Improving Segmentation of Breast Arterial Calcifications from Digital Mammography: Good Annotation is All You Need

Kaier Wang(✉), Melissa Hill, Seymour Knowles-Barley, Aristarkh Tikhonov, Lester Litchfield, and James Christopher Bare

Volpara Health Technologies, Wellington, New Zealand
kyle.wang@volparahealth.com

Abstract. Breast arterial calcifications (BACs) are frequently observed on screening mammography as calcified tracks along the course of an artery. These build-ups of calcium within the arterial wall may be associated with cardiovascular diseases (CVD). Accurate segmentation of BACs is a critical step in its quantification for the risk assessment of CVD but is challenging due to severely imbalanced positive/negative pixels and annotation quality, which is highly dependent on annotator's experience. In this study, we collected 6,573 raw tomosynthesis images where 95% had BACs in the initial pixel-wise annotation (performed by a third-party annotation company). The data were split with stratified sampling to 80% train, 10% validation and 10% test. Then we evaluated the performance of the deep learning models deeplabV3+ and Unet in segmenting BACs with varying training strategies such as different loss functions, encoders, image size and pre-processing methods. During the evaluation, large numbers of false positive labels were found in the annotations that significantly hindered the segmentation performance. Manual re-annotation of all images would be impossible owing to the required resources. Thus, we developed an automatic label correction algorithm based on BACs' morphology and physical properties. The algorithm was applied to training and validation labels to remove false positives. In comparison, we also manually re-annotated the test labels. The deep learning model re-trained on the algorithm-corrected labels resulted in a 29% improvement in the dice similarity score against the re-annotated test labels, suggesting that our label auto-correction algorithm is effective and that good annotations are important. Finally, we examined the drawbacks of an area-based segmentation metric, and proposed a length-based metric to assess the structural similarity between annotated and predicted BACs for improved clinical relevance.

Keywords: Breast arterial calcification · Deep learning · Segmentation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-27066-6_10.

Y. Zheng et al. (Eds.): ACCV 2022, LNCS 13848, pp. 134–150, 2023.
https://doi.org/10.1007/978-3-031-27066-6_10

1 Introduction

Mammography is a diagnostic imaging technique that is used to detect breast cancer and other early breast abnormalities. During mammography, each patient normally has one mediolateral oblique (MLO) and one craniocaudal (CC) projections for the left and right breasts. Breast arterial calcifications (BACs) noted on mammograms are calcium deposited in the walls of arteries in the breast [5], appearing in various structures and patterns [7]. The presence and progression of BACs have shown to be associated with coronary artery disease and cardiovascular disease (CVD) in recent clinical studies [16,20]. Prevalence of BACs in screening mammograms has been estimated at 12.7% [12]. Breast radiologists may note BACs as an incidental finding, but doing so is subjective and time-consuming, thus automation may be beneficial. Computer aided detection is not new, and several authors have reported promising BACs segmentation results using either classical computer vision algorithms [7,8,10,35] or deep learning models [3,11,28].

Deep learning models have gained increasing popularity in various medical domains [22,23] for their outstanding performance in different tasks such as lesion detection [21], tumor segmentation [27] and disease classification [30] et al. The performance of a deep learning model is typically influenced by hyper-parameters chosen during model training. These include parameters related to experimental factors such as epochs, batch size and input image size; parameters related to training strategies such as loss function, learning rate, optimiser; parameters related to model architecture such as number of layers and choice of encoder. Kaur et al. [17] and Thambawita et al. [31] demonstrated how a model's performance could be improved by properly configuring the hyper-parameters.

Apart from hyper-parameters, the effect of annotation quality on object segmentation has received little attention. In medical image segmentation especially in the scope of BACs segmentation, it is hard or impossible to conduct sufficiently sophisticated annotation due to cost and required domain expertise. Yu et al. [37] indicated that medical imaging data paired with noisy annotation is prevalent. Their experimental results revealed that the model trained with noisy labels performed worse than the model trained using the reference standard.

To address the challenge of noisy labels, in this study we propose a label correction algorithm to automatically remove false positives from manual BACs annotations. The effect of the corrected labels is evaluated along with other hyper-parameters such as image size, image normalisation methods and model architectures. Finally, we analyse the drawbacks of area-based segmentation metric, and propose a length-based metric to evaluate the structural similarity between annotated and predicted BACs for better clinical relevance.

2 Materials

2.1 Dataset

The de-identified image data were collected at a single health institution in the United States. There are 6,573 raw tomosynthesis images acquired from two

x-ray systems: 5,931 from GE Pristina system and 642 from GE Senographe Essential. GE Pristina images have a resolution of 2,850 × 2,394 and 0.1 mm per pixel; and GE Senographe Essential images have a resolution of 3,062 × 2,394 and 0.1 mm per pixel.

Digital breast tomosynthesis is a clinical imaging technology in which an x-ray beam sweeps in an arc across the breast, producing tomographic images for a better visibility of malignant structures [2]. In this study, we use the central projection image from those collected in each scan for simplicity as the source is normal to the detector for this image.

To the best of our knowledge, it may be one of the largest BACs datasets, and the first one reported to use tomosynthesis images for training and evaluating deep learning models. In comparison, other reported datasets are summarised in Table 1.

Table 1. Literature reported BACs datasets.

Literature	No. of images	BACs+ %	Modality
[34]	840	60	2D mammography
[11]	661	NA	2D mammography
[28]	5,972	14.93	2D mammography
[3]	826	50	2D mammography

2.2 Annotation

The annotation task was performed using Darwin.v7labs[1] by a third-party annotation company where the data were first split to batches then assigned to multiple annotators. Cross check or consensus reading were not performed on the annotation results, so each image was only reviewed by a single annotator.

All annotators have prior annotation experience in general domain, but little specific medical knowledge especially in radiology. An introduction to BACs and annotation guidance were supplied to each annotator. Briefly, the task is to use the brush tool to label the BACs pixels. The first 50 annotated samples were reviewed by an in-house imaging scientist with 3 years experience in mammographic image analysis. Upon the acceptance of these samples, the annotators completed the remaining images.

3 Methods

3.1 Image Pre-processing

Digital mammography generally creates two types of images: *raw* and *processed*. Raw images are the images as acquired, with some technical adjustment such as pixel calibration and inhomogeneity correction; processed images are manipulated

[1] https://www.v7labs.com/.

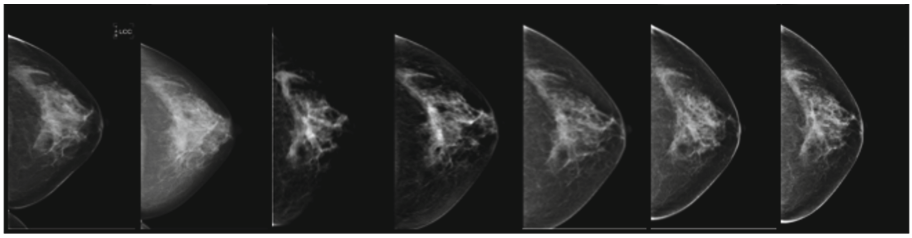

Fig. 1. Seven consecutive years of mammograms for the same patient showing as processed images by three different mammography systems (from left to right): Hologic, GE, Hologic, Hologic, Siemens, Siemens, Siemens. Presentation of the processed images varies significantly based on the mammography system characteristics and image processing.

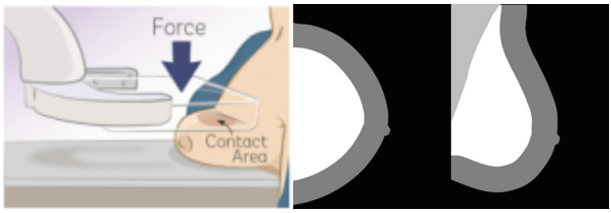

Fig. 2. Demonstration of a compressed breast during mammography (left) and its mammographic segmentation in craniocaudal (middle) and mediolateral oblique (right) views. The compressed region (contact area) is labelled in white, pectoral muscle in light gray and periphery in dark gray. The directly exposed area to x-ray beam is labelled in black as background.

from the raw images by manufacturer specific algorithm to enhance the image contrast to better fit human eye response in detecting tissue lesions. Manufacturers' preferences in image processing may result in the same breast imaged at two x-ray systems having distinctive appearances, as shown in Fig. 1. Furthermore, [36] reported a deep learning model trained from processed images from one manufacturer cannot be transferred to an unseen external dataset, possibly due to image inconsistency. In contrast, raw images record the original information of the detector response, making it ideal for further image processing to achieve a consistent contrast and visual appearance across different mammography systems. Here, we presented three different normalisation methods in ascending complexity: simple gamma correction, self-adaptive contrast adjustment and Volpara® density map. All three methods depend on a segmentation map (see Fig. 2) labelling pectoral muscle, fully compressed and peripheral regions, which were produced by VolparaDensity™ software. The segmentation accuracy of the software was validated by Woutjan et al. [4].

Simple Gamma Correction. Given a raw image I^{raw}, a logarithm transform is applied on each pixel as in Eq. (1) [24] that brightens the intensities of the breast object:

$$I^{\ln} = \ln(I^{\text{raw}} + 1.0) \ . \tag{1}$$

Then, a gamma correction Eq. (2) is applied to the log-transformed image

$$I^{\text{gamma}} = [(I^{\ln} - I^{\ln}_{\min})/(I^{\ln}_{\max} - I^{\ln}_{\min})]^{1/2} \ , \tag{2}$$

where $I^{\ln}_{\min}$ and $I^{\ln}_{\max}$ are the minimum and maximum pixel values respectively in the breast region of $I^{\ln}$.

Self-adaptive Contrast Adjustment. The algorithm brings a given raw mammogram Y_0 to a target mean intensity $\overline{Y}$ in the breast region, by iteratively applying gamma correction. At each step, the gamma value is computed as Eq. (3a), and the image is gamma corrected as Eq. (3b):

$$\gamma_{i+1} = \gamma_i \times \ln(\overline{Y} - \overline{Y_i}) \tag{3a}$$

$$Y_{i+1} = Y_i^{\gamma_{i+1}} \tag{3b}$$

where $\gamma_0 = 1$, and $\overline{Y_i}$ is the mean pixel intensity in the breast region after gamma transformation in the previous step. After a set number of iterations, or after $\overline{Y_i}$ stops converging to the target $\overline{Y}$, the process is terminated. See [19] for detailed implementation.

Volpara® Density Map. The Volpara® algorithm finds an area of the breast within a region in contact with the compression paddle that corresponds to entirely fatty tissues, referred as P^{fat}, then using it as a reference level to compute the thickness of the dense tissue h^{dt} at each pixel location (x, y) based on Eq. (4) [14,18]

$$h^{\text{dt}}(x, y) = \frac{\ln(P(x, y)/P^{\text{fat}})}{\mu^{\text{fat}} - \mu^{\text{dt}}} \ , \tag{4}$$

where the pixel value $P(x, y)$ is linearly related to the energy imparted to the x-ray detector. μ^{fat} and μ^{dt} are the effective x-ray attenuation coefficients for fat and dense tissues respectively at a particular peak potential (kilovoltage peak, or kVp) applied to the x-ray tube [13]. Equation (4) converts a raw mammographic image to a density map where the pixel value corresponds to the dense tissue thickness. The volumetric breast density is then computed by integrating over the entire breast area in the density map. The VolparaDensity™ algorithm has shown strong correlation with the ground truth reading (magnetic resonance imaging data) [33] and its density measurements are consistent across various mammography systems [9].

Figure 3 shows the normalisation results of the above three methods on raw images acquired from two x-ray systems. Despite the raw images are displayed in the same intensity range, GE Pristina tomosynthesis is clearly brighter than the GE Senographe Essential image, and the dense tissues are hardly visible in both images. The simple gamma correction and self-adaptive contrast adjustment stretch the contrast between fat and dense tissue in minor and moderate levels respectively, while the Volpara® density map is a complete nonlinear transform revealing the volumetric tissue properties.

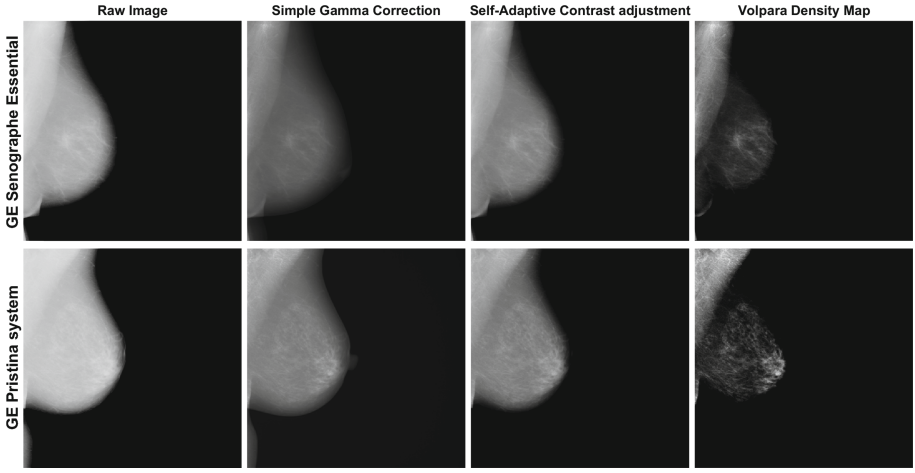

Fig. 3. Examples of normalising raw GE tomosynthesis in three different methods: simple gamma correction, self-adaptive contrast adjustment and Volpara® density map. The 16-bit raw images (first column) are displayed in the same range of [62299, 65415] for better visibility.

3.2 Other Training Variables

Apart from different image normalisation methods, we also investigated other training variables as below:

- Image size: 1024×1024 and 1600×1600[2].
- Loss function: $1/2\times$(dice loss + MSE Loss) [29].
- Model architecture: see Table 2

Table 2. Deep learning models [15] in this study.

Model name	Architecture	Encoder	Parameters, M
Unet^{R}	Unet	Resnet34	24.43
$\text{DeepLabV3+}^{\text{R}}$	DeepLabV3+	Resnet34	22.43
$\text{DeepLabV3+}^{\text{M}}$	DeepLabV3+	Mobilenetv3_large_100	4.70

[2] In this study, we used a single Tesla T4 GPU, which can accommodate a maximum of batch size 3 and 1600×1600 input image size in the training phase. The experiment of patch-based implementation on full resolution images is reported in the Supplementary Material. We found its performance is not comparable with full image implementation, and its slow inference is not practical in a busy clinical environment (GPU is normally not available on a Picture Archiving and Communication System).

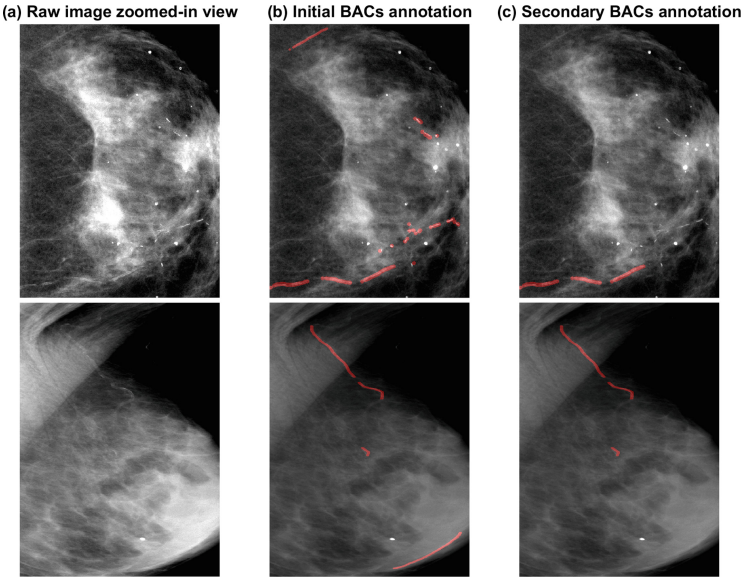

Fig. 4. Zoomed mammogram image sections for comparison between (b) third-party initial BACs annotations and our (c) in-house secondary manual annotations on the same (a) raw image (contrast enhanced for visualisation). In the top panel, micro calcifications and linear dense tissues are initially labelled as BACs; in the bottom panel, the edge of dense tissues is a false positive annotation.

3.3 Label Correction Algorithm

During evaluating the model performance, we discovered large amount of false positives in the BACs annotations. Mostly seen are mislabelling micro-calcifications and dense tissues as BACs, as shown in Fig. 4.

Instead of manually re-annotating all images, we developed a correction algorithm that automatically removes these false positive labels based on BACs' morphology. In [13], Highnam et al. derived Eq. (5) to calculate the calcification thickness (millimeter) from a Volpara® density map:

$$h^{\text{calc}}(x,y) = \frac{(\mu^{\text{dt}} - \mu^{\text{fat}})(h^{\text{dt}}(x,y) - h^{\text{dt}}_{\text{bkg}}(x,y))}{\mu^{\text{calc}}} \,. \tag{5}$$

μ^{calc} is the effective x-ray attenuation coefficient for calcification. Using the values of the linear attenuation coefficients at 18 keV: $\mu^{\text{dt}} = 1.028$ and $\mu^{\text{calc}} = 26.1$. $h^{\text{dt}}_{\text{bkg}}$ is the background tissue thickness, and can be estimated from morphological opening operation as demonstrated in Fig. 5. Thus, a Volpara® density map describing dense tissue thickness can be converted to a calcification thickness map. Then, overlay the annotation on the calcification thickness map, the labels either too short (stand-alone micro calcifications) or having insufficient mean

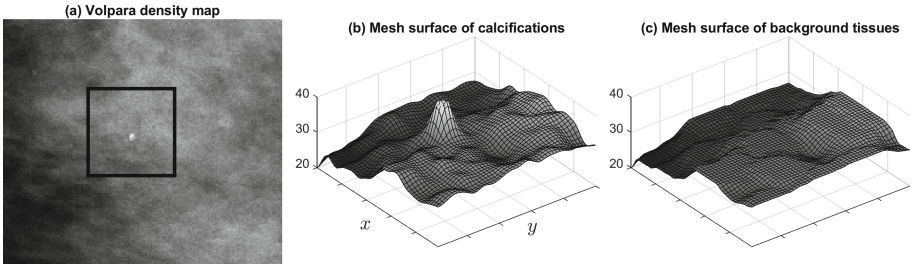

Fig. 5. Demonstration of calcifications and the corresponding background tissues in Volpara® density map. (a) Zoomed-in view of a Volpara® density map where calcifications region is highlighted. (b) Mesh surface of the tissue density (thickness in millimeters) in the highlighted region of (a). (c) Background tissues resulting from morphological opening operation of (b).

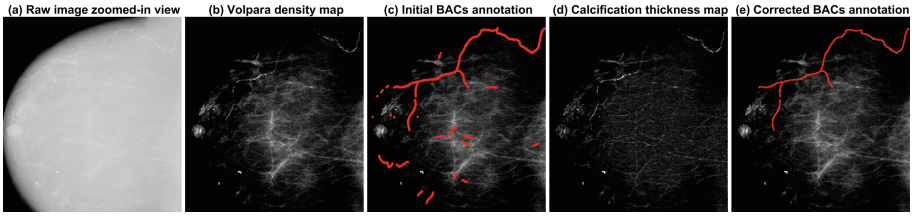

Fig. 6. Demonstration of correcting BACs annotations. First, Volpara® density algorithm converts (a) raw image (zoomed image section) to (b) density map where the pixel value represents dense tissue thickness. Then, Eq. (5) converts (b) density map to (d) calcification thickness map where the pixel value is the potential calcification thickness. The automatic label correction algorithm examines (c) initial BACs annotation. After removing the labels with insufficient length or mean calcification thickness (based on (d) calcification thickness map), the algorithm yields (e) corrected BACs annotation.

calcification thickness (dense tissues) are removed. An example of correcting BACs annotation is shown in Fig. 6.

3.4 Length-Based Dice Score

Typical semantic segmentation evaluation metrics, such as recall, precision, accuracy, F1-score/dice score, examine the overlapping area between prediction and annotation. However, BACs are small in size, and slight differences in their segmentation region may result in strong negative effects on the standard metric like dice score, despite the segmentation still capturing sufficient clinical relevant calcifications. Furthermore, the quantification of BACs is clinically measured by their length rather than area size [1,32]. Therefore, we introduced a length-based dice score to better focus on the linear trace of the BACs.

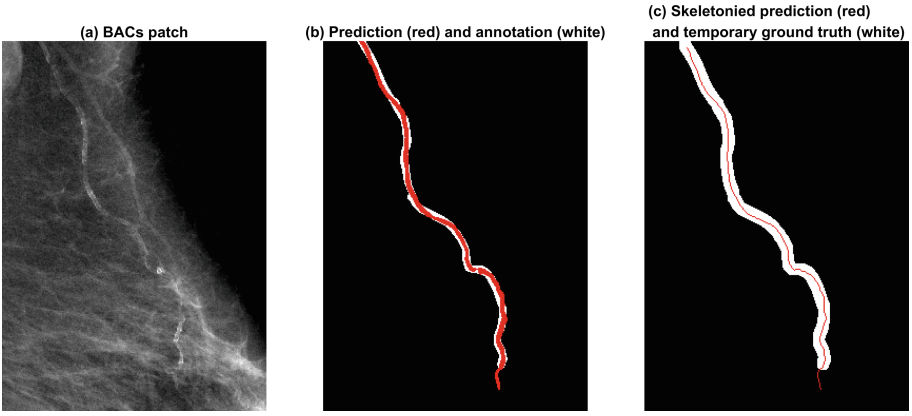

Fig. 7. An example showing (b) prediction and annotation on (a) BACs patch (the prediction and annotation were performed on the full image. We only show a patch here for better visibility). A standard area based dice similarity score from (b) is 0.7081. In (c), the temporary ground truth is derived from skeletonisation followed by dilation on the annotation of (b). The length based dice score reads 0.9750, according to Eq. (6).

Following the work of Wang et al. [35], we first derived the skeletons for both annotation and predicted BACs labels. Then we dilated the skeletonised annotation labels to a width of 2.1 mm, the typical width of a breast artery [34]. The dilated labels were taken as temporary ground truth. A length-based dice score is defined as Eq. (6)

$$\text{Dice}^{\text{L}} = 2 * \frac{\text{Length of predicted BACs within temporary ground truth zone}}{\text{Length of predicted BACs} + \text{Length of annotation BACs}} . \tag{6}$$

The numerator is simply the length of the skeletonised prediction within the region of the temporary ground truth, as seen in Fig. 7(c). In the denominator, the length of the predicted and annotation BACs can be calculated from their respective skeletons. Dice^{L} ranges between 0 and 1 where 1 being a complete match of the length between prediction and ground truth BACs.

From Figs. 7(a) and (b), clearly there are under-segmented annotation at multiple locations along the BACs trace. Wang et al. also reported similar annotation defects in Fig. 4 of [35]. Figure 7(b) shows a strong visual agreement between the prediction and annotation, and the prediction seems to have a better segmentation quality than the annotation. But their subtle mismatch in area only yields a moderate dice similarity of 0.7081. In comparison, our proposed Dice^{L} focuses on the correlation of the linear trace. As shown in Fig. 7(c), the skeleton of the prediction delineates the BACs signals in high agreement, resulting in a more clinical relevant Dice^{L} score of 0.9750.

4 Experiments

4.1 Data Split

The annotation results indicate positive BACs in 95% of the images. The images were split with stratified sampling to 80%/10%/10% for train/validation/test, so each split has a same positive rate of 95%.

4.2 Secondary Annotation on Test Data

The test data were carefully re-annotated by an in-house imaging scientist to remove the false positive segments from the initial annotation as much as possible. After the re-annotation, 93 images are unchanged, 8 images have added BACs segments, and 399 images have false positive segments removed. There are 155 images which have both added and removed segments. The re-annotated labels, noted as ExpertGT (expert ground truth), were utilised to test the deep learning models trained on the original and algorithm corrected train and validation labels.

4.3 Training Settings

The input mammograms are 16-bit, gray scale images. The image went through a contrast normalisation via the methods in Sect. 3.1 followed by a standard pixel value redistribution to achieve a mean of 0 and standard deviation of 1. We also applied common augmentation such as blurring the image using a random-sized kernel, random affine transforms and random brightness and contrast [6].

The models were trained using Adam optimiser with a initial learning rate of $1e^{-4}$. During training, the loss on the validation dataset was monitored at the end of each epoch. The learning rate was reduced by a factor of 10 once the validation loss plateaued for 5 epochs. After a maximum of 100 epochs, or after 10 epochs of no improvement on validation loss, the training was terminated. The model with the best validation loss was saved.

The deep learning segmentation model outputs a probability map. A final binary mask is obtained by applying a cut-off threshold to the probability map. In this study, such cut-off threshold was determined by a parameter sweep from 0.05 to 0.95 at a step of 0.05 on the validation dataset to achieve a highest dice similarity score.

4.4 Results

Table 3 and Fig. 8 present the BACs segmentation results from various combinations of models, image size, normalisation methods and annotation labels. For ease of interpretation, we categorised the results into two groups as illustrated in Table 3. The bottom group comprises five runs from rt0 to rt4 (r stands for run, and t is short for third-party annotated labels), examining the impact of

the image size (rt3 vs. rt4) and normalisation (rt0, rt1, rt2 and rt4) on the segmentation performance. Clearly, the model trained on larger image size yields a better dice score since higher image resolutions preserve more subtle texture information [31]. Meanwhile, simple gamma correction outperforms auto gamma correction (i.e. self-adaptive contrast adjustment) and Volpara® density map in both Dice^{TPA} and $\text{Dice}^{\text{ExpertGT}}$ scores (i.e. Dice scores from the respective third-party annotated, TPA in short, and ExpertGT labels of the test data). Comparing to the reference run rt0 (no image normalisation), image normalisation shows positive influence on improving the segmentation performance.

The top group (ra0 to ra3 where 'a' stands for algorithm corrected labels) in Table 3 investigates the impact of annotations. As mentioned in Sect. 3.3, we developed an automatic algorithm to correct false positives in TPA labels. Here, the models were trained on the corrected labels then evaluated on the original TPA and our re-annotated test labels (ExpertGT in the table). As the ExpertGT labels were corrected from TPA labels, the two kinds of labels would fall into distinct distributions. As a result, the models trained on TPA labels did not perform well on the ExpertGT labels, and vice versa. In contrast, the models trained on the algorithm corrected labels have significantly improved $\text{Dice}^{\text{ExpertGT}}$ scores (compare $\text{Dice}^{\text{ExpertGT}}$ before and after rt4 in Fig. 8). For example, ra1's Dice (0.4485) is 29% higher than rt4's Dice score (0.3477) on the ExpertGT labels of the test data, suggesting the effectiveness of our label correction algorithm.

The top group in Table 3 also probes different model structures. The runs ra2 and ra3 correspond to Unet and DeepLabV3+ architectures respectively, and they both use the same encoder structure of Resnet34, which has 5 times more parameters than $\text{DeepLabV3+}^{\text{M}}$ (see Table 2 for details). Performance-wise, $\text{DeepLabV3+}^{\text{R}}$ slightly falls behind Unet^{R}, and they both surpass $\text{DeepLabV3+}^{\text{M}}$ (ra1) in relatively large margins of 4.88% and 8.12% in $\text{Dice}^{\text{ExpertGT}}$ separately.

Shifting our attention from Dice to Dice^{L} in the last column of Table 3, the length-based Dice^{L} score provides a better intuition of how the predicted BACs clinically correlate to the ExpertGT annotation. As Dice^{L} mitigates the slight difference in BACs segmentation region or width, its score reads higher than the Dice metric. Generally, the Dice^{L} and Dice results are aligned. They both show the advantages of using larger image size, larger model and better quality annotation labels. Among these benefits, quality annotation labels perhaps play the most important role in improving segmentation performance. Figure 8 shows models rt0 to rt4 perform similarly as measured by either Dice or Dice^{L}. Models ra0 to ra3 trained on the corrected labels show consistently higher scores. The major difference between ra0 and rt0 – rt4 is whether or not the corrected labels were used during training.

Figure 9 shows the examples of BACs segmentation results from the top three performing models ra1, ra2 and ra3. The overall performance for BACs segmentation is visually very close to the annotation, and the Dice^{L} score better matches with our perception than the Dice score. In the example at the third row, we can see a surgical scar similar to BACs in appearance. The models ra1

Table 3. Comparison of BACs segmentation performance in terms of the area based Dice and length based Dice^{L} scores on various training configurations. Note the Dice score was computed against both third-party annotated (TPA) and the re-annotated (ExpertGT) test labels, while the Dice^{L} score was computed against the ExpertGT only. The highest Dice and Dice^{L} scores are highlighted.

Run ID	Model Name	Image		Label	Test Dice		Test Dice^{L}
		Normalisation	Size	Train/Val	TPA	ExpertGT	ExpertGT
ra3	DeepLabV3+$^{\text{R}}$	Simple gamma	1600	Algorithm	0.3061	0.4704	0.6261
ra2	Unet$^{\text{R}}$	Simple gamma	1600	Algorithm	0.3122	0.4849	0.6121
ra1	DeepLabV3+$^{\text{M}}$	Simple gamma	1600	Algorithm	0.3072	0.4485	0.5960
ra0	DeepLabV3+$^{\text{M}}$	Raw image	1600	Algorithm	0.3042	0.3993	0.5456
rt4	DeepLabV3+$^{\text{M}}$	Simple gamma	1600	TPA	0.3771	0.3477	0.5088
rt3	DeepLabV3+$^{\text{M}}$	Simple gamma	1024	TPA	0.3500	0.3287	0.4962
rt2	DeepLabV3+$^{\text{M}}$	Auto gamma	1600	TPA	0.3610	0.3447	0.5053
rt1	DeepLabV3+$^{\text{M}}$	Density map	1600	TPA	0.3680	0.3442	0.5084
rt0	DeepLabV3+$^{\text{M}}$	Raw image	1600	TPA	0.3590	0.3355	0.5066

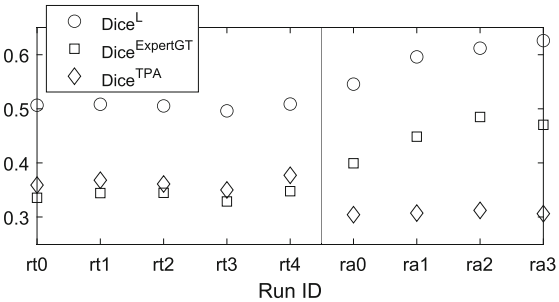

Fig. 8. Visualisation of the Dice and Dice^{L} scores in Table 3. The vertical line between rt4 and ra0, as the horizontal mid-line in Table 3, divides the data into two groups.

and ra2 incorrectly label the scar as BACs while ra3 does not have such false positive prediction. Notably, ra1 has a comparable performance with ra2 and ra3 despite significantly fewer parameters.

4.5 Additional Results

Molloi et al. [25] and Wang et al. [34] reported a symmetrical presence of BACs between the two views (CC vs MLO) and between the two breasts (left vs right). Indeed as shown in Fig. 10, by using segmentation result as a binary prediction, we find a strong correlation between the left and right breasts in a weighted F1-score of: 0.6711 for the annotation, and 0.8811 for our method. Further, a similar correlation is found between the CC and MLO views of: 0.8613 for the annotation, and 0.9139 for our method.

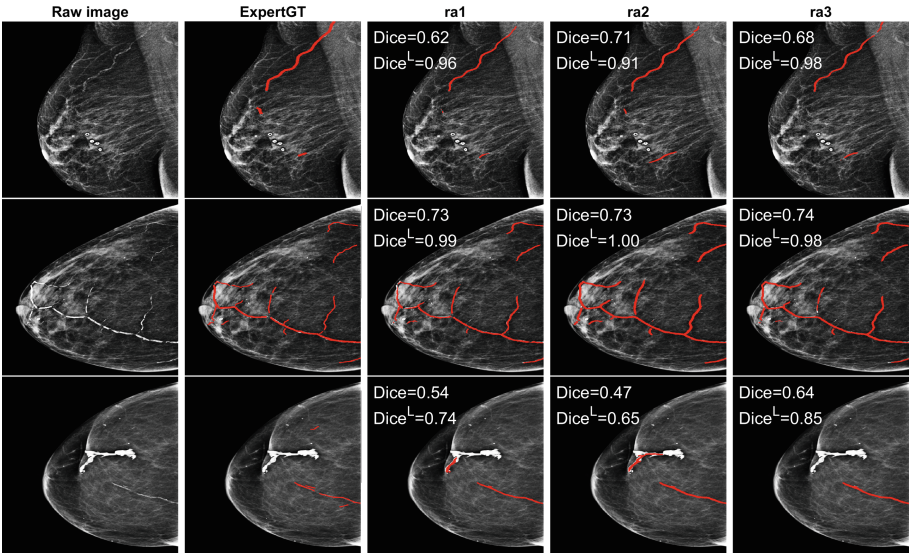

Fig. 9. Examples of BACs segmentation results for ra1, ra2 and ra3 models (see Table 3 for details) as compared to ExpertGT annotations. The raw images were zoomed and contrast enhanced for better visibility.

5 Discussion

This paper presents a comprehensive analysis of training factors and their impacts on the BACs segmentation performance. These training factors include input image size, normalisation method, model architecture and annotation quality. Firstly, we found that the segmentation accuracy benefits from using larger image size for which more subtle features are preserved, relative to down-sampled images. In the experiments of image normalisation, although the model trained on normalised images outperforms the model trained on raw images, their performance does not vary significantly where the simplest gamma correction shows a modest advantage over other more complicated methods. Further, we compared the performance of three model structures: DeepLabV3+^{M}, DeepLabV3+^{R} and Unet^{R}. DeepLabV3+^{M} has 5 times fewer parameters than Unet^{R} and DeepLabV3+^{R} but their performances are very comparable, and their segmentation results are visually close. Lastly, we revealed the importance of high quality annotation. Among other training factors, annotation seems to be the most critical factor determining segmentation performance. A good (here, algorithm corrected) annotation alone shows the highest increase in the BACs segmentation performance, relative to the other hyper-parameter settings tested. However, for hand-crafted annotation it is practically difficult to achieve a pixel-level perfection, nor consistency between readers, and delineation of fine BACs structures in the presence of imaging artefacts is particularly challenging. Expert annotation is expensive and consensus reads from many experts even more so.

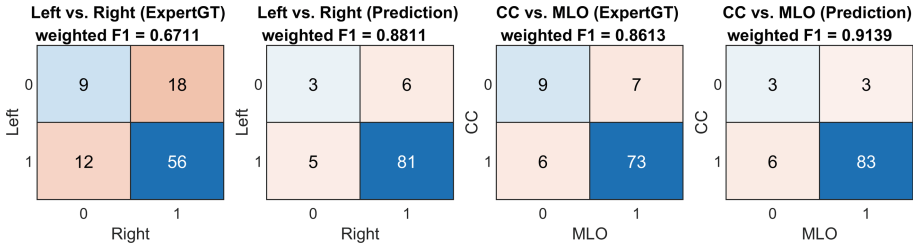

Fig. 10. Confusion matrices for BACs presence between left and right breasts, and between CC and MLO views on test data. Out of 655 test images, there are 380 images belonging to 95 patients with all 4 views (LCC, LMLO, RCC and RMLO) available. The binary prediction is achieved by examining if a segmentation mask is empty. The presented results are from ra3.

Pixel-perfect consistency is difficult even for experts. Crowd-sourced non-expert annotation is inexpensive, but imperfect. To bridge this annotation quality gap, we developed a label correction algorithm that automatically removes false positive labels from BACs annotations. The algorithm demonstrated its effectiveness by allowing a significantly improved segmentation performance.

In addition to the investigation of training factors, we also developed what may be a more clinically relevant metric to improve the evaluation of BACs segmentation. Subtle differences in artery segmentation may have a significant detrimental effect on standard evaluation metrics like Dice score, but may still be able to capture clinically important calcifications with acceptable results. To quantify such clinical relevance, we rectified Dice calculation from its focus on area similarity to trace similarity. The new metric, namely Dice^{L}, has shown to provide a more intuitive measure for the structural correlation between BACs prediction and annotation.

A limitation of this work is the lack of ground truth that is independent of human annotation, e.g. this could come from either physical or digital (simulated) phantom images [26]. We are interested to carry out a validation study where BACs segmentation can be compared to known calcification size and location.

Another important limitation is the lack of negative control images to more thoroughly train and test the model and to comprehensively validate the false positive rate. E.g., Fig. 10 is of limited value without more cases that have no BACs.

In summary, deep learning models have demonstrated promising performance on BACs segmentation. An optimal result can be achieved with higher input image resolution, appropriate image contrast adjustment and larger deep learning model. The annotation quality is found to be a key factor determining the segmentation performance. In general, a model trained with noisy labels is inferior to that trained with good annotation. We recommend other researchers conduct a comprehensive quality control over the annotation process. A thorough review would be required if the annotations were made by non-expert readers.

References

1. Abouzeid, C., Bhatt, D., Amin, N.: The top five women's health issues in preventive cardiology. Curr. Cardiovasc. Risk Rep. **12**(2), 1–9 (2018). https://doi.org/10.1007/s12170-018-0568-7
2. Alakhras, M., Bourne, R., Rickard, M., Ng, K., Pietrzyk, M., Brennan, P.: Digital tomosynthesis: a new future for breast imaging? Clin. Radiol. **68**(5), e225–e236 (2013). https://doi.org/10.1016/j.crad.2013.01.007
3. AlGhamdi, M., Abdel-Mottaleb, M., Collado-Mesa, F.: DU-Net: convolutional network for the detection of arterial calcifications in mammograms. IEEE Trans. Med. Imaging **39**(10), 3240–3249 (2020). https://doi.org/10.1109/TMI.2020.2989737
4. Branderhorst, W., Groot, J.E., Lier, M.G., Highnam, R.P., Heeten, G.J., Grimbergen, C.A.: Technical note: validation of two methods to determine contact area between breast and compression paddle in mammography. Med. Phys. **44**(8), 4040–4044 (2017). https://doi.org/10.1002/mp.12392
5. Bui, Q.M., Daniels, L.B.: A review of the role of breast arterial calcification for cardiovascular risk stratification in women. Circulation **139**(8), 1094–1101 (2019). https://doi.org/10.1161/CIRCULATIONAHA.118.038092
6. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Information **11**(2), 125 (2020). https://doi.org/10.3390/info11020125
7. Cheng, J.Z., Chen, C.M., Cole, E.B., Pisano, E.D., Shen, D.: Automated delineation of calcified vessels in mammography by tracking with uncertainty and graphical linking techniques. IEEE Trans. Med. Imaging **31**(11), 2143–2155 (2012). https://doi.org/10.1109/TMI.2012.2215880
8. Cheng, J.Z., Chen, C.M., Shen, D.: Identification of breast vascular calcium deposition in digital mammography by linear structure analysis. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, Spain, pp. 126–129. IEEE (2012). https://doi.org/10.1109/ISBI.2012.6235500
9. Damases, C.N., Brennan, P.C., McEntee, M.F.: Mammographic density measurements are not affected by mammography system. J. Med. Imaging **2**(1), 015501 (2015). https://doi.org/10.1117/1.JMI.2.1.015501
10. Ge, J., et al.: Automated detection of breast vascular calcification on full-field digital mammograms. In: Giger, M.L., Karssemeijer, N. (eds.) Medical Imaging, San Diego, CA, p. 691517 (2008). https://doi.org/10.1117/12.773096
11. Guo, X., et al.: SCU-Net: a deep learning method for segmentation and quantification of breast arterial calcifications on mammograms. Med. Phys. **48**(10), 5851–5861 (2021). https://doi.org/10.1002/mp.15017
12. Hendriks, E.J.E., de Jong, P.A., van der Graaf, Y., Mali, W.P.T.M., van der Schouw, Y.T., Beulens, J.W.J.: Breast arterial calcifications: a systematic review and meta-analysis of their determinants and their association with cardiovascular events. Atherosclerosis **239**(1), 11–20 (2015). https://doi.org/10.1016/j.atherosclerosis.2014.12.035
13. Highnam, R., Brady, J.M.: Mammographic Image Analysis. Computational Imaging and Vision, Springer, Dordrecht (1999). https://doi.org/10.1007/978-94-011-4613-5
14. Highnam, R., Brady, S.M., Yaffe, M.J., Karssemeijer, N., Harvey, J.: Robust breast composition measurement - Volpara^{TM}. In: Martí, J., Oliver, A., Freixenet, J., Martí, R. (eds.) IWDM 2010. LNCS, vol. 6136, pp. 342–349. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13666-5_46

15. Iakubovskii, P.: Segmentation Models Pytorch (2020). https://github.com/qubvel/segmentation_models.pytorch
16. Iribarren, C., et al.: Breast arterial calcification: a novel cardiovascular risk enhancer among postmenopausal women. Circ. Cardiovasc. Imaging **15**(3), e013526 (2022). https://doi.org/10.1161/CIRCIMAGING.121.013526
17. Kaur, S., Aggarwal, H., Rani, R.: Hyper-parameter optimization of deep learning model for prediction of Parkinson's disease. Mach. Vis. Appl. **31**(5), 1–15 (2020). https://doi.org/10.1007/s00138-020-01078-1
18. Khan, N., Wang, K., Chan, A., Highnam, R.: Automatic BI-RADS classification of mammograms. In: Bräunl, T., McCane, B., Rivera, M., Yu, X. (eds.) PSIVT 2015. LNCS, vol. 9431, pp. 475–487. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-29451-3_38
19. Knowles-Barley, S.F., Highnam, R.: Auto Gamma Correction. WIPO Patent WO2022079569 (2022)
20. Lee, S.C., Phillips, M., Bellinge, J., Stone, J., Wylie, E., Schultz, C.: Is breast arterial calcification associated with coronary artery disease?—a systematic review and meta-analysis. PLoS ONE **15**(7), e0236598 (2020). https://doi.org/10.1371/journal.pone.0236598
21. Li, Y., Gu, H., Wang, H., Qin, P., Wang, J.: BUSnet: a deep learning model of breast tumor lesion detection for ultrasound images. Front. Oncol. **12**, 848271 (2022). https://doi.org/10.3389/fonc.2022.848271
22. Liu, X., et al.: Advances in deep learning-based medical image analysis. Health Data Sci. **2021**, 1–14 (2021). https://doi.org/10.34133/2021/8786793
23. Lundervold, A.S., Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI. Z. Med. Phys. **29**(2), 102–127 (2019). https://doi.org/10.1016/j.zemedi.2018.11.002
24. Marchesi, A., et al.: The effect of mammogram preprocessing on microcalcification detection with convolutional neural networks. In: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), Thessaloniki, pp. 207–212. IEEE (2017). https://doi.org/10.1109/CBMS.2017.29
25. Molloi, S., Mehraien, T., Iribarren, C., Smith, C., Ducote, J.L., Feig, S.A.: Reproducibility of breast arterial calcium mass quantification using digital mammography. Acad. Radiol. **16**(3), 275–282 (2009). https://doi.org/10.1016/j.acra.2008.08.011
26. Molloi, S., Xu, T., Ducote, J., Iribarren, C.: Quantification of breast arterial calcification using full field digital mammography. Med. Phys. **35**(4), 1428–1439 (2008). https://doi.org/10.1118/1.2868756
27. Ranjbarzadeh, R., Bagherian Kasgari, A., Jafarzadeh Ghoushchi, S., Anari, S., Naseri, M., Bendechache, M.: Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. Sci. Rep. **11**(1), 10930 (2021). https://doi.org/10.1038/s41598-021-90428-8
28. Riva, F.: Breast arterial calcifications: detection, visualization and quantification through a convolutional neural network. Thesis for Master of Sciences in Biomedical Engineering, Polytechnic University of Milan, Italy (2021)
29. Savioli, N., Montana, G., Lamata, P.: V-FCNN: volumetric fully convolution neural network for automatic atrial segmentation. In: Pop, M., Sermesant, M., Zhao, J., Li, S., McLeod, K., Young, A., Rhode, K., Mansi, T. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 273–281. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_30

30. Srinivasu, P.N., SivaSai, J.G., Ijaz, M.F., Bhoi, A.K., Kim, W., Kang, J.J.: Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. Sensors **21**(8), 2852 (2021). https://doi.org/10.3390/s21082852
31. Thambawita, V., Strümke, I., Hicks, S.A., Halvorsen, P., Parasa, S., Riegler, M.A.: Impact of image resolution on deep learning performance in endoscopy image classification: an experimental study using a large dataset of endoscopic images. Diagnostics **11**(12), 2183 (2021). https://doi.org/10.3390/diagnostics11122183
32. Van Berkel, B., Van Ongeval, C., Van Craenenbroeck, A.H., Pottel, H., De Vusser, K., Evenepoel, P.: Prevalence, progression and implications of breast artery calcification in patients with chronic kidney disease. Clin. Kidney J. **15**(2), 295–302 (2022). https://doi.org/10.1093/ckj/sfab178
33. Wang, J., Azziz, A., Fan, B., Malkov, S., Klifa, C., Newitt, D., Yitta, S., Hylton, N., Kerlikowske, K., Shepherd, J.A.: Agreement of mammographic measures of volumetric breast density to MRI. PLoS ONE **8**(12), e81653 (2013). https://doi.org/10.1371/journal.pone.0081653
34. Wang, J., Ding, H., Bidgoli, F.A., Zhou, B., Iribarren, C., Molloi, S., Baldi, P.: Detecting cardiovascular disease from mammograms with deep learning. IEEE Trans. Med. Imaging **36**(5), 1172–1181 (2017). https://doi.org/10.1109/TMI.2017.2655486
35. 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), Dunedin, New Zealand, pp. 1–6. IEEE (2019). https://doi.org/10.1109/IVCNZ48456.2019.8960956
36. Wang, X., Liang, G., Zhang, Y., Blanton, H., Bessinger, Z., Jacobs, N.: Inconsistent performance of deep learning models on mammogram classification. J. Am. Coll. Radiol. **17**(6), 796–803 (2020). https://doi.org/10.1016/j.jacr.2020.01.006
37. Yu, S., Chen, M., Zhang, E., Wu, J., Yu, H., Yang, Z., Ma, L., Gu, X., Lu, W.: Robustness study of noisy annotation in deep learning based medical image segmentation. Phys. Med. Biol. **65**(17), 175007 (2020). https://doi.org/10.1088/1361-6560/ab99e5