A two-stage progressive deep segmentation network for tumor detection in breast ultrasound images

Nadeem Zaidkilani*[a,*], Mohamed Abdel-Nasser[a,b], Miguel Angel Garcia[c], Domenec Puig[a]

[a]*Department of Computer Engineering and Mathematics, University Rovira i Virgili, Tarragona, 43007, Spain*
[b]*Electrical Engineering Department, Faculty of Engineering, Aswan University, Aswan, 81542, Egypt*
[c]*Department of Electronic and Communications Technology, Autonomous University of Madrid, Madrid, Spain*

Abstract

Segmenting tumorous regions in breast ultrasound images is a challenging problem due to several factors, including the relatively low contrast of the available images, the presence of speckle noise, and the considerable variations in breast mass sizes and shapes. Current methods are not precise enough and prone to misdetections. An efficient deep neural model is proposed for automatically segmenting tumorous regions in breast ultrasound images. The model is constituted by two consecutive encoder-decoder (autoencoder) networks. The first autoencoder extracts a preliminary binary mask from the given image. The second autoencoder refines that mask after concatenating it with the original image. The encoders within each autoencoder can be defined by applying any state-of-the-art network. In addition, cost-sensitive learning has been used in order to focalize training on the segmentation errors of the minority class (tumor). Semantic segmentation based on advanced deep learning methods is thus applied in order to enhance tumor segmentation in breast ultrasound images. The proposed model offers advanced capabilities for automated segmentation with the aim of helping physicians identify and diagnose tumors using state-of-the-art techniques. This model outperforms recent tumor segmentation methods in the experi-

*Corresponding Author: Nadeem Zaidkilani
E-mail: nadeem.zaidkilani@estudiants.urv.cat, nadimkilani@gmail.com

ments conducted on two public datasets of breast ultrasound images (UDIAT and BUSI). The largest improvement for both datasets was achieved by using CoAtNet as baseline model (Dice index equal to 84.49% and 78.94%, respectively)

Keywords: Breast Cancer, Image Segmentation, Deep Neural Networks, Cost Sensitive Learning, Compound Loss Functions

1. Introduction

Breast cancer is one of the most prevalent types of cancer affecting women worldwide. Ultrasound imaging has gained significant attention due to a number of advantages, including the absence of faster imaging capabilities, the avoidance of radiation, its enhanced accuracy, sensitivity and reduced cost [1]. However, studies indicate that radiologists can potentially misdiagnose breast cancer due to the vast volume of ultrasound images generated on a daily basis. Additionally, there is a limited availability of radiologists able to analyze these medical images [2], and manual inspection is also tedious, subjective, and prone to errors.

With the rapid advancement of artificial intelligence (AI), computer-aided diagnosis (CAD) systems have become a prominent research field in modern medical imaging. These systems have been shown to help radiologists achieve precise diagnoses of breast tumors. The European Commission already provides some clear guidelines for the use of medical AI, and knowledge about it is essential for researchers in this field [3]. Traditional breast tumor classification methods follow a two-step approach, starting with a feature extractor that captures relevant characteristics, and followed by a feature classifier that utilizes these extracted features for tumor classification [4]. The extracted features are specifically designed or handcrafted. They are mainly based on texture and morphological features [5].

Breast mass segmentation plays a crucial role in CAD systems as it allows for more precise analysis of features associated with the shape of breast masses. However, automatic segmentation of Breast Ultra Sound (BUS) images presents challenges due to several factors. They include the relatively low contrast of the available images, the presence of speckle noise, and the considerable variations in breast mass sizes and shapes. These factors make the automatic segmentation of BUS images a complex task [6, 7].

A typical CAD system for breast ultrasound examination consists of four primary stages: image preprocessing, breast lesion classification, segmentation, and image feature extraction [8]. Leaving aside that the legal aspects of such preprocessing activities must be addressed [9], image segmentation is a crucial step that plays a vital role in identifying tumorous regions from the surrounding healthy tissue. This process is essential for accurate diagnosis in subsequent stages. The significance of image segmentation also holds true for low-dose computed tomography (CT) images [10]. However, the task of segmenting BUS images keeps being an open and complex problem. This is primarily due to the presence of various ultrasound artifacts introduced during the imaging process. Blurred boundaries, speckle noise, low signal-to-noise ratio, reduced contrast, and intensity variations are some of the common artifacts found in BUS images. Furthermore, the appearance of benign and malignant tumors in clinical BUS images can differ, and radiologists may interpret these images in different ways. These factors contribute to the overall difficulty of BUS image segmentation. Additionally, the quality of BUS images is strongly influenced by the specific type of ultrasonic device used to capture them. Recently, researchers have been applying deep neural networks for devising more reliable CAD systems with fewer false-positive rates [11]. Notwithstanding, these techniques still exhibit some limitations, especially regarding the lack of datasets, which makes their clinical applicability more complicated. B-Mode BUS based CAD systems have also been developed to address the variability in diagnoses among radiologists, effectively improving the performance of breast cancer diagnosis [2]. Automatically differentiating tumorous regions from healthy tissue in BUS images is an essential aspect of BUS CAD systems. By doing so, traditional subjective tumor assessments become operator-independent, reproducible, and more precise.

Figure 1 shows various examples of BUS images and their corresponding ground-truth segmentations containing either a benign tumor (top rows) or a malignant tumor (bottom rows). Both the morphology and texture of a tumor significantly vary from the surrounding healthy tissue and strongly depend on its malignancy. In addition to some of the aforementioned artifacts of BUS images, we can notice different regions with similar area and saliency values, different size, location brightness, and contrast.

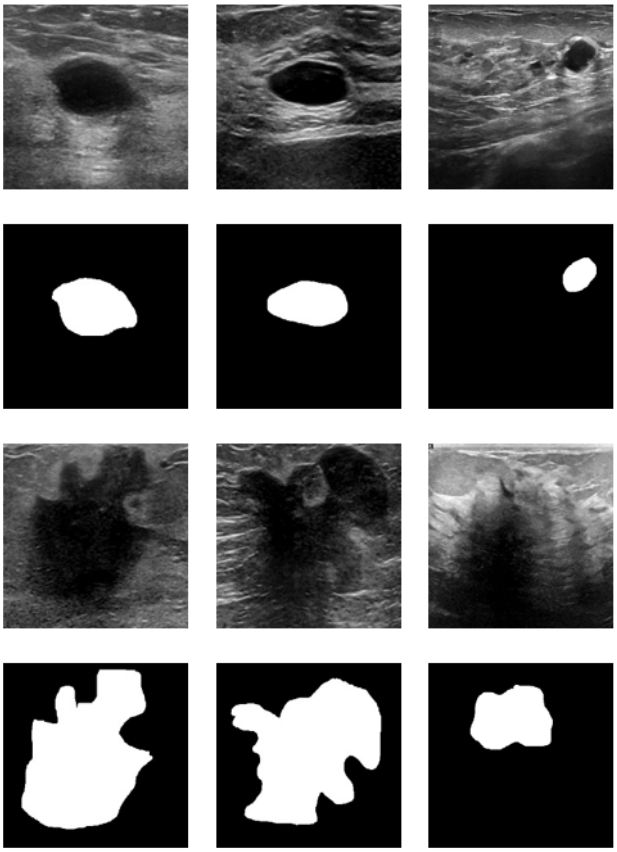

Figure 1: Examples of BUS images and corresponding ground-truth segmentations containing either benign (top rows) or malignant (bottom rows) tumors of different size, contrast and location.

This paper proposes a novel image segmentation model based on deep neural networks specifically tailored to the detection of tumors in BUS images. The model is constituted by two consecutive encoder-decoder (autoencoder) networks that are trained together. The first autoencoder extracts a preliminary binary mask from the given image. The second autoencoder refines that mask after concatenating it with the original image. The encoders within each autoencoder can be defined by applying any state-of-the-art deep network. In addition, cost-sensitive learning has been considered in order to focalize training on the segmentation errors of the minority class (tumor). The main contributions of the paper are summarized below:

- We propose a parallel refinement (local) network that enhances the segmentation mask generated by an initial (global) network.
- We investigate the application of cost-sensitive learning based on the Tversky loss and the Focal loss functions, determining their optimal hyperparameters.
- We find the optimal combination of loss functions associated with the global and local networks for each dataset.
- We apply state-of-the-art networks (CoAtNet, ConvNext, ...) to the segmentation of ultrasound images. We also propose corresponding decoders for these networks, as they were originally designed for classification tasks.

The paper is structured as follows. A comprehensive review of relevant previous work is given in Section 2. Section 3 presents a detailed description of the proposed model, including image preprocessing, network architecture, training parameters, evaluation measures, and alternative methods. Section 4 presents the experimental validation and a discussion of the attained results. Finally, Section 5 summarizes the contributions of the paper and proposes future research lines.

2. Related work

Different approaches based on deep learning have been proposed for breast mass segmentation so far. Among them, Zaidkilani et al. [12] presented a CAD system for breast ultrasound images utilizing a two-stage process with an encoder-decoder network for tumor segmentation, and fine-tuned MobileNetv2 for classifying benign and malignant tumors. Yap et al. [13] successfully utilized transfer learning with Fully Convolutional Networks (FCN) to achieve a high level of automatic breast mass segmentation. Similarly, Xu et al. [14] investigated the application of U-Net and FCNs for breast mass segmentation. They studied the impact of dilated convolutions at deeper layers of FCNs.

Han et al. [15] introduced a semi-supervised approach based on Generative Adversarial Networks (GAN) to enhance the accuracy of the segmentation maps generated by a FCN. In turn, Daoud et al. [16] proposed a superpixel-based method for lesion segmentation in BUS images. Their

two-stage pipeline decomposes the image into coarse hyper-pixels to obtain the initial tumor contour. The latter is then refined into super-pixels for improved segmentation accuracy. Zhuang [17] proposed RDAU-NET (Residual-Dilated-Attention-Gate-UNet) for tumor segmentation in BUS images. It employs residual units instead of plain neural units in the U-Net architecture to enhance edge information and mitigate performance degradation. In turn, Deng et al. [18] proposed a GAN-based approach that leveraged domain transfer learning to enhance medical images. The authors incorporated a new optimizer to generate a sample space with a continuous distribution.

Byra et al. [19] developed a selective kernel (SK) U-Net convolutional neural network, incorporating attention mechanisms to adjust the network's receptive fields, as well as fusing feature maps extracted using conventional convolutions and dilated convolutions. Zhou et al. [20] proposed LAED-Net, a lightweight encoder-decoder network for automatic ultrasound image segmentation. LAEDNet utilizes a lightweight version of EfficientNet as encoder and a Lightweight Residual Squeeze-and-Excitation (LRSE) block as decoder.

Furthermore, Tang et al. [21] introduced the MGCC framework, emphasizing semi-supervised learning for medical image segmentation. The framework tackles challenges related to collecting unlabeled medical data by leveraging synthetic medical images generated through a latent diffusion model (LDM) as unlabeled data.

Yang et al. [22] proposed a breast lesion segmentation model that combines a CNN and a Swin Transformer, utilizing a pyramid structure network for feature extraction. Key components include an interactive channel attention (ICA) module designed to emphasize crucial features. Furthermore, they introduced a supplementary feature fusion (SFF) module aimed at enhancing segmentation performance and a boundary detection (BD) module for refining lesion boundaries. By integrating CNN for feature localization and Swin Transformer for global feature extraction, along with the feature pyramid structure, the model ensures effective multiscale feature fusion.

Similarly, Ahmed et al. [23] introduced COMA-Net, a deep convolutional neural network designed for generalized medical image segmentation. The network includes dual ResNet34 backbones as encoders, two feature refinement modules (PRM and NRM), and a single decoder. Notably, a feature shifting operation on three crucial encoder feature tensors, followed by a guided process in PRM and NRM, enhances the Foreground-to-Background Ratio (FBR) via addition/subtraction. Refined features then guide the de-

coder to produce a robust segmentation map.

Additionally, Song et al. [24] introduced an innovative framework for computer-aided-diagnosis in breast ultrasound, aiming to overcome challenges prevalent in existing methods, such as dependence on masks and complex preprocessing. Employing a combination of supervised and unsupervised learning, the framework utilizes independent networks with progressively doubled feature volumes and convolutional autoencoders. This hybridization approach, tailored for diverse automated representations, seeks to address the limitations of previous methodologies.

Following a different approach, Taheri et al. [25] presented EMFSG-Net, an ensemble meta-feature space generator for breast ultrasound image classification. The methodology begins with a transfer learning approach to derive initial features from raw images. Subsequently, the application of ensemble methods generates a more efficient feature space. This approach explicitly addresses the challenges of bias and variance errors, common in ensemble methods. Additionally, by mitigating overfitting issues, the method significantly enhances classification performance through the extraction of more suitable features.

Moreover, Zhang et al. [26] addressed the limitations of feature loss and insufficient multi-scale contextual information extraction in deep learning-based medical image segmentation, particularly due to pooling operations during the encoding stage. Based on the U-shaped network (UNet), the proposed Multi-Scale Dilation Attention Network (MSDANet) is introduced. MSDANet comprises a Parallel Dilation Pooling module (PD), a Multi-Scale Channel Attention mechanism (MSCA), a Multilayer Perceptron Squeeze and Excitation module (MSE), a Big Kernel Convolution module (BC), and a Skip Feature Pyramid module (SFP). The PD minimizes the loss of subtle features during downsampling, while MSCA, MSE, and BC enhance effective feature extraction in both encoding and decoding stages. The SFP in the skip connection stage reduces the semantic gap between low-level and high-level features, accelerating network learning.

Likewise, Lu et al. [27] identified limitations in existing medical image segmentation methodologies, particularly regarding challenges in integrating multi-scale information and effectively combining local details with global contextual semantics. They proposed LM-Net, a novel, lightweight, and multi-scale architecture that leverages the advantages of both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to improve segmentation accuracy. Additionally, they introduced two critical modules, the

Local Feature Transformer (LFT) and Global Feature Transformer (GFT), which collaborate to capture local details and global semantics using multi-scale features at different levels.

Bhuyan et al. [28] addressed the problem of disease analysis using different machine learning approaches: support vector machines, K-nearest neighbors, random forests, artificial neural networks, and logistic regression. In addition, they proposed a single framework for cancer diagnosis using a Full-resolution Convolutional Network (FrCN) [29]. The latter was extended by applying a combination of deep learning approaches for cancer detection: mass detection using a You-Only-Look-Once (YOLO) approach, and the crucial aspect of segmentation by using FrCNs [30].

Finally, we recently proposed the extension of the original CoAtNet model [31] by introducing dual streams of convolution and self-attention blocks in its final layers [32]. These enhancements aggregate local and global features from initial fine-resolution maps, significantly improving segmentation accuracy in breast ultrasound images and enabling more robust tumor detection methods.

Although the methods already proposed in the literature are significant contributions to the problem, they still have difficulty in accurately segmenting tumors due to shadows and speckle noise typically present in BUS images, as well as to the high variability in shape, appearance, size, location, and texture of tumors.

3. Proposed method

We propose a new deep neural network model that significantly improves breast cancer tumor segmentation from BUS images based on state-of-the-art deep networks. The proposed network architecture is depicted in Figure 2. In particular, we apply two encoder-decoder (autoencoder) networks in sequence. The first autoencoder is fed with the given BUS image and generates a preliminary binary mask with regions of interest corresponding to candidate tumors. A second autoencoder is then fed with that preliminary mask concatenated with the given BUS image. Its goal is to refine the candidate binary regions into a final mask that represents the best tumor segmentation.

Both autoencoders have the same topology and are trained together end-to-end. The encoder network can be any state-of-the-art convolutional or attention-based deep network. In the example depicted in Figure 2, the encoder network is CoAtNet, a promising deep architecture that combines

both convolutional and transformer blocks. However, we have also run experiments with other modern architectures: ResNet, WideResNet, ConvNext and ResNext.

In turn, the decoder network is a classical deconvolutional network that generates a high-resolution binary image out of the latent space generated by the encoder. UNet's skip connections are applied between the layers of the encoder and the corresponding layers of the decoder.

As for the training stage, we define a compound loss function obtained as the weighted average of two partial loss functions. The first partial loss function compares the preliminary binary mask generated by the first autoencoder with the Ground-Truth (GT) binary mask. In turn, the second partial loss function compares the final binary mask generated by the second autoencoder with the GT binary mask. Both partial losses are combined through a weighted average. Experiments were run with different loss functions by considering cost-sensitive learning in order to compensate for the high imbalance between healthy regions and tumorous ones, this being a common problem in medical image diagnosis. Class imbalance frequently leads to high false-negative rates. More details about the evaluated loss functions are given in Section 3.1. In addition, we applied L2 regularization and dropout to avoid overfitting while keeping a good performance. Batch normalization was applied after each convolutional layer. In turn, dropout was applied after the first fully connected layer. We used a batch size of 4 and Adam optimization with a learning rate of 0.0001 and 20 epochs. In order to ensure that the proposed method is valid and generalizable, all experiments were run upon two public datasets of BUS images (UDIAT and BUSI). Those datasets were augmented by randomly applying various image transformations described in Section 4.1.

3.1. Loss functions and cost-sensitive learning

Choosing the right loss function is a crucial decision when training deep neural networks as each loss function has its own advantages and disadvantages. Furthermore, loss functions determine whether all image pixels have a similar contribution during training or, alternatively, they have a different weight (cost) depending on the class label, such that the minority class (tumorous regions in our scope) receives sufficient attention with respect to the preponderant class of healthy tissue. The latter is the basis of cost-sensitive learning, which is fundamental for medical image diagnosis.

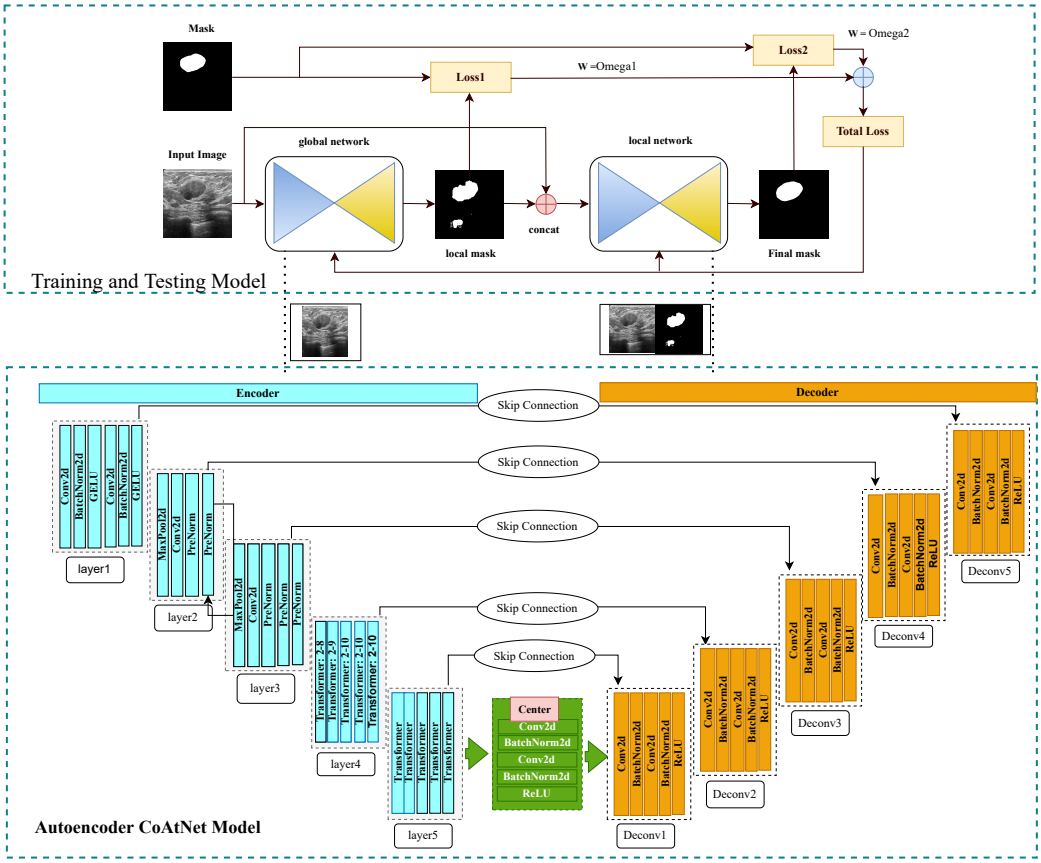

Figure 2: Architecture of the proposed network with CoAtNet being applied to the encoders.

Let us assume $\mathbf{Y}$ is the binary mask generated by one of the autoencoders, and $\mathbf{G}$ the corresponding ground-truth binary mask. Both $\mathbf{Y}$ and $\mathbf{G}$ have N binary pixels: $\mathbf{Y} = \{y_1, \ldots, y_N\}$ and $\mathbf{G} = \{g_1, \ldots, g_N\}$, such that every pixel is one if it belongs to a tumor (white in the images shown in the paper) or zero if the pixel belongs to the healthy tissue (black background). In addition, let us assume $\mathbf{P} = \{p_1, \ldots, p_N\}$ is the set of N (sigmoid-bounded) probabilities generated by the last stage of the decoders, such that p_i is a real value between 0 and 1 that expresses the probability of y_i being one (tumorous). Actually, $\mathbf{Y}$ is typically generated by thresholding $\mathbf{P}$. In our context, every loss function is computed from both $\mathbf{P}$ and $\mathbf{G}$.

Four loss functions were evaluated in this work: Binary Cross-Entropy

loss ($LBCE$), Dice loss (LD), Focal loss (LF) and Tversky loss (LT).

3.1.1. Binary Cross-Entropy Loss

The *Binary Cross-Entropy* (BCE) is a measure of classification error. It gives very high values when a GT pixel $g_i = 1$ (tumorous) but p_i is close to 0 (it is detected as healthy tissue) or, alternatively, when a GT pixel $g_i = 0$ (healthy) but p_i is close to 1 (it is detected as tumorous). The *BCE Loss* is the aggregation of the BCEs of all pixels:

$$LBCE(\mathbf{P}, \mathbf{G}) = -\frac{1}{N} \sum_{i=1}^{N} g_i \log p_i + (1 - g_i) \log(1 - p_i) \tag{1}$$

Large binary classification errors yield large values of $LBCE$. Therefore, by minimizing that loss function, the neural network is improving the pixel-wise binary classification.

3.1.2. Dice Loss

The *Dice Similarity Coefficient* (DSC), also known as *F1-score*, estimates the segmentation quality in terms of the intersection between the predicted tumorous regions, $\{y_i \in \mathbf{Y} \mid y_i = 1\}$, and the GT tumorous regions, $\{g_i \in \mathbf{G} \mid g_i = 1\}$. Let TP be the number of true positives, that is, the predicted tumorous regions that are really tumorous in the GT. Let also TN be the number of true negatives, that is, the predicted healthy regions that are really healthy. In addition, let FP be the number of false positives, that is, the predicted tumorous regions that are really healthy. Finally, let FN be the number of false negatives, that is, the predicted healthy regions that are really tumorous. These four functions can be defined in a soft-classification context (using probabilities, $\mathbf{P}$, instead of binary values, $\mathbf{Y}$) as:

$$
\begin{aligned}
TP = TP(\mathbf{P}, \mathbf{G}) &= \sum_{i=1}^{N} p_i g_i \\
TN = TN(\mathbf{P}, \mathbf{G}) &= \sum_{i=1}^{N} (1 - p_i)(1 - g_i) \\
FP = FP(\mathbf{P}, \mathbf{G}) &= \sum_{i=1}^{N} p_i (1 - g_i) \\
FN = FN(\mathbf{P}, \mathbf{G}) &= \sum_{i=1}^{N} (1 - p_i) g_i
\end{aligned}
\tag{2}
$$

The DSC is the ratio between twice the area of intersection, that is, the true positives, and both the area of predicted tumorous regions (i.e., true positives plus false positives) plus the area of GT tumorous regions (i.e., true positives plus false negatives):

$$
DSC(\mathbf{P}, \mathbf{G}) = \frac{2TP}{(TP + FP) + (TP + FN)} = \frac{TP}{TP + \frac{1}{2}FP + \frac{1}{2}FN} \tag{3}
$$

By applying (2) to (3):

$$
DSC(\mathbf{P}, \mathbf{G}) = \frac{2 \sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} \tag{4}
$$

DSC ranges between 0 (worst segmentation: no intersection at all between the predicted tumorous regions and the GT) and 1 (perfect segmentation: total coincidence between the predicted tumorous regions and the GT). Based on it, the *Dice Loss* is defined as:

$$
LD(\mathbf{P}, \mathbf{G}) = 1 - DSC(\mathbf{P}, \mathbf{G}) \tag{5}
$$

Large segmentation errors yield values of LD close to 1. Therefore, by minimizing that loss function, the neural network is improving the segmentation of tumorous regions.

3.1.3. Binary Focal Loss

The *Binary Focal Loss* (LBF) is an improvement of the Binary Cross-Entropy Loss ($LBCE$) that aims at focusing learning on hard misclassified examples and compensating for class imbalance. With $LBCE$, pixels that have correctly been classified as tumorous ($y_i = g_i = 1$) with a relatively low confidence (e.g., $p_i = 0.6$) yield a relatively high loss. By reducing that loss, the aim is to move p_i close to 1. Conversely, pixels that have correctly been classified as healthy tissue ($y_i = g_i = 0$) with a relatively low confidence (e.g., $1 - p_i = 0.6$) also yield a relatively high loss, again with the aim of maximizing $1 - p_i$. Obviously, pixels that have wrongly been classified will yield much larger losses. In the end, the training process will try to improve the segmentation of pixels that have already been correctly classified with low confidence (soft examples), as well as of wrongly classified pixels (hard examples). This means that the learning process will disperse its attention over both soft and hard examples. Focal loss introduces a factor that modifies the shape of $LBCE$ by making it sharper. In that way, the loss is slightly reduced for pixels that have already been classified correctly even with a relatively low confidence, hence focusing on the hard examples.

In addition, a weighting factor λ is introduced in the weighted sum in order to balance positive (tumorous) and negative (healthy) pixels. Since the number of healthy pixels is far larger than that of tumorous pixels, the final sum is favoring the correct segmentation of the majority healthy regions. This may lead to tumorous regions being misclassified as healthy (false negatives). This trend can be inverted by increasing the preponderance of the minority tumorous regions in the GT with a large weight, λ, while reducing the influence of the majority healthy regions with a low weight, $1 - \lambda$. This will tend to reduce the number of false negatives (undetected tumors) at the expense of increasing the number of false positives (healthy regions being wrongly classified as tumorous). Finally:

$$LBF(\mathbf{P}, \mathbf{G}) = -\frac{1}{N} \sum_{i=1}^{N} \lambda g_i (1 - p_i)^{\gamma} \log p_i + (1 - \lambda)(1 - g_i) p_i^{\gamma} \log(1 - p_i), \tag{6}$$

where λ is typically the majority-class percentage, that is $\lambda = |\{g_i \in \mathbf{G} | g_i = 0\}|/N$, where $|\cdot|$ is the cardinality operator, and $\gamma > 0$ is the hyperparameter that controls the sharpness of the loss function (the larger, the sharper). LBF reduces to $LBCE$ when $\gamma = 0$ and the class-balance weight λ is removed.

Optimal values of λ and γ in this work were found for the baseline CoAtNet model and the two tested datasets as described in Section 4.2.

3.1.4. Tversky Loss

The *Tversky Loss* is a generalization of the Dice loss that addresses the problem of class imbalance. As shown in (3), the influence of false positives and false negatives in the DSC formulation is balanced, with a constant weight of 0.5. The Tversky loss simply weighs those two terms with respective hyperparameters, α and β:

$$LT(\mathbf{P}, \mathbf{G}) = 1 - \frac{TP + \zeta}{TP + \alpha FP + \beta FN + \zeta} \tag{7}$$

Typically, $\alpha + \beta = 1$. When β is above α, the training process is more sensitive to false negatives than to false positives. This means that big values of FN will be amplified, leading to big losses. In that case, by minimizing the loss, the learning process will tend to minimize false negatives even at the expense of false positives. As a consequence, the model will tend to avoid the misclassification of tumorous regions as healthy tissue even if there is an increase of healthy regions wrongly classified as tumorous. ζ is an infinitesimal constant added to avoid division by zero. Optimal values of α and β in this work were found for the baseline CoAtNet model and the two tested datasets as described in Section 4.2.

3.1.5. Compound Loss

A compound loss function has been defined for the training stage of the proposed deep network. A first partial loss function, $Loss_1(\mathbf{P}_1, \mathbf{G})$, compares the probabilities generated by the first autoencoder, $\mathbf{P}_1$, with the GT binary mask, $\mathbf{G}$. In turn, a second partial loss function, $Loss_2(\mathbf{P}_2, \mathbf{G})$, compares the probabilities generated by the second autoencoder, $\mathbf{P}_2$, with the same GT binary mask, $\mathbf{G}$. The compound loss function for the whole segmentation network is finally defined as:

$$Loss(\mathbf{P}_1, \mathbf{P}_2, \mathbf{G}) = \omega_1 Loss_1(\mathbf{P}_1, \mathbf{G}) + \omega_2 Loss_2(\mathbf{P}_2, \mathbf{G}) \tag{8}$$

Optimal values of ω_1 and ω_2 in this work were found for the proposed model with CoAtNet and the two tested datasets as described in Section 4.2.

3.2. Evaluation measures

The proposed deep segmentation network and its alternative models were evaluated by comparing their predicted segmentation masks, $\mathbf{Y}$, with the corresponding GT, $\mathbf{G}$. In the proposed segmentation model, the final segmentation mask is the one predicted by the second autoencoder: $\mathbf{Y} = \mathbf{Y}_2$. The TP, TN, FP and FN measures were computed in a hard-classification context (using binary values, $\mathbf{Y}$, instead of probabilities, $\mathbf{P}$) as:

$$\begin{aligned} TP = TP(\mathbf{Y}, \mathbf{G}) &= \sum_{i=1}^{N} y_i g_i \\ TN = TN(\mathbf{Y}, \mathbf{G}) &= \sum_{i=1}^{N} (1 - y_i)(1 - g_i) \\ FP = FP(\mathbf{Y}, \mathbf{G}) &= \sum_{i=1}^{N} y_i (1 - g_i) \\ FN = FN(\mathbf{Y}, \mathbf{G}) &= \sum_{i=1}^{N} (1 - y_i) g_i \end{aligned} \tag{9}$$

The following evaluation measures were considered in this work:

Intersection-Over-Union (IoU) or *Jaccard index*: It is the ratio between the intersection of both the predicted and GT masks, and their union. It behaves similarly to the DSC (3):

$$Jaccard(\mathbf{Y}, \mathbf{G}) = \frac{TP}{TP + FP + FN} \tag{10}$$

Dice index: It is the Dice Similarity Coefficient (DSC) already defined in (3), also referred to as F1-score. It behaves similarly to the Jaccard index (10):

$$Dice(\mathbf{Y}, \mathbf{G}) = DSC(\mathbf{Y}, \mathbf{G}) \tag{11}$$

Recall: It is the percentage of truly tumorous regions detected by the network over the total number of tumorous regions in the GT. A recall close to 1 means that the number of false negatives (tumorous regions wrongly detected as healthy) is very low:

$$Recall(\mathbf{Y}, \mathbf{G}) = \frac{TP}{TP + FN} \tag{12}$$

4. Experimental results

4.1. Datasets and data augmentation

Two public datasets of breast cancer ultrasound images were used to validate the proposed network models in this work. The first dataset was provided by the *UDIAT Diagnostic Centre* (Sabadell, Spain) [33]. It consists of 163 BUS images containing different breast tumors. Each image is associated with a ground-truth segmentation of the depicted tumor. All images were captured using a Siemens ACUSON Sequoia C512 ultrasound system with a 17L5 HD linear array ultrasound transducer (8.5 MHz).

The second dataset (BUSI) was proposed in [34]. It contains BUS images of 600 female patients captured with a LOGIQ E9 and a LOGIQ E9 Agile ultrasound system. From its 780 images, we considered the ones that contain either a benign or malignant tumor, thus leaving 697 images. The dataset also includes ground-truth segmentations of the depicted tumors.

In both datasets, a 70% of images were used for training and a 30% for testing. All images were resized to 256×256 to be able to feed them into the evaluated neural models. We also applied different data augmentation techniques to the training splits of both datasets: rotation, width shift, height shift, shear, horizontal flip, vertical flip, shift scale rotate, median blur, hue saturation value, and random brightness contrast. Some examples of those transformations are shown in Figure 3. After data augmentation, the training set of the first dataset increased to 1,400 images, and the one of the second dataset to 7,000 images.

4.2. Hyperparameter optimization

The main hyperparameters associated with the proposed deep neural model are the coefficients of the Focal loss function, λ and γ (6), the coefficients of the Tversky loss function, α and β (7), and the coefficients of the compound loss function, ω_1 and ω_2 (8).

We first found the best coefficients for both the Focal and Tversky loss functions by only considering the first autoencoder depicted in Figure 2, with CoAtNet being used as encoder. For the Focal loss, we considered 5 different values of γ (0, 0.5, 1, 2, 5) and 5 different values of λ (0.5, 0.6, 0.7, 0.8, 0.9), and then run experiments with all 25 combinations of each γ and λ. Values of λ lower than 0.5 were discarded since λ must be above 0.5 to reduce the number of false negatives (undetected tumors) as discussed in Section 3.1, which is related to cost-sensitive learning. As for the Tversky loss, we

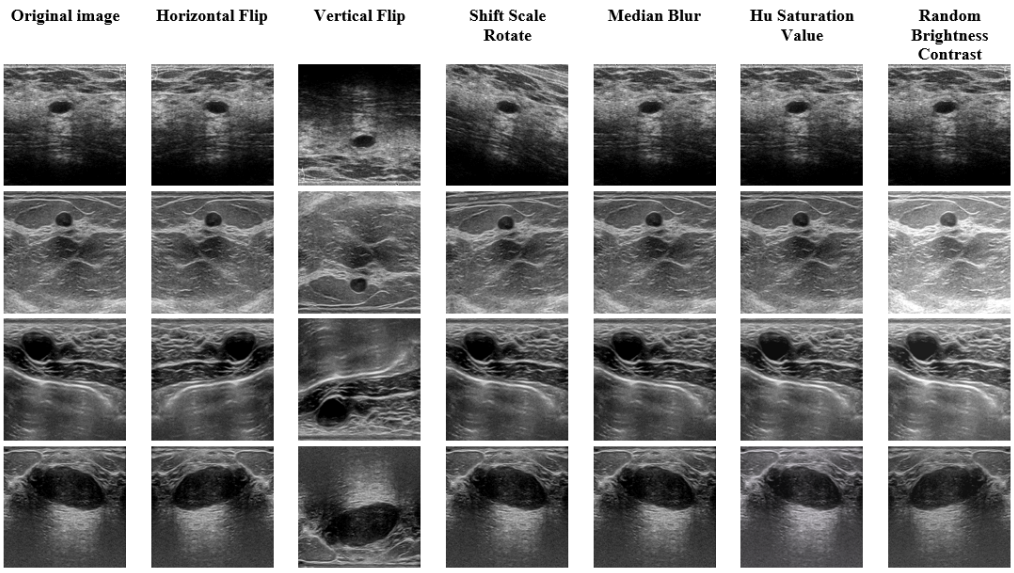

Figure 3: Examples of BUS images after applying different augmentation techniques

assumed $\alpha = 1 - \beta$ and tested 4 values of β: 0.6, 0.7, 0.8, 0.9. Coefficient β must be above α to reduce the number of false negatives (undetected tumors) as discussed in Section 3.1, which is also related to cost-sensitive learning. For each experiment, we trained the first autoencoder with the training images of the considered dataset, and then validated it with the test images of the same dataset.

Figure 4 shows the performance in terms of Jaccard (10) and Dice (11) indices for the first autoencoder based on both CoAtNet and the test images from the first dataset by considering the different combinations of coefficients for both the Tversky and Focal loss functions.

The best performance for the Focal loss was achieved with $\lambda = 0.7$ and $\gamma = 5.0$ (Dice = 77.71%). In turn, the best performance for the Tversky loss was obtained with $\alpha = 0.4$ and $\beta = 0.6$ (Dice = 77.44%). Similarly, the performance of both loss functions for the test images from the second dataset is shown in Figure 5. In this case, the best performance for the Focal loss was fulfilled with $\lambda = 0.6$ and $\gamma = 1.0$ (Dice = 77.24%), being only 0.47% below the result obtained with the best combination found for the first dataset ($\lambda = 0.7$ and $\gamma = 5.0$). In turn, the best performance for the Tversky loss was achieved with $\alpha = 0.3$ and $\beta = 0.7$ (Dice = 77.69%). These results indicate that the performance of both loss functions is dataset

dependant, mainly for the Tversky loss.

By taking into account the results corresponding to both datasets, we can conclude that, in general, the maximum Dice index for the Focal loss function is obtained when $\lambda = 0.7$ and $\gamma = 5.0$, whereas the minimum Dice is with $\lambda = 0.8$ and $\gamma = 0.5$. The larger the value of γ, the bigger the focalization of learning on the hard misclassified examples. In turn, the maximum Dice index with the Tversky loss function is achieved when $\alpha = 0.4$ and $\beta = 0.6$, whereas the worst Dice is with $\alpha = 0.1$ and $\beta = 0.9$. Very low values of α imply that false positives (healthy regions wrongly considered as tumorous) are largely disregarded in favor of reducing the false negatives (tumorous regions wrongly considered as healthy). Therefore, the segmentation yields too many false detections in that case.

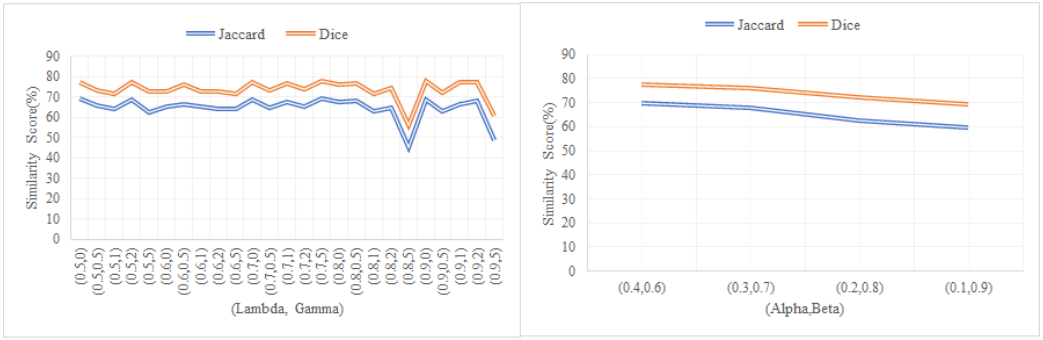

Figure 4: Performance of Focal loss (left) and Tversky loss (right) for CoAtNet-based autoencoder on first dataset (UDIAT)

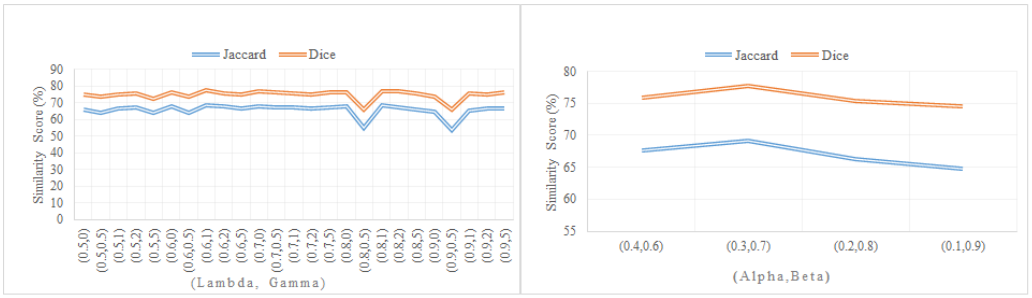

Figure 5: Performance of Focal loss (left) and Tversky loss (right) for CoAtNet-based autoencoder on second dataset (BUSI)

Once the best configurations for the Tversky loss and Focal loss were

found upon each tested dataset with only the first autoencoder, we studied the best combination of loss functions for the complete proposed model depicted in Figure 2, using CoAtNet as encoder in both autoencoders. We considered the four loss functions described in Section 3.1: Binary Cross-Entropy loss ($LBCE$), Dice loss (LD), Binary Focal loss (LBF) and Tversky loss (LT). Both the Focal and Tversky loss functions were configured with the coefficients found before for each tested dataset.

In particular, we checked 16 different combinations of $Loss_1$ and $Loss_2$ (8) by considering all variations with repetition of the four evaluated loss functions taken in twos: $LBCE/LBCE$, $LBCE/LD$, $LBCE/LBF$, $LBCE/LT$, $LD/LBCE$, LD/LD, LD/LBF, LD/LT, $LBF/LBCE$, etc. For each pair of loss functions, we run experiments for 9 different combinations of ω_1 and ω_2 (8). Actually, we set $\omega_2 = 1 - \omega_1$ and run experiments with $\omega_1 \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. In total, 144 ($16 \times 9$) experiments were run on each dataset.

For the UDIAT dataset, the best performance (Dice = 84.49%) was attained with the following configuration: $Loss_1 = LBCE$, $Loss_2 = LT_{\alpha=0.4,\beta=0.6}$, $\omega_1 = 0.3$ and $\omega_2 = 0.6$. In turn, the best performance for the BUSI dataset (Dice = 78.94%) was obtained with: $Loss_1 = LD$, $Loss_2 = LBF_{\lambda=0.6,\gamma=1.0}$, $\omega_1 = 0.2$ and $\omega_2 = 0.8$.

In the remainder of this work, the proposed model was run with the above optimal configurations of loss functions, $Loss_1$ and $Loss_2$, and weights, ω_1 and ω_2, for each tested dataset.

4.3. Experimental validation

The proposed model depicted in Figure 2 is an example where CoAtNet is the baseline deep model applied as encoder in the two autoencoders. However, the segmentation performance of other state-of-the-art deep models can also be increased by applying them as encoders of the two proposed autoencoders.

In particular, we tested the application of the proposed model to four different modern deep networks in addition to CoAtNet: ResNet50, WideResNet, ResNext101, and ConvNext. In all cases, we used the same hyperparameters previously found for CoAtNet as described in Section 4.2. Table 1 and Table 2 show the performance of the proposed model and the corresponding baseline model for the UDIAT and BUSI datasets, respectively. "Baseline Model" refers to the performance of the first autoencoder alone, whereas "Proposed Model" refers to the performance of the two autoencoders. The

first entry in those tables corresponds to the model depicted in Figure 2, using CoAtNet as the baseline in both encoders. Figure 6 depicts those Dice indices graphically.

Table 1: Performance of baseline model and proposed model for alternative modern deep networks on the UDIAT dataset

UDIAT Dataset				
Deep Network Models	**Baseline Model**		**Proposed Model**	
	Jaccard	**Dice**	**Jaccard**	**Dice**
CoAtNet	69.00	77.71	**76.31**	**84.49**
WideResNet	59.73	70.73	70.11	79.04
ResNext101	54.34	63.47	66.88	73.64
ConvNext_tiny	52.30	60.96	58.61	68.57
ResNet50	55.09	63.72	64.43	73.24

Table 2: Performance of baseline model and proposed model for alternative modern deep networks on the BUSI dataset

BUSI Dataset				
Deep Network Models	**Baseline Model**		**Proposed Model**	
	Jaccard	**Dice**	**Jaccard**	**Dice**
CoAtNet	69.11	77.69	**70.83**	**78.94**
WideResNet	64.84	73.62	67.39	75.50
ResNext101	66.42	75.21	67.72	75.77
ConvNext_tiny	58.60	68.16	61.87	70.34
ResNet50	65.62	74.01	67.64	75.84

In terms of Dice index, the proposed model yielded an improvement between 6.78% and 10.17% (average improvement of 8.47%) with respect to the baseline model for the UDIAT dataset, and between 0.56% and 2.18% (average improvement of 1.54%) for the BUSI dataset. The largest improvement for both datasets was achieved by using CoAtNet as baseline model (Dice equal to 84.49% and 78.94%, respectively).

Beyond the above quantitative results, the proposed model yields segmentation results very close to the ground truth. For example, Figure 7 shows a qualitative comparison between the segmentations generated by both CoAtNet (first autoencoder alone) and the proposed model, with CoAtNet used as encoder in the two autoencoders.

In turn, Figures 8 and 9 show examples of tumor segmentations generated by the proposed model with different state-of-the-art deep networks applied as encoders of the two proposed autoencoders for the UDIAT and BUSI datasets, respectively. On the other hand, Figure 10 shows the evolution of segmentation results generated by the proposed model with CoAtNet for the UDIAT dataset after six of the 20 training epochs: 1, 2, 3, 5, 11 and 20.

In order to assess the distribution of Dice indices, Figures 11(a) and 11(b) show the boxplots of Dice indices corresponding to the proposed model endowed with different state-of-the-art deep networks for the UDIAT and BUSI datasets, respectively. Notice that the number of outliers is acceptable and that the variance is small except for ConvNext.

The behavior of Recall (12), which measures the percentage of truly tumorous regions detected by the network over the total number of tumorous regions in the GT, is also relevant. A recall close to one implies few false negatives (tumorous regions wrongly detected as healthy). Figures 12(a) and 12(b) show the boxplots of recall for the proposed model endowed with different modern deep networks for the UDIAT and BUSI datasets. The number of outliers is low and the variance small for all variations of the proposed model. This indicates that the number of false negatives is kept low.

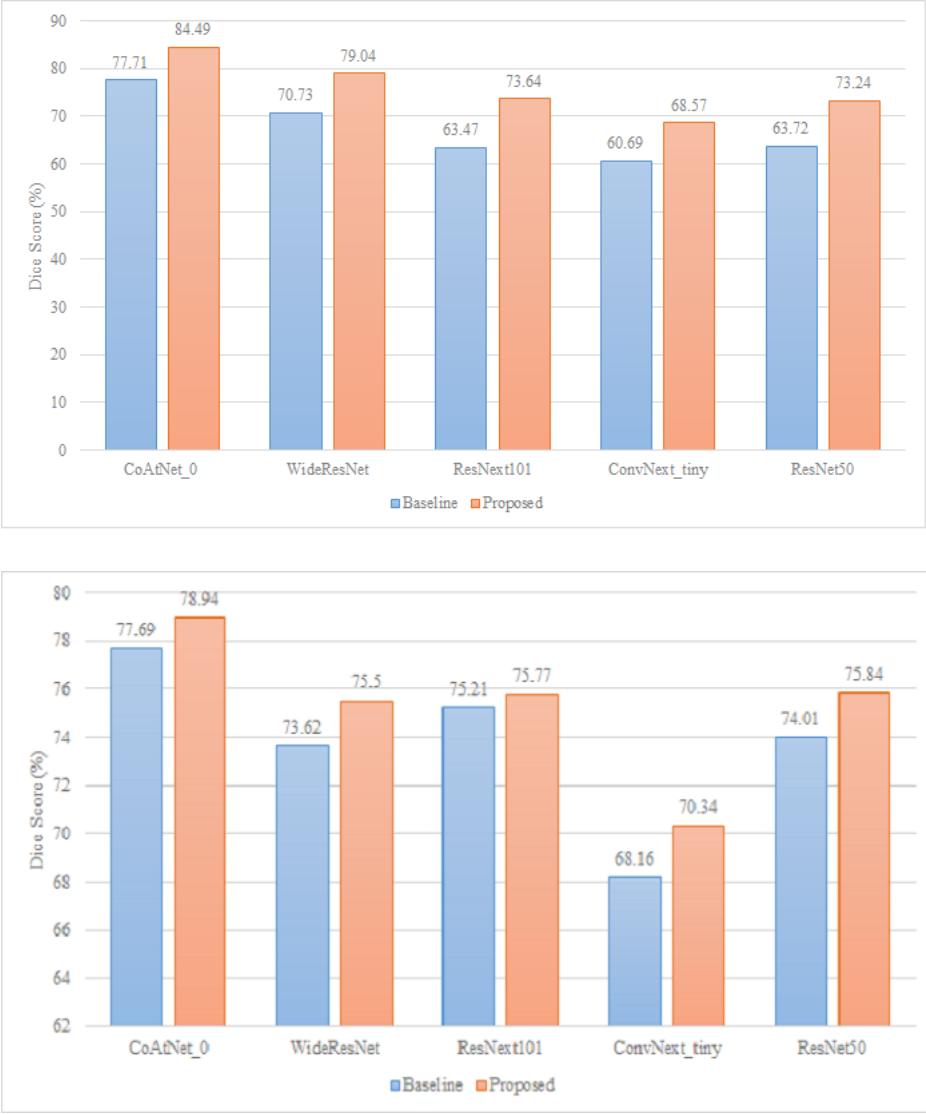

Figure 6: Performance of baseline model and proposed model for alternative modern deep networks for the UDIAT dataset (top), and BUSI dataset (bottom).

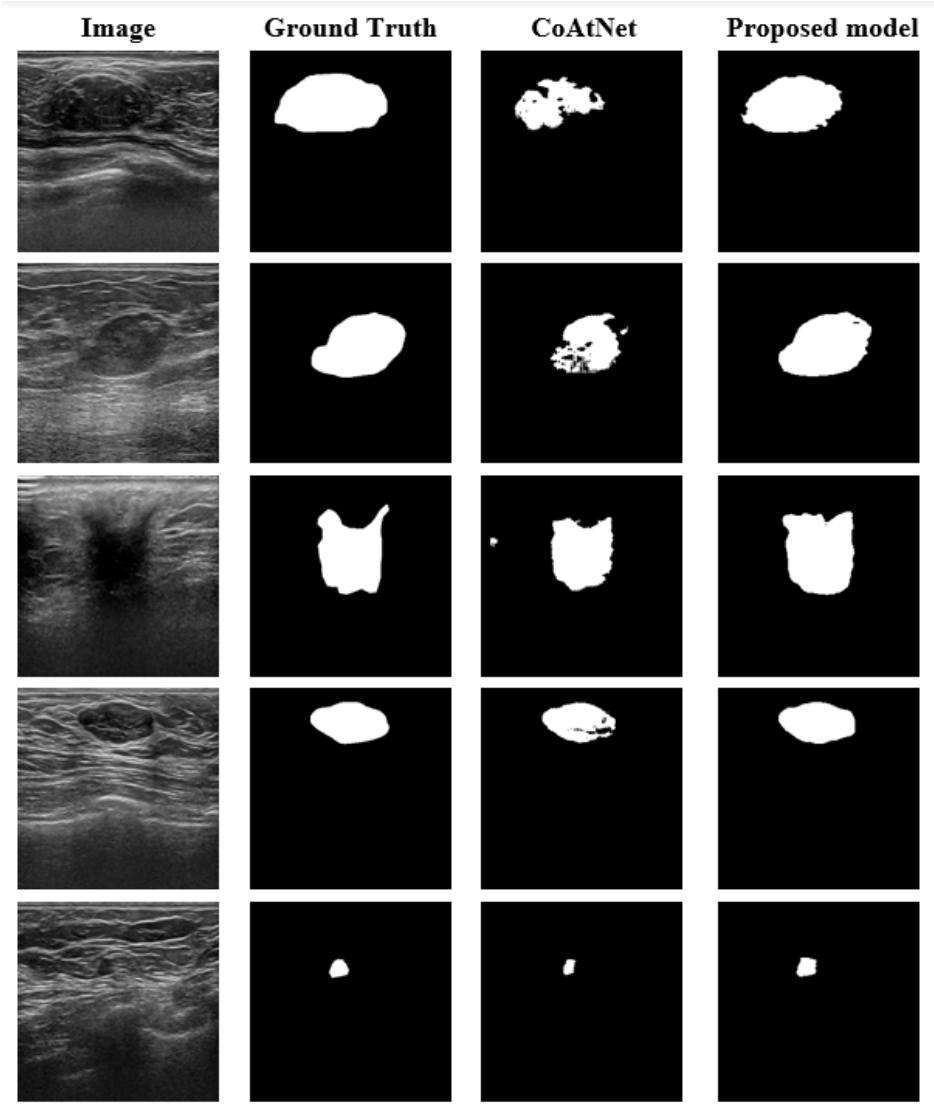

Figure 7: Examples of tumor segmentation with CoAtNet alone (preliminary mask) and with the proposed model using CoAtNet (refined mask)

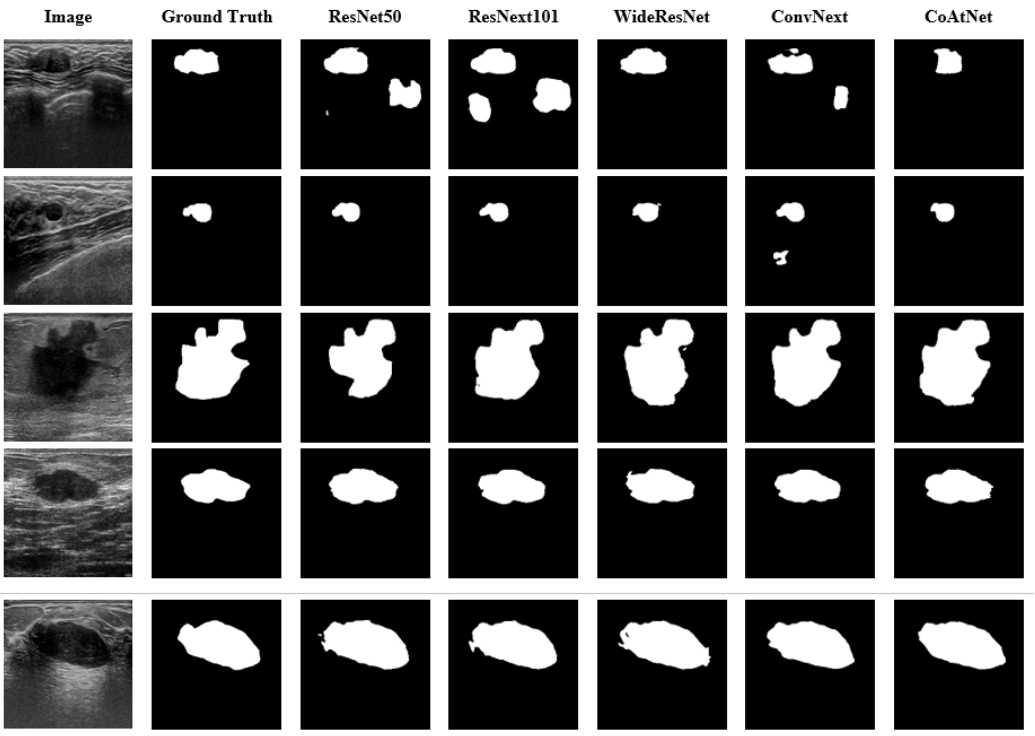

Figure 8: Examples of tumor segmentation with the proposed model configured with different state-of-the-art deep networks for the UDIAT dataset.

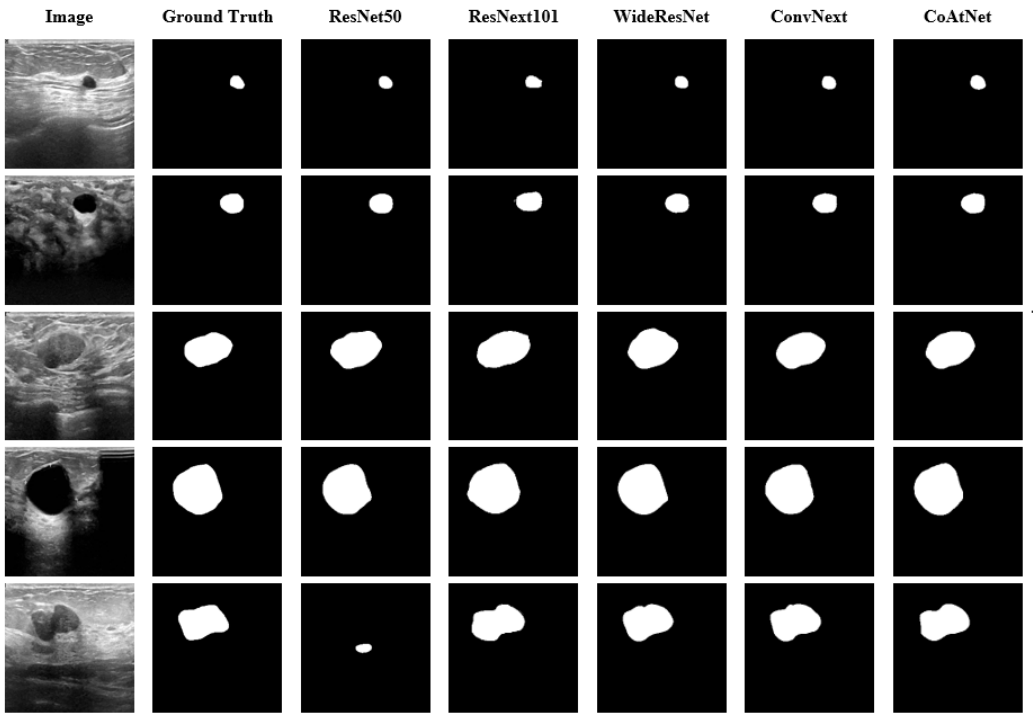

Figure 9: Examples of tumor segmentation with the proposed model configured with different state-of-the-art deep networks for the BUSI dataset

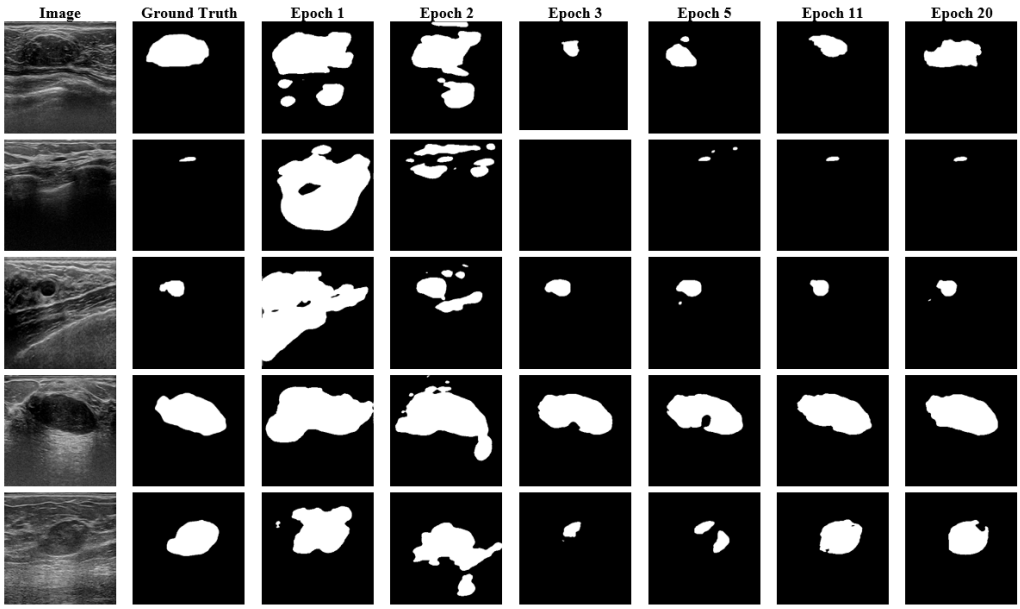

Figure 10: Evolution of segmentation results after six of the 20 training epochs with the proposed model using CoAtNet and the UDIAT dataset

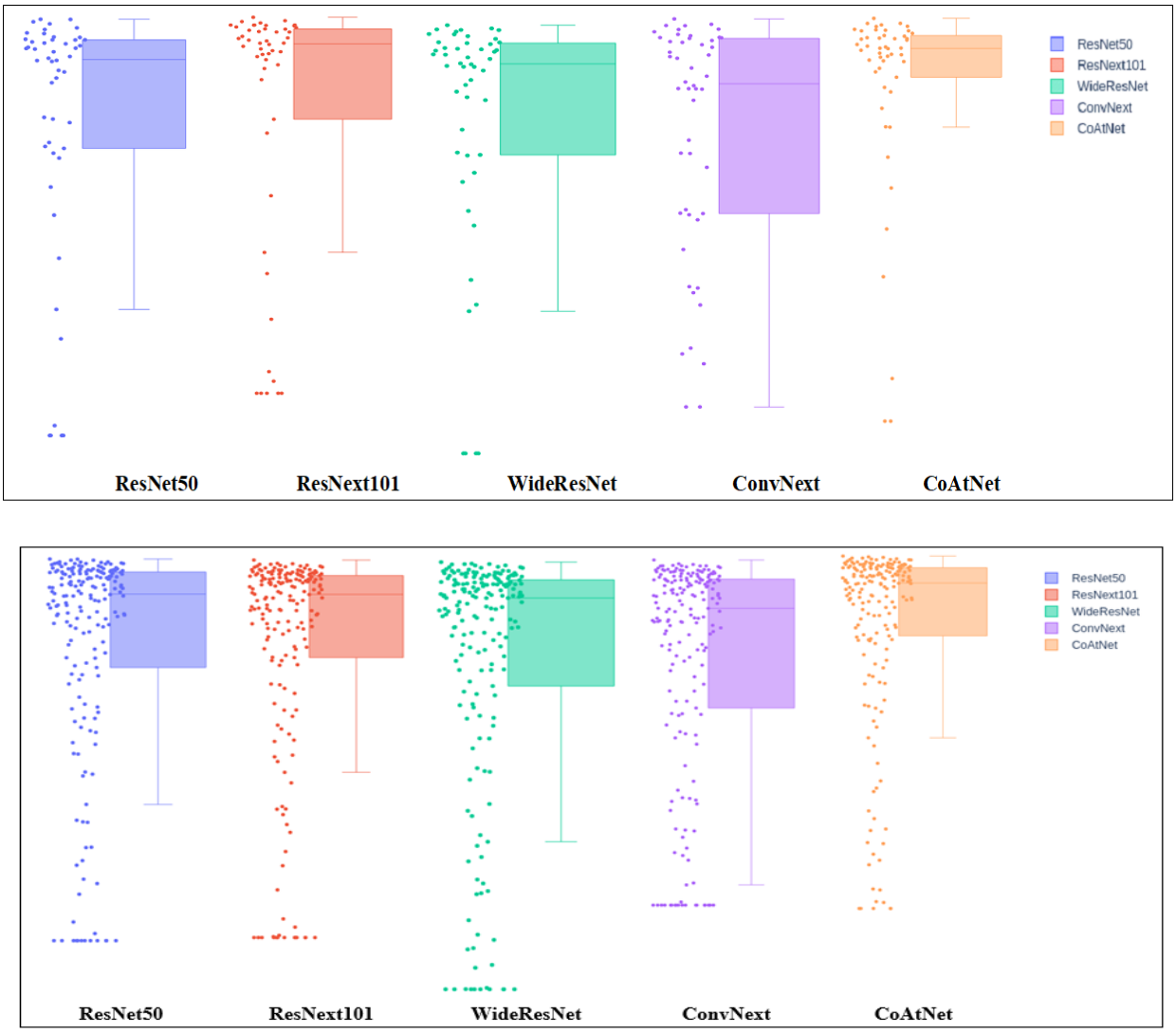

Figure 11: Boxplots of Dice indices corresponding to the proposed model endowed with different state-of-the-art deep networks for the UDIAT dataset (top), and BUSI dataset (bottom).

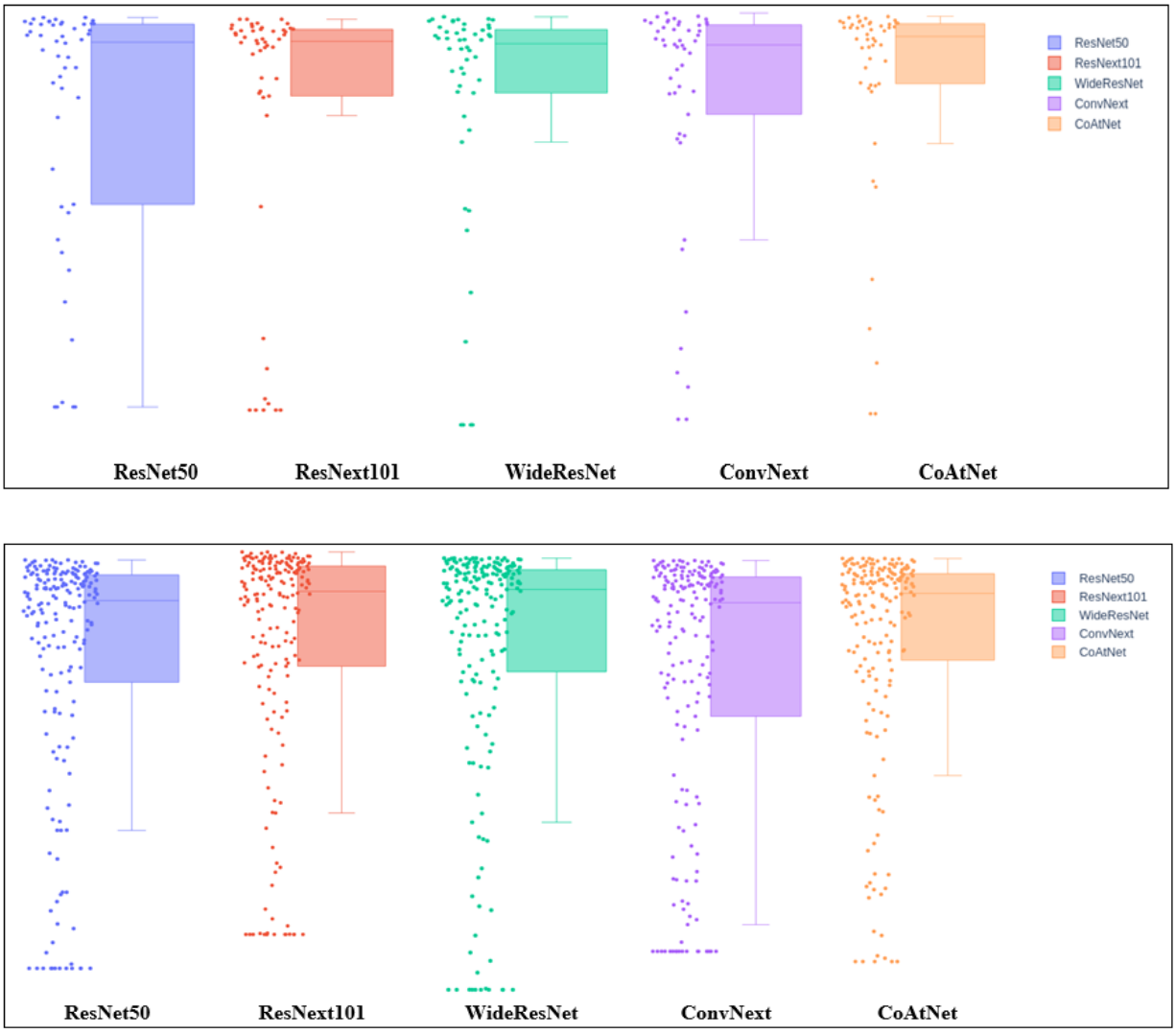

Figure 12: Boxplots of Recall indices corresponding to the proposed model endowed with different state-of-the-art deep networks for the UDIAT dataset (top), and BUSI dataset (bottom).

In sum, the above experimental results indicate that the segmentations generated by the proposed two-stage models are significantly closer to the given ground-truths than when the original single-stage baseline models are applied. This implies a more precise detection and delineation of tumorous regions, as well as a reduction of the number of misdetections (tumorous regions wrongly detected as healthy). This outcome is attributable to two main factors: the application of a two-stage model, in which an initial segmentation generated by the first stage is subsequently refined by the second stage, as well as the application of cost-sensitive learning in order to focalize training on the segmentation errors of the minority tumorous class.

Finally, we have applied the open-source package Ptflops [35] for estimating the computational complexity and number of trainable parameters of the baseline and proposed models for the alternative modern deep networks reported in Tables 1 and 2. The computational complexity is estimated in terms of GMacs (Giga Multiply-Accumulate operations). One GMac is roughly equivalent to two GFlops. Table 3 shows the results when processing 256×256 images.

Table 3: Computational complexity in GMacs (1 GMac $\approx$ 2 GFlops), and number of trainable parameters (millions) of baseline and proposed models

Model	**GMacs**	**#Parameters**
CoAtNet (baseline)	4.55	17.03
CoAtNet (proposed)	9.1	34.06
WideResNet (baseline)	29.83	126.89
WideResNet (proposed)	59.66	253.78
ResNext101 (baseline)	21.6	88.79
ResNext101 (proposed)	43.2	177.58
ConvNext (baseline)	5.86	28.59
ConvNext (proposed)	11.72	57.18
ResNet50 (baseline)	5.4	25.56
ResNet50 (proposed)	10.8	51.12

As expected, due to the addition of a second autoencoder, the computational complexity and number of trainable parameters of the proposed model are doubled compared to the baseline model. That increase is compensated by the aforementioned qualitative improvements in tumor segmentation obtained with the proposed model. Indeed, in medical applications,

and particularly in breast cancer detection, accuracy is paramount, as even minor misclassifications can lead to critical consequences for patients. Consequently, healthcare practitioners prioritize precision in diagnostics, making the enhanced detection performance of our proposed model a valuable asset that allows for better differentiation of cancerous tissues from healthy ones.

5. Conclusions and future work

This paper proposes an efficient deep network for automatically segmenting tumors in breast ultrasound images. The model consists of two consecutive encoder-decoder (autoencoder) networks. The first autoencoder is fed with the original image and generates a preliminary binary mask. The second autoencoder then generates the final binary segmentation after feeding it with the original image and the previous preliminary mask. The first autoencoder can be interpreted as an extractor of global features from the given BUS images, which are subsequently refined by the second autoencoder. The encoder within each autoencoder can be any state-of-the-art classification deep network. In this work, five modern baseline networks have been embedded and tested: CoAtNet, ResNet50, WideResNet, ResNext101, and ConvNext.

Experiments with two public BUS image datasets have shown that the proposed model with the two autoencoders significantly surpasses the application of the individual baseline networks embedded in the first autoencoder alone. Due to the significant class imbalance between healthy and tumorous regions in the available datasets, cost-sensitive learning through the application of Focal and Tversky loss functions has proven beneficial for focalizing the proposed model into the segmentation errors of the minority class (tumorous regions).

As a result of an extensive evaluation of four loss functions (Binary Cross-Entropy, Dice, Focal and Tversky), we have determined that, although their best combination in this application scope is dataset dependant, the best results are obtained when a cost-sensitive loss function (i.e., Focal or Tversky) is applied to the second autoencoder and a standard loss function (i.e., BCE or Dice) to the first one, with the weight of the second loss function being significantly larger than the one of the first function. We have also identified the hyperparameters of the Tversky and Focal loss functions that, in general, yield the largest performance.

In this work, we apply a same state-of-the-art network to both autoencoders, only changing their corresponding loss functions. It will be interesting

to evaluate the performance of a heterogeneous system that integrates networks with different architecture. We also aim to extend the capabilities of the proposed model towards multi-class semantic segmentation. In particular, we pretend to distinguish between malignant and benign tumorous regions in addition to healthy regions. In the future, we also plan to apply pixel-wise self-supervised contrastive learning to BUS images in order to compensate for the limited amount of annotated datasets in this scope. Indeed, the main weakness of neural models trained with datasets of medical images is frequently the reduced size of those datasets due to that limited amount of annotated data. Conventional data augmentation techniques alone do not address this issue effectively. In this line, we also plan to extend the training datasets with high-quality synthetic images along with their corresponding ground-truth segmentations. We will explore two promising techniques for generating those synthetic images: Deep Convolutional Generative Adversarial Networks (DCGANs), and Denoising Diffusion Probabilistic Models (DDPMs).

Acknowledgement

The Spanish Government partly supported this research through Project TED2021-130081B-C21, and Project PDC2022-133383-I00.

Competing interests

The authors declare that they have no conflict of interest.

Availability of data and materials

Both datasets used in this paper are public.

To access the UDIAT dataset, interested researchers can directly contact the UDIAT Diagnostic Centre as indicated in the citation. UDIAT-Centre Diagnostic, Corporacio Parc Taulı, Sabadell (Spain) has copyright on the data and is the principal distributor of this dataset. University of Girona and Manchester Metropolitan University are involved in an ongoing effort to develop this dataset to aid research efforts in the general area of developing, testing and evaluating algorithms for breast ultrasound lesions analysis.

The second dataset used in this paper (BUSI) is openly accessible through the provided source reference.

References

[1] Q. Huang, Y. Huang, Y. Luo, F. Yuan, X. Li, Segmentation of breast ultrasound image with semantic classification of superpixels, Medical Image Analysis 61 (2020) 101657. doi:10.1016/j.media.2020.101657.

[2] L. Singh, Z. Jaffery, Z. Zaheeruddin, R. Singh, Segmentation and characterization of breast tumor in mammograms, in: 2010 International Conference on Advances in Recent Technologies in Communication and Computing, IEEE, 2010, pp. 213–216. doi:10.1109/ARTCom.2010.60.

[3] K. Stöger, D. Schneeberger, A. Holzinger, Medical artificial intelligence: the european legal perspective, Communications of the ACM 64 (2021) 34–36. doi:10.1145/3458652.

[4] H.-D. Cheng, J. Shan, W. Ju, Y. Guo, L. Zhang, Automated breast cancer detection and classification using ultrasound images: A survey, Pattern recognition 43 (2010) 299–317. doi:10.1016/j.patcog.2009.05.012.

[5] K. M. Prabusankarlal, P. Thirumoorthy, R. Manavalan, Assessment of combined textural and morphological features for diagnosis of breast masses in ultrasound, Human-centric Computing and Information Sciences 5 (2015) 1–17. doi:10.1186/s13673-015-0029-y.

[6] M. Xian, Y. Zhang, H.-D. Cheng, F. Xu, K. Huang, B. Zhang, J. Ding, C. Ning, Y. Wang, A benchmark for breast ultrasound image segmentation (BUSIS), Infinite Study, 2018. doi:10.3390/healthcare10040729.

[7] M. Xian, Y. Zhang, H.-D. Cheng, F. Xu, B. Zhang, J. Ding, Automatic breast ultrasound image segmentation: A survey, Pattern Recognition 79 (2018) 340–355. doi:10.1016/j.patcog.2018.02.012.

[8] F. Yuan, J. Shi, X. Xia, Q. Huang, X. Li, Co-occurrence matching of local binary patterns for improving visual adaption and its application to smoke recognition, IET Computer Vision 13 (2019) 178–187. doi:10.1049/iet-cvi.2018.5164.

[9] K. Stöger, D. Schneeberger, P. Kieseberg, A. Holzinger, Legal aspects of data cleansing in medical ai, Computer Law & Security Review 42 (2021) 105587. doi:10.1016/j.clsr.2021.105587.

[10] X. Qi, L. Zhang, Y. Chen, Y. Pi, Y. Chen, Q. Lv, Z. Yi, Automated diagnosis of breast ultrasonography images using deep neural networks, Medical image analysis 52 (2019) 185–198. doi:10.1016/j.media.2018.12.006.

[11] E. Michael, H. Ma, H. Li, F. Kulwa, J. Li, Breast cancer segmentation methods: current status and future potentials, BioMed Research International 2021 (2021). doi:10.1155/2021/9962109.

[12] N. Zaidkilani, M. Abdel-Nasser, M. A. Garcia, D. Puig, Breast ultrasound cad system based on efficient tumour segmentation network and transfer-learned features, in: 2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), IEEE, 2022, pp. 1–5. doi:10.1109/IMPACT55510.2022.10029203.

[13] M. H. Yap, M. Goyal, F. M. Osman, R. Martí, E. Denton, A. Juette, R. Zwiggelaar, Breast ultrasound lesions recognition: end-to-end deep learning approaches, Journal of medical imaging 6 (2018) 011007. doi:10.1117/1.JMI.6.1.011007.

[14] Y. Hu, Y. Guo, Y. Wang, J. Yu, J. Li, S. Zhou, C. Chang, Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model, Medical physics 46 (2019) 215–228. doi:10.1002/mp.13268.

[15] L. Han, Y. Huang, H. Dou, S. Wang, S. Ahamad, H. Luo, Q. Liu, J. Fan, J. Zhang, Semi-supervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network, Computer methods and programs in biomedicine 189 (2020) 105275. doi:j.cmpb.2019.105275.

[16] M. I. Daoud, A. A. Atallah, F. Awwad, M. Al-Najjar, R. Alazrai, Automatic superpixel-based segmentation method for breast ultrasound images, Expert Systems with Applications 121 (2019) 78–96. doi:10.1016/j.eswa.2018.11.024.

[17] Z. Zhuang, N. Li, A. N. Joseph Raj, V. G. Mahesh, S. Qiu, An rdau-net model for lesion segmentation in breast ultrasound images, PloS one 14 (2019) e0221535. doi:10.1371/journal.pone.0221535.

[18] E. Deng, Z. Qin, D. Chen, Z. Qin, Y. Ding, J. Geng, N. Zhang, Engan: Enhancement generative adversarial network in medical image segmentation (2022). doi:10.21203/rs.3.rs-1219874/v1.

[19] M. Byra, P. Jarosik, A. Szubert, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, M. Andre, Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network, Biomedical Signal Processing and Control 61 (2020) 102027. doi:10.1016/j.bspc.2020.102027.

[20] Q. Zhou, Q. Wang, Y. Bao, L. Kong, X. Jin, W. Ou, Laednet: A lightweight attention encoder–decoder network for ultrasound medical image segmentation, Computers and Electrical Engineering 99 (2022) 107777. doi:10.1016/j.compeleceng.2022.107777.

[21] F. Tang, J. Ding, L. Wang, M. Xian, C. Ning, Multi-level global context cross consistency model for semi-supervised ultrasound image segmentation with diffusion model, arXiv preprint arXiv:2305.09447 (2023). doi:10.48550/arXiv.2305.09447.

[22] H. Yang, D. Yang, Cswin-pnet: A cnn-swin transformer combined pyramid network for breast lesion segmentation in ultrasound images, Expert Systems with Applications 213 (2023) 119024. doi:10.1016/j.eswa.2022.119024.

[23] S. Ahmed, M. K. Hasan, Coma-net: Towards generalized medical image segmentation using complementary attention guided bipolar refinement modules, Biomedical Signal Processing and Control 86 (2023) 105198. doi:10.1016/j.bspc.2023.105198.

[24] M. Song, Y. Kim, Optimizing proportional balance between supervised and unsupervised features for ultrasound breast lesion classification, Biomedical Signal Processing and Control 87 (2024) 105443. doi:10.1016/j.bspc.2023.105443.

[25] M. Taheri, H. Omranpour, Breast cancer prediction by ensemble meta-feature space generator based on deep neural network, Biomedical Signal Processing and Control 87 (2024) 105382. doi:10.1016/j.bspc.2023.105382.

[26] J. Zhang, Z. Luan, L. Ni, L. Qi, X. Gong, Msdanet: A multi-scale dilation attention network for medical image segmentation, Biomedical Signal Processing and Control 90 (2024) 105889. doi:10.1016/j.bspc.2023.105889.

[27] Z. Lu, C. She, W. Wang, Q. Huang, Lm-net: A light-weight and multi-scale network for medical image segmentation, Computers in Biology and Medicine 168 (2024) 107717. doi:10.1016/j.compbiomed.2023.107717.

[28] H. K. Bhuyan, V. Ravi, B. Brahma, N. K. Kamila, Disease analysis using machine learning approaches in healthcare system, Health and Technology 12 (2022) 987–1005.

[29] H. K. Bhuyan, V. Ravi, An integrated framework with deep learning for segmentation and classification of cancer disease, International Journal on Artificial Intelligence Tools 32 (2023) 2340002.

[30] H. K. Bhuyan, A. Vijayaraj, V. Ravi, Diagnosis system for cancer disease using a single setting approach, Multimedia Tools Appl. 82 (2023) 46241–46267.

[31] Z. Dai, H. Liu, Q. V. Le, M. Tan, Coatnet: Marrying convolution and attention for all data sizes, CoRR abs/2106.04803 (2021). URL: https://arxiv.org/abs/2106.04803. arXiv:2106.04803.

[32] N. Zaidkilani, M. A. Garcia, D. Puig, Dual-stream coatnet models for accurate breast ultrasound image segmentation, Neural Computing and Applications (2024) 1–17.

[33] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, R. Marti, Automated breast ultrasound lesions detection using convolutional neural networks, IEEE journal of biomedical and health informatics 22 (2017) 1218–1226. doi:10.1109/JBHI.2017.2731873.

[34] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images. data brief 28, 104863 (2020), 2019. doi:10.1016/j.dib.2019.104863.

[35] V. Sovrasov, Flops counter for convolutional networks in pytorch framework, 2019. URL: https://github.com/sovrasov/flops-counter.pytorch/.