Received: 29 March 2022 | Revised: 30 December 2022 | Accepted: 30 December 2022

DOI: 10.1002/mp.16287

MEDICAL PHYSICS DATASET ARTICLE

MEDICAL PHYSICS

BUS-Set: A benchmark for quantitative evaluation of breast ultrasound segmentation networks with public datasets

Cory Thomas[1] | **Michal Byra**[2,3] | **Robert Marti**[4] | **Moi Hoon Yap**[5] | **Reyer Zwiggelaar**[1]

[1]Department of Computer Science, Aberystwyth University, Aberystwyth, UK

[2]Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland

[3]Department of Radiology, University of California, San Diego, California, USA

[4]Computer Vision and Robotics Institute, University of Girona, Girona, Spain

[5]Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK

Correspondence
Cory Thomas
Email: cot12@aber.ac.uk

Funding information
Spanish Science and Innovation projects, Grant/Award Numbers: PID2021-123390OB-C21, RTI2018-096333-B-I00; MHY research is funded by The Manchester Metropolitan Good to Great Scheme

Abstract

Purpose: BUS-Set is a reproducible benchmark for breast ultrasound (BUS) lesion segmentation, comprising of publicly available images with the aim of improving future comparisons between machine learning models within the field of BUS.

Method: Four publicly available datasets were compiled creating an overall set of 1154 BUS images, from five different scanner types. Full dataset details have been provided, which include clinical labels and detailed annotations. Furthermore, nine state-of-the-art deep learning architectures were selected to form the initial benchmark segmentation result, tested using five-fold cross-validation and MANOVA/ANOVA with Tukey statistical significance test with a threshold of 0.01. Additional evaluation of these architectures was conducted, exploring possible training bias, and lesion size and type effects.

Results: Of the nine state-of-the-art benchmarked architectures, Mask R-CNN obtained the highest overall results, with the following mean metric scores: Dice score of 0.851, intersection over union of 0.786 and pixel accuracy of 0.975. MANOVA/ANOVA and Tukey test results showed Mask R-CNN to be statistically significant better compared to all other benchmarked models with a p-value > 0.01. Moreover, Mask R-CNN achieved the highest mean Dice score of 0.839 on an additional 16 image dataset, that contained multiple lesions per image. Further analysis on regions of interest was conducted, assessing Hamming distance, depth-to-width ratio (DWR), circularity, and elongation, which showed that the Mask R-CNN's segmentations maintained the most morphological features with correlation coefficients of 0.888, 0.532, 0.876 for DWR, circularity, and elongation, respectively. Based on the correlation coefficients, statistical test indicated that Mask R-CNN was only significantly different to Sk-U-Net.

Conclusions: BUS-Set is a fully reproducible benchmark for BUS lesion segmentation obtained through the use of public datasets and GitHub. Of the state-of-the-art convolution neural network (CNN)-based architectures, Mask R-CNN achieved the highest performance overall, further analysis indicated that a training bias may have occurred due to the lesion size variation in the dataset. All dataset and architecture details are available at GitHub: https://github.com/corcor27/BUS-Set, which allows for a fully reproducible benchmark.

KEYWORDS
breast segmentation, public datasets, ultrasound

1 | INTRODUCTION AND RELATED WORK

Breast cancer is one of the most common and deadliest types of cancer, with an estimated 55 000 women and 370 men being diagnosed in the United Kingdom every year.[1] To date, the disease's 5-year average survival rate is around 85%, with the primary preventative step for reducing mortality rates being early diagnosis.[2]

Ultrasound (US) imaging is regularly used in detection and staging of breast lesions, being cheaper and more accessible than the alternate screening method of mammography.[3] However, the main drawback of US imaging is its reliance on a radiologists experience, with scans varying greatly in complexity, image quality, speckle noise, and lesion morphology.[4] Furthermore, lesions may be indistinguishable from surrounding tissue, increasing the chances of missed detection, see Figure 1 for an example. These challenges can be mitigated using computer-aided diagnosis (CAD), that can assist and improve the accuracy of a radiologist's evaluation.[5] There are numerous examples of CAD systems, which are based on a range of traditional handcrafted and/or deep learned features.[6] Although traditional image processing methods have existed longer, recent improvements of deep learning methods have pushed convolution neural networks (CNN) and transformer neural networks (TNN) to the forefront of computer vision and image analysis.[7]

As the field of deep-learning-based segmentation has been continuously expanding since its inception, it has generated many publication within the field of breast ultrasound (BUS) segmentation. One of the main issues with these studies is that due to differences in evaluation datasets and metrics, comparisons are often difficult and open to interpretation. Therefore, it has become of importance to create a reproducible benchmark dataset for BUS segmentation. Until now, a hurdle for this research area has been the lack of publicly available data that can be used in a coordinated fashion. Currently, there are five publicly available BUS datasets; OASBUD,[8] RODTOOK,[9] UDIAT,[10,11] BUSIS,[12] and BUSI.[13] The aim of this paper is to create a reproducible benchmark from the currently available public BUS datasets and to create a baseline for future comparisons in BUS lesion segmentation. Furthermore, we evaluate the datasets based on several characteristics and calculate lesion shape features for subsequent comparisons. Several deep learning architectures are evaluated with the assembled benchmark dataset with results reported. Additional comparison of the predicted segmentation results is performed assessing whether morphological shape features have been retained, or what role training bias plays in the performance of networks. Taking into account that our analysis includes comparisons with respect to the lesion type, the most recently released BUSIS dataset has been excluded in this paper, as it does not provide lesion type labels in the released version. Additionally, the benchmark is designed for datasets to be added when they become available.

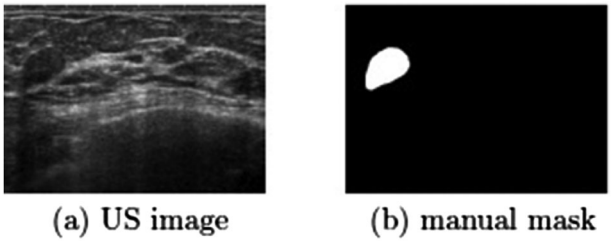

FIGURE 1 An example US image and corresponding manual mask taken from the UDIAT dataset,[10] displaying the lesion's similarity to the surrounding tissue. US, ultrasound.

In the literature, there have been several advancements in breast lesion segmentation through the means of fully convolutional networks (FCNs), a type of CNN and TNN. For instance, Yap et al. explored the use of several deep learning methods, including U-Net and transfer learning using FCN-AlexNet.[10] Additionally, they analyzed effects of key characteristics of their datasets, including lesion size and ratio of the segmented lesion region of interest (mask). In later work, Yap et al. used transfer learning models pre-trained on ImageNet for automatic segmentation of breast lesions.[11] When considering a Dice score > 0.5, they achieved 89.6 and 60.6% segmentation accuracy for benign and malignant masses, respectively. Hu et al. improved the effectiveness of FCN by addressing the issues of blurry boundaries and low contrast in images. They proposed combining a dilated FCN with a phase-based active contour model, and concluded that the dilated convolution layers improved the extraction of spatial details.[14] Chiao et al. explored the use of Mask R-CNN for the segmentation of BUS images. However, their results were reported using a validation set (no test set was defined) and no Dice evaluation was performed, making it difficult to gauge the architecture's full performance with respect to other existing approaches.[15] Byra et al. experimented with a variant of the U-Net model called the Selective kernel U-Net (Sk-U-Net), which utilizes an attention mechanism that automatically adjusts kernel sizes and the network's receptive field. Furthermore, they improved the network's ability to recognize biological objects at varying scales. They compared Sk-U-Net with a vanilla version of U-Net and found that Sk-U-Net outperformed U-net on every assessed performance metric.[16] Gomez-Flores et al. explored the application of transfer learning architectures for BUS segmentation, comparing five architectures with a variety of backbones, but they used their own private dataset.[17] They further evaluated their study by applying 10-fold cross-validation to analyze the architectures performance on smaller datasets, concluding that Deeplabv3+

performed best across the majority of their test sets. In the same year, Shareef et al. proposed the enhanced small tumor-aware network (ESTAN), a U-Net version that utilized two extraction encoders with column-wise kernels to adjust to breast anatomy, to overcome the poor segmentation performance achieved by state-of-the-art deep learning approaches on small breast lesions.[18] They compared the ESTAN model against nine state-of-the-art architectures, using a combination of the public BUSIS, BUSI, and UDIAT datasets. Zhuang et al. also proposed a U-Net variant RDAU-NET, that used a Residual Dilated Attention Gate U-Net to enhance edge information, encoder feature maps, and background suppression.[19] They evaluated their model against 10 other FCN's using BUSIS, BUSI, and UDIAT. Qu et al.[20] discussed the use of a full-resolution residual network, integrated with Global Attention Upsample and deep supervision. The model was tested on two datasets: one from Sun Yat-sen University Cancer Center and the other being UDIAT. More recently, Zhu et al.[21] published their RAT-Net model, a region aware transformer network with a U-Net backbone. This model was compared with standard U-Net configuration and four other transformer models.

A limitation of the these works is the difficulty of comparing the results due to the vast differences in datasets used. Furthermore, this is further emphasized when considering the differences in image preparation or the initial training hyper parameters. This shows the need for a reproducible benchmark to improve the current state of the art. We note that evaluation on how the architectures perform at maintaining morphological and lesion type aspects is under explored. Therefore, in this paper, we assemble a reproducible benchmark from four publicly available datasets and conduct further analysis, exploring different features of the generated prediction masks to enhance our evaluation of the architectures segmentation performances on the benchmark dataset.

2 | DATASETS

2.1 | Overview

The creation of a reproducible dataset is based on four publicly available US datasets, with each set containing US images (containing at least one lesion) and mask annotations provided by a radiologist. These datasets are OASBUD, RODTOOK, UDIAT, and BUSI, which have all been collected at various institutions. A summary of all the datasets used in this study can be found in Table 1, which includes image details and annotations, detailing field-of-view and lesion type classifications.

The OASBUD dataset was collected from patients at the Institute of Oncology, Warsaw, Poland, and consists of 100 US images (radial scans around the nipple), 48 benign and 52 malignant cases. All images contain only one lesion per image and were collected with a Ultrasonix SonixTouch Research US scanner.[8]

The RODTOOK dataset is from the Sirindhorn International Institute of Technology (SIIT), Thammasat University, Pathum Thani, Thailand. At the time of accessing the dataset, not all images were accompanied with an annotated mask. Therefore, we will exclude all the images without annotations, leaving a total of 149 US images; 59 benign and 90 malignant cases. Once again, each image contained only a single lesion and were collected using a Philips iU22 US scanner.[9]

The UDIAT dataset was collected at the UDIAT Diagnostic Centre of the Parc Tauli Corporation, Sabadell, Spain, using a Siemens ACUSON scanner. The dataset contains 163 US images: 109 benign and 54 malignant cases, with only one lesion per image.[10,11]

The BUSI dataset from the Baheya Hospital, Cairo, Egypt was collected using LOGIQ E9 and LOGIQ E9 Agile US scanners and consists of 780 US images.[13] This can be broken down into 437 benign, 210 malignant, and 133 normal cases. Since all the other datasets used in this study only contain a single lesion per image and have no normal cases, we have removed the images that contain more than one or no lesions. Therefore, reducing our total to 630 US images containing 421 benign and 209 malignant cases.

2.2 | Benchmark dataset

For our benchmark dataset, we first excluded all the images from each of the public datasets that contained more than one mass per image, which is consistent with the study of Byra et al.[16] The excluded BUS images will be used for additional analysis in Section 6.4. All the remaining US images were combined giving a total of 1154 US images for the benchmark set. The models were tested using five-fold cross-validation, with a 80/20 split for the training/testing data.

An initial preprocessing step was conducted on all datasets to remove scanner annotations contained in the US images, the details of cropped locations are available on GitHub. Last, all images were resized to 224×224 pixels using bi-cubic interpolation,[22] which is currently the standard input dimension for the CNN architectures considered.[10,16]

As one of the objectives of this paper is to make our results as reproducible as possible, we have included all the details about the images contained in each fold as comma-separated value files available on GitHub.

2.3 | Dataset comparison

Figure 2 shows examples of benign and malignant abnormalities from each of the public datasets. Qualitative analysis allows to draw various conclusions about

TABLE 1 Summary of dataset details, including the number of images for each dataset in our benchmark, benign/malignant quantities, and number of scanners used.

Dataset	Images	Benign	Malignant	Scanners	Pixel resolution
OASBUD[8]	100	48	52	1	685 × 868
RODTOOK[9]	149	59	90	1	1002 × 1125
UDIAT[10,11]	163	110	53	1	455 × 538
BUSI[13]	647	437	210	2	495 × 608

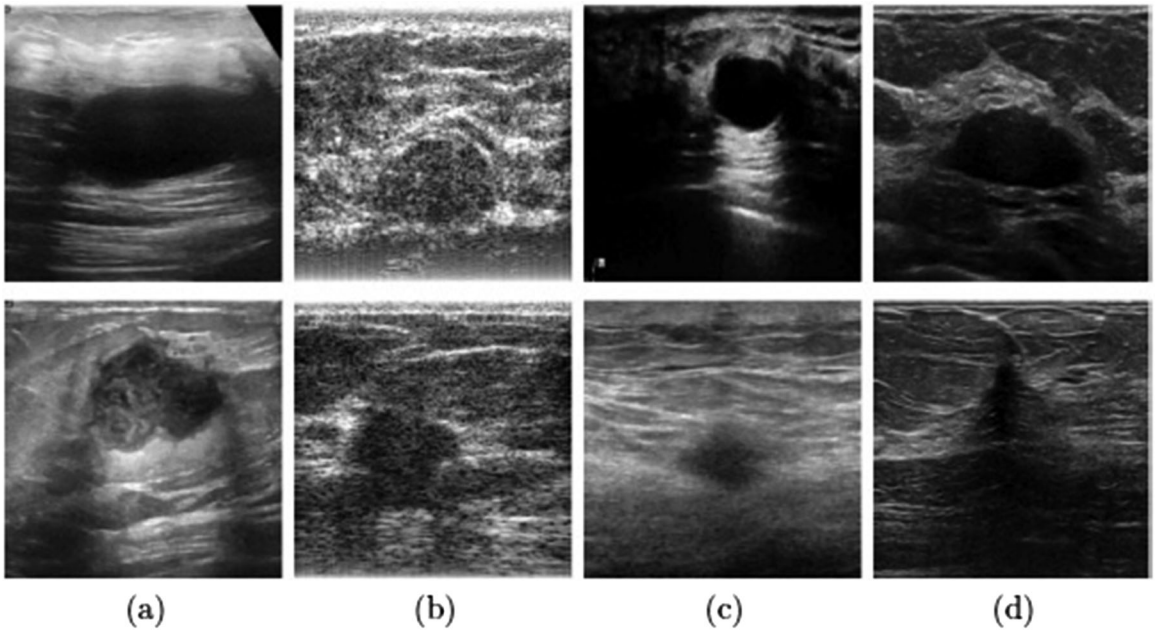

FIGURE 2 Example BUS images from BUSI (a), OASBUD (b), RODTOOK (c), and UDIAT (d), respectively. Top row: examples of benign cases; bottom row: examples of malignant cases.

differences in speckle noise and contrast in the different datasets. With regard to image quality, BUSI, UDIAT, and RODTOOK all exhibited good pixel resolution, allowing the observation of fine structures like the pectoral muscle or parenchymal tissue. However, the OASBUD dataset is of a lower resolution due to the applied image reconstruction algorithm, which was not as sophisticated as for the other scanners.[8] The appearances of lesions vary greatly in size, shape, and contrast for all datasets, with most containing clear well-defined lesions, especially compared to OASBUD. It is important to note that although the majority of the BUSI lesions are well-defined, there are several lesions that are of a low contrast and difficult to distinguish from the background. Last, speckle noise, described as granular interference caused by the environmental conditions on the imaging sensor during image acquisition[23] is also considered, with OASBUD showing this most, which is far less in all the other public datasets.

We further analyze the benchmark dataset in terms of mask area, Hausdorff distance for circularity, and moments calculation for elongation. The lesion size was estimated using pixel length from the original image as the exact metric size was not available within the meta data. All evaluations described can be seen as box plots in Figure 3, which shows that on average, the malignant lesions were larger than the benign lesions. It is worth mentioning that for both the BUSI and RODTOOK datasets, the malignant lesions are larger than the benign lesions, signifying a potential size bias in our benchmark dataset. We computed the Hausdorff distance between a circle of equivalent pixel area and annotation to analyze the irregularity of the lesions. The bar chart indicates that malignant lesions are larger than benign lesions. Furthermore, the greatest variance in terms of lesion size is displayed within the BUSI benign lesion. Last, we assessed lesions elongation using moments.[24] Once again, the BUSI dataset showed the most elongated lesions, where OASBUD showed the least elongated examples. We also found that the majority of the outliers for the Hausdorff distance are the same for the elongation estimation.

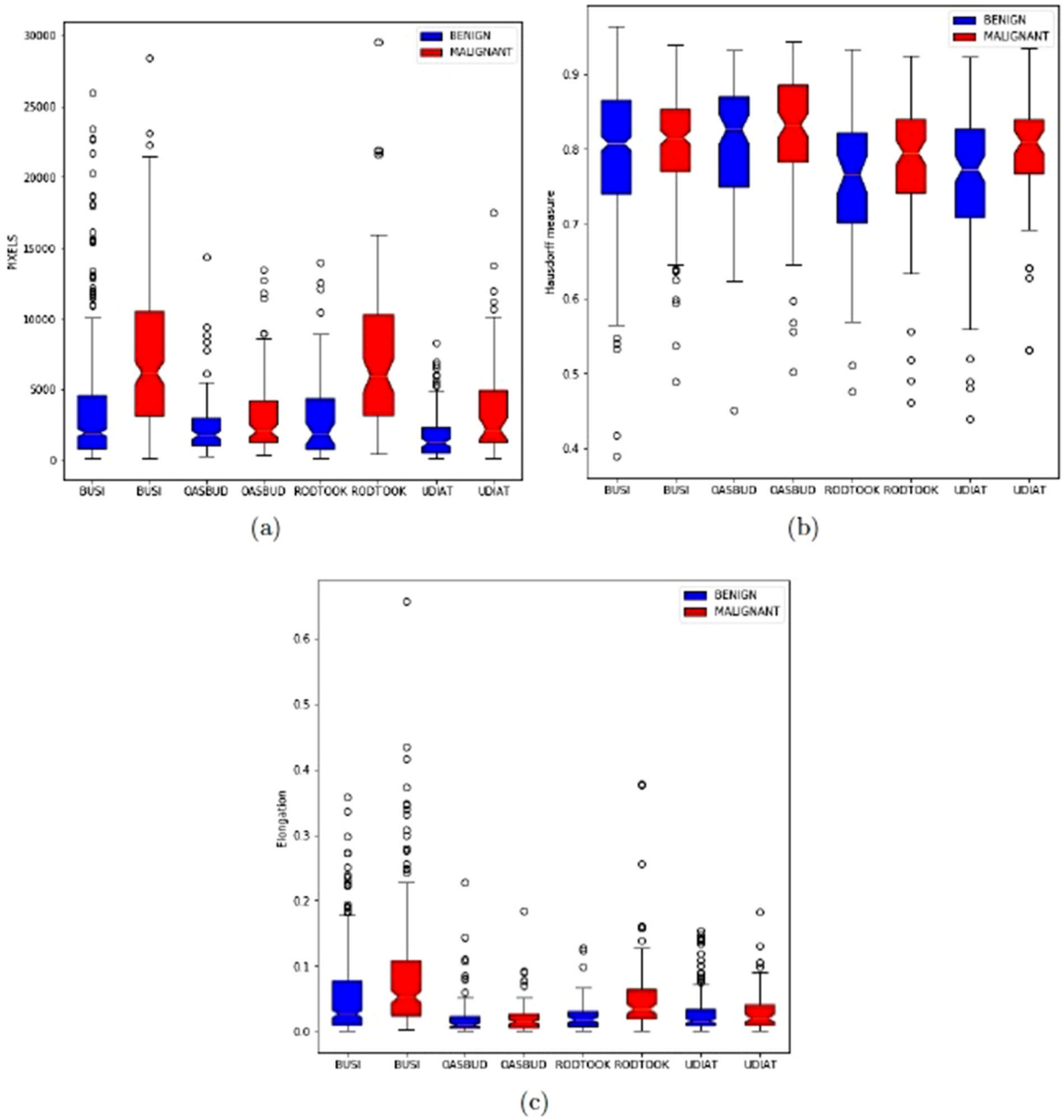

FIGURE 3 Features comparison for all four public datasets where: (a) the area of the manual mask in pixels; (b) the Hausdorff distance value compared to a circle of equivalent pixel area; and (c) elongation estimation using the lesion boundaries.

2.4 | Performance and evaluation metrics

To assess the performance, we separate our evaluation into two categories: Detection Rate (OR) (the number of correctly detected lesions, i.e., the true positive [TP], compared to the false positives [FP]) and segmentation accuracy (the number of pixels correctly identified as belonging to the lesion).

Within the related BUS and the wider literature, pixel-wise detection rate can be obtained by varying the discrimination threshold in the predicted masks, and we can then create a receiver operating characteristic curve (ROC curve) and area under curve (AUC), to get a clearer understanding of our models diagnostic ability.

For segmentation accuracy, the most frequently used evaluation metrics include pixel accuracy (Acc), Dice similarity coefficient (DSC), and the Intersection over Union metric (IoU).[10,14–17] Furthermore, true negatives (TN) and false negatives (FN) are defined here as required for the above metrics. By assessing our predicted segmentation on a pixel by pixel basis, we can calculate the overall accuracy of our models using

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}. \quad (1)$$

Although, due to such a large imbalance between mask and background pixels in some datasets, as seen in Figure 3, DSC would be considered more appropriate, due to a weighting on the TP pixels. The DSC metric is defined as two times the area of the intersection of manual and predicted mask, which is then divided by the sum of the areas of manual and predicted mask, which is given by

$$DSC(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \quad (2)$$

where X indicates the manual mask and Y is the predicted mask.

A more general case is the similar IoU metric, which calculates the intersection of the pixels found in both the manual mask and the predicted mask, whereas the union is simply comprised of all pixels found in either the manual mask or predicted mask:

$$IoU = \frac{|X \cap Y|}{|X \cup Y|}, \quad (3)$$

where X indicates the manual mask and Y is the predicted mask. Furthermore, the DSC and IoU metric results are calculated assuming that our BUS images are a single class problem (only mask), instead of a 2 class problem (background and mask). All evaluation code used within this paper can be found on GitHub.

Last, Everingham et al. suggested that the overall detection accuracy of a segmentation model can be calculated by considering predictions that achieve a DSC metric score of above 0.5.[25] Therefore, a lesion is correctly detected, or TP, if it achieves ($DSC >= 0.5$), otherwise it is classed as a FP. This method was also used by Yap et al. and then as a standalone metric by Byra et al.[11,16]

3 | METHODOLOGY

To select our benchmark architectures, we based our work on Byra et al.,[16] where they already conducted an extensive study into the capabilities of Sk-U-Net on three of the four publicly available datasets used within this study. Furthermore, they discussed several key issues within their study, including the lack of comparison with the ever popular transfer learning FCN, a subset of CNN without fully connected layers. These state-of-the-art FCN architectures have been extensively used in object detection and object-based segmentation in the literature, including medical image analysis.[16,26–28] Therefore, we include two state-of-the-art semantic segmentation FCN architectures: Matterport's implementation of Mask R-CNN[29] and TensorFlow's configuration of Deeplabv3+.[30] Furthermore, we include several state-of-the-art variations of U-Net, that is, Attention-U-Net (Att-U-Net),[31] Att-Dense-U-Net (Att-D-U-Net),[32] and U-Net++,[33] which, to our knowledge, have not been used for BUS segmentation. We also wanted to explore the capabilities of the more recent state-of-the-art vision TNNs. Dosovitskiy et al.'s vision transformer (ViT) formed the foundation of a pure transformer model for computer vision tasks.[34] Therefore, Trans-U-Net and Swin-U-Net were selected for the benchmark, being specifically built for medical image segmentation.[35,36]

3.1 | Segmentation architectures

FCN architectures differ from typical CNN as they often do not contain fully connected layers. Instead, FCNs utilize up and down sampling paths to interpret and extract image features and provide a better localization.[37]

3.1.1 | Semantic segmentation networks

Mask R-CNN[29] uses the ResNet101 backbone, a CNN that is 101 layers deep.[38] Mask R-CNN is a simple extension of the detection framework Faster-RCNN,[39] by adding a fully connected layer, achieving instance segmentation for multiple proposed masks and by predicting segmentation at a pixel level. Furthermore, Mask R-CNN utilizes a region proposal network (RPN), which generates proposed regions for the object's location within an image. The RPN can be broken down into two components; the classifier which calculates the probability of the object residing in the proposed region and the regressor, which regresses the coordinates of the proposed regions. After the RPN has proposed regions, features are then extracted from each proposed region and then bounding-box regression is performed.

Deeplabv3+[30] combines the use of Atrous Spatial Pyramid Pooling (ASPP) and encode–decoder structures to not only refine the borders of segmentation results, but also to improve the detection of small/thin objects improving fine-grained segmentation. Additionally, Deeplabv3+ typically uses the backbone model Xception-65, a 65 layer atrous CNN used to extract feature maps from the input images. An advantage of the Deeplabv3+ architecture is that it does not require large datasets and pretrained weights are readily available. A diagram of Deeplabv3+ can be found in Figure 4.

3.1.2 | FCN-based U-Net architectures

The U-Net architecture is based on a FCN model proposed by Ronneberger et al.[40] for biomedical image segmentation to overcome the need for large scale datasets, which has since been integrated with more state-of-the-art techniques. The U-Net architecture is

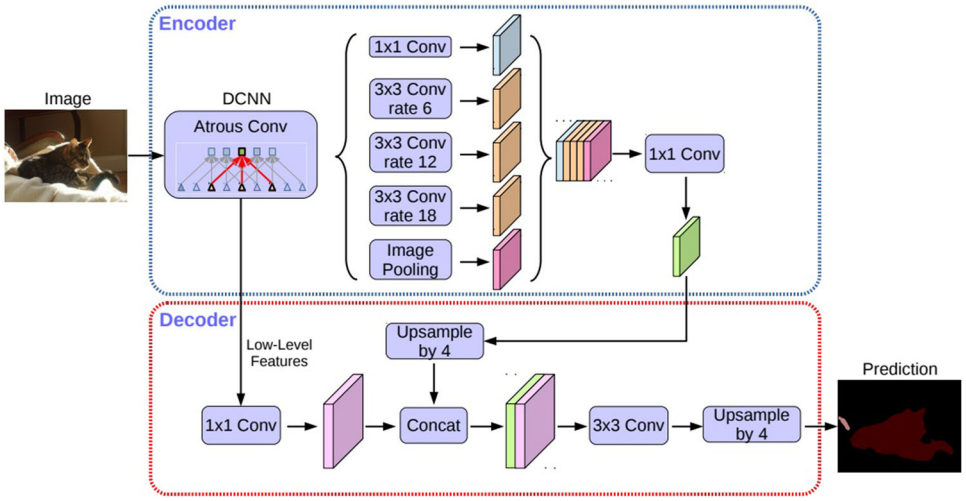

FIGURE 4 The architecture of Deeplabv3+. Reprinted with permission from [Springer Nature]: [Elsevier] [Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015] [Encoder–Decoder with Atrous Separable Convolution for Semantic Image Segmentation, Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., [COPYRIGHT] (2018).[30].

synonymous with that of an encoder–decoder architecture, containing both a contraction path (encoder) and a symmetric expansion path (decoder).[40] Att-U-Net[31] is a modified version of U-Net for tissue/organ segmentation. It employs the use of attention gates (AG) that focus on target structures, while suppressing irrelevant features. Usually, it provides a better mask localization without loss to its receptive field.[31]

Att-D-U-Net is further integrated with densely connected encoders. These dense blocks create a feed-forward fashion between each layer strengthening feature propagation, substantially reducing the number of parameters. This architecture has been applied in the field of digital mammography, achieving better segmentation results than U-Net, Att-U-Net, and Dense-U-Net without AG.[32]

Sk-U-Net[16] is another variant of the U-Net model architecture, but with the conventional blocks replaced by Sk blocks, which automatically adjusts its receptive field, resulting in a better utilization of the spatial information at varying scales, to obtain a high-level feature representation. This model architecture has shown to be more robust for classification, displaying considerable improvements over the vanilla U-Net as seen in Byra et al.[16] A diagram of Sk-U-Net can be found in Figure 5.

U-Net++[33] is a nested architecture where the encoder and decoder subnetworks are connected through a series of nested, dense skip pathways, reducing the semantic gap between the feature maps of the encoder and decoder subnetworks. Therefore, improving learning with the decoder and encoder feature maps being semantically similar. The configuration of this network can be found in Figure 6. This model has been widely evaluated for segmentation applications concluding an improved performances over U-Net in nuclei, liver, and colon polyp segmentation.[33]

3.1.3 | U-Net-based transformer networks

We have selected two transformer-based models based on the U-Net described above, each with their own unique transformer block or encoder that replaces the conventional ones. Both models utilize a modified version of the ViT architecture,[34] which splits an image up into patches with their respected linearly activated pixel-wise feature maps generated from convolutions layers. These feature maps are then transformed into a sequence of tokens and fed into the transformer. These tokens are then sequenced, outputted, and projected back to the feature maps. Allowing the analysis of low-level pixel-wise structures through tokens-wise embedding, lowering computational cost compared to CNNs.[34]

Trans-U-Net is the first architecture to utilize a modified version of the ViT architecture designed for medical image segmentation. The model uses a transformer encoder as described above, and the same decoder structure with upsampling and skip connections as in the vanilla U-Net. Their transformer uses multilayer perceptron (MLP) blocks and multihead self-attention (MSA) layers to create encoded tokenized image patches from convolutions layer feature maps. These encoded features maps are then upsampled with skip connections for accurate localization.[35]

Swin-U-Net varies from Trans-U-Net by instead of having a transformer encoder, it uses Swin-transformer blocks that replace the conventional blocks in vanilla U-Net. The encoder transformer symmetric blocks use shifted windows MSA and patch merging layers at their base to contextualize features, whereas the decoder uses patch expanding layers to upsample the extracted deep features from the bottleneck layer. Then, the symmetric blocks upsample the

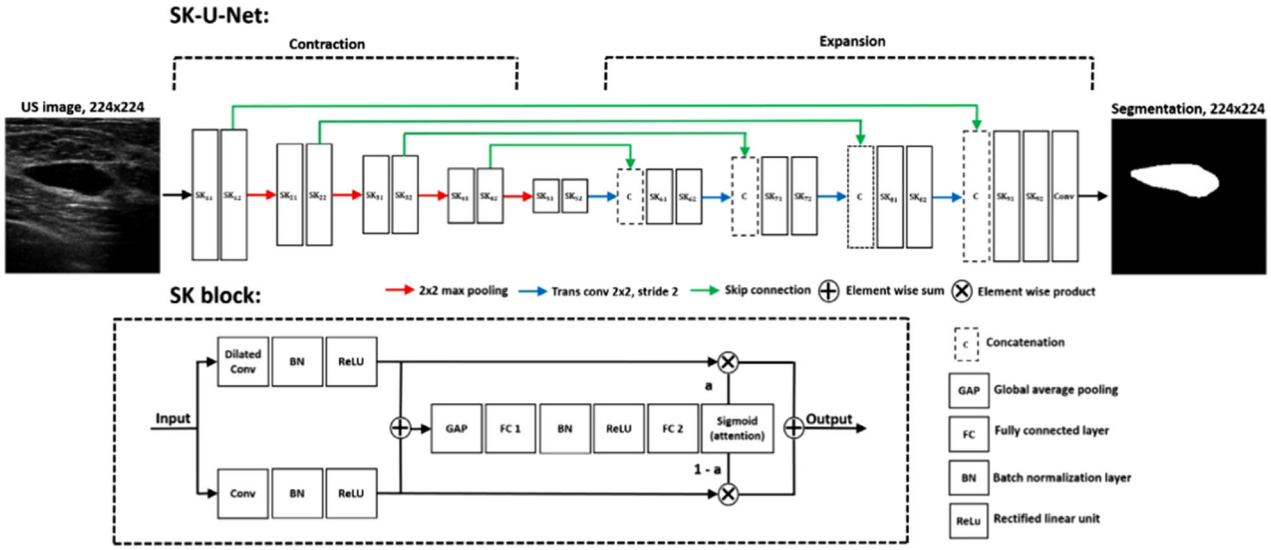

FIGURE 5 The architecture of Sk-U-Net. Reprinted with permission from [Elsevier]: [Elsevier] [Biomedical Signal Processing and Control] [Breastmass segmentation in ultrasound with selective kernel U-Net convolutional neural network, Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Ol-son, L., O'Boyle, M., Comstock, C., Andre, M. [COPYRIGHT] (2020).[16].

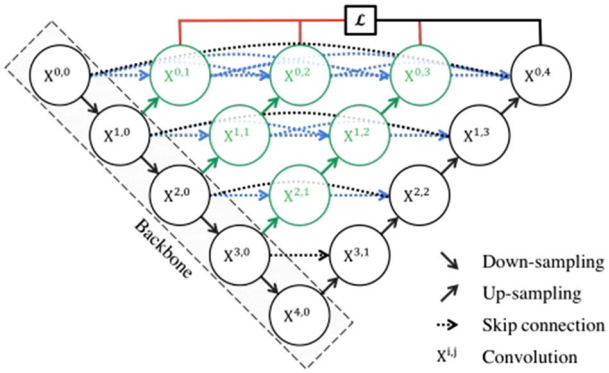

FIGURE 6 The architecture of U-Net++. Reprinted with permission from [Springer Nature]: [Elsevier] [Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015] [U-Net++:A Nested U-Net Architecture for Medical Image Segmentation, Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J. [COPYRIGHT] (2018).[33].

features maps along with concatenation from the skip connections.[36]

3.2 | Implementation

All architectures tested within the benchmark were run on color images, the format the datasets were received in. Mask R-CNN,[29] Swin-U-Net, and Trans-U-Net used pretrained weights acquired from their respected GitHub repositories. Unfortunately, we had difficulty with the original Deeplabv3+[30] source and an alternate was used without pretrained weights. The remaining models were all trained from scratch.

For training Mask R-CNN, we experimented with two backbones: ResNet50 and ResNet101, we found that ResNet101 performed the best during our preliminary experiments using only training and validation datasets. We initialized Mask R-CNN with learning rate of 0.0005 and optimized using stochastic gradient descent (SGD) with learning momentum equal to 0.9. We set the batch size to 1 due to memory constraints and trained using per-pixel softMax and a multinomial loss. Furthermore, we set the weight decay to 0.0001. We initially trained only the "head" layers of the network for 10 epochs and then trained the whole model for an additional 30 epochs. After each epoch, the model weights were saved and the model with the highest average DSC score on the validation set was selected.

For benchmarking, all the FCN U-Net-based architectures and DeepLabv3+ were trained with a learning rate of 0.001 and optimized using Adam with learning momentum of 0.9. Furthermore, we set the batch size to 16 and we decayed our learning rate exponentially by a factor of 0.1. Both models were trained using the DSC metric (see Equation 2) and early stopping was implemented such that training was stopped if there was no improvement within our validation set after 15 epochs. We used data augmentation to improve training by vertically flipping each of the training images. We have also seeded our model to facilitate as much reproducible as possible. Each different architecture was run three times and the model with the highest DSC score on our validation set was selected to remain consistent with Byra et al.[16] Additionally, we implemented Deeplabv3+ with the modified Xception71 backbone as described in Chen et al.,[30] along with atrous rates of 4, 8, 12.

Last, both TNNs were trained with a maximum epochs of 150 using a batch size of 4 because of GPU constraints. Swin-U-Net was initialized with a base learning rate 0.05, decay rate 0.0001, and SGD optimizer with

TABLE 2 Summary of the two systems used to implement the benchmark algorithms.

Configurations	Site (C.T.) (UK)	Site (M.B.) (Poland)
CPU	Intel(R) Xeon(R) Gold 6148 CPU	AMD Ryzen 7 3700X
RAM	384 GB	32 GB
GPU	Nvidia V100 GPU	Nvidia RTX 2080 Ti
Operating system	Linux	Windows 10

momentum 0.9, staying consistent with Cao et al.'s implementation.[36] Tran-U-Net had an initial learning rate of 0.01 and SGD optimizer, decay rate and momentum as Swin-U-Net, and the same as used in Chen et al.[35] Furthermore, both architectures had their own default augmentation that we utilized.

With regards to the Cudatoolkit, tensorflow, or pytorch and if required Cudnn, the following versions were used; Mask R-CNN: *Tensorflow* = 1.14.0, *Cuda* = 10.1.243, *Cudnn* = 7.6.5.32; FCN U-Net-based architectures and DeepLabv3+: *Tensorflow* = 2.4.1, *Cuda* = 10.1.243, *Cudnn* = 7.6.5; transformer networks: *Pytorch* = 1.11.0, *Cuda* = 11.3.1.

To ensure that our results were as reproducible as possible, each of the architectures was trained by one author at one site (C.T.) (UK) and then by another author at a different site (M.B.) (Poland), and vice versa. Summary of each site configuration can be found in Table 2. Both authors used the same environments, which can be obtained from GitHub.

After all our results had been collected separately, the best results (highest DSC score on each fold) from the two sites were compared and the best being selected as the benchmark. With the maximum difference between the two sites over all folds being 0.015 per DSC score.

3.3 | Statistical significance test

To statistically validate the benchmarked nine methods, we performed a MANOVA to determine if the multivariate sample means are equal. When appropriate, this was followed Anova and a Tukey's Honest Significant Difference (HSD) post hoc test to statistically compare the different methods. Furthermore, DSC and ACC metric scores are analyzed separately with IOU not being used due to its correlation with DSC. The significance threshold used within this study is 0.01.

3.4 | Further analysis methods

Qualitative results

Qualitative analysis provides a complementary perspective on the segmentation results specially related to lesion morphology and dataset characteristics. Lesions were randomly selected from each dataset where the average DSC value over all architectures was above 0.8, displaying the capabilities of the architectures. For simplicity, this analysis will only include the three best performing architectures: best semantic model, best U-Net FCN variant, best U-Net transformer variant, selected based on average DSC scores.

Lesion Accuracy/Morphology

To explore the models capabilities in terms of Lesion Accuracy/Morphology, we first calculated the Hamming distance[41] between the ground truth (manual) and prediction (automatic) masks, to obtain the percentage of pixels that were segmented correctly. Although segmentation metrics are commonly used for the evaluation of different approaches, we also propose to investigate the effects of different segmentation approaches on the lesion shape and morphology when compared to the ground truth. This was conducted to evaluate morphological properties of the obtained segmentation, assessing the level of robustness of these features. The three best average performing architectures were selected for this evaluation.

Comparing the manual and automatic masks through shape analysis

To compare the manual and automatic masks, three morphological features were analyzed: the depth-to-width ratio (DWR), circularity, and elongation. DWR was determined by calculating the major and minor axis lengths for a segmentation mask, then dividing the major by the minor we obtain a scalar estimate for the DWR. Circularity and elongation were calculated as described in Section 2.3. Both these features have been used for breast mass classification, and segmentation methods are expected to generate masks that provide accurate estimates of these basic shape descriptors.[42] Additionally, we only compared lesions that were classified as a TP. Segmentation masks were resized to the original ground truth image size in order to extract the shape descriptors to be compared. Furthermore, significance analysis was conducted transforming the correlation coefficients of DWR, circularity, and elongation to *z*-values and subsequently, we estimated the observed value of z (z_{obs}).[43] To estimate significance, we used an alpha of 0.01 and assumed a two-tailed test, meaning that the difference is significant if the calculated observed *z*-value is outside $-2.58 < z_{obs} < 2.58$.

Influence of lesion size

For this evaluation, we conduct two sets of analysis. First, we assess the size of the lesions in the benchmark and filter our results by their respective public datasets, followed by calculating their DSC and DR scores. Allowing us to making a direct comparison between predictions and ground truths.

Second, in the literature, there has been some evidence of training biases directly affecting the DSC metric scores, as described by Maier-Hein et al.[44] They mentioned the possible drawbacks of training architectures using the DSC metric when a dataset is biased towards small/large abnormalities.[44] They concluded that the DSC metric can be appropriately used for large structures, for example, organs instead of smaller pathological structures. This indicates that an architecture trained using the DSC metric for an imbalanced lesion size dataset could cause bias in the segmentation performance, favouring the dominant lesion size and exhibiting bias towards the majority class.[45] The DSC metric focuses on the segmented region during training, inducing a bias towards a specific region size (lesion size). Furthermore, we would also expect a similar bias to arise in a model trained using any of the other region-based loss function, for example, the IoU metric.

To investigate whether lesion size influences the predicted lesion size, we evaluate the same three models as in *Lesion Accuracy/Morphology*. above. The original lesion size is then plotted against its respective prediction lesion size. With the aim of finding any correlation between lesions sizes on higher or lower DSC scores.

Multiple lesions

Last, during benchmarking, the dataset was refined to include only single lesion BUS images, but this is not always the case. We evaluated the benchmarked architectures on the excluded 16 images from the BUSI dataset. Each BUS image contained two or three lesions, with 15 images being benign and one malignant. This analysis provides some insights on the robustness of the three best performing models in segmenting multiple lesions. Furthermore, the statistical evaluation described in Section 3.3 will also be applied if appropriate.

4 | RESULTS

This section presents the segmentation results with the proposed database in terms of quantitative (using different metrics) and qualitative evaluations. In addition, we show results based on lesion size and lesion elongation. Finally, we show how the various approaches deal with the presence of multiple lesions per image.

4.1 | Benchmark results

Table 3 displays benchmark segmentation results for all nine architectures over the five BUS folds. Focusing on the five U-Net-based FCN models, Sk-U-Net achieved the highest mean score for all lesions across the four performance metrics, obtaining *DSC* : 0.748, *IoU* : 0.652, and *Acc* : 0.965. Furthermore, it also achieved the highest detection rate of 0.848, obtaining 979 true detections. Looking at lesion type segmentation, Sk-U-Net still remained the highest performing U-Net-based CNN architecture, with *DSC* : 0.764, *IoU* : 0.675, *Acc* : 0.976 for benign and *DSC* : 0.723, *IoU* : 0.618, *Acc* : 0.949 for malignant BUS images. The lowest scoring U-Net-based CNN model was Att-D-U-Net, obtaining mean metric values of *DSC* : 0.640, *IoU* : 0.528, *Acc* : 0.954 for all lesion types. Comparing Att-D-U-Net to Sk-U-Net in terms of metrics, Att-D-U-Net performed 14.4% lower on DSC, 19.0% lower on IoU, and 1.1% lower on Acc. Att-D-U-Net performance can be better understood when we consider that the model only achieved a detection rate of 0.742. For semantic models, Mask R-CNN produced the highest metric scores of *DSC* : 0.851, *IoU* : 0.786, and *Acc* : 0.975. Deeplabv3+ achieved lower metric scores with *DSC* : 0.722, *IoU* : 0.621, and *Acc* : 0.963. When considering the benign and malignant breakdown, Mask R-CNN performed best on the benign subset achieving a mean DSC score of 0.863 compared to 0.831 for the malignant images. Deeplabv3+ performed similar, achieving its highest mean *DSC* score on benign lesions and remaining consistent over the other metrics.

Both transformer systems achieved similar results, with Trans-U-Net obtaining the highest metric scores of *DSC* : 0.761, *IoU* : 0.672, and *Acc* : 0.968, compared to Swin-U-Nets: *DSC* : 0.747, *IoU* : 0.642, and *Acc* : 0.964, a difference of 0.014, 0.030, and 0.004 for DSC, IOU, ACC, respectively. On the benign and malignant breakdown, Trans-U-Net obtained the highest results on the benign lesions with *DSC* : 0.793, *IoU* : 0.713, and *Acc* : 0.979. Swin-U-Net also achieved its highest mean Dice score on benign lesions, with *DSC* : 0.763, although for the malignant lesions, Swin-U-Net achieved a higher mean *DSC* score than Trans-U-Net with *DSC* : 0.721, *DSC* : 0.712. Although Trans-U-Net obtained higher means score for the other two metrics *IoU* and *Acc*.

To further validate our results, we conducted a MANOVA test using the DSC, IoU, and Acc metric scores, which indicated a statistically significant difference between all our models with a p-value <0.01. Subsequently, we employed an Anova test on the three individual metrics, which again showed a statistically significant difference between all our models with a p-value < 0.01. To distinguish which models were significantly different, we used the HSD test on the DSC and Acc metrics (we do not include IoU results at this stage as they are strongly correlated with DSC and the equivalent IoU results can be found on GitHub) separately with an alpha of 0.01. The results can be found in Figure 7.

Figure 8 shows the ROC curves and AUC values for all benchmarked architectures, displaying the models segmentation ability to distinguish between lesion and surrounding tissue. Swin-U-Net achieved the highest AUC value across all possible thresholds, closely followed by Trans-U-Net and MaskRCNN. There is no

TABLE 3 Segmentation benchmark mean metrics scores with median and standard deviation displayed within brackets.

Evaluation	Method		Settings		Metrics			Detection		
lesion type	model	Backbone	Initial LR	BS	DSC	IoU	Acc	TP	FP	DR
All	DeepLabv3+	Xception 65	1*e*−2	16	0.722 (0.848 ± 0.288)	0.621 (0.722 ± 0.282)	0.963 (0.983 ± 0.051)	952	202	0.825
	Mask R-CNN	ResNet101	5*e*−3	1	**0.851 (0.932 ± 0.238)**	**0.786 (0.870 ± 0.238)**	**0.975 (0.993 ± 0.052)**	**1059**	**95**	**0.918**
	U-Net	N/A	1*e*−2	16	0.707 (0.834 ± 0.298)	0.606 (0.709 ± 0.289)	0.961 (0.983 ± 0.053)	931	223	0.807
	Sk-U-Net	N/A	1*e*−2	16	0.748 (0.868 ± 0.282)	0.652 (0.757 ± 0.277)	0.965 (0.986 ± 0.052)	979	175	0.848
	Att-D-U-Net	N/A	1*e*−2	16	0.640 (0.747 ± 0.300)	0.528 (0.590 ± 0.287)	0.954 (0.973 ± 0.055)	856	298	0.742
	Att-U-Net	N/A	1*e*−2	16	0.709 (0.841 ± 0.297)	0.608 (0.714 ± 0.288)	0.962 (0.983 ± 0.053)	938	216	0.813
	U-Net++	N/A	1*e*−2	16	0.704 (0.835 ± 0.298)	0.602 (0.701 ± 0.289)	0.961 (0.983 ± 0.054)	927	227	0.803
	Swin-U-Net	N/A	1*e*−3	4	0.747 (0.847 ± 0.254)	0.642 (0.730 ± 0.255)	0.964 (0.984 ± 0.051)	995	159	0.862
	Trans-U-Net	N/A	1*e*−2	4	0.761 (0.887 ± 0.276)	0.672 (0.791 ± 0.277)	0.968 (0.988 ± 0.050)	981	173	0.850
Benign	Deeplabv3+	Xception 65	1*e*−2	16	0.734 (0.878 ± 0.291)	0.647 (0.764 ± 0.285)	0.975 (0.990 ± 0.042)	584	115	0.835
	Mask R-CNN	ResNet101	5*e*−3	1	**0.863 (0.943 ± 0.240)**	**0.806 (0.888 ± 0.239)**	**0.982 (0.996 ± 0.046)**	**644**	**55**	**0.921**
	U-Net	N/A	1*e*−2	16	0.730 (0.873 ± 0.303)	0.635 (0.745 ± 0.295)	0.973 (0.990 ± 0.047)	567	132	0.811
	Sk-U-Net	N/A	1*e*−2	16	0.764 (0.896 ± 0.290)	0.675 (0.803 ± 0.284)	0.976 (0.992 ± 0.044)	593	106	0.848
	Att-D-U-Net	N/A	1*e*−2	16	0.657 (0.791 ± 0.310)	0.550 (0.639 ± 0.298)	0.966 (0.983 ± 0.047)	521	178	0.745
	Att-U-Net	N/A	1*e*−2	16	0.721 (0.867 ± 0.308)	0.626 (0.750 ± 0.300)	0.972 (0.990 ± 0.048)	566	133	0.810
	U-Net++	N/A	1*e*−2	16	0.717 (0.868 ± 0.311)	0.621 (0.749 ± 0.301)	0.972 (0.989 ± 0.046)	556	143	0.795
	Swin-U-Net	N/A	1*e*−3	4	0.763 (0.870 ± 0.259)	0.664 (0.760 ± 0.259)	0.975 (0.990 ± 0.042)	604	95	0.864
	Trans-U-Net	N/A	1*e*−2	4	0.793 (0.912 ± 0.269)	0.713 (0.833 ± 0.272)	0.979 (0.993 ± 0.042)	605	94	0.866
Malignant	Deeplabv3+	Xception 65	1*e*−2	16	0.689 (0.802 ± 0.282)	0.581 (0.669 ± 0.271)	0.946 (0.968 ± 0.058)	369	86	0.811
	Mask R-CNN	ResNet101	5*e*−3	1	**0.831 (0.914 ± 0.233)**	**0.756 (0.835 ± 0.233)**	**0.965 (0.983 ± 0.060)**	**415**	**40**	**0.912**
	U-Net	N/A	1*e*−2	16	0.672 (0.789 ± 0.286)	0.561 (0.645 ± 0.274)	0.944 (0.963 ± 0.057)	362	93	0.796
	Sk-U-Net	N/A	1*e*−2	16	0.723 (0.827 ± 0.268)	0.618 (0.704 ± 0.261)	0.949 (0.969 ± 0.058)	386	69	0.848
	Att-D-U-Net	N/A	1*e*−2	16	0.613 (0.702 ± 0.283)	0.494 (0.539 ± 0.264)	0.934 (0.954 ± 0.060)	335	120	0.735
	Att-U-Net	N/A	1*e*−2	16	0.691 (0.797 ± 0.277)	0.581 (0.661 ± 0.267)	0.945 (0.965 ± 0.058)	372	83	0.818
	U-Net++	N/A	1*e*−2	16	0.685 (0.789 ± 0.276)	0.574 (0.651 ± 0.266)	0.944 (0.964 ± 0.059)	371	84	0.815
	Swin-U-Net	N/A	1*e*−3	4	0.721 (0.812 ± 0.245)	0.609 (0.686 ± 0.245)	0.947 (0.966 ± 0.059)	391	64	0.859
	Trans-U-Net	N/A	1*e*−2	4	0.712 (0.825 ± 0.279)	0.610 (0.704 ± 0.274)	0.951 (0.973 ± 0.058)	376	79	0.826

TPs and FPs values are shown, indicating whether the architectures are detecting and localizing the lesion correctly. Results were obtained over the five BUS set folds. Highest mean in every column are displayed in bold, with one per section.

Abbreviations: DSC, Dice similarity coefficient; FP, false positives; IoU, intersection over union; TP, true positive.

Acc \ DSC	Mask R-CNN	Trans-U-Net	Sk-U-Net	Swin-U-Net	Deeplabv3+	Att-U-Net	U-Net	U-Net++	Att-D-U-Net
Mask R-CNN		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Trans-U-Net	0.019		0.900	0.900	0.024	0.001	0.001	0.001	0.001
Sk-U-Net	0.001	0.900		0.900	0.399	0.025	0.015	0.007	0.001
Swin-U-Net	0.001	0.604	0.900		0.457	0.033	0.020	0.009	0.001
Deeplabv3+	0.001	0.472	0.900	0.900		0.900	0.900	0.843	0.001
Att-U-Net	0.001	0.099	0.765	0.900	0.900		0.900	0.900	0.001
U-Net	0.001	0.073	0.698	0.900	0.900	0.900		0.900	0.001
U-Net++	0.001	0.052	0.625	0.900	0.900	0.900	0.900		0.001
Att-D-U-Net	0.001	0.001	0.001	0.001	0.001	0.001	0.010	0.015	

FIGURE 7 HSD results of benchmark architectures, with top right showing *p*-value (and hence statistical significant differences) when evaluating DSC and bottom left for Acc. Values displayed in blue, highlight where the comparison between models was statistically significant, for example, *p*-value < 0.01. DSC, Dice similarity coefficient; HSD, Honest significant difference.

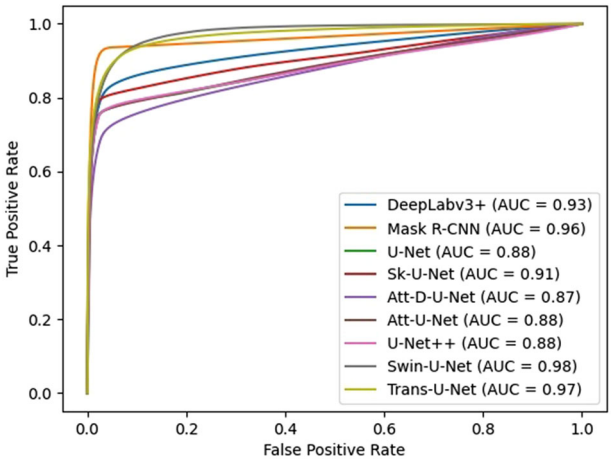

FIGURE 8 ROC curves for the benchmarked architectures, with the AUCs for all models being displayed. AUC, area under curve; ROC, receiver operating characteristic curve.

statically significant difference between the AUC values for Mask R-CNN, Swin-U-Net, Trans-U-Net, but it should be noted that the ROC for Mask R-CNN provides the closest point to (0,1) (i.e., which would be a perfect detector).

Qualitative results

Qualitative segmentation results are shown in Figures 9 and 10, for benign and malignant lesions, respectively. For benign images, most architectures provide reasonable segmentation performances and capture well-defined boundaries of the benign lesions. Furthermore, there does not seem to be a drop in the segmentation performance due to the image quality of different scanners. With respect to the malignant masses, the methods showed a good level of segmentation on lesions with DSC > 0.5.

Figures 11 and 12 show benign and malignant lesions that achieved the lowest average mean DSC scores. For benign cases, we observe that the BUSI and OASBUD lesions have not been detected by any of the architectures. Instead, choosing to detect a nearby shadow apart from Sk-U-Net, which failed to make any detection. For RODTOOK, Sk-U-Net and Trans-U-Net detected two visible lesions with one being within the ground truth, whereas Mask R-CNN, which was set to detect only one region, selected the nearby shadow over instead of lesion. Last, for the UDIAT image, Mask R-CNN managed to locate the lesion with *DSC* : 0.913, whereas, again, Sk-U-Net and Trans-U-Net obtained false detections, highlighting a nearby shadow.

In Figure 12, the BUSI malignant lesion has been undetected, with all of the architectures detecting the large posterior acoustic shadowing (PAS) region instead. Usually, a PAS overlaps with the lesion resulting in over segmentation, or the PAS is separate from the lesion causing a false detection.[46] A similar case is found for the UDIAT case, although this time the PAS overlaps with the region, resulting in over segmentation. For the RODTOOK case, only Mask R-CNN managed to partially locate the lesion, but obtained a poor mean DSC score due to over segmentation. Last, looking at the OASBUD image, all architectures failed to make an accurate prediction, only Sk-U-Net produced a prediction.

Lesion Accuracy/Morphology

For this analysis, we selected only the three best performing architectures as previously stated: Mask R-CNN (best semantic model), Sk-U-Net (best U-Net FCN variant), Trans-U-Net (best U-Net transformer variant) based on average DSC scores.

Assessing the architectures' capabilities to maintain a lesion accuracy/morphology. We first calculated the Hamming distance[41] between the manual and prediction masks, to obtain the percentage of pixels that were segmented correctly. For the three models, we found

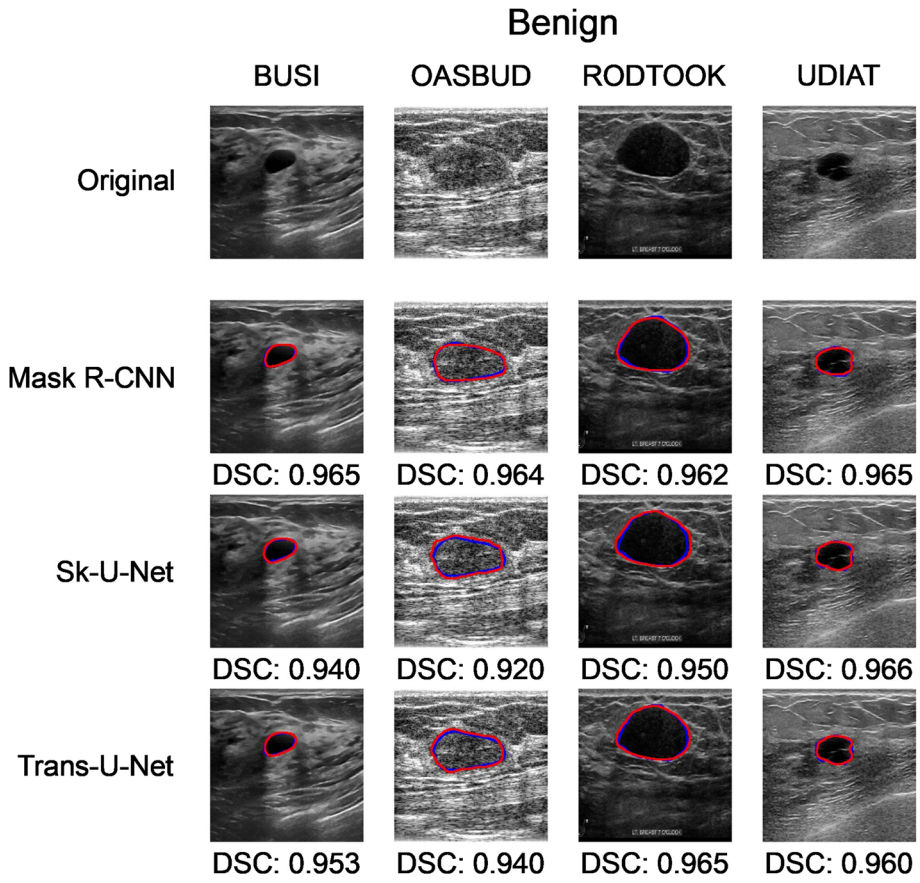

FIGURE 9 Segmentation results of the best-segmented benign lesion from the best performing model from each group are presented. The lesions were randomly selected from the segmentations with a DSC above 0.8 for all networks. Manual segmentation mask shown in"Blue" and prediction mask shown in "Red." DSC, Dice similarity coefficient.

that on average, Mask R-CNN segments 84.3% of the manual mask pixels correctly, whereas, Sk-U-Net obtained 76.8% and Trans-U-Net achieves 74.77%, showing that Sk-U-Net and Trans-U-Net achieved similar accuracy, but due to false detection's Sk-U-Net, average DSC was lower than Trans-U-Net.

Comparing the manual and automatic masks through shape analysis

Figure 13 shows the DWR, circularity, and elongation analysis for our three deep learning architectures. Regarding DWR, we found that Mask R-CNN captured the most accurate estimates and the highest linear correlation coefficient, 0.888, compared to Sk-U-Net's 0.628 and Trans-U-Net's 0.746. Mask R-CNN performed the best, most likely because of its bounding box improved lesion localization. A similar conclusion is drawn regarding circularity, with Mask R-CNN maintaining the most circularity characteristics of the lesions, with higher correlation coefficient of 0.532 compared to Sk-U-Net, 0.374, and Trans-U-Net, 0.409. Last, for elongation, we found that once again, Mask R-CNN achieved the highest correlation coefficient of 0.876. This was closely followed by Trans-U-Net with score of 0.713, whereas Sk-U-Net obtained a far lower score of 0.504, as the Sk-U-Net predictions were found to be larger due to over segmentation. For DWR, there is statistical significant difference between Mask R-CNN and Sk-U-Net with $z_{obs} = 19.650$, Sk-U-Net and Trans-U-Net with $z_{obs} = 18.353$, and no significant difference between Mask R-CNN and Trans-U-Net with $z_{obs} = 1.297$. For circularity, there is a statistical difference across all models, for Mask R-CNN and Sk-U-Net with $z_{obs} = 8.274$, Sk-U-Net and Trans-U-Net with $z_{obs} = 3.875$, Mask R-CNN and Trans-U-Net with $z_{obs} = 4.399$. With regards to Elongation, there is a statistical difference between Mask R-CNN and Sk-U-Net with$z_{obs} = 12.696$, Sk-U-Net and Trans-U-Net with $z_{obs} = 11.003$, but there is no statistical significant difference between Mask R-CNN and Trans-U-Net with $z_{obs} = 1.693$.

Influence of lesion size

The size distibution of the lesions in the benchmark is presented in Figure 3. We observed that the RODTOOK and BUSI sets contain larger malignant lesions. Furthermore, BUSI and OASBUD contain larger benign lesions.

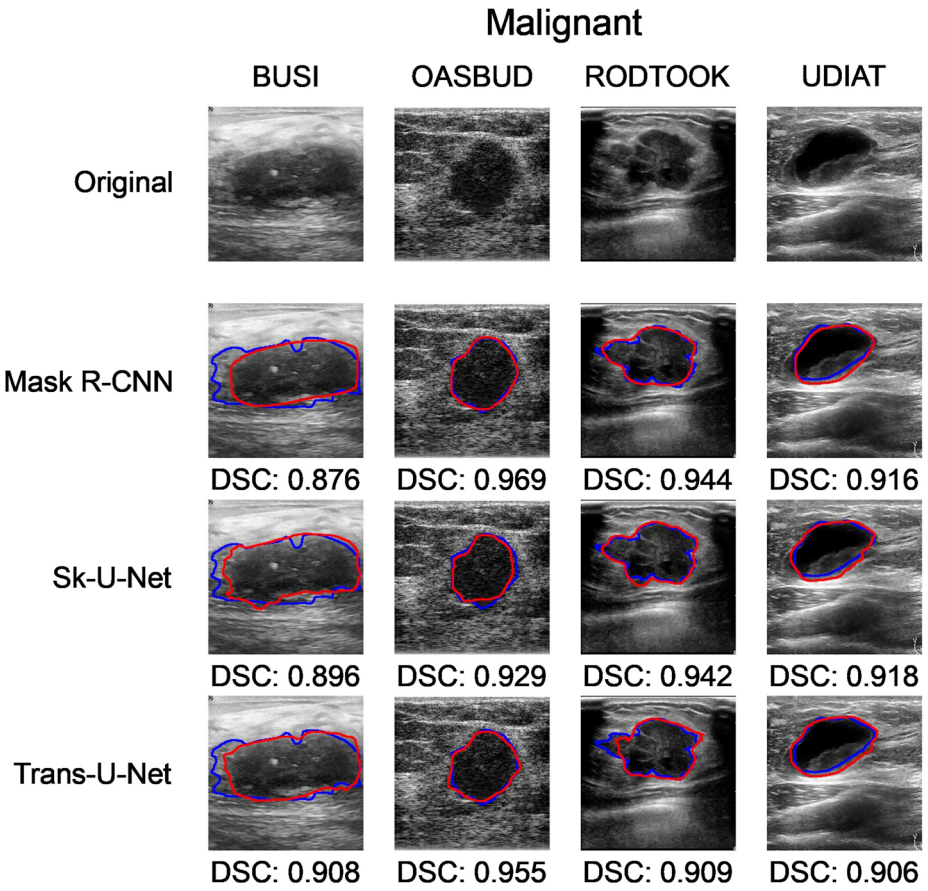

FIGURE 10 Segmentation results of the best-segmented malignant lesion from the best performing model from each group are presented. The lesions were randomly selected from the segmentations with a DSC above 0.8 for all networks. Manual segmentation mask shown in "Blue" and prediction mask shown in "Red." DSC, Dice similarity coefficient.

TABLE 4 DSC scores for different architectures, public datasets, and types of lesions.

DATASET	Type	DeepLabv3+	Mask R-CNN	U-Net	Sk-U-Net	Att-D-U-Net	Att-U-Net	U-Net++	Swin-U-Net	Trans-U-Net
BUSI	Malignant	0.701	**0.817**	0.691	0.714	0.640	0.694	0.689	0.717	0.709
	Benign	0.738	**0.854**	0.728	0.753	0.654	0.715	0.715	0.769	0.801
OASBUD	Malignant	0.544	**0.799**	0.550	0.645	0.504	0.579	0.591	0.650	0.605
	Benign	0.698	**0.872**	0.709	0.751	0.595	0.698	0.710	0.733	0.717
RODTOOK	Malignant	0.801	**0.898**	0.776	0.822	0.688	0.798	0.786	0.802	0.803
	Benign	0.759	**0.827**	0.719	0.775	0.665	0.718	0.708	0.770	0.777
UDIAT	Malignant	0.730	**0.845**	0.665	0.750	0.601	0.720	0.690	0.747	0.779
	Benign	0.795	**0.912**	0.763	0.813	0.723	0.762	0.735	0.762	0.838

Bold values indicate the best performance.

Filtering our results by their respective public datasets and calculating their DSC and DR scores are illustrated in Tables 4 and 5.

Considering benign images, on average, the architectures did the best on the UDIAT images with an average mean DSC score of 0.789, then followed by BUSI: 0.747, RODTOOK: 0.746, and OASBUD: 0.720. It is an interesting observation that UDIAT has the smallest average benign lesion size. Excluding OASBUD from our results, as its score is most likely affected by the image quality, we observed that the models performed better on smaller lesions. The models achieved near identical scores, out by 0.001, for RODTOOK and BUSI, which have a similar average lesion size. This suggests a correlation between smaller lesions and higher DSC scores.

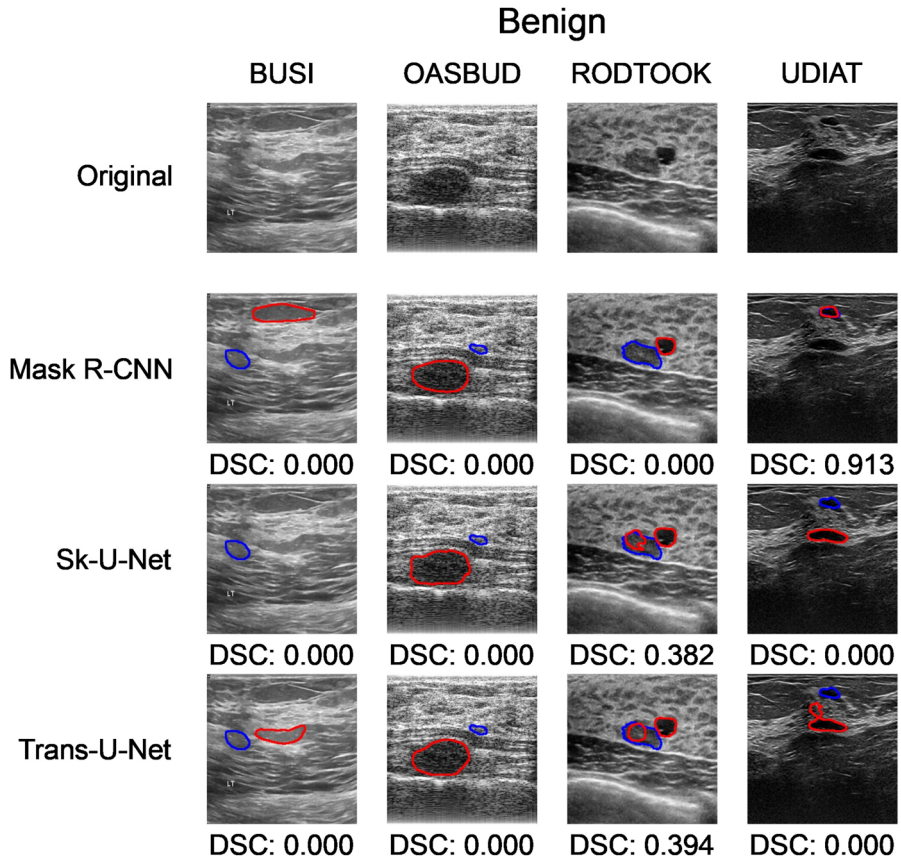

FIGURE 11 Segmentation results of the worst-segmented benign lesion from the best performing model from each group are presented. Manual segmentation mask shown in "Blue" and prediction mask shown in "Red."

TABLE 5 DR for different architectures, public datasets, and types of lesions.

DATASET	Type	DeepLabv3+	Mask R-CNN	U-Net	Sk-U-Net	Att-D-U-Net	Att-U-Net	U-Net++	Swin-U-Net	Trans-U-Net
BUSI	Malignant	0.819	**0.900**	0.833	0.843	0.776	0.819	0.838	0.862	0.824
	Benign	0.826	**0.906**	0.805	0.824	0.735	0.796	0.785	0.865	0.863
OASBUD	Malignant	0.644	**0.875**	0.635	0.760	0.596	0.702	0.683	0.760	0.712
	Benign	0.802	**0.927**	0.823	0.854	0.719	0.802	0.812	0.854	0.802
RODTOOK	Malignant	0.954	**0.977**	0.920	0.943	0.931	0.931	0.931	0.943	0.931
	Benign	0.860	**0.930**	0.825	0.912	0.772	0.825	0.825	0.912	0.895
UDIAT	Malignant	0.870	**0.926**	0.759	0.889	0.722	0.852	0.796	0.907	0.889
	Benign	0.890	**0.972**	0.835	0.908	0.798	0.853	0.807	0.844	0.917

Bold values indicate the best performance.

For malignant images, the highest average mean DSC score, 0.797, was obtained on the RODTOOK dataset, which contains the second largest average malignant lesion size. Moreover, the UDIAT dataset, which contains a lower than average lesion size, obtained a score of 0.725. The two other datasets obtained the following scores BUSI: 0.708 and OASBUD: 0.605, indicating a bias towards larger lesions size for the malignant BUS images.

To investigate whether lesion size influences the predicted lesion size, we evaluated the best three models as described, with the aim of analyzing if there is any correlation between lesions sizes and higher or lower DSC scores. Figure 3 details lesions size profiles in the datasets. Comparing this with the graph on the left in Figure 14, we can see that smaller lesions are being of higher concentration, with most of the larger lesions being malignant. This is also supported

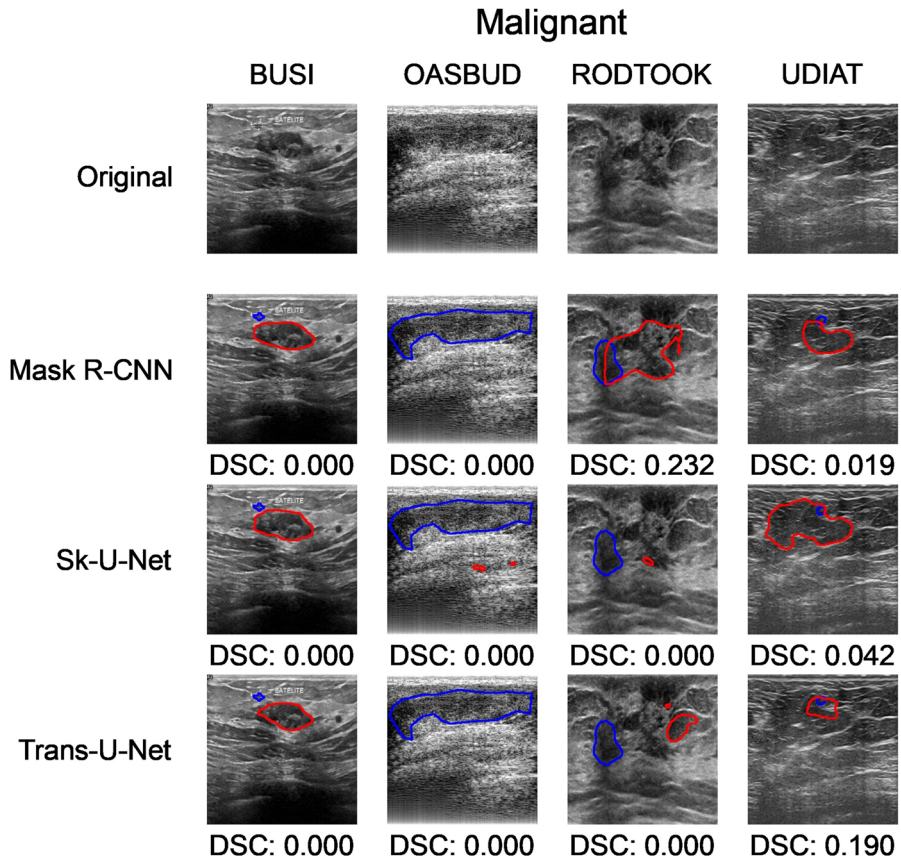

FIGURE 12 Segmentation results of the worst-segmented malignant lesion from the best performing model from each group are presented. Manual segmentation mask shown in "Blue" and prediction mask shown in "Red."

TABLE 6 Multiple lesions segmentation results, displaying mean metrics scores, median and standard deviation within brackets.

Method	Metrics		
model	DSC	IoU	Acc
Mask R-CNN	**0.839** (**0.916** ± **0.118**)	**0.740** (**0.845** ± **0.169**)	**0.977** (**0.986** ± **0.023**)
Sk-U-Net	0.707 (0.849 ± 0.268)	0.574 (0.713 ± 0.278)	0.956 (0.986 ± 0.061)
Trans-U-Net	0.592 (0.677 ± 0.305)	0.482 (0.511 ± 0.287)	0.953 (0.984 ± 0.058)

Results were obtained on 16 images (15 benign and one malignant) from the BUSI dataset, where there was more than one lesion per image; highest mean for every column is displayed in bold.
Abbreviations: CNN, convolution neural network; DSC, Dice similarity coefficient; IoU, intersection over union.

when considering the group means for each model. The means and standard deviation for ground truth lesion sizes are: 617.12 ± 251.35, 1738.66 ± 399.39, 3860.99 ± 960.95, 10630.76 ± 4518.07. Then, if we consider Mask R-CNN with prediction size means and standard deviation for ground truth lesion sizes are: 996.56 ± 1709.15, 2066.17 ± 2135.08, 4022.75 ± 1787.79, 9859.00 ± 4262.98. This trend stays the same when looking at the mean for the other two models.

Multiple lesions

Multiple lesion detection results can be found in Table 6. Mask R-CNN performed the best, achieving the highest mean scores across all metrics, followed by Sk-U-Net and then Trans-U-Net. Splitting the results on lesion type, Mask R-CNN, Sk-U-Net, and Trans-U-Net obtained the DSC scores 0.759, 0.844, and 0, respectively, for the single malignant case. For the remaining 15 benign cases, they produced mean dice scores of 0.844, 0.696, and 0.632. Surprisingly, Mask R-CNN received a higher mean score on multiple lesions compared to its benchmark performance, displaying the capabilities of semantic segmentation on multiple lesions of BUS images, although Sk-U-Net did achieve a higher average performance than Mask R-CNN on the single malignant lesion. Statistically using MANOVA based on DSC, IoU,

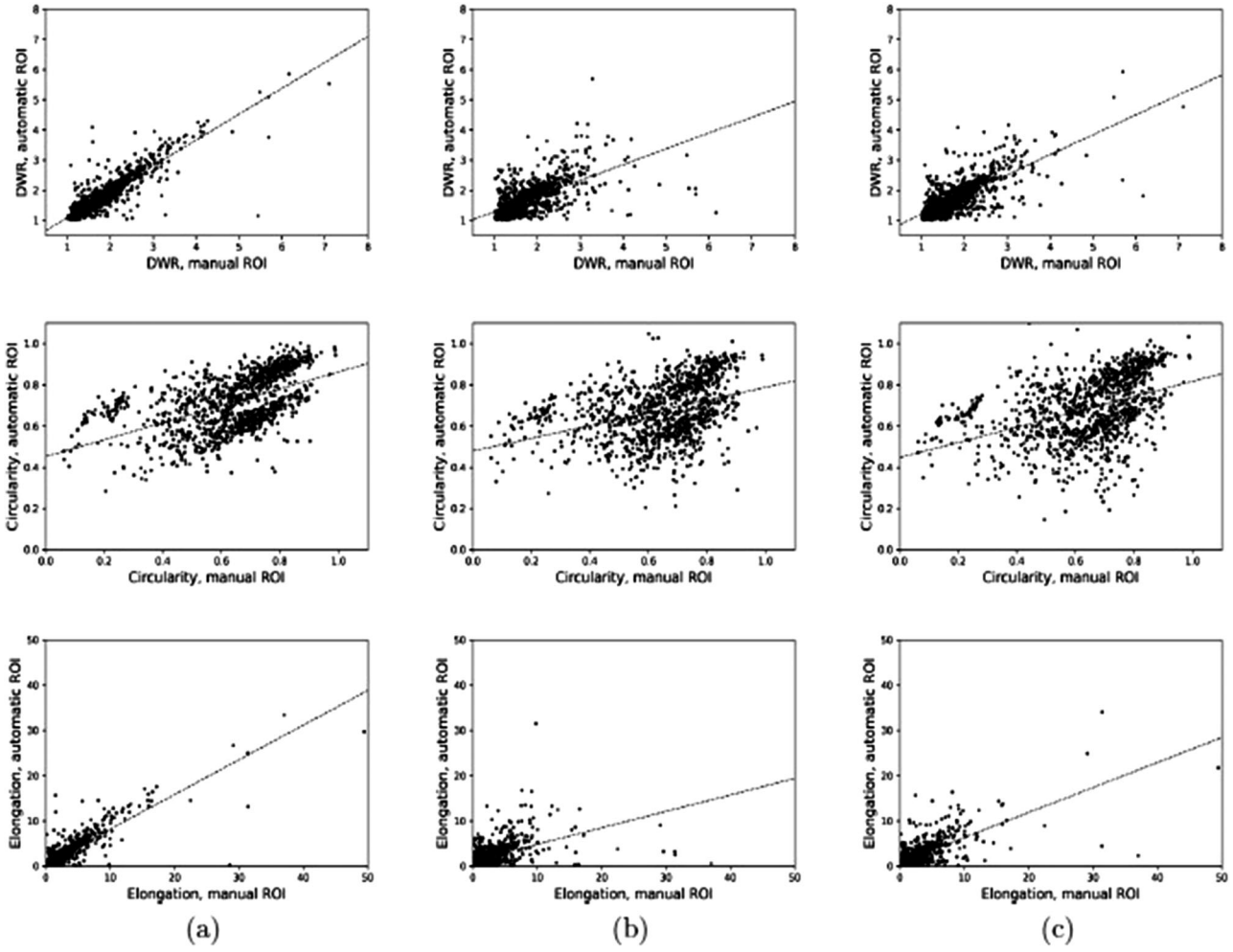

FIGURE 13 The top row displays the depth-to-width ratio of the segmentation, the middle row measures the circularity ratio of the segmentation and the bottom shows elongation for Mask R-CNN, Sk-U-Net, and Trans-U-Net, shown as (a), (b), and (c), respectively. Furthermore, the line represents the linear approximation. DWR, depth-to-width ratio.

and Acc, we find a p-value of 0.065, which is greater than 0.01, and as such not indicating a statistically significant difference between the three models.

5 | DISCUSSION

Benchmark discussion

Mask R-CNN achieved the highest segmentation results compared to all the other architectures shown in Table 3. Followed by Trans-U-Net and then Sk-U-Net, where both achieved lower segmentation performance with respect to all metrics. For the benign and malignant breakdown, once again Mask R-CNN achieved the highest metrics scores and this was consistent for both benign and malignant masses.

Looking at the statistical analysis displayed in Figure 7. For DSC, the HSD indicated a significant statistical difference for Mask R-CNN and Att-D-U-Net and all other models, confirmed by p-value < 0.01 (to note that Mask R-CNN is better than all other models, and that Att-D-U-Net performs worse than the other models). For Acc, HSD indicated that there was a significant statistical difference between Mask R-CNN and all other models apart from Trans-U-Net. Furthermore, it also showed that there was a significant statistical difference between Att-D-U-Net and other models, apart from U-Net and U-Net++. Highlighting that when TN are included in the metric calculations, it becomes more ambiguous in regards to clear distinction between models.

Benchmarks comparison

The capabilities of deep-learning-based methods for segmenting breast masses can be seen clearly in the results in Table 3, which showed that on average, 8/9 of the explored architectures achieved a DSC score above 70%, with only Att-D-U-Net achieving an average score below 65%.

Considering benign and malignant types separately within the benchmark, on average, the models

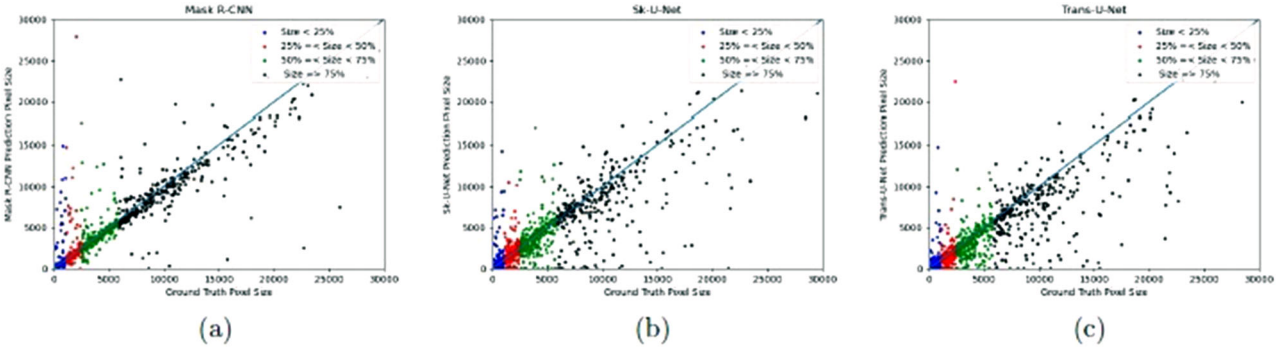

FIGURE 14 Manual mask lesion pixel area against automatic mask pixel area. All lesions were split into groups based on the median and 25 and 75 interquartile range of the ground truth shown in the charts legends, and with Mask R-CNN, Sk-U-Net, and Trans-U-Net, shown as (a), (b), and (c), respectively.

performed better on benign cases, with the two best performing models being Mask R-CNN and Trans-U-Net. In contrast, the worst performing method, the Att-D-U-Net, obtained its best scores on the malignant set. This could be explained by the fact that the ratio in the benchmark set is biased towards benign images. Additionally, benign lesions usually exhibit well-defined boundaries, compared to malignant lesions. Overall, the benchmark results highlight the importance of Trans-U-Nets transformer block and Mask R-CNN bounding box localization for improving detection rates. Interestingly, when considering the U-Net-based architectures that are integrated with AG and dense blocks, we find that there does not seem to be any improvement with regards to the metric scores.

Our obtained benchmarks express similar results to other papers on BUS segmentation, although direct comparison is difficult, due to a variety of datasets and methodologies. However, there are two papers that were based on a combination of some of the public datasets used in this study. Byra et al. achieved similar results with Sk-U-Net and U-Net obtaining 0.826 and 0.778 for DSC scores, respectively, a difference of 0.048. Whereas, in our benchmark, Sk-U-Net managed a mean DSC of 0.748 and U-Net 0.707, a difference of 0.041. With such a small variation in performance between each architecture, we observed that similar performances were achieved, although they used a dataset of 882 lesions compared to 1154 in this work.[16] Another comparable study was conducted by Gomez-Flores et al., where DeepLabv3+ achieved a median DSC score of 0.902, whereas in our benchmarks, we obtained a median DSC 0.848.[17] However, there were some important differences as they used a private dataset consisting of 3061 BUS images. Interestingly, they also applied 10-fold cross-validation, to assess the architectures based on different US machines, with the lowest median DSC being 0.81.[17] Many examples of the predicted segmentations across several models were also shown, which their qualitative results show similar morphological aspects of our own results. A thorough comparison is difficult as they only displayed examples that obtained a good segmentation for all models tested.

Lesion Accuracy/Morphology

For the three models, Mask R-CNN obtained the highest accuracy of 84.3% when considering the Hamming distance.[41] Furthermore, Figure 13 showed that Mask R-CNN was found to maintain the most morphological features associated with lesions, followed by Trans-U-Net and then Sk-U-Net. Overall, Sk-U-Net obtained far lower scores for DWR, circularity, and elongation when compared with the other networks. It is also interesting to see that the Mask R-CNN and Trans-U-Net prediction masks display a far-more circular appearance, reinforced by the high elongation correlation coefficient as both networks seem to miss the finer details of the manual masks producing a more circular prediction. Then from the statistical analysis, we could only draw that over the three methods, Mask R-CNN is statistically significantly different from Sk-U-Net.

Influence of lesion size

Results presented in Tables 4 and 5 show that our models performed the best on the RODTOOK dataset, achieving an average mean DSC of 0.777 over both lesion types, followed by UDIAT: 0.767, BUSI: 0.735, and OASBUD: 0.625. The poorer performance of the architectures on the OASBUD set is likely because of the image quality and smaller lesions, as indicated in Figure 3. This is reinforced by Table 3, indicating that the UDIAT images have the lowest average detection rate, especially for the benign images. This could indicate that the performance drop on the UDIAT images is because of the lesions size, resulting in fewer features for architectures to learn from, making detection difficult. We expected the RODTOOK benign images to also have low detection rates, which is different to what can be found in Table 5, which shows UDIAT having a lower detection rate. Each of the public datasets was also obtained on a variety of different scanners, leading to differences in BUS image resolution.

Furthermore, it is generally difficult to compare performance score obtained based on different datasets. Specific biases present in the training data may result in better performance of specific networks. Therefore, lesion size may have generated a training bias, as the overall set contains a large number of bigger lesions, as seen in Figure 3. This could directly affect the size of the generated masks and in turn the DSC metric scores.

In Figure 14, where for three models, the original lesion size is plotted against its respective prediction lesion size, there seems to be a distinct level of bias with large lesions being predicted smaller and small/medium lesions being predicted larger. Considering that Mask R-CNN was trained using IoU loss compared to Sk-U-Net and Trans-U-Net, which were trained using DSC loss, it performed poorer on smaller masses. Note: these results align well with the work by Maier-Hein et al.[44]

Multiple lesions

It was hard to draw any conclusion from the multiple lesion benchmark in Table 6. Although Mask R-CNN did obtained the highest performance, the MANOVA expressed that there was no statistical significance difference between models likely probably due to the insufficient number of testing samples.

5.1 | Limitations and future work

Based on our investigations, we identified several issues related to this study. The four public datasets contained within this study were annotated by different radiologists, which could create variations, especially along the boundaries where it becomes more personal interpretation. Furthermore, image acquisition protocols were most likely different between centers leading to more variations between datasets. An evaluation of the agreement between radiologists could help to improve the consistency of the masks. Second, we did not enhance our segmentation with post processing methods, like region growing or watershed,[47,48] which might help to reduce the amount of features lost along the boundaries of the lesions. We also did not compare different loss functions and optimization methods for training.

Our future work will focus on feature differences between benign and malignant lesions and the semantic segmentation in 2D and 3D US. Additionally, we also plan to develop our own segmentation networks, building on the research performed within this study and conduct further investigation on different scanners, lesions types, and how these affect network performance.

6 | CONCLUSIONS

We have proposed a reproducible benchmark, tested across two different system configurations using publicly available data. Our benchmark results found that Mask R-CNN achieved the best overall results on the benchmark dataset (1154 BUS images) achieving metric scores of $DSC: 0.851$, $IoU: 0.786$, and $Acc: 0.975$ using five-fold cross-validation. Furthermore, our results were further analyzed using MANOVA that indicated a statistically significant difference between models with a p-value < 0.01. Further evaluation using one-way ANOVA and Tukey test across the DSC metric scores highlighted a significant difference between Mask R-CNN and Att-D-U-Net and all other models. This became more ambiguous when considering the Acc metric score, which indicated significant statistical difference between Mask R-CNN and all other models apart from Trans-U-Net. Additionally, it showed a significant statistical difference between Att-D-U-Net and other models, apart from U-Net and U-Net++.

From evaluating the prediction masks, we found that Mask R-CNN presented the highest linear correlation with DWR, circularity, and elongation, therefore, it maintained the morphological features of the lesions most accurately. We further found that models missed similar structures along the mask boundaries. Further statistical analysis based on the linear correlation coefficients for the three methods separately, indicated that Mask R-CNN is statistically significantly different from Sk-U-Net.

Possible challenges of the benchmark have been highlighted, with reasonable evidence to conclude that there is a training bias in the benchmark results. As there are a smaller number of malignant images causing the architectures to over segment or incorrectly detect lesions, affecting the DSC metric score. For this work, we can see evidence of a DSC training bias but more research would be required to evaluate the effects. Finally, three architectures were assessed on 16 images with multiple lesions, where Mask R-CNN performed the best with a mean DSC score of 0.895. Although the difference between models were not statistically significant. All dataset details and evaluation code used within this study are available on GitHub.

ACKNOWLEDGMENTS

This work was supported by the SCW (Super Computing Wales), CDT AIMLAC, RM research is funded by Spanish Science and Innovation projects PID2021-123390OB-C21 and RTI2018-096333-B-I00 and MHY research is funded by The Manchester Metropolitan Good to Great Scheme.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

1. Breast Cancer Now. Facts and statistics 2020. Last accessed December 2020. https://breastcancernow.org/about-us/media/facts-statistics

2. Breast Cancer UK. Survival for breast cancer | Breast Cancer | Cancer Research UK. December 2020. Last accessed December 2020. https://www.cancerresearchuk.org/about-cancer/breast-cancer/survival
3. Golemati S, Gastounioti A, Nikita KS. Toward novel noninvasive and low-cost markers for predicting strokes in asymptomatic carotid atherosclerosis: the role of ultrasound image analysis. *IEEE Trans Biomed Eng*. 2013;60(3):652-658.
4. Clement GT, Hynynen K. A non-invasive method for focusing ultrasound through the human skull. *Phys Med Biol*. 2002;47(8):1219-1236.
5. Samulski MRM, Snoeren PR, Platel B, et al. Computer-aided detection as a decision assistant in chest radiography. In *Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment*. Vol 7966. International Society for Optics and Photonics; 2011:796614.
6. Ruschin M, Tingberg A, Båth M, et al. Using simple mathematical functions to simulate pathological structures' input for digital mammography clinical trial. *Radiat Prot Dosim*. 2005;114(1-3):424-431. https://academic.oup.com/rpd/article/114/1-3/424/1596018
7. Altaf F, Islam SMS, Akhtar N, Janjua NK. Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access*. 2019;7:99540-99572.
8. Piotrzkowska-Wróblewsk Ha, Dobruch-Sobczak K, Byra M, Nowicki A. Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. *Med Phys*. 2017;44(11):6105-6109. doi:10.1002/mp.12538
9. Rodtook A, Kirimasthong K, Lohitvisate W, Makhanov SS. Automatic initialization of active contours and level set method in ultrasound images of breast abnormalities. *Pattern Recognit*. 2018;79:172-182.
10. Yap MH, Pons G, Marti J, et al. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE J Biomed Health Inf*. 2018;22(4):1218-1226.
11. Yap MH, Goyal M, Osman FM, et al. Breast ultrasound lesions recognition: end-to-end deep learning approaches. *J Med Imaging*. 2018;6(1):011007.
12. Zhang Y, Xian M, Cheng HD et al. BUSIS: a benchmark for breast ultrasound image segmentation. *Healthcare (Basel)*. 2022;10(4):729. doi:10.3390/healthcare10040729
13. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data Brief*. 2020;28:104863.
14. Hu Y, Guo Y, Wang Y, et al. Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model. *Med Phys*. 2019;46(1):215-228.
15. Chiao J-Y, Chen K-Y, Liao K, Hsieh P-H, Zhang G, Huang T-C. Detection and classification the breast tumors using mask R-CNN on sonograms. *Medicine*. 2019;98(19):e15200. insights.ovid.com
16. Byra M, Jarosik P, Szubert A, et al. Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. *Biomed Signal Process Control*. 2020;61:102027. http://www.sciencedirect.com/science/article/pii/S174680942030183X
17. Gómez-Flores W, Coelho de Albuquerque Pereira W. A comparative study of pre-trained convolutional neural networks for semantic segmentation of breast tumors in ultrasound. *Comput Biol Med*. 2020;126:104036.
18. Shareef B, Xian M, Vakanski A. Stan: small tumor-aware network for breast ultrasound image segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). 2020:1–5.
19. Zhuang Z, Li N, Joseph Raj AN, Mahesh VGV, Qiu S. An RDAU-NET model for lesion segmentation in breast ultrasound images. *PLoS ONE*. 2019;14(8):e0221535.
20. Qu X, Shi Y, Hou Y, Jiang J. An attention-supervised full-resolution residual network for the segmentation of breast ultrasound images. *Med Phys*. 2020;47(11):5702-5714. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7905659/
21. Zhu X, Hu H, Wang H, et al. Region aware transformer for automatic breast ultrasound tumor segmentation. In: Medical Imaging with Deep Learning. 2022. https://openreview.net/forum?id=2bVDHzy7xwV
22. Parsania P, Virparia DV. A review: image interpolation techniques for image scaling. *Int J Innov Res Comput Commun Eng*. 2015;02:7409-7414.
23. Maity A, Pattanaik A, Sagnika S, Pani S. A comparative study on approaches to speckle noise reduction in images. In: 2015 International Conference on Computational Intelligence and Networks. 2015:148-155.
24. Stojmenović M, Žunić J. Measuring elongation from shape boundary. *J Math Imaging Vision*. 2008;30(1):73-85. doi:10.1007/s10851-007-0039-0
25. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The Pascal visual object classes (VOC) challenge. *Int J Comput Vision*. 2010;88(2):303-338. http://link.springer.com/10.1007/s11263-009-0275-4
26. Vuola AO, Akram SU, Kannala J. Mask-RCNN and U-Net ensembled for nuclei segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). 2019:208-212.
27. Wang K, Khan N, Highnam R. Automated segmentation of breast arterial calcifications from digital mammography. In: 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ).
28. Johnson JW. Adapting mask-rcnn for automatic nucleus segmentation. Proceedings of the 2019 Computer Vision Conference, Vol. 2. *arXiv preprint arXiv:1805.00500* 2018.
29. He K, Gkioxari G, Dollár P, Girshick R, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.322
30. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, vol 11211. Springer, Cham; 2018. doi:10.1007/978-3-030-01234-2_49
31. Oktay O, Schlemper J, Le Folgoc L, et al. Attention U-Net: learning where to look for the pancreas. *Med Imaging Deep Learn*, 2018. https://openreview.net/forum?id=Skft7cijM
32. Li S, Dong M, Du G, Mu X. Attention dense-U-Net for automatic breast mass segmentation in digital mammogram. *IEEE Access*. 2019;7:59037-59047.
33. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. UNet++: a nested U-Net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA ML-CDS 2018 2018. Lecture Notes in Computer Science*, vol 11045. Springer, Cham; 2018. doi:10.1007/978-3-030-00889-5_1
34. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognitionat scale. 9th International Conference on Learning Representations, ICLR 2021,Virtual Event, Austria, May 3—7, 2021. OpenReview.net, 2021. https://openreview.net/forum?id=YicbFdNTTy
35. Chen J, Lu Y, Yu Q, et al. TransUNet: transformers make strong encoders for medical image segmentation. *CoRR*. abs/2102.04306, 2021. https://arxiv.org/abs/2102.04306
36. Cao H, Wang Y, Chen J, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky L, Michaeli T, Nishino K, eds. *Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science*, vol 13803. Springer, Cham; 2023. doi:10.1007/978-3-031-25066-8_9
37. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Computer Vision and Pattern Recognition (cs.CV). 2014. doi:10.48550/arxiv.1411.4038

38. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778. doi:10.1109/CVPR.2016.90
39. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks, 2015, MIT Press, Cambridge, MA, USA, Proceedings of the 28th International Conference on Neural Information Processing Systems. Volume 1, pages 91–99, Montreal, Canada, NIPS'15.
40. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, vol 9351. Springer, Cham; 2015. doi:10.1007/978-3-319-24574-4_28
41. Norouzi M, Fleet DJ, Salakhutdinov RR. Hamming distance metric learning. *Adv Neural Inf Process Syst*. 2012;25:1061–1069.
42. Yap MH, Yap CH. Breast ultrasound lesions classification: a performance evaluation between manual delineation and computer segmentation. *Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment*. Vol. 9787. SPIE, 2016.
43. Pallant J. *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS*. 7th ed. Routledge; 2020.
44. Maier-Hein L, Reinke A, Kozubek M, et al. BIAS: transparent reporting of biomedical image analysis challenges. *Med Image Anal*. 2020;66:101796.
45. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *J Big Data*. 2019;6(1):27.
46. Zhou Z, Wu S, Chang K-J, et al. Classification of benign and malignant breast tumors in ultrasound images with posterior acoustic shadowing using half-contour features. *J Med Biol Eng*. 2015;35(2):178-187. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4414937/
47. Fan H, Meng F, Liu Y, Kong F, Ma J, Lv Z. A novel breast ultrasound image automated segmentation algorithm based on seeded region growing integrating gradual equipartition threshold. *Multimedia Tools Appl*. 2019;78(19):27915-27932.
48. Lo C, Chen R, Chang Y, et al. Multi-dimensional tumor detection in automated whole breast ultrasound using topographic watershed. *IEEE Trans Med Imaging*. 2014;33(7):1503-1511.

How to cite this article: Thomas C, Byra M, Marti R, Yap MH, Zwiggelaar R. BUS-Set: A benchmark for quantitative evaluation of breast ultrasound segmentation networks with public datasets. *Med Phys*. 2023;50:3223–3243. https://doi.org/10.1002/mp.16287