Received: 3 March 2020 | Revised: 9 October 2020 | Accepted: 18 October 2020

DOI: 10.1002/ima.22516

WILEY

RESEARCH ARTICLE

Deeply supervised U-Net for mass segmentation in digital mammograms

N Ravitha Rajalakshmi[1] | **R Vidhyapriya**[2] | **N Elango**[3] | **Nikhil Ramesh**[1]

[1]Department of Information Technology, PSG College of Technology, Coimbatore, India

[2]Department of Biomedical Engineering, PSG College of Technology, Coimbatore, India

[3]Department of Radiology, PSG Medical Sciences and Research, Coimbatore, India

Correspondence
N Ravitha Rajalakshmi, Assistant Professor (Senior Grade), Department of Information Technology, PSG College of Technology, Coimbatore, India.
Email: myravithar@gmail.com, nrr.it@psgtech.ac.in

Abstract

Mass detection is a critical process in the examination of mammograms. The shape and texture of the mass are key parameters used in the diagnosis of breast cancer. To recover the shape of the mass, semantic segmentation is found to be more useful rather than mere object detection (or) localization. The main challenges involved in the mass segmentation include: (a) low signal to noise ratio (b) indiscernible mass boundaries, and (c) more false positives. These problems arise due to the significant overlap in the intensities of both the normal parenchymal region and the mass region. To address these challenges, deeply supervised U-Net model (DS U-Net) coupled with dense conditional random fields (CRFs) is proposed. Here, the input images are preprocessed using CLAHE and a modified encoder-decoder-based deep learning model is used for segmentation. In general, the encoder captures the textual information of various regions in an input image, whereas the decoder recovers the spatial location of the desired region of interest. The encoder-decoder-based models lack the ability to recover the non-conspicuous and spiculated mass boundaries. In the proposed work, deep supervision is integrated with a popular encoder-decoder model (U-Net) to improve the attention of the network toward the boundary of the suspicious regions. The final segmentation map is also created as a linear combination of the intermediate feature maps and the output feature map. The dense CRF is then used to fine-tune the segmentation map for the recovery of definite edges. The DS U-Net with dense CRF is evaluated on two publicly available benchmark datasets CBIS-DDSM and INBREAST. It provides a dice score of 82.9% for CBIS-DDSM and 79% for INBREAST.

KEYWORDS

conditional random fields, deep supervision, mammograms, mass segmentation

1 | INTRODUCTION

According to Global Cancer Observatory Database (GLOBOCAN) released in the year 2018, breast cancer is contributing to over 25.4% of all the new cases diagnosed.[1] Screening programs are organized by the government to aid in the early diagnosis of the disease. Mammograms are widely accepted as the primary screening tool for breast cancer. A mammogram is an X-ray image of the breast, which capture the changes in the breast tissue. Presence of mass and microcalcification in the mammograms characterize the disease.[2,3] The

detection of these regions is difficult as their pixel intensities often correlate with the normal tissue.

Computer-aided detection and computer-aided diagnosis tools[3] are used to lessen the burden of the radiologist in the process of diagnosis. They interpret digitally captured medical images and provide useful information about the suspicious regions. They employ artificial intelligence and image processing techniques for the task of detection and analysis. CAD systems are constantly evolving with the advent of new techniques in the domain to provide accurate results. This paper deals with mass segmentation: a key step in the diagnosis of breast cancer. Earlier studies on mass segmentation employ region growing[4] or contour-based techniques.[5] The shortcomings of the above-mentioned approaches are the initial seed selection and the initial contour selection.[2] With the recent success of the deep neural networks in most of the vision-oriented challenges, there has been a spurt of research into their applicability for medical diagnosis.[6-9] There are two different approaches applied for lesion identification: Object Detection and Semantic Segmentation. In object detection, mass locations are approximated by a bounding box and the network is trained to predict its coordinates. Cascade of belief networks[10] and region-based convolutional neural networks (CNNs)[11] have previously been explored for mass detection. The limitation of the object detection technique is that it requires additional processing to recover the object boundary. In the semantic segmentation, every pixel is classified either as a background or mass region. Fully CNN[12] is popularly used for the task wherein the fully connected layers of CNN are replaced with up sampling layers to restore the spatial context of the lesion. Though FCN flared well in most of the recognition tasks, it provides a coarser output. U-Net[13] employed a symmetrical encoder and decoder path with the concept of skip connections to recover the region of interest (ROI) context in the image. U-Net outperformed other models significantly when the image possess higher signal to noise ratio (SNR). The proposed work improves the U-Net model with deep supervision and channel attention to prevent the loss of information at the encoder.

The main contributions of the work include:

1. A lightweight U-Net architecture is proposed with improved attention to the boundaries of the suspicious regions. The output of the encoder layer is monitored for proper attention of the mass boundaries and the output of the decoder layer is monitored for proper attention of the mass regions. This leads to faster convergence and lesser false positives.
2. A learnable fusion layer is used to combine the output of the intermediate layers with output of the last layer. This prevents the loss of information at the encoder.
3. Exponential logarithm dice loss function is used to improve the segmentation of the smaller objects. It is an unbiased metric, which treat the objects equally irrespective of their size.
4. Unlike other works, the proposed model uses dedicated preprocessing and post processing modules to improve the segmentation result. CLAHE is used to enhance the contrast of the input image and dense conditional random field (CRF) is used to recover the boundary of the detected regions from the final segmentation map of deeply supervised U-Net model (DS U-Net).

2 | METHOD

2.1 | Overview

The proposed model employs an end-to-end architecture for mass segmentation in mammograms. The model is trained over the images from CBIS-DDSM and INBREAST dataset (Figure 1). It starts with preprocessing the images using CLAHE to improve SNR. Then, it segments the mass regions in the preprocessed image using DS U-Net. Finally, it applies dense CRF to recover the shape of the mass. Figure 2 shows the various components involved in the proposed architecture.

2.2 | Preprocessing

Preprocessing is often neglected when the deep neural networks perform the task of segmentation. Since mammogram images exhibit low SNR, the detection becomes infeasible without proper preprocessing. From the literature, it is found that the histogram equalization

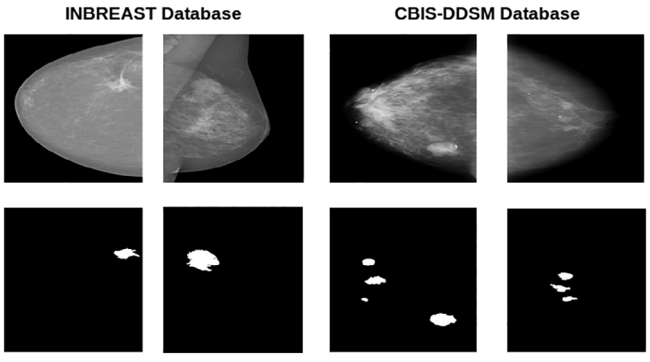

FIGURE 1 Mammogram images with ground truth indicating mass. First row: full field of view mammogram images in INBREAST and CBIS-DDSM dataset. Second row: manually segmented mass regions

technique and its variants are beneficial for the mammogram images.[14,15]

In the proposed work, CLAHE is used to improve the contrast in the mammogram image.[11] It splits the image into nonoverlapping blocks of size N * N and computes the transformation function for each of these blocks separately using histogram equalization. Then, for every pixel in the image, four nearest blocks are considered and their respective transformation mapping is used to find the resultant pixel intensity. Applying histogram equalization to homogenous regions can introduce noise in the resultant image. To remove the noise, Intensity values with frequency count greater than clip limit (threshold) are redistributed across other bins with lesser frequency count. This technique is found to improve the edges in the image. Figure 3 shows the sample images from INBREAST and CBIS-DDSM which are enhanced using CLAHE.

2.3 | Deeply supervised U-net

Deep supervision[16-18] brings in transparency to the intermediate hidden layers of deep neural networks. It uses the error factor of intermediate feature maps with respect to the ground truth in the training objective criterion. This improves the robustness of the neural network in both the segmentation and the classification tasks.[16] In the segmentation network, the output of intermediate layers is up-sampled and compared with ground truth to quantize their error margin.

U-Net[13] has become the de-facto standard for biomedical image segmentation. It employs an encoder decoder-based architecture. Encoder consists of convolution layers for retrieving contextual information and pooling layers for down-sampling the images. Downsampling aids in the retrieval of higher-level contextual information and also offers translational invariance. U-Net has a symmetrical decoder path which up-samples the feature map so as to recover the spatial context of the detections. Skip connections are pathways which carry the spatial information from the encoder to the decoder. The proposed work has extended over the idea of Chen et al[19] and Mishra et al[13] to use the intermediate layers of the encoder and the decoder to focus on the boundaries and regions simultaneously. In addition, the proposed model utilizes Squeeze Excitation blocks[12,20] to reinforce the channels of higher importance before combining the feature maps from the skip connection and the decoder. Figure 4 shows the complete architecture of proposed model. The DS U-Net is used to improve the recovery of the object boundary and significantly reduce the number of false positives.

The encoder block applies repeated 3*3 convolution on the input tensor as shown in Figure 5A. After every convolution, batch normalization is applied followed by ReLU activation. The decoder block (Figure 5C) up-samples the high-level features from the previous layer using bilinear interpolation and combines it with spatial context (low level features) from the encoder block. Subsequently, 3*3 convolution is applied before passing the tensor to the next layer. After every stage of the decoder,

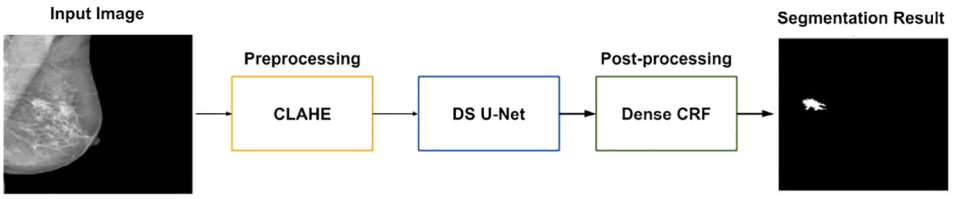

FIGURE 2 Proposed architecture for mass segmentation in mammograms [Color figure can be viewed at wileyonlinelibrary.com]

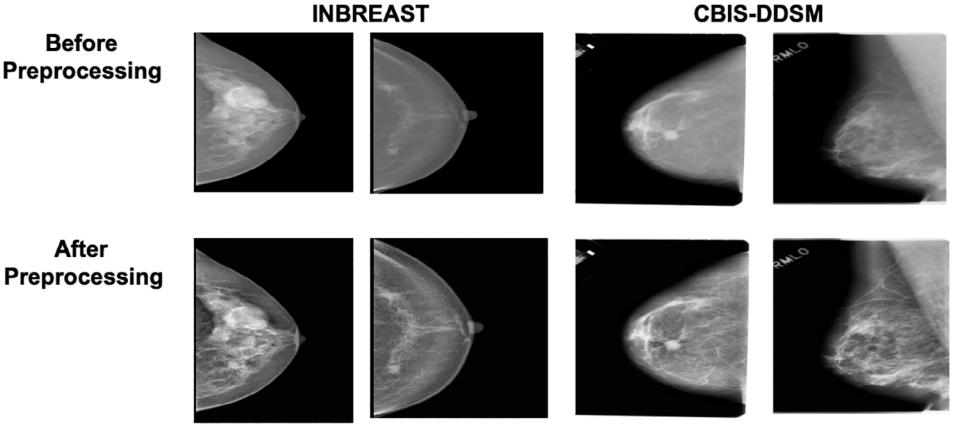

FIGURE 3 Preprocessed mammogram images using CLAHE

a drop out layer with the probability of 0.5 is used to avoid overfitting. The high-level features and the low-level features are subjected to channel attention before being fed as the input to the decoder block. Channel attention reinforces the feature maps with salient features using two-layered network. The first layer contains nodes equal to 1/16th of the number of input feature maps and the second layer contains nodes equal to the number of input feature maps. The network operates over the average intensity value of the input channels as shown in Figure 5B. Blocks A1 to A6 are attention blocks, which supervises the encoder and decoder layers as shown in Figure 4. It applies 1*1 convolution to reduce the dimensions and bilinear upsampling to enlarge the size of the intermediate tensor to match the output dimensions (Figure 5B). Fusion layer produces the end segmentation result as the linear combination of the outputs from the attention blocks and the output from the final decoder layer (D_4) as shown in Equation (1).

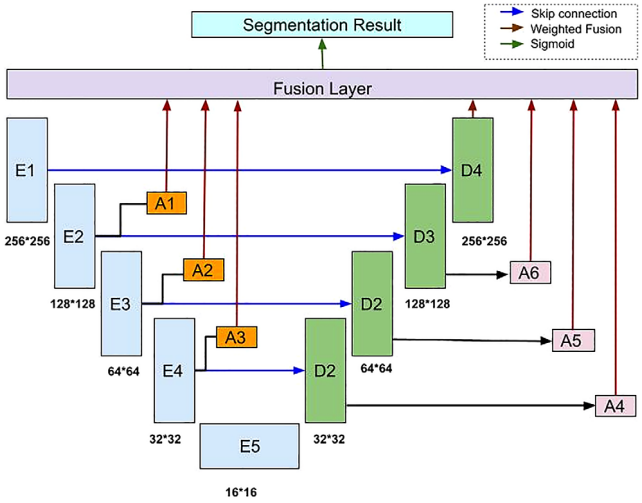

FIGURE 4 Deeply supervised U-Net: auxiliary blocks A1, A2, and A3 are providing attention to the boundaries, and auxiliary blocks A4, A5, and A6 are used to discriminate the entire mass region from the background [Color figure can be viewed at wileyonlinelibrary.com]

$$\hat{y} = \text{sigmoid}\left(\sum_{i=1}^{6} h_i O(A_i) + h_D O(D_4)\right) \quad (1)$$

Here, $\hat{y}$ represents the final segmentation result, O(.) represents the output of the block provided in the parenthesis and h $\{h_D, h_1, h_2 \ldots . h_6\}$ represents the weights associated with the output of final decoder block and the attention blocks. The objective criterion is modified as in Equation (2) so that the attention blocks A1 to A3 can steer the network toward boundary identification and blocks A4 to A6 can steer the network toward region identification.

$$L(\hat{y}, y_B, y_M; \theta) = L_f(\hat{y}, y_M) + \text{Auxiliaryloss} \quad (2)$$

$$\text{Auxiliary loss} = \left(1 - \frac{t}{T}\right)^d \left(\sum_{i=1}^{3} L_{A_i}(O(A_i), y_B) + \sum_{i=4}^{6} L_{A_i}(O(A_i), y_M) + L_{D_4}(O(D_4), y_M)\right) \quad (3)$$

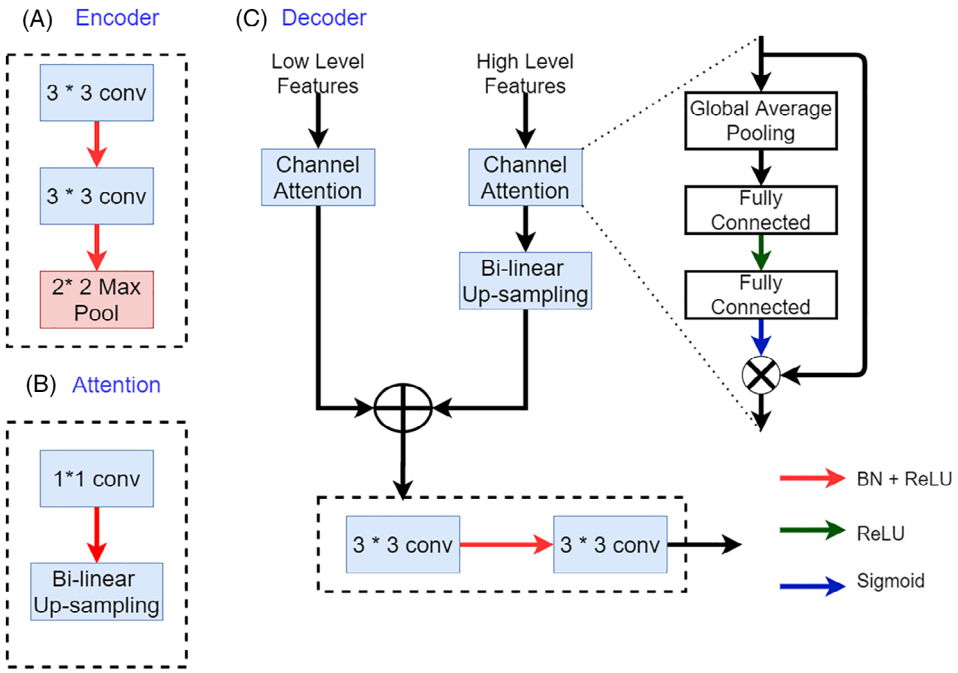

FIGURE 5 Various blocks of DS U-Net architecture, A, encoder, B, attention, and, C, decoder [Color figure can be viewed at wileyonlinelibrary.com]

Here, L represents the loss function which is discussed in section D, θ represents the parameters of the network which include (W_{encoder},W_{decoder}, $W_{\text{attention}}$,h), Y_B represents the ground truth labeled with only the boundary pixels of the suspicious mass and Y_M represents the ground truth labeled for the entire region of the suspicious mass.

A factor of $\left(1-\frac{t}{T}\right)^d$ is used for auxiliary loss to reduce the impact of attention layers as the number of iterations (t) approach the total number of epochs T.[17] It ensures that the intermediate activations are not modified to a larger extent at the later stage of the training process. d is a constant which is set to a value of 2. The boundary pixels are extracted by performing morphological erosion on the region mask with disk shaped structuring element (value of radius is set to 3) and subtracting the resultant from the original image. The boundary mask obtained after the morphological operation is shown in Figure 6.

2.4 | Loss function

Sorensen dice coefficient is the popularly used metric to measure the similarity between two images. To gauge the differences between the prediction and ground truth, error value of (1-Dice coefficient) is used. The size of the object can influence this value as it just takes into account the percentage of misclassified pixels. In particular, this error function results in higher value even with fewer number of misclassifications for smaller objects. Thus, the network may be biased toward the object of larger sizes upon using this loss function. To achieve a balanced loss whereby the network focuses on both the objects with low and high prediction accuracy equally, exponential logarithm dice loss is used.[21]

Let y_i represents the ground truth of i^{th} pixel which takes either a value of 0 (or) 1. A value of 0 indicates the background (normal region) and a value of 1 indicates the mass (abnormal region). Let p_i represents the prediction probability of i^{th} pixel belonging to mass. The exponential logarithm dice loss is computed as in Equation (4), where M denotes the total number of training samples in the mini-batch.

$$\mathcal{L}=\frac{\sum_{k=1}^{M}\left(-\ln(\text{Dice}_k)\right)^{\gamma}}{M} \tag{4}$$

The dice score for image k is calculated as $\text{Dice}_k=\frac{\sum_{i=1}^{N}2y_i*p_i+\varepsilon}{\sum_{i=1}^{N}y_i+\sum_{i=1}^{N}p_i+\varepsilon}$. ε(epsilon) value of 1e-6 is used for numerical stability. γ value in the loss is used to adjust the non-linearity between the dice score value and the error. In the experiments, γ value of 0.3 is used[21] as it offers a decreasing gradient with dice score values less than 0.5 and an increasing gradient with dice score greater than 0.5. This helps the loss function to behave differently for objects of varying sizes.

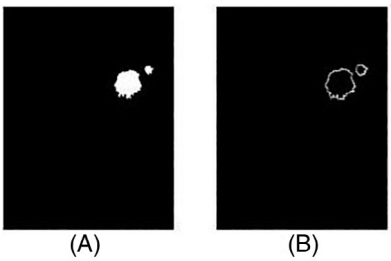

FIGURE 6 Segmentation masks. A, Region mask. B, Boundary mask

2.5 | Post-processing

DS U-Net precisely identifies the location of the mass in a given mammogram image. As the shape of the detected lesions can aid in the diagnosis of the disease, post-processing is applied to recover the entire context of the detected lesions. Probabilistic graphical models have proven to be effective in modeling the neighborhood dependencies. Therefore, it possesses higher discriminatory power to differentiate the foreground pixels from the background pixels. Hence, CRFs is used to fine-tune the resulting segmentation map of Deep Supervised U-Net model. Works by Chen et al[22] and Zheng et al[23] showed that state of the art segmentation results can be obtained by incorporating the dense CRF model. The model accepts the raw intensity values of input image and the unary potentials (sigmoid probabilities of DS U-Net) to generate the final probability map as shown in Figure 7.

CRF model treats the individual pixels X_u as random variables which can take values from the set $L = 0, 1$ where 0 indicate background and 1 indicate foreground (suspicious mass region). CRF inference assigns a configuration $X=X_1,X_2,X_3.....X_{N^2}$ for every random variable in the image I such that the energy given in Equation (5) is minimized.

$$\begin{aligned}E(X,I)&=\sum\nolimits_{uV}\Phi_u(X_u=x\,|\,I)\\&+\sum\nolimits_{\{u,v\}\xi}\Psi_{u,v}(X_u=x,X_v=y|I)\end{aligned} \tag{5}$$

The term $\Phi_u$denotes the unary cost associated with assigning label x for the variable X_u. This cost is usually obtained from the classifier. The interdependencies

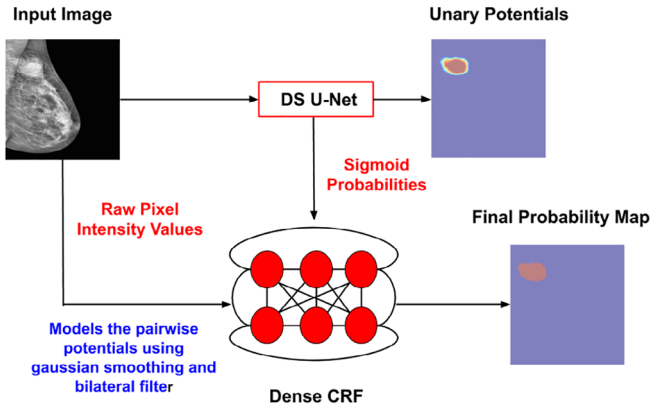

FIGURE 7 Dense CRF model process flow [Color figure can be viewed at wileyonlinelibrary.com]

between the pixels are modeled using pairwise cost $\Psi_{u,\,v}$ which accounts for the cost of assigning labels x and y for the pixels X_u and X_v, respectively. The common pairwise cost used is Pott's model which provides a cost of 0 when the neighboring pixels are assigned with same label and a value of 1 otherwise. Further, certain assumptions should also be made on how the pixels connect with each other. There are two popular configurations used: Grid CRF and fully connected CRF. Fully connected CRF assume that every pixel in the image is connected with every other pixel whereas Grid CRF assumes that every pixel is connected only to its adjacent pixels. Dense CRF leverages on fully connected CRF to model the neighborhood dependencies. CRF has been formulated on the basis of the Gibbs distribution as given in Equation (6).

$$P(X\,|\,I) = \frac{1}{Z(I)}\exp(-E(X,I)) \tag{6}$$

The learning phase of CRF is defined as the computation of marginal probabilities for every random variable X_u and for every pair of random variables (X_u, X_v) which are connected by edges, that is, all pixels except itself $u \neq v$ in a fully connected graph. Detection of true distribution for the unary Φ and pairwise potentials Ψ are infeasible. Mean field inference approximates the p(X|I) using a simpler distribution Q(X|I) and iteratively reduces the Kullback Leibler divergence between Q(X|I) and p(X|I). In the proposed model, unary potentials Φ are obtained from DS U-Net and the pairwise potentials Ψ are modeled using the sum of two Gaussian kernels as suggested in Reference 24: Gaussian smoothing kernel $\exp\left(-\frac{||p_u-p_v||^2}{2\sigma_\alpha^2}\right)$ and Gaussian preserved bilateral filtering kernel $\exp\left(-\frac{||p_u-p_v||^2}{2\sigma_\alpha^2}-\frac{||I_u-I_v||^2}{2\sigma_\beta^2}\right)$. Here, $p_u$and p_V denote the position of the pixels. I_u and I_v denote the intensity of the pixels. Weights are associated with the kernels to control the significance of the filters.

Thus, the pairwise potentials in CRF is formulated as $\Psi_{u,\,v}(X_u, X_v) = \mu(X_u, X_v)k(f_u, f_v)$ where μ denotes the label similarity using common Pott's model and k denotes the sum of Gaussian kernels. Mean Field updates for assigning a label l for the random variable X_u is provided in Equation (7).

$$Q_u(X_u = l\,|\,I) = \frac{1}{Z_u}\exp -\Phi_u(I) - \sum_{l'L}\mu(l,l') \sum_{m=1}^{2} w^{(m)}\sum_{v\neq u}k^{(m)}(f_u,f_v)Q_u(l') \tag{7}$$

Here, $w^{(m)}$ indicate the strength of the filter m. In the experiments, weights for Gaussian smoothing and bilateral filters are considered as 1 and 5, respectively. Number of iterations T_0 for convergence is set as 50.

III Experiment

2.6 | Datasets

The model is evaluated on two popularly available benchmark datasets: INbreast[25] and CBIS-DDSM.[26] INbreast contains 107 full field digital mammography (FFDM) images with mass findings. These images are available in DICOM format and are of size 3328 * 4084 (or) 2560 * 3328 with 14-bit resolution. The boundary pixels of the mass are manually annotated and provided in XML format. Because of the smaller size of the dataset, a 5-fold cross-validation is performed on the dataset to gauge the model's performance.

CBIS-DDSM dataset contains curated mammogram images from DDSM, a largest available mammogram database. They provide separate training and test set for mass detection and microcalcification detection. The dataset comprises of FFDM images along with the segmentation mask and the cropped ROI for every suspicious finding in DICOM format. For the experiments, a curated set of 689 images is used for training and a set of 168 images is used for testing.[27] There are multiple ROI detections in an image which are provided as separate masks in the dataset. They are merged into a single mask before training. Both the datasets contain significant amount of benign cases[27] and they also include masses of varying sizes as shown in Figure 1.

2.7 | Experimental framework

All the experiments are carried out using NVIDIA GTX 1080 Ti GPU and the models are implemented in keras. Adam[28] optimizer with warm restart is used for the training

Algorithm 1

Mean field inference for dense CRF

Input:

X_u Grayscale intensity of the Input image at pixel position u.
$U_u(l)$ Output of last decoding layer of DS U-net for the random variable u and label l.
Output:
$Q_u(l)$ Probability of the pixel X_u computed using Dense CRF
Initialize parameters
$\boldsymbol{W}$ - a vector of size 1 x m where m denotes the number of filter kernels $\boldsymbol{\mu}$ - compatibility matrix
Initialize $\boldsymbol{Q_u(l)}$ with the softmax on unary potentials $U_u(l)$
while T< T_0 (T_0 denotes the number of iterations)
For every filter m in the set {Gaussian smoothing filter, Gaussian bilateral filter}
$\bar{Q}_u(l)$ + = w[m] * sum of the filter coefficients
when operated on pixels other than u
$\bar{Q}_u(l)$ = sum the potentials of the labels l' that are consistent with l according to $\boldsymbol{\mu}$
$\bar{Q}_u(l) = -U_u(l)$-$\bar{Q}_u(l)$
$\bar{Q}_u(l)$ = Compute softmax over $\bar{Q}_u(l)$ considering pixel value of all the other labels
end while

process with the batch size as 4. The models are trained for about 200 epochs with 0.001 as the initial learning rate. Here, the learning rate is decreased using cosine annealing for a defined number of iterations known as a cycle and it is reinitialized to a maximum value after every cycle. The parameters of the model are randomly assigned as the other initializers failed to converge. The intensity values in the image are scaled to a fixed range between 0 and 1 before being fed as the input to the network.

2.8 | Metrics

To measure the correctness of the prediction, two types of metrics are used: pixel level metrics and region level metrics. Pixel level metrics include accuracy, sensitivity, and specificity. The **Accuracy** measures overall pixel accuracy, that is, percentage of correctly classified pixels, **Sensitivity** measures the percentage of correct predictions corresponding to pixels constituting the mass region in the ground truth, **Specificity** measures the percentage of correct predictions corresponding to pixels constituting the normal region in the prediction map.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{9}$$

Here, true positive (TP) denotes the count of pixels which are classified correctly as belonging to mass. False positive (FP) denotes the count of normal pixels classified as mass and false negative (FN) denotes the count of pixels which are incorrectly classified as normal pixels. Sensitivity will be low when there are more FN. It happens due to under-segmentation wherein the mass occupies a very small region that cannot be properly captured by the model. Specificity is affected when there are large number of FP. False positives are the consequence of over-segmentation wherein the normal parenchymal regions are detected as mass.

Object level metrics use the size and the boundary of positive predictions for evaluation. Jaccard coefficient, dice score, relative area difference (ΔA), and Hausdorff distance are used for evaluating the regions in the resulting prediction map.

$$\text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{10}$$

$$\text{Dice score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{11}$$

$$\Delta A = \frac{|(TP + FP) - (TP + FN)|}{(TP + FN)}. \quad (12)$$

Jaccard and **dice score** measure the amount of overlap between ground truth (GT) and prediction map. A value of 0 indicates that there is no overlap between prediction and GT. A value of 1 indicates an exact match between GT and prediction. Jaccard is an estimate of average performance of the model and Dice score is an estimate of worst-case performance of the model. **Relative area difference** is used to measure the differences in the size of prediction against the ground truth. To measure the maximum deviation along the boundary between GT and predictions, **Hausdorff distance** metric is used. It measures the longest distance between any two closest points that lie in the boundary of ground truth and prediction.

$$H = \max(h(\mathrm{GT}, \mathrm{pred}), h(\mathrm{pred}, \mathrm{GT})) \quad (13)$$

Here, $h(\mathrm{A}, \mathrm{B}) = \max(\min(\|a - b\|))a \in A, \quad b \in B$ where A and B represent the set of boundary pixels. The lower values of H indicate accurate segmentation.

2.9 | Experimental results

The proposed work is aimed at improving the segmentation of lesions in whole mammograms. An ablation study is conducted to determine the effectiveness of pre-processing and post-processing techniques. To measure the performance of the model, experiments are repeated thrice and the average value of the metrics are considered.

2.9.1 | Impact of preprocessing

Mammograms exhibit low contrast when compared with other medical images. To improve the contrast in the mammograms, CLAHE is used. Since the images in CBIS-DDSM are obtained from four different scanners, they exhibit varying contrast. Hence, selective preprocessing is used, that is, CLAHE is applied only to those images which exhibit poor global contrast. The global contrast in an image is estimated using root mean square (RMS) metric. RMS measures the variance of the normalized gray level values in an image as specified in Equation (14).

$$\mathrm{RMS} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (14)$$

where x_i is a normalized gray level value $0 \le x_i \le 1$, $\bar{x}$ is the mean normalized gray level and n denotes the total number of pixels in an image. The block size and the clip limit for CLAHE are considered as (128128) and 0.01, respectively.

Table 1 depicts the performance of DS U-Net for both the datasets w/ & w/o preprocessing. There is a significant improvement in all the metrics for both the datasets when preprocessed images are used in the model construction. In particular, the dice score is increased by a factor of 4% and 1.6% for CBIS-DDSM and INbreast, respectively.

2.9.2 | Impact of postprocessing

Postprocessing is applied on the prediction maps obtained from the DS U-Net model using dense CRF. It is used to recover the boundary of the abnormal region. This stage is critical as the region boundary can offer crucial information about the malignancy of the region. Dense CRF uses the probability of pixels from DS U-Net as unary potential and models the pairwise potential as the sum of Gaussian smoothing and bilateral filters. After a grid search over different kernel sizes and filter strength, ideal values are determined. The kernel size of (7,7) for Gaussian smoothing filter and (3,3) for bilateral filter provided the optimal performance.

Table 2 depicts the performance of DS U-Net after postprocessing. Hausdorff distance is the only metric which accounts for the shape of the object out of all the considered metrics. A significant reduction of 0.05 is observed for Hausdorff distance metric after applying dense CRF over the prediction maps of DS U-Net. As the

TABLE 1 Evaluation metric values observed for DS U-Net model trained using images w/ and w/o preprocessing

Dataset	Pre-processing	Dice score	Jaccard coefficient	Accuracy	Sensitivity	Specificity	Hausdorff distance	ΔA
CBIS-DDSM	No	78.7	83.5	99.7	78.1	99.8	2.7	29.2
	Yes	82.7	85.7	99.7	83.1	99.8	2.48	21.9
INBREAST	No	77	83.2	98.7	78.01	99.47	1.97	31.5
	Yes	78.6	83.4	99	81.1	99.3	1.91	33.5

TABLE 2 Evaluation metric values observed for DS U-Net model w/ and w/o postprocessing (post processing is carried out on the model output over the preprocessed images)

Dataset	Models	Dice score	Jaccard coefficient	Accuracy	Sensitivity	Specificity	Hausdorff distance	ΔA
CBIS-DDSM	DS U-Net	82.7	85.7	99.7	83.1	99.8	2.48	21.9
	DS U-Net with dense CRF	82.7	85.7	99.7	84.1	99.8	2.43	22.7
INBREAST	DS U-Net	78.6	83.4	99	81.1	99.3	1.91	33.5
	DS U-Net with dense CRF	79	83.4	98.11	81	98.4	1.86	31.1

TABLE 3 Comparison of proposed model (DS U-Net with dense CRF) with the other popular segmentation architectures for whole mammograms

Dataset	Models	Dice score	Sensitivity	ΔA	Hausdorff distance
CBIS-DDSM	U-Net[13]	73.6	79.4	42.7	3.38
	AU-Net[27]	80.9	84.5	29.9	2.97
	DS U-Net + Dense CRF	82.7	84.1	22.7	2.43
INBREAST	U-Net[13]	69.3	70.4	44	4.54
	AU-Net[27]	79.1	80.8	37.6	4.04
	DS U-Net + Dense CRF	79	81	31.1	1.86

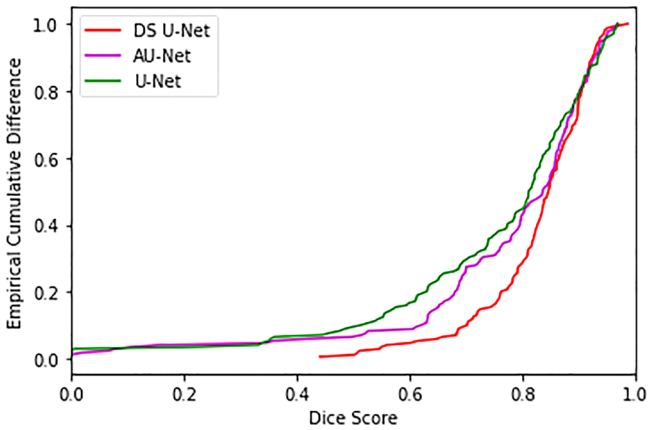

FIGURE 8 Empirical cumulative difference plot of dice score metric for the test images belonging to the CBIS-DDSM dataset [Color figure can be viewed at wileyonlinelibrary.com]

technique only changes the class label for a smaller percentage of pixels which lie along the boundary of the detected regions, it does not affect the value of other metrics as reported in the table.

2.9.3 | Comparison with other state-of-art models

In order to compare the effectiveness of the proposed model, AU-Net[27] and U-Net[13] models are considered. The U-Net model uses skip connections to carry information from the encoder stage to the decoder stage for recovering the spatial context. This information is then combined with the feature maps of the decoder through concatenation. But it is found to be ineffective and also attributes to certain information loss.[27] This problem is addressed in AU-Net by utilizing an asymmetrical network backbone and attention guided dense up-sampling. AU-Net offers better accuracy than the U-Net model at the cost of increased computational complexity. DS U-Net addresses the loss of information by employing deep supervision over the intermediate activations. It also uses channel attention to retrieve important information from the feature maps. The computational complexity of DS U-Net is reduced by employing feature addition instead of concatenation. Table 3 depicts the performance of all the three models. It clearly shows that the DS U-Net provides substantial improvement in all the key metrics when compared with other models. In particular, the dice score of DS U-Net shows an improvement of 1.8% in comparison with the AU-Net model for the CBIS-DDSM dataset. There is also a huge reduction in the Hausdorff distance and ΔA by 0.54 and 7.2, respectively, for the CBIS-DDSM dataset when compared with the AU-Net model. Dice score values of all the models are inspected further using CDF (Figure 8).

The qualitative improvement of the proposed model over the others is shown in Figures 9 and 10. The

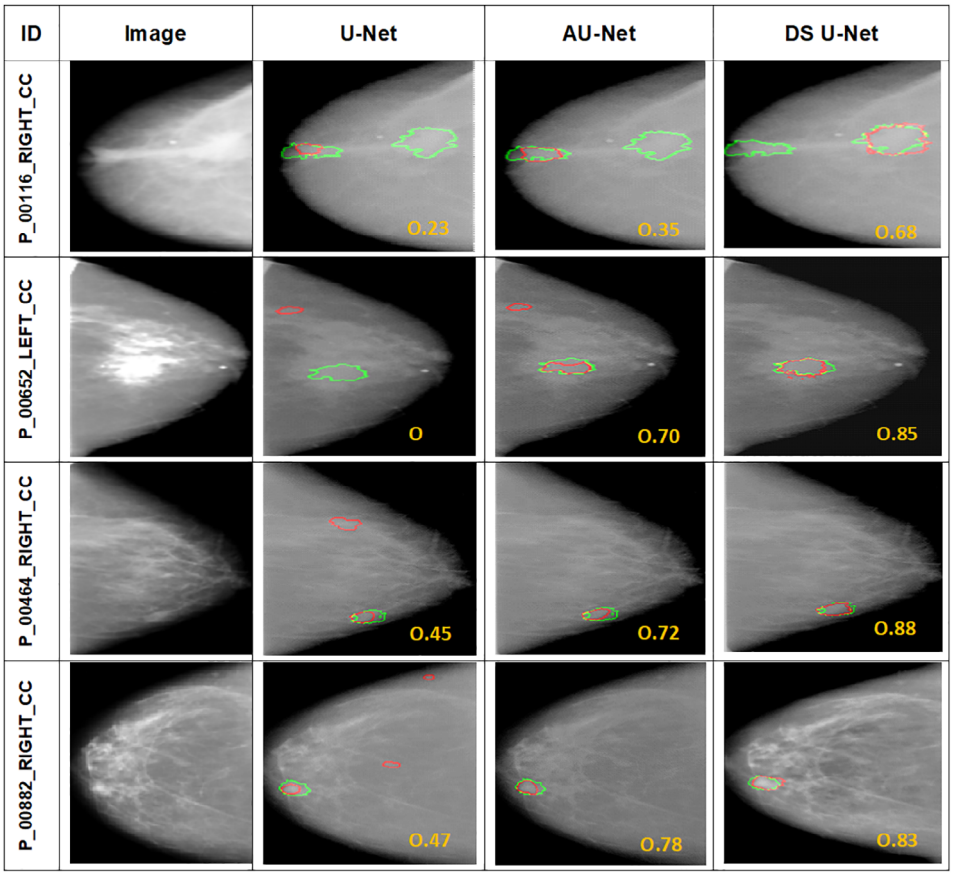

FIGURE 9 Prediction results for sample images from the CBIS-DDSM dataset. (The green line marks the boundary of ground truth and the red line marks the boundary of prediction. Dice score for predictions of the model is specified at the bottom right corner of the image.) [Color figure can be viewed at wileyonlinelibrary.com]

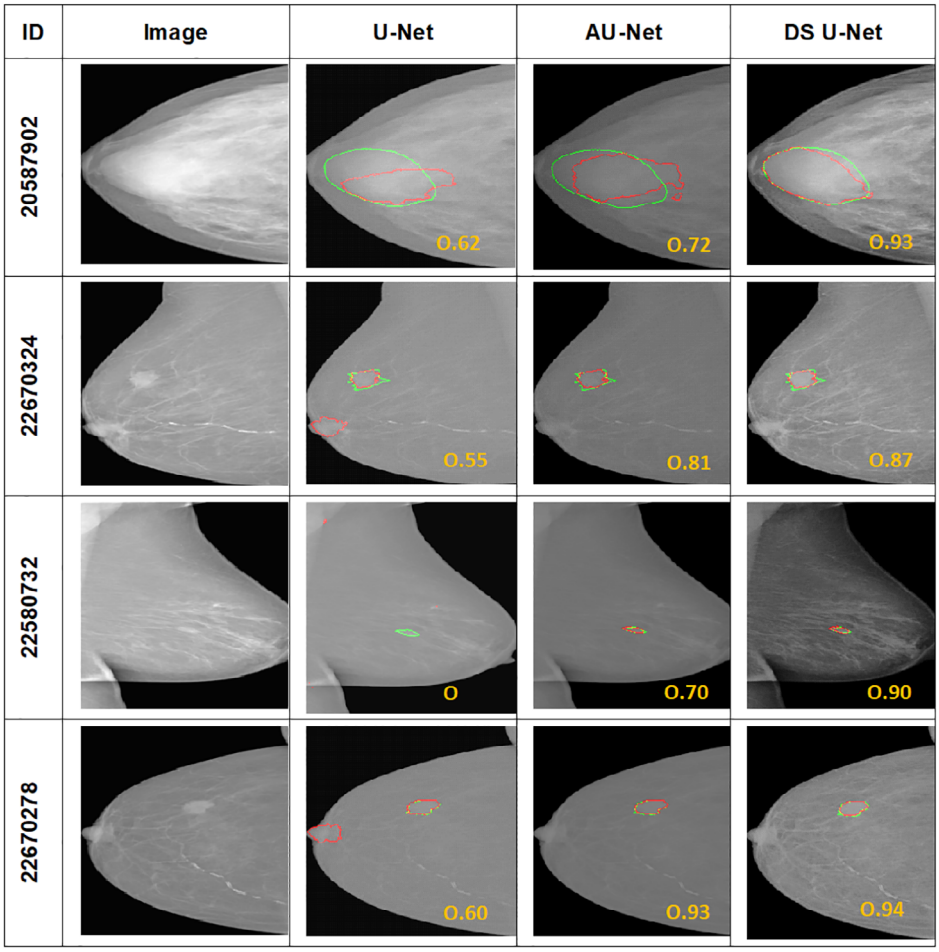

FIGURE 10 Prediction results for sample images from the INbreast dataset. (The green line marks the boundary of ground truth and the red line marks the boundary of prediction. The dice score for predictions of the model is specified at the bottom right corner of the image.) [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Computational complexity of various models

Model	U-net	AU-Net	DS U-Net
Number of trainable parameters (in millions)	31	75.4	31.6

TABLE 5 Overview of semantic segmentation architectures used in the literature for segmentation of mass in digital mammograms

Segmentation models	Architectural modifications	Dataset	Images	Dice score	Sensitivity
PSPNet[29,30]	Spatial pyramid pooling	Private dataset	380	67.85	72.02
DeepLabV3+[29,31]	Atorus convolution and an efficient decoder	Private dataset	380	68.27	70.72
FC-DenseNet[29]	Dense connections	Private dataset	380	73.55	79.68
ASPP-FC-DenseNet[29]	Atorus convolution with dense connections	Private dataset	380	76.97	79.83
Attention Net[32]	Attention gates	DDSM dataset	400	77.32	68.01
Attention DenseNet[29]	Dense blocks and Attention Gates	DDSM dataset	400	81.36	76.19
AU-Net[27]	Complex up-sampling with asymmetrical encoder and decoder blocks	CBIS-DDSM	857	80.9	84.5

prediction result of all the models for sample images from both the datasets are depicted. The boundary of the ground truth is delineated using lines in green color and the boundary of the predicted regions are delineated using lines in red color. From the figures, it is apparent that the U-Net model fails to detect lesions of smaller size.

To better understand the overlap of predictions and ground truth across the entire set of images in CBIS-DDSM dataset, empirical cumulative difference of dice score is plotted (Figure 8). It shows the frequency count of every possible dice score in the obtained predictions. The DS U-Net locates lesion in the images quite comfortably as the graph starts with dice score of 0.4, that is, the model does not have any predictions with a dice score less than 0.4. The graph of the DS U-Net also shows a steeper increase after the dice score 0.8. This is indicative of the fact that most of the model predictions are precise.

Table 4 specifies the number of trainable parameters of the considered models. The DS U-Net uses a smaller number of convolution layers in the encoder and the decoder blocks compared with AU-Net. This subsequently reduces the training and inference time. It also performs better than AU-Net with half the number of parameters involved in it.

Table 5 provides an overview of other popular semantic segmentation architectures that are used for mass segmentation in the digital mammograms. Most of the works were carried out on a dataset with not more than 400 images. This work uses 857 images from the CBIS-DDSM dataset.

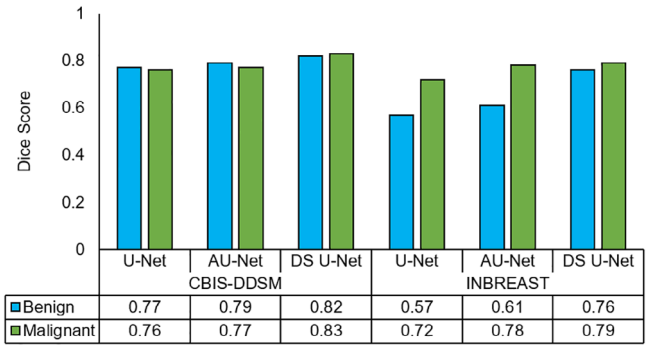

FIGURE 11 Average dice score of different models with respect to lesion type [Color figure can be viewed at wileyonlinelibrary.com]

Effectiveness of the model in the segmentation of benign lesions

The existing CAD system fails to detect lesions at benign stage, as the lesions are more subtle in nature. To assess the effectiveness of the model in detecting the lesions at benign stage, the model's average dice score for benign and malignant lesions are determined. Figure 11 depicts the performance of AU-Net, U-Net and DS U-Net model for both the datasets. The DS U-

Net provides a similar dice score for both the lesion categories. There is also a significant improvement in the average dice score of DS U-Net for benign lesions when compared with AU-Net. This strengthens the inference that DS U-Net works better for mammograms when compared to other models.

3 | CONCLUSION

In this work, an end-to-end framework is created for the mass segmentation in digital mammograms. It uses CLAHE to improve the contrast of the input mammogram, deep supervised U-Net to segment the lesions and dense CRF to recover the entire context of the lesions. The DS U-Net has improved the U-Net model with channel attention and deep supervision for faster convergence and increased performance. Several pixel level and region level metrics are used to evaluate the performance of the model. The proposed architecture is found to perform better when compared with AU-Net and U-Net. In future, this model can also be explored for the segmentation of other medical images.

ACKNOWLEDGMENTS

Research was conducted as part of the study (Ref:19/359) approved by Institutional Human Ethics Committee of PSG Institute of Medical Sciences and Research. We appreciate Cheng Li, Shenzhen Institutes of Advanced Technology, Chinese Academy of Science, for sharing the representative set of images from CBIS-DDSM to enable validation of the model.

DATA AVAILABILITY STATEMENT

The CBIS-DDSM data that support the findings of this study are openly available in Mass-Training and Mass-Test folders at https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY, reference number [26]. The INBREAST data used in the study can be obtained on request to the authors of the article DOI: 10.1016/j.acra.2011.09.014, reference number [25]

ORCID

N Ravitha Rajalakshmi https://orcid.org/0000-0003-2006-3773
R Vidhyapriya https://orcid.org/0000-0003-0665-6820
Nikhil Ramesh https://orcid.org/0000-0003-4125-6275

REFERENCE

1. Cancer Facts & Figures (GLOBOCAN). https://gco.iarc.fr/. Accessed September 2020.
2. Matthias E, Alexander H. CADx of mammographic masses and clustered microcalcifications: A review. *Medical Physics*. 2009; 36(6Part1):2052–2068. http://dx.doi.org/10.1118/1.3121511.
3. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. *Acad Radiol*. 1999;6:22-33.
4. Kupinski MA, Giger ML. Automated seeded lesion segmentation on digital mammograms. *IEEE Trans Med Imaging*. 1998; 17(4):510-517.
5. Brake GM, Karssemeijer N. Segmentation of suspicious densities in digital mammograms. *Med Phys*. 2001;28(2):259-266.
6. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88. https://doi.org/10.1016/j.media.2017.07.005.
7. Zhu Q, Du B, Yan P. Boundary-weighted domain adaptive neural network for prostate MR image segmentation. *IEEE Trans Med Imag*. 2020;39(3):753-763. https://doi.org/10.1109/TMI.2019.2935018.
8. Zhu Q, Bo D, Turkbey B, Choyke P, Yan P. Exploiting interslice correlation for MRI prostate image segmentation, from recursive neural networks aspect. *Complexity*. 2018;2018:1-10. https://doi.org/10.1155/2018/4185279.
9. Zhu Q, Du B, Wu J, Yan P. A deep learning health data analysis approach: automatic 3D prostate MR segmentation with densely-connected volumetric convnets. Paper presented at: 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro; 2018. pp. 1-6. https://doi.org/10.1109/IJCNN.2018.8489136.
10. Dhungel N, Carneiro G, Bradley AP. Automated mass detection in mammograms using cascaded deep learning and random forests. *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference*. New York, NY: IEEE; 2015:1-8.
11. Joseph J, Sivaraman J, Periyasamy R, Simi VR. An objective method to identify optimum clip-limit and histogram specification of contrast limited adaptive histogram equalization for MR images. *Biocybernet Biomed Eng*. 2017;37(3):489-497. https://doi.org/10.1016/j.bbe.2016.11.006.
12. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Paper presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018. pp. 7132-7141. https://doi.org/10.1109/CVPR.2018.00745
13. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015. pp. 234-241.
14. Makandar A, Halalli B. Breast cancer image enhancement using median filter and CLAHE. *Int J Sci Eng Res*. 2015;6(4):462-464.
15. Sundaram M, Ramar K, Arumugam N, Prabin G. Histogram modified local contrast enhancement for mammogram images. *Applied Soft Computing*. 2011;11(8):5809–5816.
16. Lee C-Y, Xie S, Gallagher P, Zhang Z, Zhuowen T. Deeply supervised nets. *Proc Machine Learning Res*. 2015;38:562-570.
17. Mishra D, Chaudhury S, Sarkar M, Soin AS. Ultrasound image segmentation: a deeply supervised network with attention to boundaries. *IEEE Trans Biomed Eng*. 2019;66(6):1637-1648.
18. Zhu Q, Du B, Turkbey B, Choyke PL, Yan P. Deeply-supervised CNN for prostate segmentation. Paper presented at: 2017 International Joint Conference on Neural Networks (IJCNN); (2017. https://doi.org/10.1109/ijcnn.2017.7965852
19. Chen H, Qi X, Cheng J-Z, Heng P-A, et al. Deep contextual networks for neuronal structure segmentation. Paper presented at:

13th AAAI Conference on Artificial Intelligence; 2016. pp. 1167-1173.
20. Zhu W, Huang Y, Tang H, Qian Z, Du N, Fan W, Xie X. AnatomyNet: deep 3D squeeze and excitation UNets for fast and fully automated whole volume anatomical segmentation. arXiv: 1808.05238; 2018. pp. 1-14.
21. Wong KCL, Moradi M, Tang H, Syeda-Mahmood T. 3D segmentation with exponential logarithmic loss for highly unbalanced object sizes. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science*. Vol 11072. Switzerland: Springer; 2018. https://doi.org/10.1007/978-3-030-00931-1_70.
22. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Paper presented at: ICLR; 2015.
23. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C Torr P. Conditional random fields as recurrent neural networks. Paper presented at: IEEE ICCV; 2015.
24. Krahenbuhl P, Koltun V. Efficient inference in fully connected crfs with Gaussian edge potentials. Paper presented at: NIPS; 2011.
25. Moreira Inês C., Amaral Igor, Domingues Inês, Cardoso António, Cardoso Maria João, Cardoso Jaime S. INbreast. *Academic Radiology*. 2012;19 (2):236–248. http://dx.doi.org/10.1016/j.acra.2011.09.014.
26. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*. 2017;4:170177.
27. Sun H, Li C, Liu B, et al. AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms. *Physics in Medicine and Biology*. 2020;65(5): 055005. https://doi.org/10.1088/1361-6560/ab5745.
28. Reddi SJ, Kale S, Kumar S. *On the convergence of adam and beyond*. Paper presented at: ICLR 2018 Conference.
29. Hai J, Qiao K, Chen J, et al. Fully convolutional densenet with multiscale context for automated breast tumor segmentation. *Journal of Healthcare Engineering*. 2019;2019:1-11. https://doi.org/10.1155/2019/8415485.
30. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR; January 2017, pp. 6230-6239. https://doi.org/10.1109/CVPR.2017.660
31. Chen L, Zhu Y, Papandreou G, Schroff F, Aug CV. Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv; 2018. http://arxiv.org/abs/1802.02611.
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. https://doi.org/10.1109/ACCESS.2019.2914873.

How to cite this article: Ravitha Rajalakshmi N, Vidhyapriya R, Elango N, Ramesh N. Deeply supervised U-Net for mass segmentation in digital mammograms. *Int J Imaging Syst Technol*. 2020; 1–13. https://doi.org/10.1002/ima.22516