive field difficulties and extract semantic feature maps to achieve greater variability [15]. GG-Net comprises a global guidance block that combines non-local features in both spatial and channel domains under the guidance of multi-layer integrated features to learn powerful contextual information [16].

B. Dataset

To perform the comparative study, a wide multi-center dataset was created. Here, three public BUS image datasets were considered. The first one is the BUSI dataset [19], from the Baheya Hospital for Early Detection and Treatment of Women's Cancer (Cairo, Egypt), which comprises 210 images with benign lesions, 437 with malignant lesions, and 133 without lesions. The images have an average size of 500×500, where lesions vary in size. The second dataset, UDIAT [20], contains 110 images with benign lesions and 53 with cancerous masses. This dataset was collected from the UDIAT Diagnostic Center of the Parc Taulí Corporation (Sabadell, Spain). It has a mean image size of 760×570, and most contain small tumors. The third dataset contains 42 BUS images of 128×128 pixels with unclassified lesions, provided by the Imaging Department of the First Affiliated Hospital of Shantou University [21]. All the datasets provided ground truth segmentation masks manually delineated by observers.

Since only the segmentation task is addressed in this study, only images with lesions were considered. Thus, a dataset with 852 BUS images was created, where 320 contain benign lesions, 490 include malignant masses, and 42 are unclassified lesions. These were randomly distributed and divided into groups of 682 (80%), 85 (10%), and 85 (10%) images for training, validation, and testing, respectively. The validation set is used to evaluate the models during training, while the testing set is used to compute the final performance of the models. Each method was assessed using the same dataset distribution, allowing the direct comparison of results.

C. Preprocessing

After combining the three datasets, preprocessing was executed to uniformize the images from the different sources. Here, since the original images have different sizes, a resizing technique followed by padding was applied to obtain a final image size of 384x512. The resizing was performed ensuring that the relation between the width and height of the image was maintained. Afterward, normalization was performed by converting the intensity values into the range of [0, 1].

III. Implementation Details

A. Training Procedure

Each network training was performed using the Adam optimizer with a learning rate of 1e-4 and a batch size of 10 for 5000 epochs. A weight-balanced loss function was adopted by combining the cross-entropy and Dice loss functions to

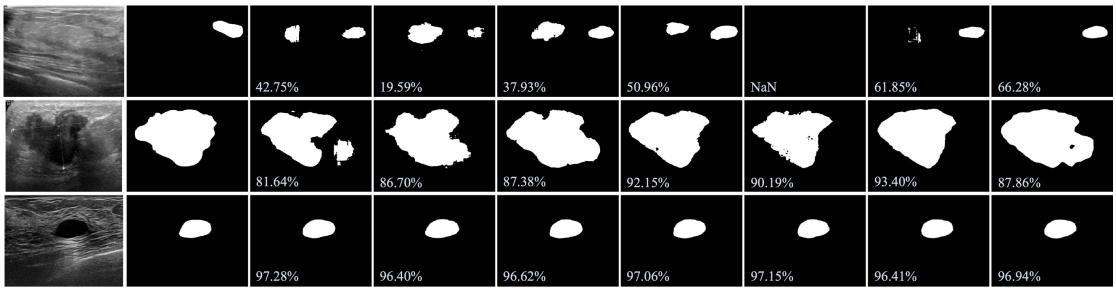

Figure 2. BUS image segmentation results. The first, second and third rows of images represent the original image and segmentation masks for the test images with the worst, the median, and the best segmentation quantitative results with a mean DC of 46.56%, 88.47% and 96.84%, correspondingly.

adapt to the various tumor sizes and overcome the imbalances between the lesion regions and background. Also, to avoid overfitting problems and improve performance, data augmentation techniques, including spatial and intensity-based transformations, were applied to training images. Spatial transformations included flips, zooming in, and grid distortion, whereas intensity transformations included intensity scaling and shifting, Gaussian noise and smooth addition, and contrast adjustment. Finally, the architectures were implemented in Python 3.9.7 and PyTorch 1.10.0 and executed using NVIDIA A100 with Ubuntu 20.04 operating system, CUDA 11.3, cuDNN 8.2, and 40GB of GPU memory.

B. Evaluation Metrics

Six widely used segmentation metrics are used to assess the performance of the models, namely the Dice coefficient (DC), Jaccard index (JI), accuracy (Acc), recall (Rec), specificity (Spe), and precision (Pre). DC and JI measure the similarity between the segmentation result and the ground truth, while Acc, Rec, Spe, and Pre compute pixel-wise classification accuracy [22]. Also, a second experiment was performed using a new metric, denominated "bad contours" (BC), where the unsuccessfully segmented cases are identified. Here, the number of results with a JI inferior to 25% was determined and used to compute the percentage of unsuccess accordingly.

IV. RESULTS

The test images were input into the models to obtain the predicted labels. The output labels and the ground truth segmentations were then used to compute the seven metrics and assess the performance of the networks. Table I summarizes the mean values of the calculated metrics for each network using the complete test set. Analyzing Table I, it is possible to identify that the GG-Net delivered the best results, reaching a DC of 82.56%, a JI of 74.95%, a Spe of 98.40%, and a Pre of 85.25%. However, SegResNetVAE outperformed the GG-Net in terms of Acc, Rec, and BC, obtaining 97.34%, 84.55%, and 5.88%, respectively. Concerning BC, we verified that the same 6 of the 85 test images obtained a JI lower than 25% for all networks. Therefore, these 6 images were excluded, to prevent biases in our analysis. The methods' performances computed using the successfully segmented 79 images are presented in Table II. Fig. 2 illustrates examples of bad, median, and good segmentation outputs.

Table I. QUANTITATIVE COMPARISON OF THE DIFFERENT METHODS.

Networks	Evaluation Metrics						
	DC	***JI***	***Acc***	***Rec***	***Spe***	***Pre***	***BC***
U-Net	76.9± 24.0	67.3± 25.6	96.7± 5.4	77.4± 25.5	97.8± 5.1	82.0± 23.1	8.2
UNETR	78.4± 25.0	69.7± 25.9	96.6± 5.7	79.4± 26.7	97.8± 5.1	81.7± 24.3	10.6
DynUNet	81.9± 23.4	74.1± 24.7	97.2± 5.0	83.5± 24.6	98.2± 4.6	83.5± 22.8	8.2
SRNetV[a]	81.9± 20.1	73.0± 22.1	**97.3± 4.4**	**84.6± 20.6**	98.1± 4.3	83.9± 21.1	**5.9**
RF-Net	76.2± 25.3	66.8± 26.4	96.4± 6.1	82.6± 32.2	96.4± 6.3	63.3± 33.7	22.4
MDA-Net	82.0± 23.0	74.0± 23.6	97.2± 5.1	83.9± 22.7	98.4± 4.4	84.8± 22.4	7.1
GG-Net	**82.6± 23.3**	**75.0± 24.1**	97.3± 5.4	83.6± 24.2	**98.4± 4.6**	**85.3± 21.9**	7.1

a. SRNetV refers to SegResNetVAE

V. DISCUSSION

This study proposed a comparative analysis of current deep learning networks for breast lesion segmentation in ultrasound imaging. U-Net presented one of the lowest performances as expected since most of the evaluated networks are extensions of this architecture that significantly boost performance. Moreover, UNETR delivered lower results compared to those shown in [14] for other imaging modalities. This may be due to the high degree of speckle noise present in the BUS images, and tumor variability in shape, size, texture, and location. To overcome these difficulties UNETR needs to be optimized when applied to BUS. DynUNet and SegResNetVAE, like UNETR, were not developed specifically for BUS, so their tuning can lead to better results. Concerning RF-Net, it revealed lower results than anticipated, potentially related to the usage of a multi-center dataset, hampering the network's fitting to the data. Thus, the network did not show the same robustness as it would when applied to a less variable dataset. MDA-Net achieved one of the best performances, as it implements attention mechanisms, which potentially enhanced the

Table II. QUANTITATIVE COMPARISON OF THE DIFFERENT METHODS, EXCLUDING IMAGES WITH A JI UNDER 25%.

Networks	Evaluation Metrics					
	DC	***JI***	***Acc***	***Rec***	***Spe***	***Pre***
U-Net	82.0± 15.7	72.0± 19.6	97.7± 3.5	81.4± 19.4	98.5± 3.3	86.5± 13.9
UNETR	83.5± 17.02	74.5± 19.7	97.6± 4.0	83.7± 20.2	98.6± 3.4	86.1± 16.2
DynUNet	87.0± 13.9	79.1± 17.1	98.3± 3.0	88.1± 15.9	98.9± 2.8	87.9± 13.9
SegResNetVAE	86.4± 11.0	77.5± 14.9	98.2± 2.8	87.6± 14.2	98.7± 3.0	87.9± 12.5
RF-Net	77.6± 23.8	68.1± 25.3	97.1± 5.0	88.1± 25.3	97.1± 5.3	67.4± 30.8
MDA-Net	87.7± 10.3	79.3± 14.0	98.3± 2.6	87.9± 14.1	99.2± 1.5	89.8± 10.1
GG-Net	**88.3± 10.8**	**80.3± 14.4**	**98.4± 3.0**	**88.8± 14.8**	**99.2± 2.1**	**90.1± 9.3**

segmentation results. GG-Net performance met expectations since it was developed more recently and specifically for BUS image segmentation. In particular, the authors considered that other approaches, such as U-Net, conducted convolutional operations in local regions to learn deep discriminative features, producing unsatisfactory results. Thus, they propose the integration of all CNN layers to produce multi-level integrated features as a global guidance block to supplement more breast lesion boundary details [16]. Since the best-performing methods implement attention mechanisms, we can infer that this has a significant impact on the breast lesion segmentation task, helping locate the regions of interest with clear boundaries and accurately segment the lesion.

Table II exhibits the mean values of the metrics when the 6 images that revealed a low JI are removed. Compared to Table I, the overall performances increased significantly, with GG-Net delivering the best performance, achieving 88.25%, 80.33%, 98.38%, 88.78%, 99.22%, and 90.07% of DC, JI, Acc, Rec, Spe, and Pre, respectively. Moreover, when analyzing the DC metric, statistically significant differences were found between the GG-Net and three of the studied networks, namely U-Net, UNETR, and RF-Net (p-value < 0.05 in a two-tailed paired t-test). To improve the results and to account for the variation in results on challenging images, a semi-automatic method that allows the specialist to accept or reject the outcome could be beneficial.

Analyzing Fig. 2, we can qualitatively evaluate the results. The first row of Fig. 2 shows a small benign mass, with unclear boundaries, resembling the background, which all methods failed to segment accurately. RF-Net was unable to detect a lesion, predicting an empty mask, unlike other methods that always locate a lesion. The second row shows a large malignant lesion with irregular boundaries, which is the median of the DC for all models. The third row shows a well-defined benign lesion with fewer shadows in the background, which could be why most of the evaluated methods achieved the highest DC (96.84%). Overall, the results corroborate the added value of deep learning strategies for the segmentation of challenging BUS images.

As future work, it is intended to go further in this study by comparing the importance of different modules, namely loss functions, optimizers, data augmentation, and adaptive learning rate strategies. Additionally, the performance comparison of deep learning strategies for breast lesion classification is also envisioned.

VI. Conclusion

A comparative analysis of seven state-of-the-art deep learning networks for the segmentation of breast lesions on 2D ultrasound imaging was developed. In summary, all networks demonstrated their potential to be used for the task of breast lesion segmentation. GG-Net and SegResNetVAE architectures showed to outperform the remaining networks. However, additional experiments should be conducted to improve the methods' comparison, namely studying different loss functions and hyperparameters.

References

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] Y. Hu *et al.*, "Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model," *Med. Phys.*, vol. 46, no. 1, pp. 215–228, 2019, doi: 10.1002/mp.13268.

[3] H. R. Torres *et al.*, "A review of image processing methods for fetal head and brain analysis in ultrasound images," *Comput. Methods Programs Biomed.*, vol. 215, p. 106629, 2022, doi: 10.1016/j.cmpb.2022.106629.

[4] P. Morais *et al.*, "Semiautomatic estimation of device size for left atrial appendage occlusion in 3-D TEE Images," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 66, no. 5, pp. 922–929, 2019, doi: 10.1109/TUFFC.2019.2903886.

[5] K. Wang, S. Liang, and Y. Zhang, "Residual Feedback Network for Breast Lesion Segmentation in Ultrasound Image," in *Lecture Notes in Computer Science*, vol. 12901 LNCS, 2021, pp. 471–481.

[6] Y. Jiménez-Gaona, M. J. Rodríguez-Álvarez, and V. Lakshminarayanan, "Deep-Learning-Based Computer-Aided Systems for Breast Cancer Imaging: A Critical Review," *Appl. Sci.*, vol. 10, no. 22, p. 8298, 2020, doi: 10.3390/app10228298.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*, 2015, pp. 234–241.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in *Computer Vision - ECCV 2016*, 2016, pp. 630–645.

[9] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv Prepr. arXiv1804.03999*, 2018.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[11] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Lecture Notes in Computer Science*, vol. 11045 LNCS, 2018, pp. 3–11.

[12] F. Isensee *et al.*, "nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation," 2018, doi: 10.1007/978-3-658-25326-4_7.

[13] A. Myronenko, "3D MRI Brain Tumor Segmentation Using Autoencoder Regularization," in *Lecture Notes in Computer Science*, 2019, pp. 311–320.

[14] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3D Medical Image Segmentation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1748–1758, doi: 10.1109/WACV51458.2022.00181.

[15] A. Iqbal and M. Sharif, "MDA-Net: Multiscale dual attention-based network for breast lesion segmentation using ultrasound images," *J. King Saud Univ. - Comput. Inf. Sci.*, 2021, doi: 10.1016/j.jksuci.2021.10.002.

[16] C. Xue *et al.*, "Global guidance network for breast lesion segmentation in ultrasound images," *Med. Image Anal.*, vol. 70, p. 101989, 2021, doi: 10.1016/j.media.2021.101989.

[17] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[18] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv Prepr. arXiv2010.11929*, 2020.

[19] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Br.*, vol. 28, p. 104863, 2020, doi: 10.1016/j.dib.2019.104863.

[20] M. H. Yap *et al.*, "Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 4, pp. 1218–1226, 2018, doi: 10.1109/JBHI.2017.2731873.

[21] Z. Zhuang, N. Li, A. N. J. Raj, V. G. V. Mahesh, and S. Qiu, "An RDAU-NET model for lesion segmentation in breast ultrasound images," *PLoS One*, vol. 14, no. 8, pp. 1–23, 2019, doi: 10.1371/journal.pone.0221535.

[22] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imaging*, vol. 15, no. 1, 2015, doi: 10.1186/s12880-015-0068-x.