Segmenting breast ultrasound scans using a generative adversarial network embedding U-Net

Abraham Etinosa Enobun[1], Uche Henry Anakwenze[1], Aboozar Taherkhani[1 [0000−0002−3627−6362]], Zacharias Anastassi[1 [0000−0001−9190−2816]], Fabio Caraffini[2 [0000−0001−9199−7368]], and Hassan Eshkiki[2 [0000−0001−7795−453X]]

[1] School of Computer Science and Informatics, De Montfort University, Leicester, UK
etinosa.enobun@gmail.com,p2688618@my365.dmu.ac.uk,
{aboozar.taherkhani,zacharias.anastassi}@dmu.ac.uk
[2] Department of Computer Science, Swansea University, Swansea SA1 8EN, UK
{fabio.caraffini,h.g.eshkiki}@swansea.ac.uk

Abstract. Breast ultrasound imaging, due to its noninvasive nature and cost-effectiveness, has become an indispensable instrument in the early detection of breast cancer, highlighting the importance of early detection of lesions for timely intervention. In this study, we discuss possible problems deriving from using deep learning techniques on such images and propose novel solutions towards achieving a segmentation tool based on a generative adversarial network architecture. As a proof-of-concept, we build on existing methods to develop our system by modifying a U-Net known as Residual-Dilated-Attention-Gate with the addition of skip modules and dilated convolutional neural networks after the decoder stage. Compared with other state-of-the-art methods in established evaluation metrics, the results indicate that the proposed model achieves the highest accuracy of 98.11%, despite being trained on a limited number of epochs. However, it still requires further tuning and optimisation to enhance precision, ensuring that it is more balanced, robust, and thus competitive with the state-of-the-art.

Keywords: generative adversarial network · U-Net, · dilated convolution · breast ultrasound

1 Introduction

Breast cancer has a high global mortality rate. For instance, in China, iter constitutes 7.82% of the total mortality associated with female malignant tumours, establishing itself as one of the most lethal diseases.

Patients diagnosed with metastatic breast cancer typically face a poor prognosis, characterised by an average 5-year survival rate of approximately 27%. Upon metastasis (cancer that has spread to other parts of the body), the malignancy often progresses to a more severe tumour stage [19]. Although there have been improvements in the methods used to detect breast cancer, patients diagnosed with metastatic breast cancer still tend to have poor outcomes. This

poor prognosis is largely due to the fact that cancer is often diagnosed at a later stage, when it is more visible and painful, but also more difficult to treat effectively. This situation underscores the critical need for early detection and diagnosis for timely intervention and risk mitigation. The potential benefits of using AI to aid in these processes are numerous.

The prospective advantages of employing artificial intelligence in these diagnostic and therapeutic processes are numerous, given advances in the manipulation of medical images with AI-driven solutions in the last decades [13], [3]. Note that non-invasive breast cancer diagnosis modalities include X-ray, and Magnetic Resonance Imaging (MRI), etc., which results in imagery data. Among these diagnostic modalities, Breast Ultrasound (BUS) imaging, due to its noninvasive nature, absence of ionising radiation, and cost effectiveness, has emerged as an indispensable instrument in the early detection of breast cancer [5].

BUS segmentation facilitates the precise identification and analysis of tumours, thereby increasing diagnostic accuracy. Although segmentation methodologies in various imaging modalities, such as MRI and computed tomography (CT), often employ analogous techniques, BUS segmentation (i.e., the extraction of the tumour region from the image) presents significant challenges. These challenges are primarily attributed to the inherently low quality of ultrasound images, which are marred by speckle noise and low contrast.

In this project, we focus on tumour segmentation in BUS imaging by modifying a GAN to employ U-Net within its architecture. We use dilated convolution to enlarge the receptive field after multiple down-sampling in the encoder and decoder, therefore boosting the classifier's accuracy. In addition, the model uses the residual block which replaces the basic neural units and an attention mechanism.

2 Background and Design Motivations

Segmentation constitutes a pivotal component of image analysis in the context of breast cancer diagnosis, encompassing critical processes such as detection, feature extraction, classification, and treatment. Nevertheless, BUS images are characteristically low-resolution and monochromatic, in contrast to other imaging modalities. Hence, ROIs of malignant regions are often uneven in shape, blurred, and have an indistinct border [9].

Different Deep Learning (DL) methods have been proposed to address the aforementioned challenges in the segmentation of BUS images, from semantic segmentation models such as SegNet [20] to multiple kinds of GANs [8, 18, 16]. For example, the study in [17] introduces an improved conditional GAN segmentation algorithm to learn tumour features, which employs an atrous convolution layer. To equalise the influence of high-level encoded characteristics, they adopted a channel-wise weighting block. The model undergoes training using a composite loss function comprised of the Structural Similarity Index (SSIM), the L1-norm, and adversarial loss. An interesting approach proposed in [1] uses the You-Only-Look-Once (YOLO) model [14]. They segment large datasets using

full-resolution convolutional networks (FrCN), and a deep convolutional neural network (CNN) determines whether the mass is benign or cancerous. Notably, the DeepLabv3+ semantic segmentation model from [4] is based on an encoder-decoder architecture, wherein the encoder is responsible for extracting both shallow and high-level features, and the decoder integrates these low-level and high-level features to enhance segmentation accuracy. DeepLabv3+ leverages ResNet architectures as its foundational backbone, integrating Atrous Convolution and the Atrous Spatial Pyramid Pooling (ASPP) module. The ASPP module encompasses a global average pooling operation and convolutional layers with varying dilation rates (specifically 1, 6, 12, and 18).

The U-Net model is another widely recognised and favoured approach for mammogram image segmentation. ERU-Net has U-shaped architecture, is designed like an auto-encoder. It contains two pathes, an encoding path (contracting) and a decoding path (expanding). Its capability in training on a relatively limited dataset of annotated images, coupled with the capabilities of high-performance GPU computing, renders it a viable and efficient option for this application [6]. In [10], a novel deeply supervised U-Net model (DS U-Net) integrated with dense conditional random fields (CRFs) is introduced, whereas [23] delineates a Residual-Dilated-Attention-Gate-UNet (RDAU-NET) derived from U-Net. This model substitutes neural units with residual units and incorporates an attention gate (AG) to enhance edge delineation and mitigate network performance degradation issues. These examples show the success and versatility of U-Net.

In this context, we investigate the use of a hybrid system that employs the Wasserstein Generative Adversarial Network [2] to perform the segmentation task. This variant ensures robust convergence and minimises the Wasserstein distance between real and generated data distributions. We used RDA-NET as the generator within the GAN architecture, as shown in Figure 1. It should be noted that the RDA-NET employed, an improved variant of the fundamental U-Net architecture referenced in [11], incorporates specific modifications to rectify the limitations inherent in the U-Net model and increase the efficacy of the generative framework. The generator aims to produce data with the same distribution as the original to deceive the discriminator, and these generated data will be lesion-segmented maps of Breast Ultrasound Images in our system. With reference to Figure 1, it can be observed a conventional encoder-decoder architecture. The input is subjected to successive down-sampling stages until it attains a bottleneck layer after which the process is inverted. Within this architecture, information traverses each hierarchical level, encompassing the bottleneck, to effectively capture both high-level and low-level features from the input data. It is ideal to transfer this data directly through the network because many image translation tasks share a significant amount of low-level information between the input and output. To provide the Generator with a way to bypass the bottleneck, we included skip-connections between layers of the same size in the encoder and decoder.

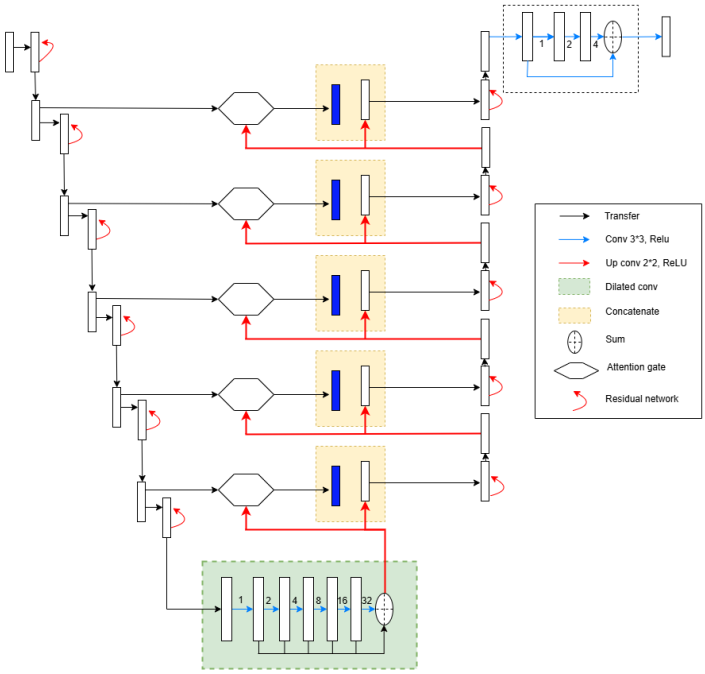

Fig. 1. RDA-U-Net architecture.

We implemented additional changes to the original architecture. Specifically, the original six neural units along the encoder path are replaced with six residual units, employed to prevent accuracy saturation (vanishing gradients) during training. However, smaller feature maps can reduce the accuracy of semantic segmentation. To address this, the outputs of the encoder pipeline are fed into a series of dilated convolution modules. These modules use 3×3 convolution kernels and have dilation ratios of $1, 2, 4, 8, 16$, and 32, respectively. The feature maps from the dilated convolution are summed, forming the output of the dilated convolution which is added to the architecture to broaden the receptive field. The output from this module is directed into the decoder pipeline, which consists of an up-sampling mechanism and five residual networks. Each layer within the decoder facilitates the upsampling process by integrating the intricate feature outputs derived from the decoder with the corresponding semantic information procured from the encoder. In contrast to the traditional U-Net's skip-connection components, we replaced the cropping and copying units with attention-gate modules (one for each residual net in the decoder), thus enabling the model to focus more on the lesion region and less on the unnecessary background.

Note that convolution and pooling in CNNs reduce spatial information, affecting segmentation accuracy. A Fully Convolutional Network (FCN) executes

these operations to reduce the spatial dimensions of the image and extract abstract features, subsequently enlarging this output through upsampling. The convolution-pooling phase can be conceptualised as a downsampling operation that may induce information loss, thereby compromising the accuracy of the process and substantially diminishing the transferability of data details.

Since the U-Net encoder is a Fully Connected CNN (FC-CNN), dilated convolutions are used to insert 'holes' (i.e., zeros) into the kernel to achieve a larger receptive field than traditional convolution without a loss of resolution [22]. By design, they are suitable for dense prediction tasks, differing structurally from image classification by computing a label for each pixel. Their use enables us to obtain high-level multiscale contextual information while reducing the number of parameters and computational costs while performing segmentation.

The dilated convolution operator is characterised by a hyperparameter known as the dilation rate, which specifies the extent to which the kernel intervals are expanded[3]. Figure 2 (a) and (b) illustrate the visual field of a 3×3 convolution kernel with $r = 1$ and $r = 2$, respectively. When $r = 2$, the receptive field increases to 7×7 (shown as the orange and blue parts in (b)) compared to traditional convolution ($r = 1$, as shown in the blue part of (a)). Therefore, the dilation process increases the size of the receptive field and compensates for the subsampling.

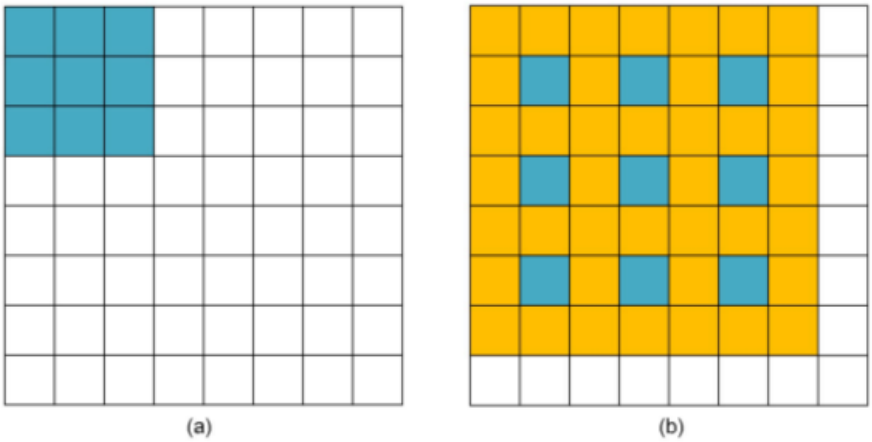

Fig. 2. Illustration of receptive field for $r = 1$ and $r = 2$ [23].

In the RDAU-NET model, the feature maps of size 4×4 obtained at the end of the encoder pipeline are fed into a series of dilated convolution modules with $r = 1, 2, 4, 8, 16, 32$ and $N = 3 \times 3$, 7×7, 15×15, 31×31, 63×63 and 127×127 respectively. The outputs of the six convolutions are added, upsampled (by a

[3] A dilation rate of 1 results in a classic convolution.

factor of 2), and then fed into the decoder pipeline as shown in Figure 1. In the dilated convolution module, output feature maps match input sizes but capture information from a wide range of receptive fields, enhancing feature learning.

The proposed system incorporates a greater number of layers to enhance learning capability. Given that this may result in decelerated or halted learning, attributed to the phenomenon known as the 'vanishing gradient', we implement the residual learning correction technique as delineated in [7] to sustain the efficacy of gradient updates throughout the training process. Furthermore, to address common CNN issues like reduced spatial awareness from shared weights and redundant channels in U-Net-like networks [15], we add an attention module in the skip connection and concatenate low-level and high-level features to emphasise relevant channels and suppress irrelevant ones. The inclusion of attention modules in our model is motivated by successful studies, such as the one presented in [12], which integrates an Attention Gate (AG) module into a U-Net framework to facilitate spatial location and subsequent segmentation.

To enhance stability, we have selecetd a pre-trained model from [11] and used Adam to optimise the loss function with a learning rate of $1 - e4$.

3 Resources & Methods

3.1 Dataset

The experimental dataset is the BUS images in [21], containing a total of 645 low-contrast breast ultrasound images with evenly distributed benign and malignant lesions. Samples in this dataset are partitioned into training, validation, and test subsets. The training and validation subsets comprised 538 and 50 samples, respectively. The test subset included 57 samples, each with corresponding ground truth masks. The model's segmentation performance was evaluated using the test subset. Figure 3 illustrates several sample images along with their ground truth annotations.

3.2 Training the Model

The training procedure is delineated in Algorithm 1. It consists of iteratively alternating training phases for the discriminator and the generator over e epochs until the total allowed number of epochs is reached.

Initially, the discriminator is rendered trainable, whereas the generator remains untrainable. The generator is employed to produce image predictions, enabling the discriminator to classify these images and subsequently update its parameters accordingly. Conversely, the second phase is performed similarly but with an untenable discriminator and a trainable generator. This completes one training iteration and multiple iterations are performed according to the prefixed computation budget (total number of epochs). Following each training iteration, the model undergoes evaluation on the validation dataset. With each successive cycle, the segmentation accuracy in relation to the ground truth

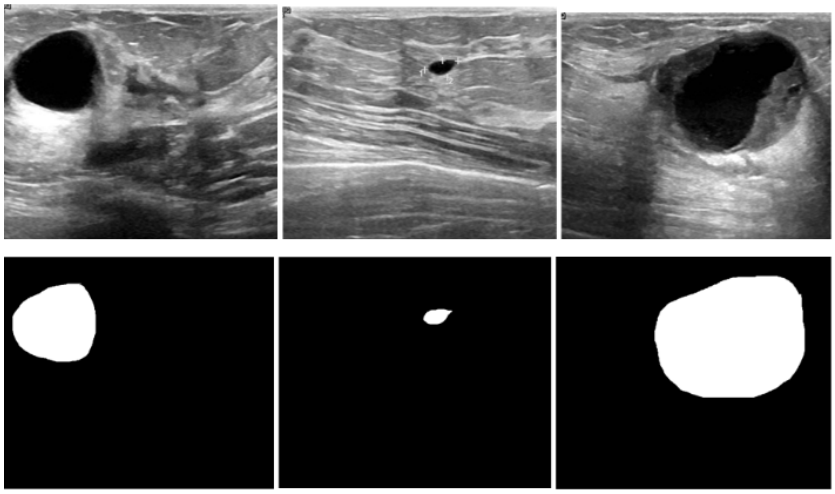

Fig. 3. Sample images from the dataset with their corresponding ground truth directly under

is expected to demonstrate improvement. Optimal performances are achieved when the discriminator's accuracy asymptotically approaches 50%. Indeed, as the generator's performance enhances (i.e., it produces increasingly realistic images), the discriminator's efficacy deteriorates, because it becomes incapable of differentiating between authentic and synthetic data.

Algorithm 1 Training process

Fetch X ▷ Training set of BUS images
Initialise e ▷ Number of epochs ($e = 50$)
Initialise n ▷ Batch size ($n = 32$)
$\sigma = \frac{|X|}{n}$ ▷ Steps per epoch
for e times **do**
    for σ times **do**
        Make the discriminator trainable and the generator untrainable
        Use the Generator to predict an image
        Prepare batches and train the discriminator
    end for
    for σ times **do**
        Make the discriminator untrainable and the generator trainable
        Train the Generator
    end for
end for
return Trained GAN model

3.3 Experimental Setup

Owing to computational constraints, the experimental phase is conducted with a limitation of 50 epochs to train the model. The results are compared with those derived from state-of-the-art methodologies. Specifically, we have chosen the U-Net, SegNet, and RDAU-NET models for comparative analysis. The input images for all models are standardised to a resolution of 128×128, and the segmentation outputs are generated at the same resolution.

3.4 Results

Consistent with the design principles articulated in Section 2, we refer to our model as the RDA-NET-GAN. The segmentation outcomes obtained with the setup in Section 3.3 are evaluated using established metrics, as detailed in Table 1.

Model	Loss	Acc	Dice	Precision	Sensitivity	Specificity	M-IOU	F1
U-Net	17.95	97.57	82.04	81.85	84.66	98.91	79.83	82.11
SEGNET	18.29	97.52	81.70	81.41	83.95	98.83	79.14	81.71
RDAU-NET	15.30	97.91	84.69	88.58	83.19	99.34	80.67	84.78
RDA-NET-GAN	25.03	98.11	85.84	84.78	75.14	99.07	79.97	66.16

Table 1. Segmentation performance of models across multiple evaluation metrics.

Overall, RDAU-NET appears to offer the most balanced and robust performance across multiple metrics, while RDA-NET-GAN, despite a high accuracy of 98.11, may need further tuning to improve sensitivity and M-IOU. U-Net and SEGNET offer decent performance but are outperformed by RDAU-NET in key areas. The fact that our model improves the accuracy as well as perceived qualitative analysis of the segmented images (show an example of the results in figure 4) shows that the idea behind the algorithm is promising, even though it required more investigation to make it competitive with RDAU-NET.

The proposed model has more loss as this is a combined system and the adversarial losses is from the combination of the generator and the discriminator. Below are the outputs while testing the segmentation performance of the proposed model during training.

The suboptimal precision observed for the proposed model can be attributed to the limited number of epochs for which the model was trained, especially in comparison to other models. To a certain degree, this result was foreseeable given the low number of epochs used during the training phase. Strategies to increase the number of epochs without precipitating overfitting will be a focal point of future research activities.

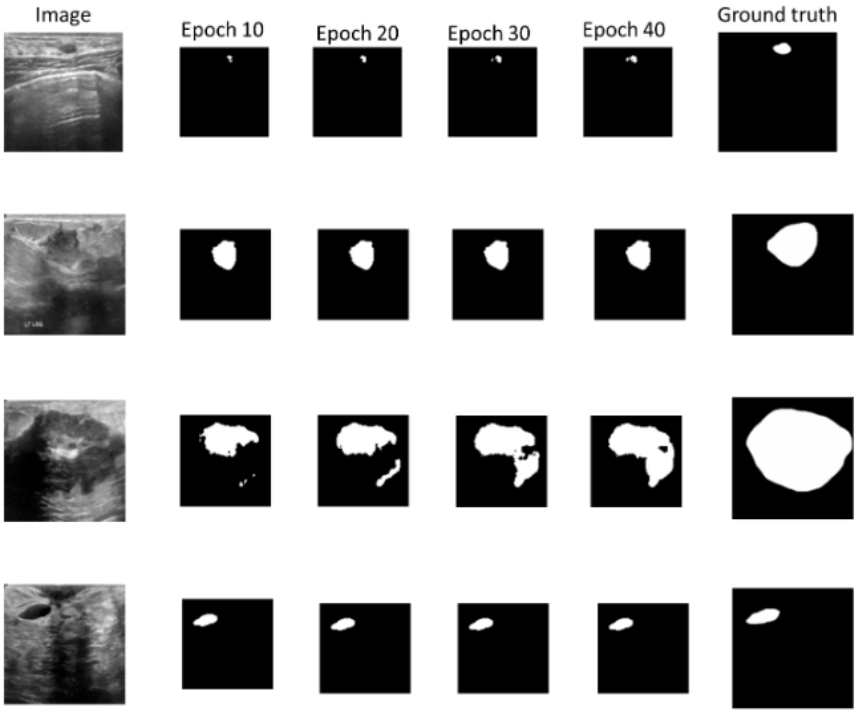

Fig. 4. Segmentation results of the proposed model

4 Conclusion

This study summarises relevant problems and literature gaps in using GAN for segmentation tasks and proposes ways to overcome these problems. When applied to the BUS images at dashed, it shows good visual results and displays the highest accuracy. It can be inferred that the GAN architecture holds significant promise for the segmentation of noisy datasets, and our proof-of-concept study indicates substantial potential for future advancements. Specifically, the incorporation of dilated CNNs after the decoder stage represents a novel methodology for the segmentation of breast lesions. This approach enhances the receptive field, thereby increasing the accuracy compared to directly applying the U-Net architecture. By comparing the model with the state-of-the-art, we are aware that there is room for improvement and that the model is not yet competitive with established methods such as RDAU-NET which is more rosubst and performing across various metrics, while our model performs very poorly in terms of pr precision. This aspect needs to be significantly improved.

During the training phase, the vanishing gradient problem manifested despite the algorithmic design precautions implemented to address the complexities inherent in the deep learning architecture of the proposed method. Therefore, the

model requires further optimisation to be more competitive with the state-of-the-art.

It is worth noting that despite the suboptimal precision of the model, this outcome was somehow expected due to the limited number of epochs employed. This also demonstrates that acceptable results can still be achieved with a reduced number of training iterations (which can be advantageous in preventing overfitting) and with low-resolution images.

Next, the model will be subjected to additional optimisation to enhance its precision. Furthermore, an extended and more rigorous training phase will be conducted and we will apply this model to various datasets, thereby validating its segmentation capabilities. Additionally, considering that speckle noise is an intrinsic characteristic of ultrasound images, we will examine the impact of filtering techniques to refine our segmentation pipeline.

References

1. Al-antari, M.A., Al-masni, M.A., Choi, M.T., Han, S.M., Kim, T.S.: A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification. International Journal of Medical Informatics **117**, 44–54 (2018). https://doi.org/https://doi.org/10.1016/j.ijmedinf.2018.06.003, https://www.sciencedirect.com/science/article/pii/S1386505618302880
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
3. Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N.C., Sardanelli, F.: Ai applications to medical images: From machine learning to deep learning. Physica Medica **83**, 9–24 (2021). https://doi.org/https://doi.org/10.1016/j.ejmp.2021.02.006, https://www.sciencedirect.com/science/article/pii/S1120179721000946
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation (2018)
5. Cheng, H., et al.: Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recognition **43**(1) (2010), https://doi.org/10.1016/j.patcog.2009.05.012
6. Harrison, P., Michael, E., Ma, H., Li, H., Kulwa, F., Li, J.: Breast cancer segmentation methods: Current status and future potentials. BioMed Research International **2021**, 9962109 (2021). https://doi.org/10.1155/2021/9962109, https://doi.org/10.1155/2021/9962109
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
8. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019)
9. Kelly, K.M., Dean, J., Comulada, W.S., Lee, S.J.: Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts. European Radiology **20**(3), 734–742 (2010). https://doi.org/10.1007/s00330-009-1588-y, https://doi.org/10.1007/s00330-009-1588-y

10. N, R.R., R, V., N, E., Ramesh, N.: Deeply supervised u-net for mass segmentation in digital mammograms. International Journal of Imaging Systems and Technology **31**, 59 – 71 (2020), https://api.semanticscholar.org/CorpusID:228916143
11. Negi, A., Raj, A.N.J., Nersisson, R., Zhuang, Z., Murugappan, M.: Rda-unet-wgan: An accurate breast ultrasound lesion segmentation using wasserstein generative adversarial networks. Arabian Journal for Science and Engineering **45**(8), 6399–6410 (2020). https://doi.org/10.1007/s13369-020-04480-z, https://doi.org/10.1007/s13369-020-04480-z
12. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas (2018)
13. Rakic, M., Wong, H.E., Ortiz, J.J.G., Cimini, B., Guttag, J., Dalca, A.V.: Tyche: Stochastic in-context learning for medical image segmentation (2024)
14. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. CoRR **abs/1506.02640** (2015), http://arxiv.org/abs/1506.02640
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR **abs/1505.04597** (2015), http://arxiv.org/abs/1505.04597
16. Saffari, N., Rashwan, H.A., Abdel-Nasser, M., Kumar Singh, V., Arenas, M., Mangina, E., Herrera, B., Puig, D.: Fully automated breast density segmentation and classification using deep learning. Diagnostics (Basel, Switzerland) **10**(11), 988 (2020). https://doi.org/10.3390/diagnostics10110988
17. Singh, V.K., Rashwan, H.A., Abdel-Nasser, M., Sarker, M.M.K., Akram, F., Pandey, N., Romani, S., Puig, D.: An efficient solution for breast tumor segmentation and classification in ultrasound images using deep adversarial learning (2019)
18. Singh, V.K., Rashwan, H.A., Romani, S., Akram, F., Pandey, N., Sarker, M.M.K., Saleh, A., Arenas, M., Arquez, M., Puig, D., Torrents-Barrena, J.: Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network. Expert Syst. Appl. **139**(C) (jan 2020). https://doi.org/10.1016/j.eswa.2019.112855, https://doi.org/10.1016/j.eswa.2019.112855
19. Tashk, A., Hopp, T., Ruiter, N.v.: An innovative practical automatic segmentation of ultrasound computer tomography images acquired from usct system. Iranian Journal of Science and Technology - Transactions of Electrical Engineering **43**(2) (2019), https://doi.org/10.1007/s40998-018-0098-9
20. Vianna, P., Farias, R., de Albuquerque Pereira, W.C.: U-net and segnet performances on lesion segmentation of breast ultrasonography images. Research on Biomedical Engineering **37**, 171–179 (2021)
21. Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., Marti, R.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics **22**(4), 1218–1226 (2017)
22. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions (2016)
23. Zhuang, Z., Li, N., Joseph Raj, A.N., Mahesh, V.G., Qiu, S.: An rdau-net model for lesion segmentation in breast ultrasound images. PloS one **14**(8), e0221535 (2019)