POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Breast Arterial Calcifications: detection, visualization and quantification through a convolutional neural network

Master Of Science In Biomedical Engineering

AUTHOR: Francesca Riva
ADVISOR: Professor Giuseppe Baselli; Professor Francesco Sardanelli, MD
ACADEMIC YEAR: 2020-2021

1. Introduction

Breast arterial calcifications (BACs) are common findings in mammograms acquired for breast cancer screening. Unlike coronary arterial calcifications, they do not cause clinical signs of vessel restriction or occlusion, therefore are not traditionally mentioned on medical reports. Recently BACs presence and intensity have been considered as a risk factor of cardiovascular disease (CVD) [1]. CVD risk in women is often underestimated, and the rate of decline of deaths by CVD is lower in woman than in men. This could be caused by lack of sex-specific risk factors, thus the inclusion of BACs severity in preventive risk assessment might improve upon the reduction of CVD burden in female population.

Despite 80.7% of radiologists declare to be aware of the correlation between BACs and CVD, only 61.9% report BACs findings and 20% quantify the calcifications severity [1]. This low rate of reports is caused by both the lack of a robust method for BACs quantification and by the absence of an adequate automatic support.

This work aims at addressing the latter issue by developing and validating the technical steps needed for BACs automatic detection and quantification: a deep convolutional neural network (CNN) is trained for the detection of BACs presence. Next, in the framework of AI explainability, a visualization method is applied to map the CNN response. Finally, an automatic procedure for quantifying BACs severity is proposed based on such maps. Similar workflows are reported in literature [2,3]; nonetheless, the training of all state-of-the-art quantification tools rely on pixel-wise images annotations to produce an accurate BACs segmentation. This requires time-consuming manual segmentation of the calcifications performed by clinicians, which causes difficulties in training and testing the algorithm with a sufficient number of images. Moreover, this increases the rate of human errors in the annotation used as ground-truth. On the other hand, the proposed CNN performs a binary classification, so it is trained on image-wise annotations that report only BACs presence (BAC+ image) or absence (BAC- image), which are easier to produce. The dimensions of the dataset used are therefore higher, increasing reliability of results. Moreover, BACs severity assessment doesn't require a training dataset with manual BACs segmentation since it is based on the extraction of geometrical features from the heatmaps produced to visualize network's results. Only a small subset with manual annotations of BACs lengths is needed to assess the correlation between the automatic severity prediction and the manual reference.

2. Methods

2.1. Mammographic dataset

Four views mammograms of retrospectively enrolled patients were collected. Images were acquired by full-field digital mammography devices at IRCCS Policlinico San Donato and labelled by three human readers as positive (BAC+) or negative (BAC-) to BACs both at patient level and at image level. For privacy protection, all patients were anonymized, and data associated with each image were discarded except for age, mammographic view and acquisition device.

Patients' age was analyzed and an a-posteriori inclusion criteria was fixed: patients with age<45 were excluded from the study, since no BAC+ case younger than 45 years was found.

Images were preprocessed by extracting the breast region of interest (ROI): Otsu thresholding was applied to each image, separating breast tissue over threshold from the dark background. Pixels corresponding to background were fixed to a value of -20, while breast pixels were normalized to obtain zero-mean distribution and variance equal to 1. Breast ROI was cropped and resized by rigid rescaling, until reaching dimensions of 1536x768 pixels, that coincide with the input shape of the CNN.

The dataset was split into three subsets: training, validation, and test subsets, containing respectively 70%, 15% and 15% of data. Considering the correlation of BACs incidence with age, the splitting strategy was focused on maintaining age distribution of the original dataset across the three subsets. The age quartiles of BAC+ population were used to define four age classes (Class1=minimum-Q_1, Class2=Q_1-Q_2, Class3=Q_2-Q_3, Class4=Q_3-Q_4), that were used to divide the dataset based on patients' age. For each age class, the splitting in training, validation and test subsets was performed, and the resulting four classes for each subset were further merged.

Taking into account the low prevalence of BAC+ patients (14.93%), reducing data unbalance in the training set was needed to improve CNN training. BAC+ prevalence in each age class of the training dataset was therefore evaluated, performing undersampling of BAC- images to reach 30% BAC+ prevalence in each class. Validation and test sets were not undersampled, to reflect the real BAC+ prevalence.

2.2. Convolutional neural network

The neural network architecture used to classify BACs is the one developed by Ienco et al. for this task, based on VGG16 architecture [4]. The first 13 convolutional layers and are organized into five blocks: the first two are composed of two layers, the remaining ones of three layers; after each block a max pooling over a 2x2 window is performed. Convolutional layers are followed by fully connected layers of 256 neurons and an output fully connected layer of 1 neuron. All layers present leaky ReLU activation function, except for the output layer that uses a sigmoidal activation. The training strategy developed by Ienco et al. relies on transfer learning from VGG16 for the first 8 convolutional layers, which parameters were frozen, and initializes the remaining trainable layers with Glorot uniform function. The fully connected layers were trained with 0.3 dropout rate. A cosine annealing strategy was applied, setting the learning rate as:

$$lr_{eph} = lr_{start} * \frac{cos(\pi * eph/eph_{max}) + 1}{2} \qquad (1)$$

where, at each epoch eph, learning rate is lr_{eph}; learning rate's starting value before the decay is lr_{start}, and eph_{max} is the number of epochs after which the learning rate goes to zero.

Briefly, the network considered by Ienco et al. presented these parameters: lr_{start}= $10^{-5,}$ eph_{max}=100, number of training epochs n_{eph}=50, dropout rate=0.3. This network was trained by 7-fold cross validation on a small dataset, producing 7 different models. In the current work, the best performing model was referred to as MG-Net and was used as starting point to improve hyperparameters tuning, further training and independent testing, to finalize the actual clinical validation of the CNN, thanks to the larger data-base available.

Considering the unbalanced dataset, metrics used to evaluate results were precision, recall and F1, along with area under ROC curve (ROC AUC) and area under precision-recall curve (PR AUC).

The initialization of trainable layers both with Glorot uniform function and with MG-Net weights was explored.

Tuning of the most relevant network's hyperparameters was then performed by gradually modifying them with respect to MG-Net. Learning rate decay was evaluated firstly by varying lr_{start}, assigning it values of 10^{-n}, with n=

[4,5,6]. Subsequently the decay rate was explored by changing eph_{max}, assigning it values of 200, 400, 600 and 800. The number of epochs n_{eph} was analysed within a range from 25 to 300 epochs, and the dropout rate for the fully connected layers was studied for values between 0.2 and 0.5.

The classification threshold used to produce a binary result from the sigmoidal output was fixed at 0.5 for all models tested. Results were compared over the validation subset allowing to extract the best performing network, BAC-Net.

BAC-Net performances were further tested on the independent test subset, using different classification thresholds between 0 and 1.

Classification thresholds resulting in the best precision were referred to as P-th, the one maximising recall as R-th and the one maximising F1 as F1-th. Obviously, such thresholds are related to the actual dataset, still provide useful general indications.

An ultimate classification threshold τ was computed by averaging F1-th assessed over the test and the validation sets. Classification with τ was performed to evaluate results both image-wise and patient-wise, considering a patient as BAC+ if at least one of the four mammographic views was classified as BAC+ image.

2.3. Results visualization

To explore BAC-Net behavior, state-of-the-art visual explanation methods developed for neural networks (Saliency maps, SmoothGrad, GradCAM, GradCAM++) were compared. Their ability to provide an explanation of network's results was evaluated along with radiologists. The best performing method was found to be GradCAM++, that presented lower noise and higher accuracy in BACs location and delineation. GradCAM++ produces a heatmap of the activation of each pixel by assigning it a weight proportional to the derivative of the output score with respect to the feature maps activation of the selected convolutional layer. The behavior of all convolutional layers was explored, and the last convolutional layer was the one considered for final heatmaps generation, as it contained high-level information and showed higher accuracy.

2.4. Severity scoring

A small dataset of BAC+ patients previously included in a manual BACs semiquantitative score (BAC-SS) study [5] was used to perform an assessment of the possible correlation between manual evaluation of BACs length (l_{BAC}) and automatically extracted scores based on GradCAM++ heatmaps thresholding.

Two mammographic views per patient, one for each breast, were selected, to reflect the procedure applied for manual scoring, and preprocessed as described in section 2.1. The dataset was then fed to BAC-Net, and sigmoidal outputs were evaluated by generating R-th, P-th and F1-th specific to this set of predictions. Since precision maximization provides a classification with the minimum number of false positives, P-th was considered to proceed in automatic scores evaluation. GradCAM++ heatmaps were generated and, for each heatmap, binary thresholding was performed with threshold $T_{heatmap}$ varying from 0 to 1 with step 0.1.

Three continuous severity scores were considered for automatic extraction (Figure 1): the heatmap's area with intensity above $T_{heatmap}$ (A_{BAC}), the sum of pixels' intensities inside this area (I_{BAC}), and an estimation of BACs length obtained by skeletonization of the over-threshold objects (L_{BAC}). In case of BAC+ images, these three scores were computed for each $T_{heatmap}$; for BAC- images, all scores were set to 0.

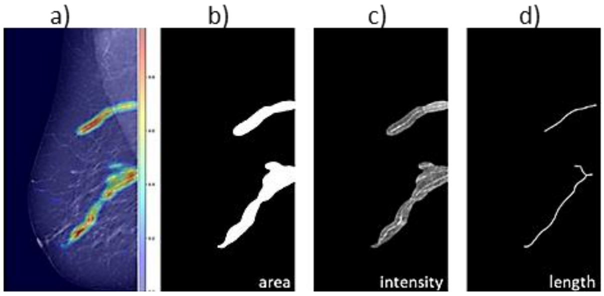


Figure 1. a) Example of GradCAM++; b) thresholding with $T_{heatmap}$=0.5 and A_{BAC} extraction; c) pixels summed to compute I_{BAC}; d) skeletonization to extract L_{BAC}

For each $T_{heatmap}$, l_{BAC} was compared with A_{BAC}, I_{BAC} and L_{BAC} through linear regression and by computing Spearman correlation coefficient. For each score, the optimal $T_{heatmap}$ value was considered as the one maximising correlation. Optimal thresholds for area, intensity and length are indicated respectively as $T_{opt\text{-}A}$, $T_{opt\text{-}I}$ and $T_{opt\text{-}L}$.

Since BAC-SS evaluated BACs length also with a quartile-based length score (l_Q) ranging from 0 to 4, three ordinal scores were generated for area (A_Q), pixels intensity (I_Q) and predicted length (L_Q). They were computed by assessing the quartiles of A_{BAC},

I_{BAC} and L_{BAC}, using them as thresholds to generate values ranging from 1 to 4; as for continuous scores, value 0 was assigned to BAC- image.

The quartiles-based length l_Q was compared with A_Q, I_Q and L_Q obtained by thresholding the heatmap with T_{opt-A}, T_{opt-I} and T_{opt-L}. The scores correlation was assessed by producing a confusion matrix comparing A_Q, I_Q and L_Q predictions with l_Q ground truth. Accuracy of predictions was computed as the sum of true positive predictions over the total number of predictions.

Classification performed with R-th and F1-th was finally evaluated and compared with previous results.

3. Results

3.1. Dataset

Application of inclusion criteria removed 64 BAC- patients; the resulting dataset composed of 1493 female subjects (5972 images), of which 194 BAC+ (14.93%).

Patients' ages followed a non-normal distribution (Shapiro-Wilk test resulted in W= 0.96, p-value< 0.01). Quartiles of the BAC+ age distribution were computed (minimum=45years, Q_1=60y, Q_2=70y, Q_3=73y, Q_4=87y), and used as age classes during data splitting. The training subset resulted of 1042 patients, of which 908 negatives and 134 positives to BACs (12.85% BAC+ prevalence); the validation subset contained 222 patients, of which 194 BAC- and 28 BAC+ (12.61%); lastly the test set was composed of 229 patients, 197 BAC- and 32 BAC+ (13.9%).

Regarding the training set, since Class3 and Class4 were already characterized by 30% BAC+ prevalence, undersampling was performed only for Class1 and Class2. This resulted in randomly removing 474 BAC- patients from Class1 and 158 BAC- patients from Class2. The final training dataset was therefore composed of 410 patients, of which 276 BAC- and 134 BAC+ (32.68% BAC+ prevalence).

3.2. Network tuning and evaluation

Evaluation over the validation set of the best initialization for the trainable layer resulted in F1=0.178 for initialization with Glorot uniform function, and F1=0.406 for initialization with MG-Net, therefore the latter strategy was chosen.

The network behaved randomly for lr_{start}= 10^{-4}, and overfitted the training set for lr_{start}= 10^{-5}. For these reasons, lr_{start}= 10^{-6} was chosen. Value of eph_{max}= 800 resulted in the best F1 performances over the validation set, and reduced overfitting. The best number of training epochs was found to be n_{eph}= 25: despite the absence of overfitting, when increasing training epochs, the results over validation set did not improve due to output neuron's saturation, that caused it to behave like a binary classifier reducing its discrimination potential. Dropout rate was maintained at 0.3; lower or higher values produced worse results both over validation and training set.

The best performing network, BAC-Net, was used to classify the test set images, allowing the evaluation of the classification thresholds maximizing precision, recall and F1, that resulted respectively in: P-th=0.99, R-th=0.13 and F1-th=0.88. Applying P-th to classification of test set resulted in F1=0.565, precision=1.0, recall=0.394. Conversely, predictions with R-th resulted in F1=0.232, precision=0.131, recall=1.0.; classification with F1-th resulted in F1=0.767, precision=0.802, recall=0.734. The ultimate optimal threshold τ was computed averaging F-th for test set and F-th for validation set (0.83), resulting in τ=0.85.

Results of images classification by applying τ over training, validation and test sets are reported in Table 1; patient-wise results are reported in Table 2.

Dataset	Precision	Recall	F1
Training	0.963	0.723	0.723
Validation	0.9	0.707	0.792
Test	0.831	0.680	0.748

Table 1. Image-wise BAC-Net results

Dataset	Precision	Recall	F1
Train	0.914	0.873	0.893
Validation	0.813	0.928	0.866
Test	0.831	0.680	0.748

Table 2. Patient-wise BAC-Net results

BAC-Net classification of mammographic images reported good results over the test set, and the possibility to vary the classification threshold allows for future adaptability of the CNN to the scope of the prediction: for BACs screening amongst women, a low threshold favoring recall will guarantee a low number of false negatives, including all subjects with a possible CVD risk in the BAC+ category; on the other hand, for research purposes (such as testing of the scoring procedure proposed in this thesis), a high threshold favoring precision can be used to avoid false positive predictions, allowing to extract BAC+ images with high confidence. BAC-Net future improvements should be focused on reducing the output neuron saturation, allowing for a higher number of training epochs. Moreover, a larger mammograms database might increase the variability of training data, ultimately producing better predictions.

3.3. GradCAM++ visualizations

GradCAM++ heatmaps were able to highlight presence and position of one or multiple BACs when computed for true positive predictions (TP) (Figure 2a). Severe calcifications were easily detected, while in case of small multiple BACs the heatmap wasn't always able to highlight all of them. False positive (FP) cases were generated mainly by presence of fibrous tissue (Figure 2b) or benign calcifications with linear shape. The presence of round microcalcifications was not misleading when their shape was well defined and they were not superimposed to dense tissue, but in some less defined cases represented a confounding factor as well.

GradCAM++ of negative predictions (TN) highlighted the whole breast (Figure 2c) and allowed to understand how medical devices (as pacemakers, cardiac loop recorders or breast implants) do not bias the network outcomes, therefore they don't represent a confounding factor. False negative predictions (FN) were usually related to small BACs over dense breast tissue (Figure 2c).

Overall, GradCAM++ heatmaps of BAC-Net predictions allowed to start to open the black box of the network and explore its behavior; moreover, the possibility of visualizing BAC position predicted by the CNN encouraged a discussion among engineers, physicists, and radiologists about possible improvements and increased the clinicians' confidence in prediction results.

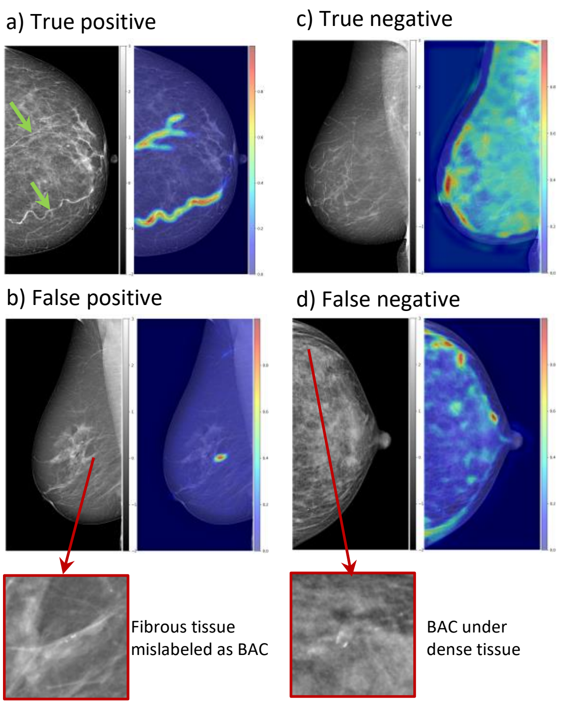


Figure 2. a) TP case of severe BACs correctly identifying the calcified vessels; b) TN case highlighting the whole breast; c) FP case, fibrous tissue mislabeled as BAC; d) FN case, mislabeling is caused by tissue density

3.4. Severity scoring

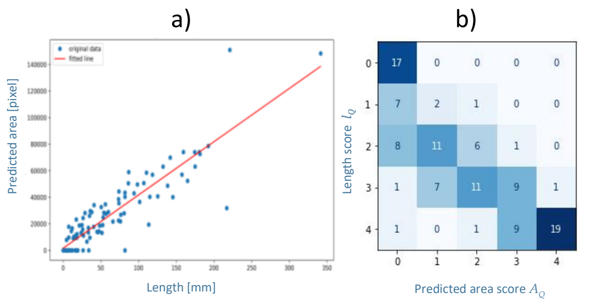
The scoring dataset was composed of 56 BAC+ patients; for each patient the two mediolateral oblique views were considered, for a total of 112 mammograms, of which 95 BAC+ and 17 BAC- images.

BAC-Net sigmoidal outputs for this set of mammograms allowed to compute P-th=0.7, F1-th=0.6 and R-th=0.1.

By using P-th, BAC-Net predicted 78 images as BAC+, 34 images as BAC-, of which 0 false positive predictions and 17 false negative predictions. Correlation between A_{BAC}, I_{BAC} and L_{BAC} and l_{BAC} was assessed for variable binarization threshold $T_{heatmap}$. The $T_{heatmap}$ maximising Spearman's correlation coefficient between l_{BAC} and A_{BAC} was $T_{opt\text{-}A}$= 0.2, the same value resulted for I_{BAC}, so that $T_{opt\text{-}I}$= 0.2, while for L_{BAC}, $T_{opt\text{-}L}$= 0.3. These optimal thresholds were also the one minimizing p-value for Spearman's coefficient.

By using the respective binarization threshold, correlations of l_{BAC} with A_{BAC} ($R_{spearman}$=0.90, p-value=$6.33e^{-41}$), with I_{BAC} ($R_{spearman}$=0.90, p-value=$4.36e^{-41}$), and with L_{BAC} ($R_{spearman}$=0.89, p-value=$1.64e^{-39}$) were compared. The best predictor for BACs real length was found to be A_{BAC}. A linear regression between l_{BAC} with A_{BAC} is shown in Figure 3a.

The comparison of l_Q with quartiles-based scores resulted in identical performances for A_Q and I_Q (accuracy=0.47) while L_Q predictions were slightly worse (accuracy=0.46). The confusion matrix comparing l_Q to A_Q can be found in Figure 3b.



Length score l_Q \ Predicted area score A_Q	0	1	2	3	4
0	17	0	0	0	0
1	7	2	1	0	0
2	8	11	6	1	0
3	1	7	11	9	1
4	1	0	1	9	19

Figure 3. a) Linear regression between real length l_{BAC} and predicted area A_{BAC} ($R_{spearman}$=0.90, p-value=6.33e^{-41}); b) Confusion matrix displaying real length score l_Q on vertical axis, predicted area score A_Q on horizontal axis (accuracy=0.47)

Evaluation of linear regression for scores extracted by using F1-th and R-th resulted in lower correlations, due to the increase in number of false positives caused by lower classification thresholds. Nonetheless, performances of A_{BAC} were always better than the ones of I_{BAC} and L_{BAC}. Quartiles-based scores computed with F1-th provided better results with respect to the ones computed with P-th, while R-th worsened the predictions. A_Q resulted the best predictor for l_Q both when using F1-th and R-th as classification thresholds: F1-th provided best results with respect to P-th (accuracy=0.53) while R-th worsened the predictions (accuracy= 0.36).

It must be considered that preliminary results here reported for the scoring procedure are tested on a small dataset, which required manual BACs segmentation. So, further validation with a larger dataset is needed to provide a more robust correlation and to fix continuous (S_{BAC}) and ordinal (S_Q) final BACs scores. Nonetheless, this work demonstrates the feasibility of predicting BACs severity without requiring the manual segmentation of the training set images.

4. Conclusions

All technical steps needed to develop an automated procedure for BACs analysis have been studied in this thesis, demonstrating the possibility to classify mammograms based on BACs presence by using a convolutional neural network, and to quantify calcifications severity extracting geometrical scores from network's heatmaps.

Once the scoring procedure will be finalized, it will be possible to actuate the workflow proposed in Figure 4.

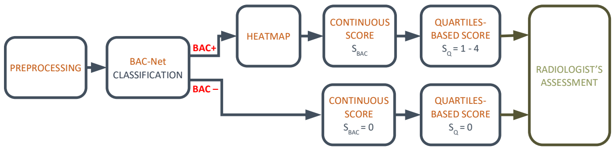

PREPROCESSING
BAC-Net CLASSIFICATION
BAC+
BAC –
HEATMAP
CONTINUOUS SCORE S_{BAC}
QUARTILES-BASED SCORE $S_Q = 1 - 4$
CONTINUOUS SCORE $S_{BAC} = 0$
QUARTILES-BASED SCORE $S_Q = 0$
RADIOLOGIST'S ASSESSMENT

Figure 4. Possible workflow for automatic detection and quantification of BACs

The clinicians' workload for BACs detection and quantification will be reduced by this procedure, since all steps are automatized. Ultimately, clinicians will be supported in their decision about the need to further investigate patient's CVD risk. This would help increasing the number and quality of BACs reports during screening mammography, and ultimately improve CVD stratification for women. Moreover, a higher amount of data quantifying BACs severity could be produced, encouraging further clinical tests for BACs correlation with cardiovascular pathologies such as coronary heart disease or cerebrovascular disease, but also with other CVD risk factors.

References

[1] Trimboli RM, Capra D, Codari M, Cozzi A, di Leo G, Sardanelli F. Breast arterial calcifications as a biomarker of cardiovascular risk: radiologists' awareness, reporting, and action. A survey among the EUSOBI members. European Radiology 2020. https://doi.org/10.1007/s00330-020-07136-6.

[2] Alghamdi M, Abdel-Mottaleb M, Collado-Mesa F. DU-Net: Convolutional Network for the Detection of Arterial Calcifications in Mammograms. IEEE Transactions on Medical Imaging 2020;39:3240–9. https://doi.org/10.1109/TMI.2020.2989737.

[3] Guo X, O'Neill WC, Vey B, Yang TC, Kim TJ, Ghassemi M, et al. SCU-Net: A deep learning method for segmentation and quantification of breast arterial calcifications on mammograms. Medical Physics 2021;48:5851–61. https://doi.org/https://doi.org/10.1002/mp.15017.

[4] Ienco MG, Codari M, Baselli G, Sardanelli F. Breast arterial calcifications on mammograms: deep learning detection for women's cardiovascular risk stratification. 2018.

[5] Trimboli RM, Codari M, Cozzi A, Monti CB, Capra D, Nenna C, et al. Semiquantitative score of breast arterial calcifications on mammography (BAC-SS): intra- and inter-reader reproducibility. Quantitative Imaging in Medicine and Surgery 2021;11:2019–27. https://doi.org/10.21037/qims-20-560.