POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

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

THESIS FOR MASTER OF SCIENCES IN BIOMEDICAL ENGINEERING

Author: ***Francesca Riva***

Advisors: *Professor Giuseppe Baselli; Professor Francesco Sardanelli, MD*

Co-advisors: *Doctor Nazanin Mobin; Doctor Davide Capra, MD*

Academic Year: 2020-2021

Student ID: 945008

Abstract

Cardiovascular diseases (CVD) are the world's leading death cause; their risk is often underestimated in female patients, indicating lack of sex-specific factors in current risk stratification. Breast arterial calcifications (BACs) are common findings in mammograms acquired for breast cancer screening and have recently been considered as a CVD risk factor specific to women. Although they might improve upon risk stratification, only 61.9% of radiologists report their presence during breast cancer screening, and annotations about their severity are reported just by 20% of radiologists. An automatic support to BACs analysis might encourage the use of BACs as CVD risk marker, by reducing the radiologists' workload for their assessment.

This thesis studies the technical steps needed to develop an automated workflow for BACs assessment, by training a deep convolutional neural network (CNN) for the detection of BACs presence, testing a visualization method, and proposing an automatic BACs severity scoring procedure. The CNN architecture considered was developed in previous works for BACs binary classification (present vs absent). It consists of 16 layers, of which the first 8 trained with transfer learning from VGG-16. The remaining layers were trained in this work with a set of 1640 images; validation was performed with 888 images. Hyperparameters analysis was conducted and the best performing model (BAC-Net) was tested with 916 images, resulting in precision=0.831, recall=0.68, F1=0.748, ROC AUC=0.95. BAC-Net outputs were analysed through GradCAM++, producing heatmaps of the last convolutional layer. BACs position and extent were highlighted in the heatmaps with good precision, allowing to increase clinicians' confidence in BAC-Net predictions. A smaller dataset of 112 images, annotated also with BAC quantification, was used for the preliminary test of a scoring procedure based on heatmaps thresholding. This method allowed to extract three different continuous scores related to area (A_{BAC}), pixel intensities (I_{BAC}) and predicted length (L_{BAC}) of over-threshold regions. Linear regression between BACs length (l_{BAC}) assessed by human readers and these scores was performed. A_{BAC} resulted in the highest correlation with l_{BAC} (R_S=0.90, p-value=6.33e^{-41}). Lastly, ordinal scores (A_Q, I_Q, L_Q) ranging from 1 to 4 were extracted from A_{BAC}, I_{BAC}, and L_{BAC} based on their distribution quartiles, and correlated with ground-truth length ordinal score l_Q (based on l_{BAC}). A_Q showed the highest prediction accuracy for l_Q (accuracy= 0.53).

Key-words: breast arterial calcifications, cardiovascular diseases stratification, convolutional neural network, explainable artificial intelligence

Abstract in italiano

Le malattie cardiovascolari (CVD) sono la principale causa di morte al mondo; il loro rischio è spesso sottostimato nelle donne, indicando una mancanza di fattori specifici per il genere femminile nell'attuale stratificazione del rischio di CVD. Le calcificazioni arteriose mammarie (BAC) sono ritrovamenti frequenti nelle mammografie acquisite a scopo di screening per i tumori al seno, e sono state recentemente considerate un fattore di rischio specifico per le donne. Nonostante le BAC abbiano la potenzialità di migliorare la prevenzione del rischio cardiovascolare, appena il 61.9% dei radiologi riporta la loro presenza, e la gravità delle calcificazioni viene indicata solo dal 20% dei radiologi. Un supporto automatico all'analisi delle BAC potrebbe incoraggiare il loro uso come marker per le CVD, riducendo il lavoro a carico dei radiologi per il loro studio.

Questa tesi si occupa dello studio dei passaggi tecnici necessari allo sviluppo di una procedura automatizzata per l'analisi delle BAC, tramite il training di un neural network convoluzionale (CNN) dedicato alla detezione della presenza di queste calcificazioni. È stato inoltre testato un metodo di visualizzazione dei risultati del CNN, ed infine è stata proposta una procedura di valutazione della gravità delle calcificazioni tramite assegnazione di un punteggio automatico. L'architettura del CNN considerata è stata sviluppata precedentemente per la classificazione binaria delle BAC (presenti/assenti). È costituita da 16 layer, di cui i primi 8 allenati tramite transfer learning basato su VGG-16. I restanti layers sono stati allenati in questa tesi tramite un dataset di 1640 immagini; la validazione è stata effettuata con 888 immagini. È stato eseguito il tuning degli iper-parametri al fine di produrre il modello con le performances migliori (BAC-Net), che è stato infine testato con 916 immagini (precisione=0.831, recall=0.68, F1=0.748, ROC AUC=0.95). I risultati di BAC-Net sono stati analizzati tramite visualizzazioni di tipo GradCAM++, producendo delle heatmaps dell'ultimo layer convoluzionale. La posizione e le dimensioni delle BAC sono state correttamente individuate da questo metodo, permettendo di aumentare la fiducia dei radiologi nelle predizioni fornite da BAC-Net. Un dataset ridotto di 112 immagini è stato usato per testare in modo preliminare la procedura di assegnazione di un punteggio basandosi sul tresholding delle heatmaps prodotte in precedenza. Questo metodo ha permesso di estrarre tre punteggi continui correlati all'area (A_{BAC}), all'intensità dei pixels (I_{BAC}) ed alla previsione della lunghezza (L_{BAC}) delle regioni sopra-soglia. La regressione lineare tra la lunghezza delle BAC misurata dai radiologi (l_{BAC}) ed i tre punteggi sopracitati è stata calcolata. A_{BAC} ha mostrato la correlazione più alta con l_{BAC} (R_S=0.90, p-value=6.33e-41). In ultimo, tre punteggi ordinali (A_Q, I_Q, L_Q) con valori da 1 a 4 sono stati estratti da A_{BAC}, I_{BAC}, e L_{BAC} basandosi sulla rispettiva distribuzione in quartili. È stata quindi studiata la correlazione tra A_Q, I_Q e L_Q ed un punteggio ordinale di riferimento l_Q (derivante da l_{BAC}); A_Q ha mostrato l'accuratezza maggiore per la predizione di l_Q (accuratezza=0.53)

Parole chiave: calcificazioni arteriose al seno, stratificazione del rischio cardiovascolare, neural network convoluzionale, intelligenza artificiale spiegabile

Extended summary

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' ages were analyzed and an a-posteriori exclusion criteria was fixed: patients with age<45 were left out 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:

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

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 (Figure 1) were considered for automatic extraction: 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.

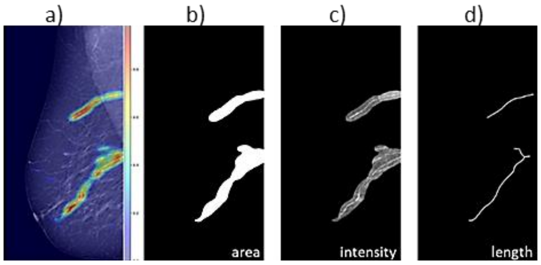


Figure 1. a) Example of GradCAM++; b) thresholding with $T_{heatmap}$=0.5 and A_{BAC} extraction; c) pixels summed to compute I_{BAC}; d) skeletonization to extract L_{BAC}

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 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 fixed. 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.

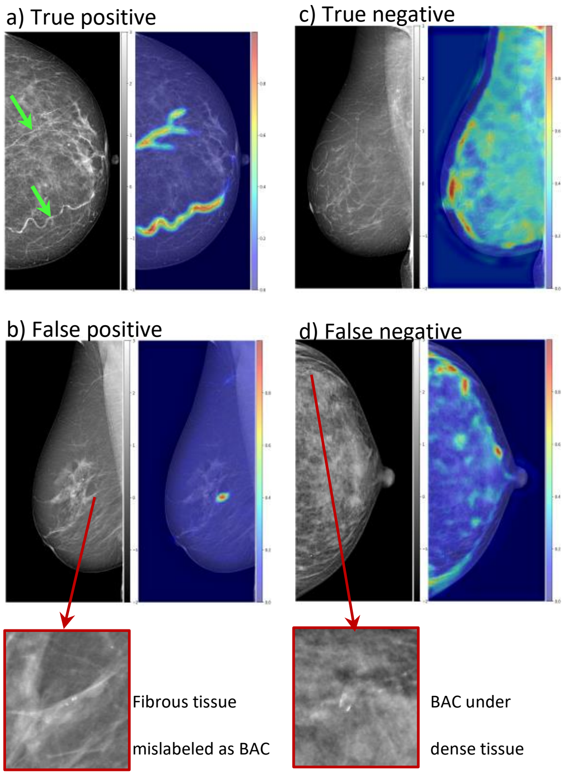


Figure 2. a) TP case of severe BACs correctly identifying the calcified vessels; b) TN case highlighting the whole breast; c) FP case, fibrous tissue mislabeled as BAC; d) FN case, mislabeling is caused by tissue density

Severity scoring

The scoring dataset was composed of 56 BAC+ patients; for each patient, the two mediolateral 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-A}= 0.2, the same value resulted for I_{BAC}, so that T_{opt-I}= 0.2, while for L_{BAC}, T_{opt-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.

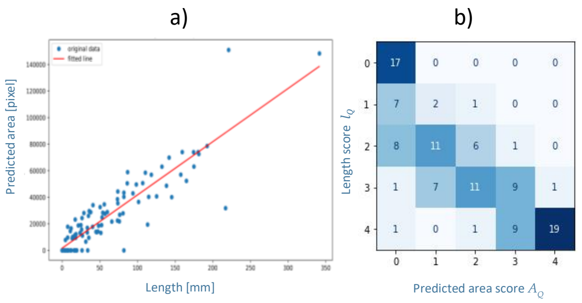


Figure 3. a) Linear regression between real length l_{BAC} and predicted area A_{BAC} ($R_{spearman}$=0.90, p-value=6.33e^{-41}); b) Confusion matrix displaying real length score l_Q on vertical axis, predicted area score A_Q on horizontal axis (accuracy=0.47)

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.

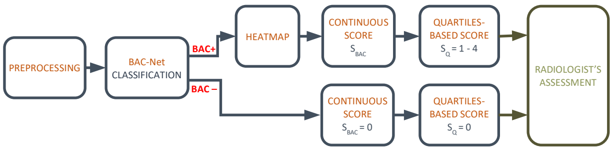

PREPROCESSING
BAC-Net CLASSIFICATION
BAC+
BAC –
HEATMAP
CONTINUOUS SCORE S_{BAC}
QUARTILES-BASED SCORE $S_Q = 1 - 4$
CONTINUOUS SCORE $S_{BAC} = 0$
QUARTILES-BASED SCORE $S_Q = 0$
RADIOLOGIST'S ASSESSMENT

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.

Contents

1 Introduction

1.1. General aim of the study

Breast Arterial Calcifications (BACs) appear in mid-age women with a prevalence slightly above 10% (increased by ageing) and are shown in the mammography of one or both breasts as hyperintense tracts (order of a few centimetres) in one or more arterioles. The search for BACs is foreseen as a valuable screening of cardiovascular (CV) risk, secondary to the primary oncological screening, at zero cost both economic and relevant to RX exposure since mammography is anyway recommended to all the women population, with ageing. Importantly, this population is subject to a high incidence of CV diseases, so, the detection of BAC is believed a precious alert to prescribe further clinical exams addressing the main vascular districts (e.g., coronaries, brain, etc.).

A major problem in BAC detection is the high heterogeneity of the breast images (mainly, size and density) and of the BACs themselves (position, lengths, tortuosity, etc.). Furthermore, similar structures such as laciferous ducts and healthy blood vessels represent consistent confounding factors. Finally, the radiologist inspection is primarily direct to the spotty shapes of small oncological lesions and an additional search specific to BACs would consistently increase the workload.

For the above reasons, an Artificial Intelligence (AI) approach is proposed to assist radiologist in the detection of BACs, also providing maps pointing the BAC (or BACS) localization and quantifying their size.

The first aim of this thesis is the validation of a Convolutional Neural Network (CNN) able to discriminate the presence or absence of Breast Arterial Calcifications (BACs). The architecture of the CNN adopted is the one proposed in a previous thesis by Ienco [1]. In the present progression, a larger population was address, thus permitting to split the dataset into: i) a large training-set; ii) a validation set for a finer tuning of hyperparameters; iii) a test set to quantify performances on independent data. Consistent preliminary work was devoted to the randomized composition of the subsets, which implied a segmentation into age classes to correct (at least partially) the prevalence trend with age.

A further advancement developed by the present thesis was in the direction of AI explainability, with the obvious aim to increase the clinicians' confidence in the automatic predictions. To this purpose, a visualization method highlighting regions of interest found by CNN in the mammograms was proposed. Lastly, such localization maps permitted to develop a protocol to automatically extract BAC quantification parameters relevant to size and intensity. This was preliminary tested on a smaller database annotated with the manual quantifications. This procedure was motivated by the knowledge that BACs intensity and size are correlated with CV risk. This is foreseen as an effective refinement of screening, based on a quantitative threshold rather than a binary detection indicator. Finally, quantitation is believed to play a core role in the prospective follow-up of women undergoing many mammographies through the years.

Overall, this thesis lays the technical bases for future development of a fully automated workflow, able to process of mammograms focusing on BACs: the CNN allows a fast detection of positive cases; this is followed by the visualization of heatmaps highlighting BACs predicted position, and finally by the assessment of a severity score that correlates with CVD risk.

1.2. Digital mammography

Breast Arterial Calcifications detection is done by means of mammography. An analysis of this imaging technique is necessary to understand the challenges of BACs recognition on radiological images. To this end, mammographic technical equipment will be described in this chapter, followed by a depiction of breast and BAC appearance on mammograms.

1.2.1. Digital mammography equipment

Mammography is an X-ray-based two-dimensional breast imaging technique. It is used mainly for breast cancer detection and is widely applied in cancer screening programmes. It presents several challenges and requires highly specialized X-ray equipment, as well as a specific acquisition technique different from other radiological methods.

The major challenge is the need to provide contrast for small high-density calcifications (ranging from 20 to 100 μm) and for ill-defined masses, against a background of mixed densities. The achievement of differentiation between glandular tissue and tumoral mass is possible only at low energies, because of their similar densities [2]. Therefore, low-energy X-rays are employed, with the side benefit of minimization of the dose of radiations delivered to the patient, allowing repeated image acquisitions for screening purposes (the EU recommends one mammography every two years for women over 50 [3]). Breast compression must also be provided correctly by the mammographic equipment. This is required to improve image resolution and homogeneity, as well as to reduce movement-induced blurring. For compression, the breast is placed between a support and a plate while paying great attention to minimize the patient's discomfort.

The digital mammographic X-ray unit is designed to overcome these challenges; it is composed of a specific C-shaped arm adjustable in height and orientation. At one end of the arm, the X-ray tube generates a photon beam with an energy of around 30keV that is directed at the patient’s breast. The ray must pass through the compression plate, the breast, and the breast support before reaching the part of the X-ray unit dedicated to detection (Figure 1.1).

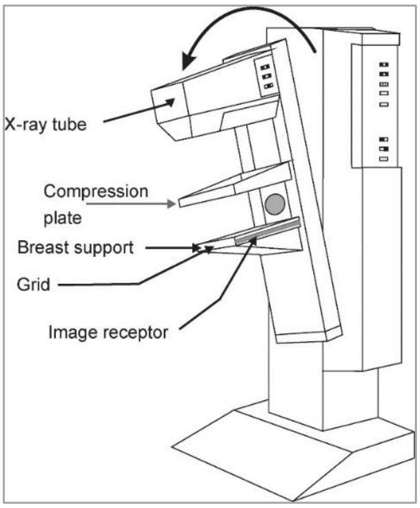


Figure 1.1 Digital mammographic X-ray unit components [http://elektroarsenal.net/x-ray.html]

The image acquisition is performed using scintillators that convert X-rays into visible light and Flat Panel Detectors (FPDs) that are composed of a matrix of light-sensitive elements, each capturing the image intensity related to a single pixel. An anti-scatter collimation grid is also used before the scintillator: more than 40% of the X-rays directed toward the detector may scatter, producing noise in the image if not correctly stopped by the grid [4].

Nowadays, the beam energy, the type of the X-ray beam filtration, and the exposure time are automatically adjusted considering the breast thickness and position over the support. The image acquisition is commonly followed by pre-processing, display, and post-processing to increase the image readability.

During a screening mammography, two views are acquired per patient's breast: craniocaudal (CC) and mediolateral oblique (MLO) (Figure 1.2). Thus, four images are collected for each subject (Figure 1.3).

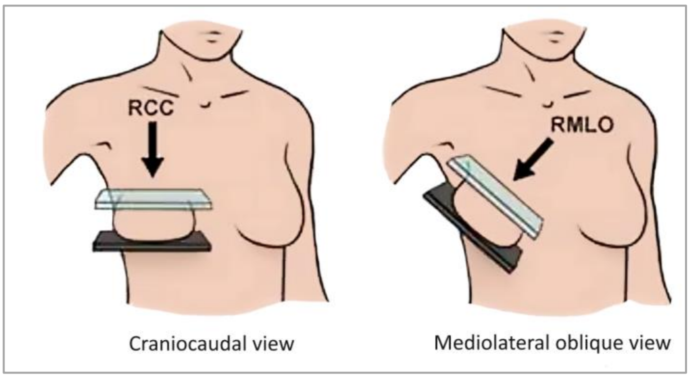


Figure 1.2 Breast support and compression plate position to acquire CC and MLO views of the right breast

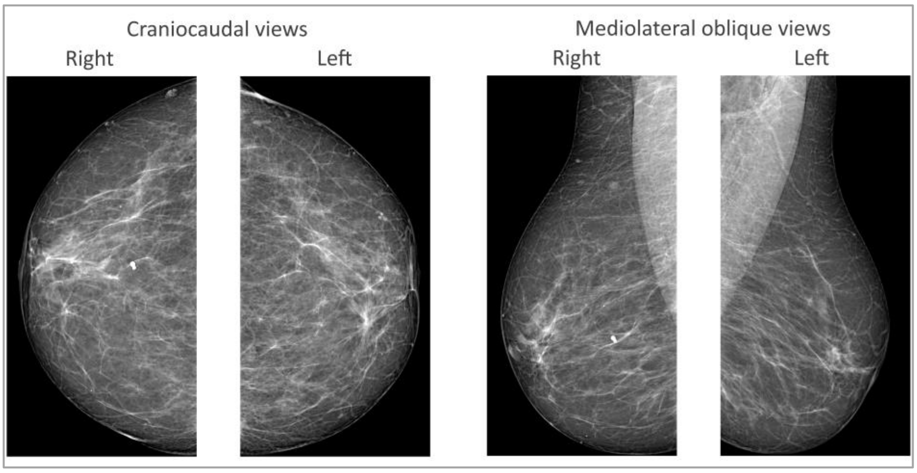


Figure 1.3 Visualization of right and left CC and MLO views

1.2.2. Image characteristics

Pixel sizes on digital mammograms usually range from 50 μm to 100 μm. They usually have a field size of 24x30cm and are shown on a scale of 4096 grey levels, using 12 bits per pixel [5]. Spatial resolution of mammograms is affected by the dimensions of the focal spot of the system. Movements of the patient with respect to the detector, detector's structure, and spatial sampling also have an influence on resolution.

Different noise sources affect mammographic images, compromising their interpretability and ultimately the radiologist's medical report [4]. The production of X-rays and their interaction in the detector follows the Poisson distribution. This phenomenon is reflected as poissonian noise in the image, also known as quantum noise, which is the predominant noise in digital mammograms. Scattering of rays inside the patient's breast, which reduces up to 80-90% by the anti-scatter grid [4], can still decrease the contrast of the images, and add random noise that lowers the signal to noise ratio. Moreover, the electronic noise generated by the instrumentation can be modelled as a gaussian noise. Salt and pepper noise is detected on the images as random black or white pixels caused by sudden fluctuations in signal intensity during the sampling of FPD data. The quality of images is also affected by size, shape, and density of the breast.

1.2.3. Breast and BAC representation in mammograms

Breast is sited on the anterior chest wall and overly the pectoralis major muscle. Its anatomy can be divided into two main parts: a glandular component and a supporting structure. The glandular component consists of 15-20 lobes that radiate from the nipple; each one is made of tens of lobules containing multiple acini, where milk is produced and stored. A network of ducts allows for the milk to reach the surface of the nipple. The supportive structure includes connective tissue, fat tissue, and Cooper's ligaments that allow the connection between the adipose tissue and the skin (Figure 1.4).

A gradual regression of breast tissues, called involution, is noticeable starting from the end of the fourth decade of life causing replacement of connective tissue with adipose tissue [2]. Moreover, progressive lobular atrophy and reduction of glandular components occur; post-menopausal breast might therefore appear entirely fatty, as seen in Figure 1.5.

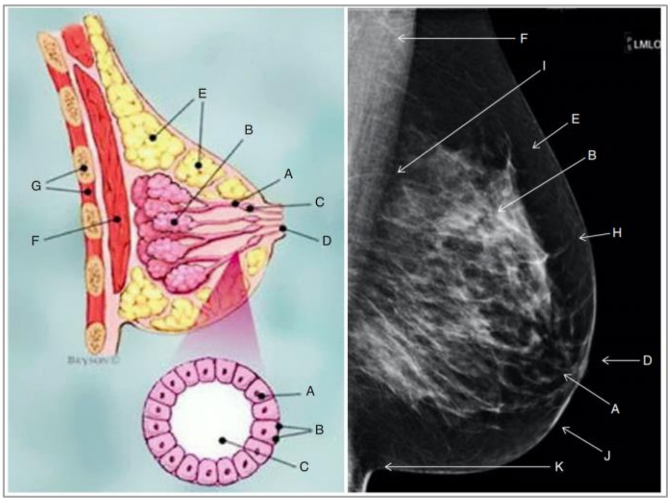


Figure 1.4 A. Lactiferous duct, B. Lobules, C. Cross-section of lactiferous duct, D. Nipple, E. Adipose tissue, F. Pectoralis major muscle, G. Chest wall/ribs, H. Cooper's ligaments, I. Retromammary space, J. Skin, K. Inframammary fold [2]

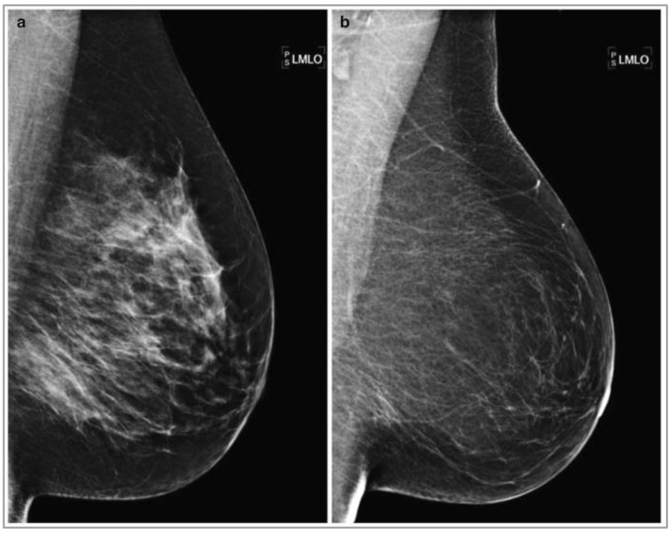


Figure 1.5 (left) mature breast; (right) involuted breast [2]

Breasts that are predominantly fatty appear as low-density since adipose tissue is less dense and consequently more transparent to X-rays. As the prevalence of glandular tissue increases, so does the density of the image. Density is important particularly when considering the sensitivity to detection of tumoral masses or calcifications on mammograms. In fact, X-ray beam attenuation is similar in dense tissue, cancer tissue, and calcifications. Several studies have shown that women with dense breast tissue have higher risk of developing breast cancer because it can be obscured in the dense tissue, and so not diagnosed [6]. Furthermore, there is high variability in the breast vascular system visualization, given both by the variable number of branches of the vessels, and by the 2D projections of the highly complex 3D vascular system. Around 60% of the blood supply to the breast comes from the perforating branches of the internal mammary artery; additional supply is derived from the thoracoacromial artery, the lateral thoracic artery, and the intercostal arteries. Venous drainage is mainly through the axillary vein. When BACs are well recognizable on mammography, they appear as linear and parallel opacities on both sides of a vessel's lumen. However, they can have various representations due to different directions of the vessel's projection. They can either involve only one side of the artery, appear as small intense dots, or be highly fragmented due to variances in calcium deposition. (Figure 1.6).

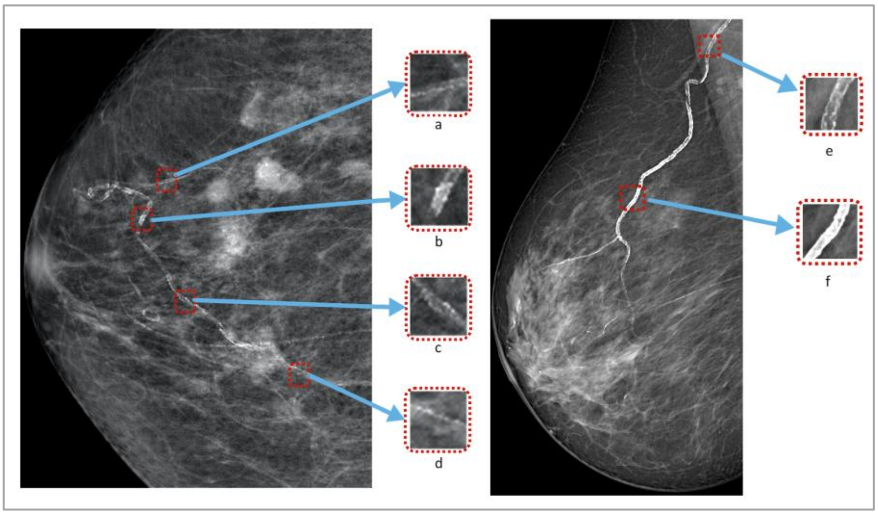


Figure 1.6 a. Mammogram of two patients positive to BACs. a-d) different BACs appearance due to different calcium depositions and to 2D projections, adapted from [7]; e) railroad track appearance of BACSs; f) complete calcification of the artery rendering a full opacification

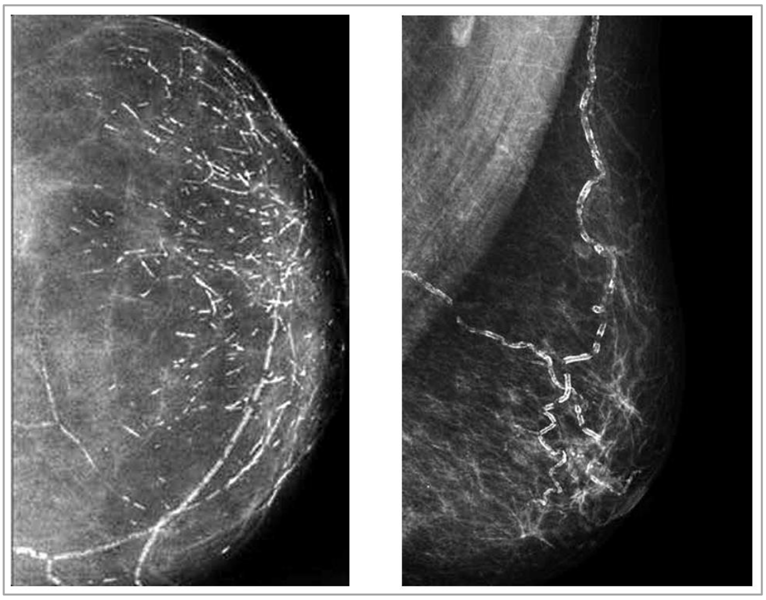


Figure 1.7 Left: linear milk ducts calcifications, right: BACs [8]

The complexity of BACs detection is even higher, considering that BACs are not the only type of calcifications visible on mammographic images. Calcifications may have different distributions inside the breast, different origins, and different shapes. Round or punctate calcifications can originate at lobular level. Skin calcifications also appear to have a round shape, but they are bigger and located on the surface of the breast. Milk ducts can show large linear calcifications, usually bilateral, or thin linear calcifications when they are compromised by carcinoma [8]. Any of these distributions and shapes can have a confounding effect on BACs recognition, but the linear calcifications originated in the milk ducts are more likely to be misclassified as arterial calcifications (Figure 1.7) which are not rare, especially in dense breasts with lower visibility. Additionally, it should be considered that many radiologists are not used to reporting BACs, therefore are not highly trained for their recognition. Automatic detection could make up this difficulty (see chapter 1.6) and reduce the false positives resulting from different types of calcifications misclassified as BACs in medical reports.

1.3. Cardiovascular disease and Breast Arterial Calcifications

Cardiovascular disease (CVD) is a class of diseases involving either the heart or blood vessels. CVDs include a wide range of conditions, such as ischaemic heart disease (IHD), stroke, heart failure, and heart rhythm disturbances. CVDs are the main global death cause, constituting 32% of overall deaths in 2019 according to the World Health Organization. IHD and stroke occupy the first two places in the global cause of death ranking (Figure 1.8*)*, both for men and women, and account for 84% of deaths caused by CVD [9].

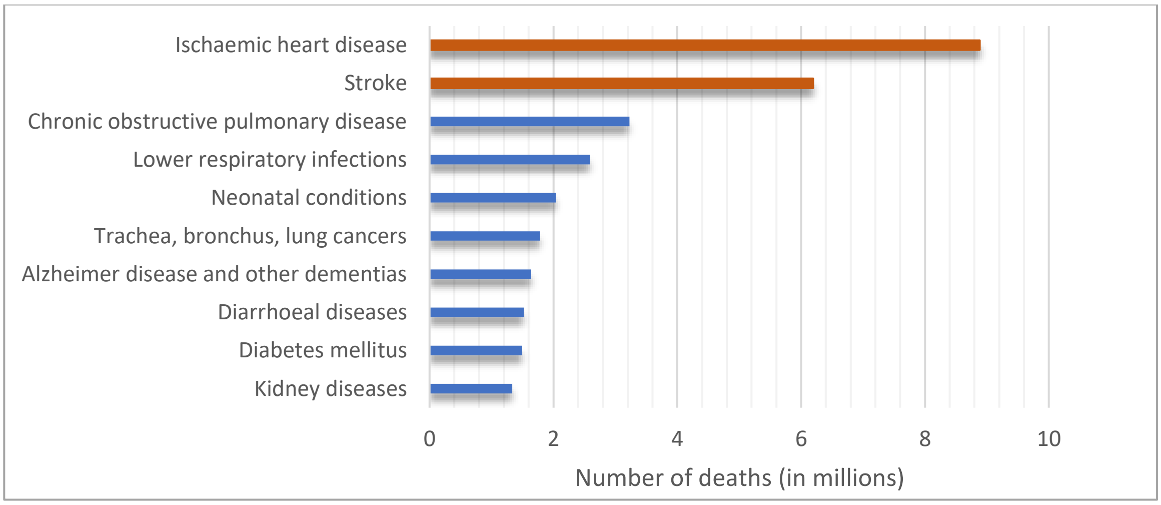


Figure 1.8 IHD and Stroke as the leading causes of death in 2019. Adapted from [9]

CVDs mortality has been decreasing since 1980 within the European Region [10] and in America [11], nonetheless, several studies report signs of stagnation in the reduction of CVD cases [Martinez et al., 2021]. The rate of the decline is historically lower for women than for men [11], especially when considering IHD (Figure 1.9). Therefore, development of strategies to improve the decline of CVDs is a major source of concern in the medical field; screening and risk stratification are required to allow prevention and early treatment.

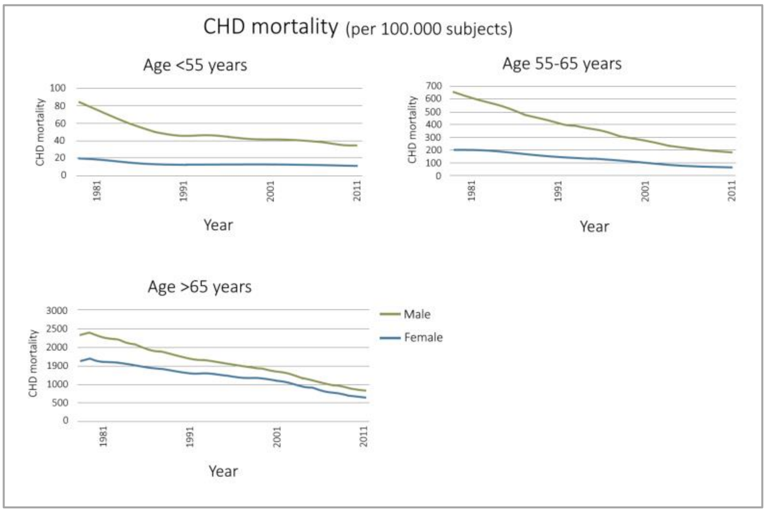


Figure 1.9 Coronary heart disease (CHD, synonym of IHD) mortality per 100.000 subjects, evaluated for male and female patients of different ages. Stagnation in the number of cases since the early 2000' can be noted. Women's mortality decline rate is lower with respect to men's one in all age classes. Adapted from [11]

1.3.1. Cardiovascular Disease: traditional and non-traditional risk factors

Risk stratification is defined as the estimation of the probability of dying from a disease or benefitting from medical treatment. It is used to identify the best suited level of care for each individual, therefore increasing disease prevention and improving patient treatment. Being CVD the leading cause of mortality at the global level, this procedure is of the main importance in this field.

The main traditional CVD risk factors considered for stratification are cholesterol, high systolic blood pressure, cigarette smoking, diabetes mellitus, and adiposity [3]. Typically, these features are collected for each patient and serve as input of algorithms used to estimate the CVD risk. The European Society for Cardiology (ESC) Guidelines on cardiovascular disease prevention suggest the use of the Updated Systematic Coronary Risk Estimation (SCORE2) algorithm. SCORE2 combines CVD mortality and morbidity (non-fatal myocardial infarction or stroke) estimating the 10-year risk of both fatal and non-fatal CVD events based on age, sex, and the main CVD risk factors which are mentioned above [3]. Pooled Cohort Equations (PCE) developed by the American College of Cardiology [12], Framingham risk score (FRS) [13], largely used in clinical practice, and Reynolds Risk Score (RRS) [14] are alternative methods to estimate the 10-year risk of fatal CVD. These algorithms may provide various outcomes but are all based on the same factors considered for SCORE2.

Further research demonstrated that the outlined algorithms predicted the 10-years CVD risk inaccurately. As an example, a prospective epidemiologic study on a multi-ethnic population compared the calibration of PCE, FRS and RRS [15], reporting a high overestimation of the expected risk when compared to the observed one, both for men and women. The authors claimed that this overestimation was mostly due to the incomplete capture of cardiovascular events through traditional risk factors. As a result, it has been suggested that non-traditional risk variables are added to the existing traditional factors. According to a literature review by Van Bussel et al. [16], homocysteine, coronary calcium score (see chapter 1.3.2), patient's frailty, and number of medications used can improve accuracy of risk scoring. The ESC is recommending non-traditional variables for the risk assessment as well; amongst them, female-specific factors such as menopause, pregnancy disorders, and gynaecologic conditions [3], have now been introduced. Despite not being incorporated in the risk predicting algorithms yet, a correlation between these factors and CVD has been proven [17], and they might contribute to the increase in predictions accuracy. Moreover, some studies have focused on the presence of a sex-bias during risk assessment. Abuful et al. claimed the presence of a lower perceived risk for women during diagnosis [18], and Hyun et al. reported that women are less likely to be assessed for CVD risk [19]. Another study by Tabenkin et al. found no difference in patient treatment during risk evaluation [20]. Nonetheless, the three studies agree on the presence of a lower number of pharmacological prescriptions aimed at preventing CVD for women patients. The introduction of sex-specific factors might therefore improve on this aspect, supporting an equal evaluation for men and women.

1.3.2. Coronary Artery Calcification as a risk factor

Along with traditional and non-traditional risk classification based on patient's history and blood samples, European and American guidelines recommend using Coronary Artery Calcification (CAC) score [3], [21] which has been demonstrated to improve clinical risk-assessment. CACs are calcifications of the inner layer (tunica intima) of the coronary arteries resulting from inflammatory atherosclerotic processes (Figure 1.10a). Their detection is based on non-contrast chest computed tomography (CT), currently acquired by means of multidetector CT (MDCT). Even though the CAC scoring is accurate and cost-effective [22], it requires exposure to radiation. State of the art research allows to reach 1 mSy of radiation [22], but the dose can reach up to 7mSy in clinical practice [23]. The high radiation dose makes CAC an unadvisable measure, especially if the intention is to use it for screening. CAC scoring is therefore only recommended when the risk assessment is uncertain [24].

1.3.3. Breast Arterial Calcifications as a risk factor

Mammography is another imaging technique that can improve CVD risk evaluation, while at the same time reduce the sex-bias of predictions. Breast Arterial Calcifications (BACs) are common incidental findings on mammograms. BACs are caused by sheet-like deposit of calcifications in the tunica media of breast arteries, known as Mönckeberg medial calcific sclerosis. This phenomenon causes thickening of the vessels that make them stiffer and less compliant (Figure 1.10b). BACs do not occlude the vessels like coronary artery calcifications, hence are not considered as a risk for the patient; they are usually ignored when observed on mammograms and rarely mentioned instated on the medical report.

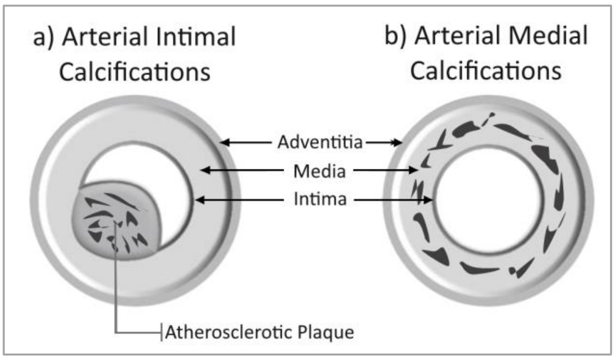

a) Arterial Intimal Calcifications
b) Arterial Medial Calcifications
Adventitia
Media
Intima
Atherosclerotic Plaque

Figure 1.10 a) arterial intimal calcifications that constitutes CACs; b) arterial medial calcifications typical of BACs [25]

In research by Hendriks et al., BACs were detected in 12.7% of women undergoing mammography [26]. The prevalence of BAC highly depends on age (Figure 1.11), ranging from lower than 10% for women under 50 years, to around 50% for women in their 80 [27].

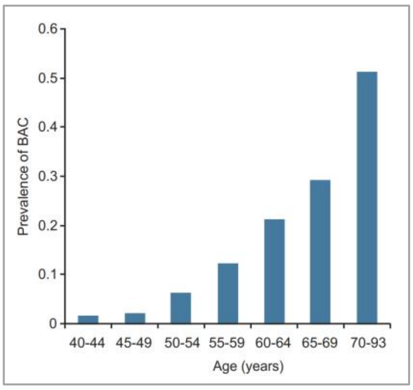


Figure 1.11 Correlation between age and BAC prevalence [27]

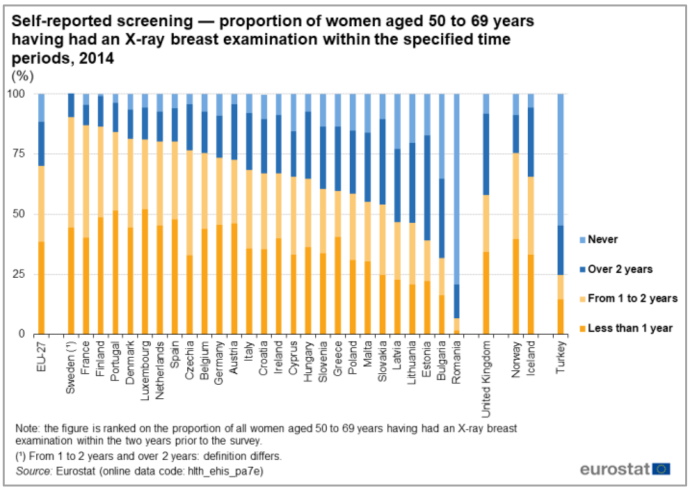
In a recent study, Bui et al. reviewed twenty studies on the correlation between BACs and CVD, finding that most of them confirm this association with an adjusted odds ratio between 0.96 and 8.13. The authors claimed that variability of the results was due to the different way the outcome of CVD was defined, as self-report by patients, CAC score, angiography, or risk stratification by traditional factors [28]. A meta-analysis of 59 studies from Lee et al. [29] highlighted the correlation of BACs with Coronary Artery Disease (CAD), one of the major CVDs. The study reported a summarized odds ratio of 2.61 between BAC and CAD.

The correlation of BACs with the traditional risk factors has also been investigated. As reported by a systematic review of 52 articles [26], diabetes is associated with BAC with an odds ratio of 1.72. Hyperlipidaemia and hypertension were not considered by the authors because of the heterogeneity of different studies. Interestingly, the odd ratio is below 1 for smoking, implying that BACs prevalence is lower amongst smokers; this is justified by the effects of smoking on weight and metabolism and by the survival rate of smokers without BAC after the age of 50. Margolies et al. retrospectively studied the association between BAC (assessed by quantitative scoring from 0 to 12) and CAC score, in women that underwent both mammography and chest CT [30]. BAC scoring was evaluated as a predictor for CAC, resulting in an adjusted odd ratio of 2.3 for mild BAC (with scores 1-3) and of 3.2 for marked BAC (scores 4-12). Moreover, the predicting capability of BAC was assessed by studying the area under the ROC curve (AUC); it's demonstrated that BACs score >0 had an AUC of 0.73 for identification of women with CAC. This result is equivalent to the predicting ability of both the established FSR and PCE algorithms for CAC score.

The proven correlation between CAC and BAC, along with the predictive capability of BAC for CVDs, suggests the possibility to use BACs presence instead of CAC score for women when encountering an uncertainty in risk assessment. This is further motivated by the lower radiation dose needed to perform the analysis: mammography's effective dose is 0.64 mSy [23], ten times lower than the one required for chest CT dedicated to CAC scoring. Moreover, mammography is a routine exam for women with more than 50 years which is employed for breast cancer detection: according to Eurostat, considering European states, in 2014 the share of women between 50 and 69 years that never underwent mammography is only between 5.0% and 10.0% (Figure 1.12) [31]. Similarly, in 2015, 71.3% of United States' women between 50 and 64 years had at least one mammography within the past two years [32]. One of the important issues to consider is the age of screened women, since younger patients are excluded from screening mammographic exams. Nonetheless, women younger than 50 have a BAC prevalence lower than 10% [27]; age is also associated with the increased likelihood of development of other CVD risk factors, and most CVD occurs starting from 50 years [33]. Therefore, under the hypothesis of developing a CVD screening for women using BACs as an indicator, the population group that should be screened is the same as the population group that is already being tested for breast cancer Therefore, BACs analysis would not require any additional radiation exposure or further sanitary expenses. Despite this evidence, a survey amongst the European Society of Breast Imaging members demonstrates that only 60% of radiologists report the presence of BACs, even though 80% of them are conscious of the association between BAC and cardiovascular risk. Moreover, 64.8% of reports are a simple annotation of presence, and the remaining are based on quantitative or semiquantitative notes with different scoring systems [34].

The correlation between BACs and CVDs intensities has also been demonstrated [35], proving the need for the development of a unified system for BACs quantification. Several quantitative and semiquantitative scores have been proposed (see chapter 1.4) but there are still no official guidelines to direct radiologists' annotations.

Increasing the physicians' awareness and number of BAC reports, as well as defining a quantitative classification method are the most challenging aspects of using BACs as a risk factor for CVDs. Another problem is the absence of commercially available software able to assist BACs detection during a possible screening program and to help radiologists throughout the current highly time-consuming scoring procedure. Overcoming these issues will lead to reducing the gender-bias and to increasing risk assessment accuracy, ultimately helping in the battle against CVDs amongst women population.


Self-reported screening — proportion of women aged 50 to 69 years having had an X-ray breast examination within the specified time periods, 2014

Note: the figure is ranked on the proportion of all women aged 50 to 69 years having had an X-ray breast examination within the two years prior to the survey.

(¹) From 1 to 2 years and over 2 years: definition differs.

Source: Eurostat (online data code: hlth_ehis_pa7e)

Figure 1.12 Frequency of mammographic examination for women between 50 and 69 years in EU states. Note how in most states more than 50% of the considered population is examined at most every two years [31]

1.4. Quantitative and semiquantitative methods for BAC classification

Most of the available literature focused on association between BACs and cardiovascular diseases relates the two only by studying the presence or absence of BACs. Further studies investigated the relationship between BACs and CVDs severities, finding a positive correlation [30], [36]. For this reason, quantitative BAC scores could provide a stratification of patients' CVD risk based on mammograms and at the same time, allow deeper study of the phenomenon. Several approaches are presented in different studies, proposing a scoring procedure either quantitative or semiquantitative which are summarized in Table 1.1.

Author	Type of score	Scale	Variables considered
Molloi et al. [37]	Quantitative	Continuous scale 0 to 100 mg of calcium	Densitometry (prediction of BACs calcium mass from BACs pixel value)
Mostafavi et al. [35], Ružičić et al. [36]	Semiquantitative	Score 1-4	Number of vessels Shape (punctuate or coarse BACs)
Margolies et al. [30]	Semiquantitative	Score 0-12	Number of vessels Density of calcifications Length of calcifications
Trimboli et al.[38]	Semiquantitative	Score 0 to (5+Nv)	Number of vessels (Nv) Opacification of the vases Length of calcifications

Table 1.1 Summary of the state-of-the-art BACs scoring methods

In the quantitative study by Molloi et al. [37], the correlation between BACs and the calcium mass of each BAC segment extracted by densitometry was investigated. After the acquisition of the mammogram, a region of Interest (ROI) was manually determined around all the noticeable BACs. In each region, an estimation of the background of pixels showing a calcification was made through a linear interpolation algorithm based on the surrounding pixel values. This estimate is subtracted from the original ROI, yielding an approximation of only the calcium elements (Figure 1.13). The pixel values corresponding to BACs were then summed to give the total calcium mass [37], [39]. This technique was not tested for correlation with CVD risk stratification or CAC scoring but showed a good inter-reader reproducibility (correlation coefficient 0.97 between two observers' measurements).

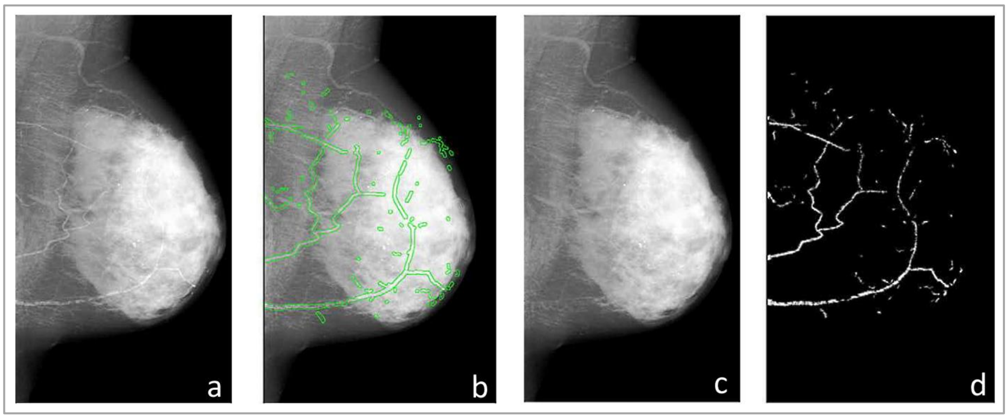


Figure 1.13 a) original mammogram; b) regions of interest around the calcified arteries; c) removal of arteries' pixels, substituted with estimated background; d) subtraction of c from a, showing the calcified arteries. Adapted from [37]

Mostafavi et al. focused on BACs predictive ability for Coronary Artery Disease (CAD) diagnosed with computed tomography angiography (CTA) [35]. The group evaluated mammograms ad CTA scans on a semiquantitative scale from 1 to 4, where 1 corresponds to no BAC, 2 to few punctate BAC, 3 to coarse BAC with tram track appearance in less than three vessels, and 4 to coarse BAC in ≥3 vessels (Figure 1.14).Amongst the patients positive to BACs, 83% resulted positive also to CAD, and moderate to severe BACs (with scores 3 and 4) resulted associated with moderate to severe CAD. The same semiquantitative score has been used by Ružičić et al. [36] to search for a correlation with CAD severity (evaluated through the SYNTAX score [40]), proving that intermediate-to-high SYNTAX group had a significantly higher prevalence of severe BAC.

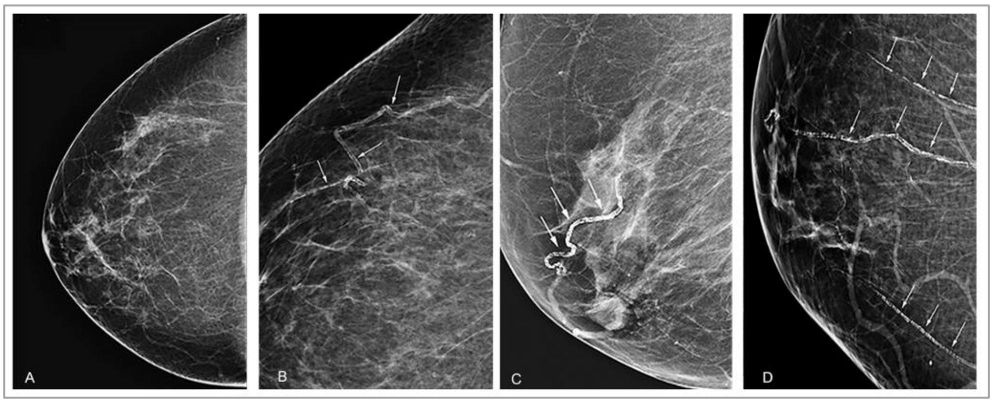


Figure 1.14 BAC scoring according to the studies by Mostafavi et al., and Ružičić et al. A) Grade 1: No vascular calcifications, B) Grade 2: Few punctate vascular calcifications, no tram track or ring calcifications. C) Grade 3: Coarse or tram track calcifications affecting <3 vessels. D) Grade 4: Coarse or tram track calcifications affecting ≥3 vessels [35]

Margolies et al. evaluated the relationship between BAC and coronary artery calcifications (CAC) on non-contrast CT scans, as well as the correlation of BAC score with the Framingham Risk Score (FRS), clinically in use for risk stratification of CVDs [30]. Both BAC and CAC were evaluated on a semiquantitative scale from 0 to 12; for BACs, the number of calcified vessels, the length of the vessels, and the density of calcium deposits were considered to assign the score. The study proved the correlation, showing that in the severe CAC group (score 4 to 12), 56% of patients were also classified as severe BACs. Furthermore, BAC and FRS provided similar results for the identification of women with severe CAC.

Similarly, in a recent study Trimboli et al. [38] defined the BAC Semiquantitative Score (BAC-SS) as:

$$BAC\_SS = Nv + Ov + L \tag{1.1}$$

where Nv refers to the number of calcified vessels and Ov is the vessel opacification (assigned with 0 if the arterial walls are visible, 1 if they are calcified). L is based on the sum of the lengths of all the calcifications in a single mammography, resulting in scores from 1 to 4 based on length quartiles defined over the studied population (score L=0 was assigned in case of BAC absence) (Figure 1.15). The study only tested medio-lateral oblique mammograms and reported a 77% intra-reader and a 64% inter-reader reproducibility, without analysing the possible correlation between the score and CVDs severity.

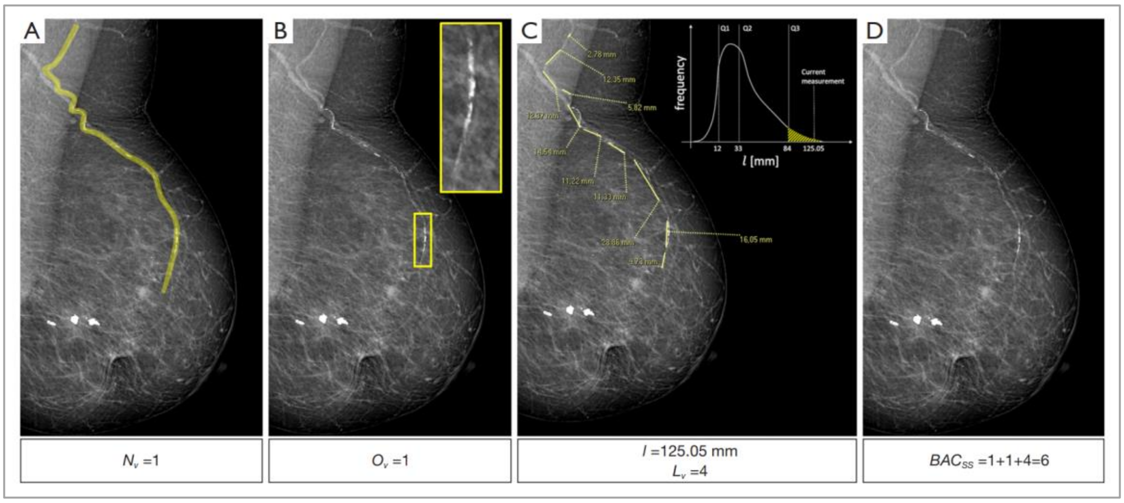


Figure 1.15 Analysis of a MLO view for assigning BAC-SS score: A) mammogram with a single calcified vessel, Nv= 1 ; B) full opacification of the vessel, giving Ov= 1; C) demonstration of the length computation and scoring based on length quartiles, resulting in L= 4 ; D) resulting BAC-SS of 6 [38]

In case a standard BACs scoring procedure was to be fixed, the efficacy of state-of-the-art scores must be assessed by taking into consideration their ability to correctly predict CVD risk; moreover, inter-reader reproducibility of results and time needed to perform the scoring have to be considered. The definition of a ROI for densitometry measurement or the segmentation of BACs is needed for all the proposed scores. These procedures can reduce reproducibility and increase the time needed for the analysis. On the other hand, they are necessary to increase the accuracy of scoring. The proposal of automatic detection and segmentation methods is growing, aiming at reducing the burden of the scoring procedure, as well as increasing its reproducibility. Nonetheless, none of these methods is presently used in clinical practice.

1.5. Importance of Artificial Intelligence and Machine Learning in radiology

Over the years, Artificial Intelligence (AI) and Machine learning (ML) are becoming more and more relevant in the medical field, especially in radiology. The number of scientific publications in radiological applications of AI and ML has increased from about 100 articles in 2016, to more than 900 in 2020 [41]. In 2019, The interest in the topic was also manifested at the European Congress of Radiology (ECR), where AI was reported amongst the top five trends [42]. Commercially available applications are also growing in number. Considering 100 products on the market analysed by Van Leeuwen et al., two-thirds of them were commercialized between 2018 and 2020 [43].

To better understand the reason for such interest in the field, the definitions of Artificial Intelligence and Machine Learning must be understood. Considering a cognitive modelling approach [44], a machine based on AI can be defined as able to "think humanly", mimicking human abilities such as learning or solving problems.

ML is the branch of AI focused on producing algorithms that allow computers to learn from data, improving their performances through experience. When working with images, classic Machine Learning exploits some features that need to be manually extracted, pre-processed, and fed into the model to perform training and ultimately to obtain the desired results [45]. The most recent technique in the ML field employs artificial neural networks (NN) to process data (see chapter 2.1) and is referred to as Deep Learning (DL). The use of NN allows to train algorithms avoiding manual features extraction, obtaining a higher robustness of the results [45]. The main drawback of this technique is the request of higher amount of training data when compared to classical ML methods; on the other hand, DL can be efficiently applied to image-based operations of classification, segmentation, or detection, since NN are able to process a high number of data [46]. For this reason, DL algorithms are especially suited to the development of applications for radiology, where the ability to extrapolate information from images is of the main importance.

Furthermore, advances in medical imaging technologies have allowed for the generation of an increasing amount of data in recent years, which require technological improvements in their analysis. According to a study based in Mayo Clinic, Minnesota, the number of images that each radiologist must interpret per minute during his workday increased four times from 2.9 in 1999 to 16.1 in 2010 [47]. A similar workload changes the image analysis modality for radiologists, shifting it from clinical interpretation to a mere detection task [48].

AI is a tool that can support the radiologists’ work, taking care of time-consuming jobs such as [49]:

- images screening, detecting with high sensitivity the negative studies, leaving the problematic ones for radiologists to inspect
- segmentation of structures of interest
- comparison of current and previous images taken on the same patient, to allow follow-ups

The growth in the production of medical images is also pushing radiology to move from a subjective perceptual skill to an objective science[48]. Reproducibility of radiological results requires a reduction of intra- and inter-reader variability in image interpretation. This can be easily obtained by AI applications that could be used as a unified tool for the extraction of results, ultimately aiming at standardization in radiology.

The development of radiomics is also pushing toward a more quantitative approach for image interpretation. The term radiomics refers to the derivation of a large number of features from medical images, their storage in databases and the subsequent mining of the data for knowledge extraction. The process is focused on improving medical decision-making based on patients’ data [50]. This method heavily relies on ML algorithms that are currently being tested with promising results in the oncological [51], [52] and cardiovascular fields [53], [54], but have not been translated into clinical use yet.

Following the possible applications of AI technology outlined above, some commercial applications have recently been introduced to the market. The Diagnostic Image Analysis Group (DIAG) from the Radboud University Medical Center in the Netherlands has performed a market analysis in 2020, finding 100 AI commercial applications that satisfied the requirement of being approved by both the Food and Drug Administration (FDA) and marked as conform with European health, safety, and environmental protection standards (CE-mark) [43]. Of these, the majority addresses neuroradiology and chest radiology, followed by breast and musculoskeletal radiology. The main tasks executed by the software are quantification, detection, and diagnosis (Figure 1.16). The group found evidence for lack of independent validations of the tools developed, and for absence of a standard model to evaluate the efficacy of AI applications.

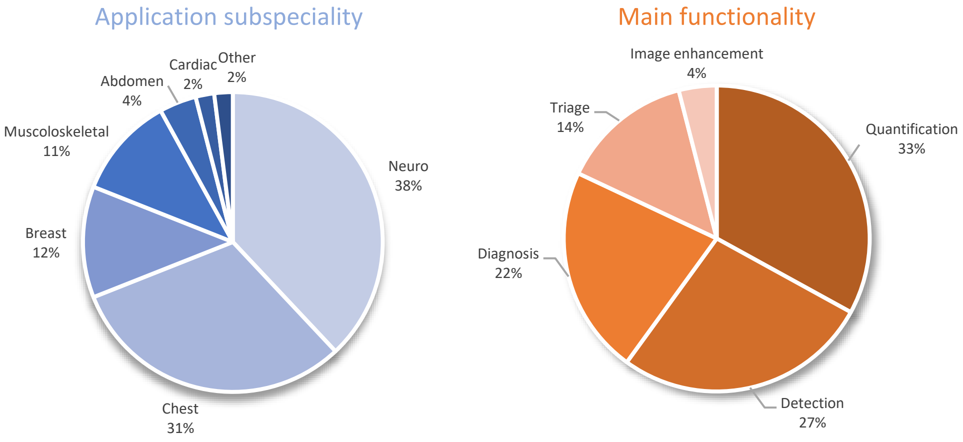


Figure 1.16 AI commercial applications divided by subspecialty (left) and main functionality (right). Adapted from [43]

Medical device regulation is still evolving and just starting to include the notion of AI-based medical software [55]. For this reason, some gaps are still present in the regulatory frameworks. As reported by Larson et al.: examining FDA, International Medical Device Regulators Forum and EU frameworks, the study highlighted a lack of mechanisms to compare similar algorithms, an insufficient characterization of safety and performance elements, and a low number of resources to assess performance at the installation sites [56].

These are symptoms of a market still in its infancy but growing at a fast pace: the DIAG study considered only 100 applications in 2020, but in the past year 80 new products satisfying the same inclusion criteria were added on the group's official website. Without any doubt, AI applications on the market and under research will continue to grow in the future and will become an integral part of radiology. They will help physicians to work in a more quantitative and accurate way, at the same time allowing them to focus more on the clinical interpretation of results and patients' wellbeing.

1.6. Artificial intelligence in BAC detection: state of the art

AI is the new frontier of radiology, and its use is spreading in the field: Machine Learning algorithms and numerous Neural Networks are being developed to facilitate radiologists' tasks (see chapter 1.5). Both BACs detection and severity assessment present some challenges that could be solved with AI: the calcifications are sometimes hidden by dense breast tissue, can present themselves in many different shapes and be confused with other types of breast calcifications (see chapter 1.3.3). Moreover, all the scoring procedures proposed so far for BACs quantification require a time-consuming process for length evaluation or for the definition of region of interest around BACs (see chapter 1.4).

State of the art literature is focused on BACs segmentation, using both Machine Learning (ML) and Deep Neural Networks (DNN) methods. One of the first algorithms dedicated to BAC detection was developed by Cheng et al. It is based on a Machine Learning procedure that is able, when provided with seeding points, to track a calcified vessel. The segmentation algorithm started with the pre-processing of the mammography that allowed to highlight the vessels, generating a vesselness map [57]. A global thresholding of this map produces the seeding points that were used as starting pixels for vessel tracking. The latter was done through the random-walk technique, that generated multiple paths for each seed by searching points with a high value in the vesselness map, within a predefined distance range around the seed. After their definition, paths needed to be further processed and linked to obtain the final vessels segmentation (Figure 1.17). Length and diameters of resulting calcified vessels were compared with the ground truth given by two readers, finding only small differences [7]. Other ML-based methods have been proposed by the same research group [58], but the advancement of deep neural network techniques has provided a more efficient mean to segment BACs.

Wang et al. developed a deep Convolutional Neural Network (CNN) (see chapter 2.1) for BAC segmentation based on a pixel-wise binary classification: for each pixel, a patch of its surroundings is extracted from the image and fed to the network. The CNN's output was a classification of the central pixel of the patch as one if it belonged to the BAC class, as zero otherwise. The operation was repeated for all the pixels in the image, yielding a segmentation of BACs. Results were evaluated by extracting the calcium mass from the pixel classified as positive (through the densitometry technique, see chapter 1.4) and comparing it to the ground truth mass, computed over manual segmentation. Linear regression analysis resulted in only a small deviation from the perfect correlation[59].

U-Nets (see chapter 2.1) have been proposed for BACs segmentation, inspired by their application to many biomedical tasks with good results [60]. lghamdi et al. proposed a modification to the U-net structure, adding a dense layer after each convolutional layer to prevent the model from learning redundant features [61] The resulting network, called DU-Net, has shown better results (F1=92.19) compared to the CNN used by Wang et al. (F1=56.8), and outperformed the segmentation done by two experts with three years of experience. To reduce the computational burden given by U-Nets, Guo et al. introduced the Simple Context U-Net (SCU-Net) [62]. This network has a similar structure to the U-Net, but thanks to input dimensions reduction, its parameters number is two orders of magnitude smaller than the one of U-Net. The input images used in the study are in fact cropped into partially overlapping patches, and each patch is separately processed by the network; the results for all patches are further merged. The segmentation results (F1=0.729) are comparable to U-Nets output (F1=0.735). Moreover, the ability to track BACs over time was proven (Figure 1.18). To study the time evolution, the group evaluated the severity of BACs summing the area of pixels labelled as belonging to a BAC and with an intensity over a fixed threshold, generating the Sum of Mask Area with Threshold x Metric (TAMx score). TAMx scores were then normalized by breast area to overcome differences in breast positioning between scans done in different years. Finally, the normalized TAMx was summed for all the mammographic views of a patient, generating a score able to track the increase of BACs in time. All the described methodologies demonstrate that BAC detection and segmentation is possible with great accuracy, especially through Deep Learning techniques.

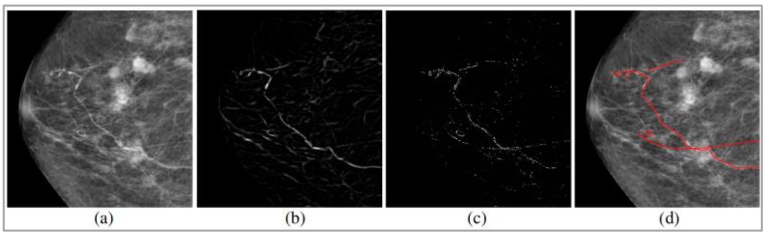


Figure 1.17 Chang et al. ML procedure: a. original image, b. vesselness map, c. seed points found by thresholding, d. segmentation results. Adapted from [7]

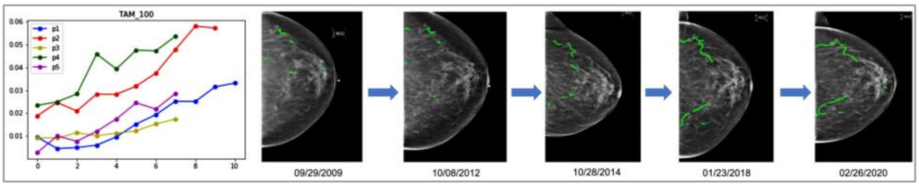


Figure 1.18 BACs longitudinal quantification by TAMx score. Left: TAMx score for five different patients over subsequent years. Right: segmented mammograms for a single patient evaluated from 2009 to 2020 [62]

2 Method

2.1. Neural Networks and Deep Learning

Artificial Neural Networks (ANN, often simplified as NN) are computing systems modelled on biological neural networks. They are one of the many available Machine Learning techniques, and the one that's showing the most promising results in a wide application range. Their structure is based on a fundamental unit called artificial neuron, that was first theorized in 1943 by McCulloch and Pitts [63] and represents the computational model of a neuron. At a biological level, the electrical signal that travels through the neuron's dendrite is perceived by a receiving neuron thanks to synapses, whose efficiency is defined as the ability of the presynaptic input to influence the postsynaptic output. Efficiency depends on the frequency of synaptic activity. The receiving neuron can intercept multiple stimuli, both excitatory and inhibitory, that are integrated inside the soma and compared with an activation threshold. If the sum of stimuli is greater than the threshold, the receiving neuron activates producing an action potential, which in turn will stimulate further neurons (Figure 2.1a). This process is reflected by the artificial neural structure (Figure 2.1b): it combines several inputs, that are weighted modelling synaptic strength with excitation or inhibition properties. If the signal resulting from input integration, called action potential (P), is higher than an activation threshold, it is passed as argument to the activation function, that produces a proportional output.

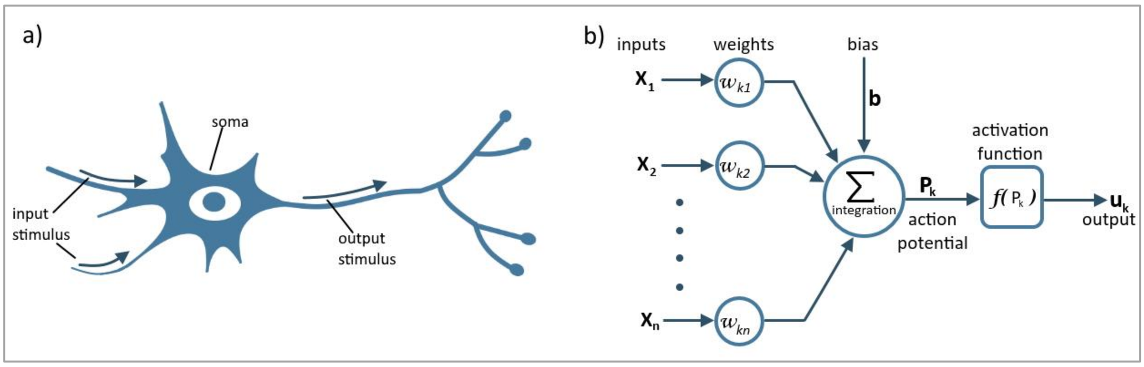


Figure 2.1 a) schematic biological neuron; b) artificial neuron model

In a NN several artificial neurons are arranged in layers: the input layer reads the processed data and produces an output that is fed into the next layer; this procedure, called forward propagation, is repeated up to the output layer, where a score is obtained. The layers between input and output are called hidden layers (Figure 2.2a). The score produced can consist of a binary classification (0 or 1) or it can be more complex, for instance when the desired output is a multilabel classification or an image segmentation. These different outputs can be obtained by modifying the network's structure. Importantly, the operation in a single layer is a linear weighted sum combining the synapse weights with the output of the previous layer. Conversely, the thresholding of activation is a non-linear operation described by the so called "activation function". As a result, NNs are providing a vast set of non-linear I/O functions, if properly trained.

The NN's architecture, especially the number of hidden layers, plays a key role on the quality of the results. The presence of multiple hidden layers allows to have a high flexibility in the interpretation of input signals, and permits to reach high levels of generalization, because each passage from one layer to the next can extract information with superior level of abstraction. A NN that presents a high number of hidden layers (Figure 2.2b) is called Deep Neural Network (DNN).

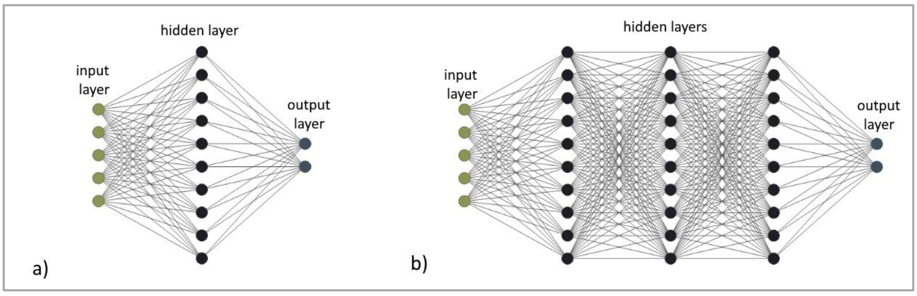


Figure 2.2 a) neural network; b) deep neural network

The majority of DNNs are trained with the supervised learning technique: the dataset to be analysed is labelled by experts in the field of interest; these labels represent the reference value that the network should learn to predict. The learning ability of neural networks resides in the possibility to tune their weights and the neuron's activation thresholds, aiming at minimizing the cost function, that represents the error between the predicted output and the reference. This can be done by backpropagation: the cost function's gradient is transmitted from the output layer to the input one, updating the weights through an optimization algorithm (see chapter 2.1.2).

DNN architectures with different purposes have been proposed in literature. Structures such as illustrated in Figure 2.2b are also called Feed Forward NN (FFNN); if their hidden layers are of convolutional type, they can also be called Convolutional NN; these kinds of structures are generally used for images classification. Recurrent NN present connections between non-consecutive layers, and have been introduced to solve the vanishing gradient problem (see chapter 2.2) often present in CNN and FFNN. Structures such as U-Nets [60] are able to provide outputs of the same dimensions as their input, and are often used for images segmentation (see chapter 1.6); their architecture contains a contracting path (encoder) that allow to extract features from the whole image; its output is fed into an expanding path (decoder) that generates the segmentation; the two paths are linked also by long-skip connections, that connect opposing layers from the contracting to the expanding path (Figure 2.3).

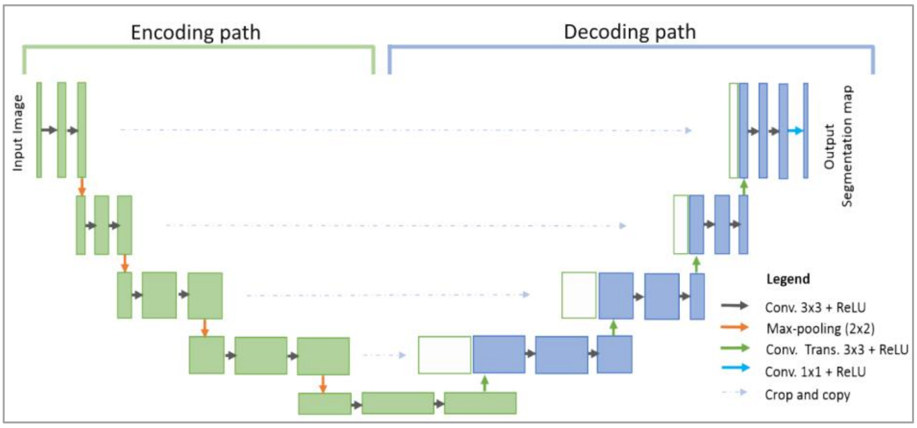


Figure 2.3 U-net structure. The green layers represent the contracting (encoding) path, the blue layers the expanding (decoding) path [60].

All the cited structures are used for supervised learning, and require an annotated training set. DNNs also allows for unsupervised learning, that can be developed for example through Generative Adversarial Networks (GANs) [64], and for solving temporal problems such as natural language processing through Recurrent architectures (RNN) [65].

Since this thesis deals with the interpretation and classification of mammographic images, a CNN based on VGG16 architecture [66] was considered. For this reason, the next paragraph lays the foundations of Convolutional Neural Networks (CNN) architectures, explaining how they are able to process images and extract positional information. Further details about VGG16 and the specific network used in the present work can be found in chapter 3.3.2.

2.1.1. Convolutional Neural Networks

Recognizing an object inside an image is a basic ability of the human brain, but it becomes a challenging computational task when performed by artificial intelligence. Physiologically, each neuron of the visual pathway is sensitive to a small sub-region of the visual field, called receptive field; therefore, neurons act as local filters over the input space. The image that hits the retina is fed to a series of visual cortices, that allow to encode the image's meaning throughout their layers [67]. From one layer to the next, the neuronal population size decreases, while the generalization ability grows because the spatial receptive field of each cell increases. Despite this, spatial correlations present in the image are maintained. Moreover, specific neurons are devoted to encoding different image properties, such as lines orientation, vertices, colours and light intensities [67]. Thanks to this neuronal division of tasks, any visual item is encoded by the brain into a visual feature map, that is generated by stacking several layers, each representative of a specific image characteristic.

Convolutional Neural Networks are deigned to mimic the previously described neuronal organization; they are therefore able to exploit local spatial correlation. CNNs architectures may vary according to the task they are focused on, but they are usually deep neural networks composed of convolutional and subsampling layers (called pooling layers) stacked on top of each other. These modules are always followed by one or more fully connected layer, the last of which provides a prediction of the input image class label (Figure 2.4). In the following paragraphs, an in-depth description of the structure and role of each type of layer can be found.

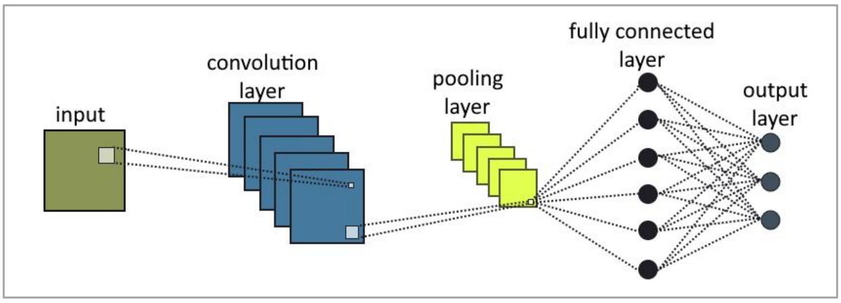


Figure 2.4 Basic convolutional neural network structure. The input image is fed to a single convolution layer composed of multiple feature maps, represented by the blue squares. Subsequently, the output of each feature map is fed to a pooling layer (green squares) that reduces the signal's dimensions. The pooled data are the input of a fully connected layer, whose outcomes are fed to the output layer that provides the wanted results.

Convolutional Layers

Convolution is the main processing step of a CNN, and takes place in layers devoted to this operation, where the layer's input is convolved by a square set of weights smaller than the image's dimensions, called kernel. The kernel slides along the image with a predefined stride, is multiplied point by point with the data, and the results of this multiplication are summed generating a feature map that forms the input of the following layer (Figure 2.5). The same convolutional layer might contain multiple filtering kernels, representing a high dimensional feature space. Importantly, the convolutional structure gains position invariance relevant to object detection or segmentation. Such spatial invariance is also mirrored by the training of convolutional weights, which dramatically reduces the degrees of freedom in CNNs presenting a very high number of synaptic weights to be adapted in the training process.

According to the convolution procedure, each neuron of one layer is connected only to a neighbourhood of neurons in the previous layer thanks to the kernel's weights. The presence of a shared kernel over the whole input image allows features to be detected regardless of their position in the visual field. Moreover, it reduces the number of weights to be trained, decreasing the computational burden.

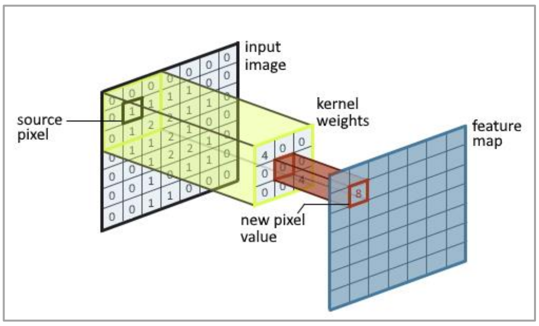


Figure 2.5 Convolution of an input image with a kernel, producing a feature map as output.

Considering a kernel with dimensions LxM, the input of a single neuron in position i,j, is determined by the sum of the input signals $a_{i+l,j+m}$ in its receptive field, weighted by the kernel's values $w_{l,m}$; this input is summed to the neuron's threshold b. The resulting value is the action potential P_{ij}, that is the argument of the neural activation function $f(P_{ij})$:

$$u_{ij} = f\left(b + \sum_{l=1}^{L}\sum_{m=1}^{M} w_{l,m}a_{i+l,j+m}\right) = f(P_{ij}) \tag{2.1}$$

The activation function $f(P_{ij})$ is traditionally either a sigmoid (Figure 2.6a) or a hyperbolic tangent (Figure 2.6b). Recently Rectified Linear Unit (ReLU) activation function [68] became popular due to its simplicity and efficiency (Figure 2.6c). ReLU is expressed as:

$$f(P_{ij}) = \begin{cases} P_{ij} & if\ P_{ij} \geq 0 \\ 0 & if\ P_{ij} < 0 \end{cases} \tag{2.2}$$

It was demonstrated that this activation function leads to fast convergence and avoids the vanishing gradient problem (see chapter 2.2), but when the neuron is not active (Pij=0) the ReLU derivative is equal to zero and this might lead to suboptimal training and slow convergence. Leaky ReLU [69] was introduced to solve this problem. It is a function with a small non-null gradient when the neuron is inactive (Figure 2.6d), defined as:

$$f(P_{ij}) = \begin{cases} P_{ij} & if\ P_{ij} \geq 0 \\ \lambda P_{ij} & if\ P_{ij} < 0 \end{cases} \tag{2.3}$$

with λ between [0,1]. Xu et al. compared the results of a CNN with ReLU and with Leaky ReLU as activation functions, finding that the network performance improves when lambda is big enough [70].

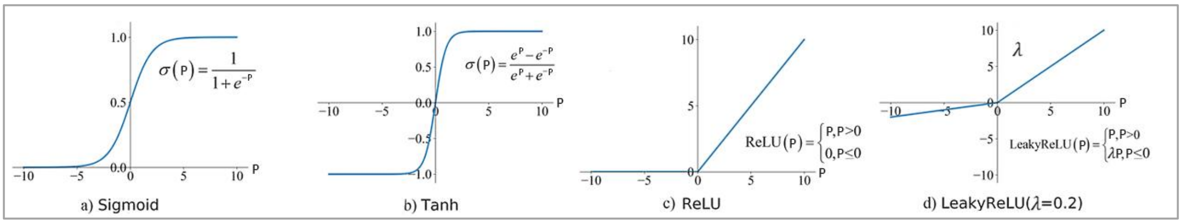


Figure 2.6 Commonly used activation functions (P= neuron's action potential). a) Sigmoid; b) Hyperbolic tangent; c) Rectified Linear Unit; d) Leaky Rectified Linear Unit

Pooling Layers

Convolutional layers are typically followed by pooling layers, that execute a down sampling on the feature map coming from the convolution, producing a new feature map with reduced resolution. This procedure lessens the computational cost of training, and irrelevant details are discarded allowing an analysis of the image's invariant features. It also allows the combination of local features in a larger scale, which eventually leads to object recognition.

There are two strategies to develop a pooling layer: the max-pooling operator applies a window function to the feature map of the previous layer and computes the maximum in the neighbourhood resulting from the windowing (Figure 2.7a); alternatively, average-pooling computes the average value of the set of pixel that results from the window function (Figure 2.7b). Pooling layers do not require training, as their input is not weighted.

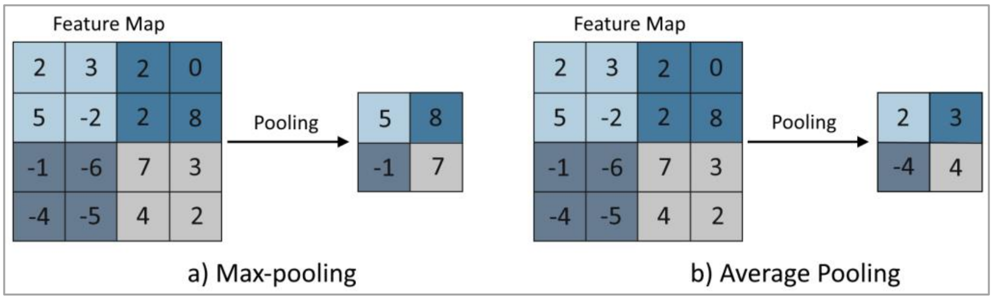


Figure 2.7 a) Max pooling; b) Average pooling [71]

Fully connected and output Layers

Fully connected layers are used after the stack of convolutional and pooling layers to allow the production of an output. Each neuron of a fully connected layer receives as input the weighted output of all the neurons of the previous layer (Figure 2.4). As opposed to convolutional layers, the weights in fully connected layers are numerous, and the spatial features are lost during the processing.

For multi-class classification problems, the activation function of the last fully connected layer is the Soft Max. It maps the output into a vector of values in range (0,1) that add up to one, representing the predicted probability of the input image to belong to each class. If the problem is binary, a single neuron with a sigmoid activation function (Figure 2.6a) is added at the network's output after the fully connected layer. A threshold is fixed to perform a crisp classification. The resulting value will be 0 or 1 according to the predicted class.

2.2. Neural network training

The training of a neural network is based on supervised learning: the desired output is known thanks to the reference that is associated with each input; therefore the objective is to minimize the cost function (that represents the error) between the ground truth and the prediction.

The cost function originally proposed for neural networks was the mean square error, that caused stagnation in learning, influencing the training times. To solve this problem, the cross-entropy (CE) cost function was introduced by Badr et al. [72]. It accelerates the training algorithm, reducing stagnation periods. Considering C classes for the classification, cross entropy of the i-th neuron is expressed as:

$$CE_i = -\sum_{c=1}^{C} t_{c,i} \log\left(u_{c,i}\right) \tag{2.4}$$

Where for each class *c*, t_c is the reference value and $u_{c,i}$ is the prediction. Note that, differently from the notation previously used, neurons are here considered as arranged in vectors and indexed with *i*, but the following equations hold, with slight modifications, also for convolutional layers, where neurons are arranged in 2D matrices with indexes *i,j*.

When the classification problem is binary, the binary cross entropy (BCE) is used, and the previous equation becomes:

$$BCE_i = -t_i \log(u_i) - (1 - t_i) \log(u_i) \tag{2.5}$$

The cost function can be computed only for the last layer, since the output of neurons belonging to hidden layers cannot be compared to any reference. For this reason, the training procedure aims at backpropagating the error from the output layer to the hidden ones, updating the weights according to an optimization procedure. The most basic optimization is the gradient descent, that updates each weight proportionally to the gradient of the BCE cost function with respect to the weight itself.

Prior to backpropagation, the outputs for all the possible N input data, with k ranging from 1 to N, must be computed through forward propagation. Subsequently, for each i-th neuron of the output layer L, the derivative of the cost function BCE is computed with respect to the neuron's output u_i^k and multiplied by both the derivative of the activation function $f(P_i^k)$ with respect to the neuron's action potential P_i^k, and the output of the previous layer's j-th neuron $y_j^{L-1,k}$.

The product $\frac{\partial BCE_i^k}{\partial u_i^k} f'(P_i^k)$ is also called $\delta_i^{L,k}$. This operation is equivalent to the derivative of the cost function of the i-th neuron with respect to the weight w_{ij}^L, computed for the k-th input data:

$$\frac{\partial BCE_i^k}{\partial w_{ij}} = \frac{\partial BCE_i^k}{\partial u_i^k} f'(P_i^k) y_j^{L-1,k} = \delta_i^{L,k} y_j^{L-1,k} \quad (2.6)$$

Finally, for each k, the BCE derivatives with respect to w_{ij} are summed and multiplied by a factor η called learning rate, that modulates the amplitude of the update:

$$\Delta w_{ij}^L = \eta \sum_{k=1}^{N} \delta_i^{L,k} y_j^{L-1,k} \quad (2.7)$$

The next backpropagation step updates the weights of layer L-1; its result is passed to layer L-2 and so on until reaching the input layer. Weight updating for the i-th neuron in the l-th hidden layer follows the equations:

$$\delta_i^{l,k} = \sum_{r=i}^{M_{l+1}} (\delta_r^{l+1,k} w_{ri}^{l+1}) f'(P_i^k) \quad (2.8)$$

$$\Delta w_{ij}^l = \eta \sum_{k=1}^{N} \delta_i^{l,k} y_j^{l-1,k} \quad (2.9)$$

Where $M_{l+1}$is the number of neurons connected to the i-th neuron in the (l+1) hidden layer.

Backpropagation technique has a high computational cost because of the high number of neurons in a NN. Moreover, since DNNs have many hidden layers, and the error’s gradient is smaller in each backpropagation step, the vanishing gradient problem might arise, bringing the layers closer to the input to learn slowly or to have an erratic update of weights that reduces the learning ability. Optimization based on gradient descent also comes with some challenges; the choice of learning rate can be hard, and the minimization can get trapped in local minima or saddle point of the cost function without reaching the global minima. Several alternative optimizations algorithms have thus been proposed, ranging from batches gradient descent, that processes subsets of the training dataset, to stochastic gradient descent that updates weights considering one input data at a time [73].

One of the most used algorithms is the Adaptive Moment Estimation (Adam), a stochastic optimization method that presents an adaptive learning rate for each network weight and adds to the weights update equations data related to the first and second momentum of the gradient of the loss function. This improves the speed of convergence and requires less memory for the computation with respect to gradient descent [74].

During the training procedure, only a part of the whole dataset is fed to the network. This training dataset usually consists of 70% of the whole data, while 15% is used as validation set, to monitor the network results during training. The remaining 15% is the test dataset, used only when the network is fully trained, to test the generalizability of the results on an independent set. The training set is cyclically scanned during the learning procedure: it is typically subdivided into n batches that are used to update the weights n times. One scanning of the whole dataset is called epoch; several epochs are needed to minimize the loss function.

The main reason for the use of a validation set is the need to avoid overfitting, that occurs when the network adapts too much to the training data, so that its generalization ability decreases. Through the validation set is possible to evaluate how each training step affects the generalization: if the results of the monitored metric start to worsen on the validation data, early stopping is performed, interrupting the training procedure and saving the network at the epoch that shows the lowest overfitting. The validation set is also used to optimize core hyperparameters such as the learning rate and the number of training epochs.

2.2.1. Transfer Learning

A major assumption in many machine learning algorithms is that train and test dataset belong to the same feature space, and are both independent and identically distributed (iid) [75]. When the distribution of data changes (for example, due to new data added to the test set), most of the ML models need to be rebuilt.

This could represent a problem when the number of data available is lower than what is required to effectively train the algorithm. Indeed, for deep neural networks the presence of a high number of parameters (weights connecting one layer to the next) requires extremely large datasets to be trained. In literature, during the developing of deep convolutional neural networks, the ImageNet dataset is usually exploited for training: it contains over 14 million images belonging to 1000 classes. E.g., VGG16 [66], AlexNet [76], [77] and ResNet [78] are state-of-the-art convolutional networks based on ImageNet. In many real-world applications, it is highly expensive, if not impossible, to obtain such a huge amount of data. This is especially true when considering problems such as diseases identification from radiographic images: the low prevalence of a disease amongst the population is the first reason for the scarcity of data, along with the time-consuming process of image labelling that has to be done by an expert radiologist. Moreover, experiments are often based on single-centre studies, meaning that the dataset is collected in a single hospital, and is therefore in the order of hundreds of images, an exiguous number when compared to ImageNet.

A specific training technique has been developed to solve the problem of low data availability and to develop networks able to provide a high accuracy even if trained with a small number of images. This solution is inspired by physiological learning: human brain can store knowledge gained while solving one problem and apply it to a related issue that has never been presented to it. A similar process can be followed for CNNs, if the idea that the train and test dataset are iid is relaxed [79] by using transfer learning. This technique is based on the hypothesis that is possible to transfer knowledge from a source domain to a target one with a different distribution, if they share a set of common features. A network can therefore be trained with a big dataset such as ImageNet (source domain) and subsequently adapted, to be used for the study of medical images (target domain) that are not belonging to the 1000 labels that ImageNet has learnt to classify.

The process of modifying the output structure of an already trained CNN and retraining only some of its layers with a target dataset is called fine tuning. The first convolutional layers are dedicated to learning general image characteristics such as borders or lines orientation, consequently what they learn can be applied to any kind of image. During the fine tuning these layers are frozen, so that their weights are not modified. The last convolutional layers and the fully connected output of a CNN adapt to the specific task that is required during the classification; therefore they need to be fine-tuned with data coming from the target domain. Overall, the re-raining of the network requires to optimize a lower number of parameters with respect to the number of weights that had to be learnt while training from scratch, producing good results even if with a small dataset. Several studies have proven that the fine-tuning strategy is effective when dealing with medical images. Amongst them, Tajbakhsh et al. based their CNN on AlexNet, and studied classification, detection and segmentation tasks for three different medical imaging modalities. Their results suggest that fine-tuned CNNs should be preferred to network trained from scratch, especially if the dataset contains less than 100 samples [80].Lastly, a recent study analysed eight different pre-trained networks (VGG16, VGG19, AlexNet and ResNet amongst them) fine-tuned with mammographic images for breast cancer detection. It finds an accuracy of prediction ranging from 82.5% to 94.3%, and compares these results with a network trained from scratch, that gives an accuracy of only 74.2%[81].

2.3. Explainable AI

The explainability of an AI model's results is a problem intrinsic in all artificial intelligence applications: an inverse proportion exists between the predictive accuracy and the explainability of models[82].

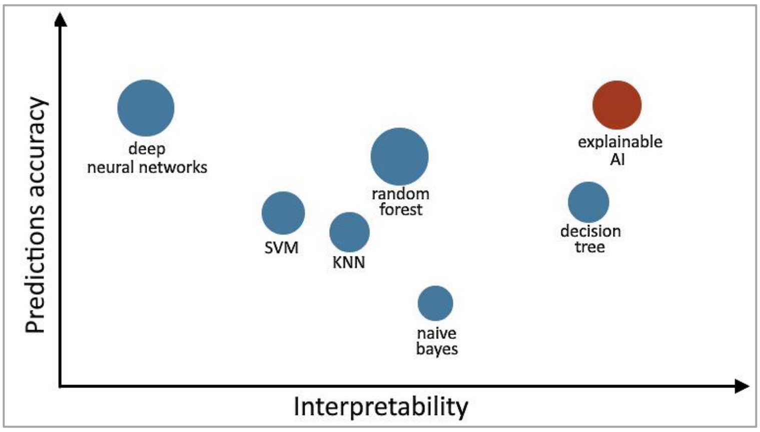


Figure 2.8 Relationship between accuracy and interpretability for several AI techniques. Blue dots represent classical ML techniques The red dot represents explainable AI that presents both high interpretability and prediction accuracy. Adapted from [82]

As can be seen in Figure 2.8, neural networks are the AI technique with the maximum accuracy, although with the minimum explainability. This is justified by the fact that they do not exploit a-priori knowledge about the behaviour of a system: they don't require to define a mathematical model for the extraction of information from data. This approach makes them extremely flexible and simplify their application, but their trained parameters do not have any intelligible meaning. The "reasoning" that allows from the input signal to produce an output is not understandable, and for this reason NNs are defined as "black boxes". This holds for any NN architecture but is especially true for deep neural networks because of their multilayer nonlinear structure, along with the high number of weights.

This lack of explainability and interpretability is problematic in sensitive areas such as criminal justice, autonomous driving [82] and especially healthcare [83]. Lots of medical applications of CNNs showing a great predictive ability have been proposed, but to be applied in the field they must gain the trust of physicians, giving them the possibility to understand why and how a prediction has been made.

Explainable AI (XAI) is a new research branch of AI focused on solving this problem, whose possibility are explored especially for neural networks given their high complexity. It is possible to distinguish two types of explainability [83]: ante-hoc explainability that incorporates the interpretability of results directly in the structure of the NN by adding specific layers [84], and post-hoc explainability that explains what the model predicts only after the generation of the results. The latter allows to exploit traditional NN structures, implementing new algorithms to extract information about the interpretation of results. This procedure is particularly suitable for convolutional neural networks: their input signal is usually an image, to which it is possible to superimpose information about the importance of each pixel for the prediction, that can be extracted from the convolutional layers.

To this aim, several methods have been proposed in literature, and well synthesized in the review by Angelov et al. [82]. In the next paragraph, an in-depth examination of the Saliency maps and Grad-CAM methods can be found.

2.3.1. Saliency maps and Grad-CAM visualizations

Saliency maps [85] were one of the first techniques developed for the visual explanation of CNNs. Convolutional NN are queried about the spatial support of a class c in a specific image I_0 by ranking the pixels of I_0 based on their influence on the score $S_c(I)$ (with I representing any possible input image). This score is the network's output, therefore is a highly non-linear function, but it can be approximated by first order Taylor expansion as:

$$S_c(I) \approx w_c^T I + b_c \tag{2.10}$$

Where w_c is the weights matrix that represents the pixels influence on $S_c(I)$, and can be derived for the image I_0 from the previous equation:

$$w_c = \left.\frac{\partial S_c(I)}{\partial I}\right|_{I_0} \tag{2.11}$$

The saliency map is a representation of w_c's intensities, therefore it renders the partial derivative of the output score for class c, computed for a selected image. The w_c matrix can also be interpreted as an indication of which pixels need to be changed the least to maximally affect the output score: these pixels correspond to the object's location in the image [85]. An example of saliency map is displayed in Figure 2.9b. This visualization has been used by Ienco et al. for the analysis of the output of a CNN for the classification of mammograms based on the presence of breast arterial calcifications.

Saliency maps are usually considered too noisy and other later developed methods are preferred in literature. Amongst them, Grad-CAM (Gradient-weighted Class Activation Mapping) and its modifications are widely applied [86]. Grad-CAM uses the gradient information entering one of the convolutional layers of a CNN to determine the importance of each pixel for the class c. The weights w_c^k are proportional to the gradient of $S_c(I)$ computed respect to the k-th feature maps activations A^k of the selected convolutional layer considered:

$$w_c^k \propto \left.\frac{\partial S_c(I)}{\partial A^k}\right|_{I_0} \qquad (2.12)$$

The partial derivative is computed by backpropagation from the output layer (see chapter 2.2). Once the weights are estimated, the class-specific heatmap to be visualized is expressed as:

$$L_c = ReLU\left(\sum\nolimits_k w_c^k A^k\right) \qquad (2.13)$$

The last convolutional layer is generally the one considered for heatmap generation, as it contains high-level information about the class prediction, as well as precise spatial details (Figure 2.9c). This method has been improved in its localization accuracy and in the ability to localize multiple occurrences of the same class in an image by the Grad-CAM ++ algorithm [87]. Examples of the application of Grad-CAM and Grad-CAM++ can be found in literature for the detection of breast cancer [88], Covid-19 [89], and for the examination of histologic images [90].

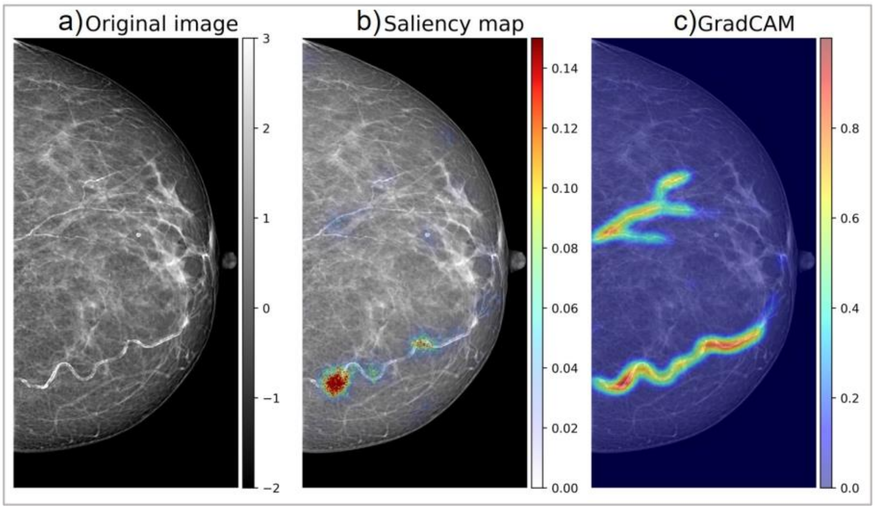


Figure 2.9 a. mammogram, containing one severe BAC (classified as class=1), b. saliency map for class 1, c. Grad-CAM map for class 1

2.4. Heatmaps thresholding for segmentation of medical images

GradCAM and saliency maps results can be expected to highlight image regions corresponding to the position of the element that a neural network aims at classifying. This is due to their mathematical definition (see chapter 2.3.1) and is usually exploited to better analyse the behaviour of the network and the reasons for a given result. However, since these heatmaps highlight an area of the image that the network considers as region of interest (ROI), they can also be used for the extraction of the ROI's contours and for quantitative measures. This could be of interest particularly for radiological applications, where the outline of the object under analysis (denoted as image segmentation), along with interesting measurements as area, mean intensity or length, are extracted manually for the assessment of the pathology under exam.

Neural networks dedicated to segmentation have already been developed with excellent accuracy results (see chapter 1.6), and they fall into the category of supervised segmentation algorithms. Their main drawback is the need for high number of images with pixel-wise annotations about presence or absence of the object of investigation: since segmentation is a time-consuming practice for radiologists, it's often difficult to have access to enough data to perform the training of such networks. On the other hand, networks for the classification of presence or absence of a pathology are easier to train given the higher availability of classified images. From these networks, GradCAMs or saliency maps can be generated and used for the extraction of the ROI without the need of a segmented ground truth. This possibility is referred to as unsupervised segmentation and has been explored in a limited number of studies in the medical field.

Amongst them, a paper by Nunnari et al. [91] compares automatically generated and manually segmented ROIs of a publicly available skin cancer dataset. The automatic extraction is done by thresholding the GradCAM generated both by VGG16 and by RESNET50 (a CNN with lower number of layers with respect to VGG16 [78]): the pixels of the GradCAM with a value higher than a predefined threshold are considered as part of the ROI. Several thresholds (τ) are evaluated, and the quality of the segmentation is assessed with the Jaccard coefficient (J-coefficient) that counts the number of common pixels between manual and automatically generated ROIs. VGG16 has shown better results than RESNET50, and the threshold that produces the highest J-coefficient is τ=0.5. The results of this study prove that the segmentation obtained through this method is possible but has a J-coefficient that is less than half the one obtained with networks dedicated to pixel-level classification of the same database [92]

Guan et al. [93] applied the same approach on chest X-rays but analysing the overlap of bounding boxes enclosing the thresholded ROI with ground truth bounding boxes, therefore delivering only a localization and not a true segmentation. This method proves to be more accurate than other unsupervised localization methods [Wang Xiaosong and Peng, 2019] but, as for the Nunnari et al. study, the results are less accurate than the ones generated by using information about the ground truth bounding boxes during the network's training [95].

In conclusion, unsupervised segmentation based on heatmaps has yet to be developed to its full potential, but even if it's true that this approach might never reach the accuracy of neural networks developed for supervised segmentation purposes, it can be a valid alternative especially for the detection of pathologies that do not require extreme precision in their localization. An example of possible application is BACs detection: what is of the highest importance in this field is the severity of the calcifications, that has been proven to be correlated with cardiovascular disease risk. The exact location and extent of the calcified vessels are less clinically relevant, as can be noticed by analysing the prevalence of semiquantitative over quantitative scores for BACs classification available in literature (see chapter 1.4). Nonetheless, no study has been found in literature about heatmap-based segmentation of BACs; an ad-hoc strategy has therefore been developed in this thesis, along with a proposal for the application of the unsupervised segmentation to the prediction of BACs severity (see chapter 3.6).

3 Protocol

3.1. Protocol Overview

The protocol followed during the development of this thesis aimed at producing an algorithm able to automatically classify mammograms based on the presence and severity of breast arterial calcifications.

First, data were collected and preprocessed (chapter 3.2). This included an anonymization procedure to protect patients' privacy and an image-processing protocol to prepare the images for being analysed by the neural network. The dataset was then split into three subsets, namely the training, validation, and test set. To perform this splitting two factors were taken into consideration: the age distribution of patients in the subsets needed to reflect the original one, since BAC is an age-dependent phenomenon; moreover, considering that patients positive to BACs (BAC+) are a minority in the database compared to the negative ones (BAC-), the data unbalance needed to be tackled to avoid the training of a network too focused on BAC- patients.

The convolutional neural network was developed relying on transfer learning from a network previously built by Ienco et al. for BAC+ classification (see chapter 3.3.2). Considering the presence of unbalanced data, particular attention was paid to the metrics used to evaluate the performances. Traditional measures such as accuracy are biased by the low BAC+ prevalence, therefore precision, recall and F1 metrics were preferred. Tuning of network was performed by manual optimization of the most influential hyperparameters (chapter 3.4), with the aim of improving the CNN prediction ability. Results were evaluated image-wise on the training and validation sets during this phase. The best-performing network (BAC-Net) was used for an independent evaluation of the metrics over the test set (chapter 3.5), allowing to analyse the tradeoff between precision and recall. Moreover, since each patient is associated with four mammograms (two views per side, see chapter 1.2.3), patient-wise metrics were computed as well. Network classification of the test set was also used to generate heatmaps displaying the influence of each image's pixel on the final prediction: this allowed to start to open the "black box" of the CNN in the general framework of AI explainability, facilitating a discussion with the clinical team about the network performances.

Lastly, a preliminary study was conducted to define a method for the automatic classification of the severity of BACs (chapter 3.6). The proposed technique is based on the extraction of geometrical and intensity scores from the heatmaps previously generated. The correlation between these scores and BACs length as measured by a radiologist was finally assessed for a subset of patients.

3.2. Mammographic dataset

3.2.1. Collection, annotation, anonymization

To perform this study, a series of consecutive patients aged 45 years and over were retrospectively enrolled. They underwent mammography for oncological screening purposes at IRCCS Policlinico San Donato between January 2nd and March 14th, 2018. The project was preliminarily approved by the local Ethical Committee (Ethics Committee of IRCCS Ospedale San Raffaele; protocol code SenoRetro, authorized on November 9th, 2017 and updated on July 18th, 2019).

For each patient enrolled, four-view mammograms were acquired using two different full-field digital mammography devices produced by IMS Giotto s.p.a. The labelling procedure involved three readers; two of them worked at patient level, labelling patients as positive to BACs presence (BAC+) or negative to it (BAC-). The third reader performed second level screening in all BAC+ patients, labelling each of the four views. Every image was labelled as positive to BAC if at least one calcified vessel was visible, while women were classified as BAC+ if at least one view was BAC+.

For privacy protection, all patients were pseudo-anonymized by coding their identity; the code was assigned randomly and is known only to one of the clinicians, allowing to perform back tracing of women for clinical purposes. For each patient, single images were labelled using the anonymised code along with an encoding of the view they were representing; this permitted to perform both image-wise and patient-wise analysis of the results after the classification done by the network. If more than four images were acquired per patient, which might occur due to imaging artifacts or mispositioning of the breast, clinical opinion was requested to remove the mammograms with lower quality from the dataset, ensuring to maintain one image per view. All demographic data deemed as non-useful such as birth date, acquisition date, and performing physician’s name, were removed from the files. All labels, along with patients' codes and views’ codes, were collected in a database and served as the ground truth for training and testing the neural network’s models.

3.2.2. Images preprocessing

Mammograms were preprocessed by following two stages, as proposed by Ienco et al. [1]. The whole dataset was prepared to enhance the wanted mammographic characteristics: region of interest (ROI) extraction and pixels intensity normalization were performed. After the dataset splitting, online image number augmentation was implemented only on the training set to reduce overfitting.

Data preparation

An analysis of mammographic images shows histograms characterised by a bimodal distribution of grey levels (Figure 3.1a): the peak with lower values represents background pixels, while the second peak refers to the breast tissues. The automatic generation of a threshold with the Otsu thresholding method [96] allowed to separate these two peaks and to generate a binary image: over-threshold pixels belonging to biological tissues were labelled as 1, under-threshold pixels as 0. Since multiple objects might be identified in each image, the wanted breast ROI was extracted by selecting the object with the biggest area; the bounding box around it was used to crop the image (Figure 3.2b).

The desired image size of 1536x768 pixels was based on the standard measures used as input for the network initialized by Ienco et al. To reach the correct measures, a rigid rescaling of the cropped image was performed (Figure 3.2c): the longest size of the ROI was matched with the corresponding standard dimension, while the other size was scaled maintaining proportions. If the resulting image didn't have the wanted dimensions, it was padded with background pixels (Figure 3.2d). To determine if the breast was on the right or on the left side of the original image its centroid was used. This information was needed to fix the ROI position on the top left or top right of the image before the padding procedure. The resizing has the main consequence of producing higher magnification in smaller breasts; this doesn't represent a confounding factor for the classification through deep convolutional network, but a posterior rescaling was necessary when dealing with BACs severity scoring.

Pixels corresponding to the background of cropped and resized images were fixed to a value of -20. Biological tissues (with values higher than Otsu's threshold) were normalized to obtain a zero-mean distribution, with variance equal to 1 (Figure 3.1b). Each pixel $p(x,y)$ with over-threshold value was normalized as:

$$p(x,y) = \frac{p(x,y) - M}{\sigma}$$

where M is the mean and σ the standard deviation of the over-threshold pixels.

All the preprocessed images were saved in a single Hierarchical Data Format version 5 (HDF5) file, containing information about the anonymization code and of the view code for each mammogram.

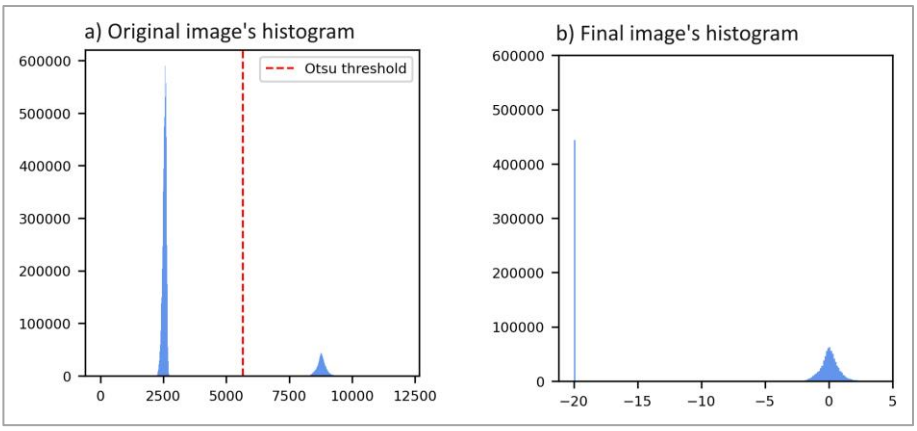


Figure 3.1 a) original image histogram, highlighting the bimodal distribution and the position of the binarization threshold found by Otsu's thresholding; b) final image's histogram, highlighting the background pixels fixed at -20 and the biological tissues normalized around 0.

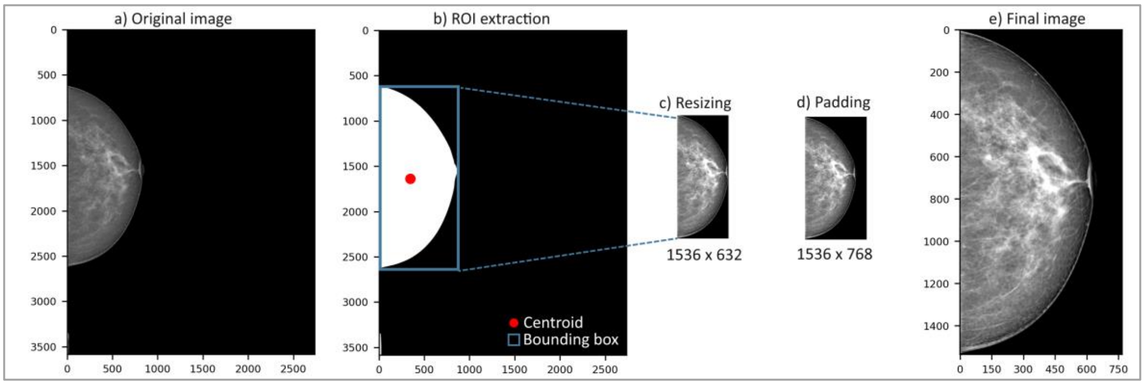


Figure 3.2 Image preprocessing steps. a) original image; b) binarized image, the bounding box around the breast and position of breast's centroid are shown; c) ROI resized by matching the height with the standard dimension (1536 pixels) and scaling the width; d) padding with background pixels to reach the wanted dimensions (1536 x 768 pixels); e) enlargement of the final image, resized and normalized

Image number augmentation

One of the major problems that can be encountered during neural network training is overfitting, that might be caused by lack of data variability. Data augmentation provides a possible solution to this [97]. The training set is augmented by surrogate data derived from the original ones. In imaging, augmentation is obtained by geometric transformations, adding noise, and filtering. In this way, the CNN is trained to fix salient features and, conversely, overlook features randomly related to the image acquisition process. More details about the transformations applied are reported in Table 3.1[1].

Transformation		Details
Geometrical	Vertical flip	Applied with a probability of 50%
	Horizontal flip	
	Zoom	Selected in a uniform distribution between -30% and +5%
	Width shift	Randomly selected in a uniform distribution between $-0.001n$ and $0.001n$ pixels (n = number of columns or rows in the image)
	Height shift	
	Rotation	Randomly selected in a uniform distribution between -3 and +3 degrees, with step equal to 10^{-16}
Noise	Gaussian noise	With probability density function is defined by a mean value randomly selected in the range [0, 0.5] with step equal to 10^{-16} and standard deviation in range [0.01, 0.4]
	Salt and pepper noise	Covering randomly from the 0.01 to 1 % of breast pixels. Salt pixels have random intensity in the range [$Imax$, 1.2$Imax$] (Imax= maximum image intensity). Pepper pixels intensities belong to the range [1.2$Imin$, $Imin$] (Imin= minimum image intensity). The ratio between bright and dark noisy pixels ranges from 0 to 100 %
Filtering	Gaussian filtering	Filter with a randomly selected kernel size in range [3, 7] pixels
	Average filtering	

Table 3.1 Transformations applied during image augmentation [1]

3.2.3. Splitting strategy

A study of normality of age distribution in the dataset was performed through Shapiro-Wilk test [98]. The distribution was found to be not normal, so that for all the subsequent analysis the median was used as measure of central tendency, and the interquartile range as measure of age dispersion [99]. Since no patient BAC+ was found with an age lower than 45 years, and considering that mammographic screening is recommended by the EU for women older than 50 [3], an exclusion criterion was established: patients of age <45 were removed from the dataset, and the final dataset was called truncated dataset.

After the removal of younger patients, data-splitting was performed. The splitting strategy was based on the knowledge that age is a critical factor when dealing with breast arterial calcifications: BAC+ prevalence is correlated with it (see chapter 1.3.3), and mammograms acquired at different ages show several differences, especially in tissue density (see chapter 1.2.3). Age distribution of patients was therefore preserved across subsets created from the splitting. As a first step, the age distribution quartiles of BAC+ population were used to define four age classes. The truncated dataset was split following these classes; median and BAC+ prevalence were computed for all classes. Next, aiming at maintaining the original age distribution, the splitting was performed for each class independently by random patients' extraction. It was decided to include 70% of data in the training subset, 15% in the validation subset and 15% in the test subset, complying with the most used proportion found in literature. The four classes for each subset were then merged, generating the three complete subsets. Median and BAC+ prevalence for each subset were then compared with the original ones for every class.

After the splitting, since BAC+ patients are the minority of the population, the problem of training a network with unbalanced data was considered. Unbalance might lead to biased predictions since it causes the CNN to be more focused on BAC- cases (majority class) and less able to recognize BAC+ (minority class). To solve this problem, two opposite approaches are proposed in literature [100]: oversampling, which requires to apply a larger data augmentation for the minority class, or, conversely, undersampling, based on the reduction of samples in the majority class to balance the data distribution. Maintaining the strategy used by Ienco et al., undersampling was applied to the training subset aimed at rising the BAC+ prevalence from the native 10% to 30% patient-wise, in each age class. Based on a technique inspired by the work of Veni et al. [101]., the BAC+ prevalence in each age group was checked and, if needed, the BAC- class was randomly undersampled to reach the wanted prevalence. Validation and test set were instead kept with the original data unbalance, to reflect the real distribution of BACs amongst patients. It must be noted that the described procedure was based on data analysed per patient, but the dataset is composed of four images per subject that might have different labels (see chapter 3.2.1), therefore the BAC+ prevalence results are different when computed image-wise.

3.3. Convolutional Neural Network architecture

The neural network architecture used to accomplish the classification of BACs is the one developed by Ienco et al. [1], that was built specifically for this task. It relies on transfer learning, starting from the VGG16 network (see Par. 3.3.2 Network structure). The previous study was carried on a smaller dataset, which permitted only a cross-validation, while data were insufficient for the real testing. That work addressed the selection of the best number of initial layers to be frozen (transfer learning) and the later ones to be retrained addressing BACs, the design of the fully connected output layers, and the validation of the optimal hyperparameters. This thesis aims at further training the cited network with a wider database, while fine-tuning some of the hyperparameters. Moreover, results are examined over a test dataset, independent from training and validation sets.

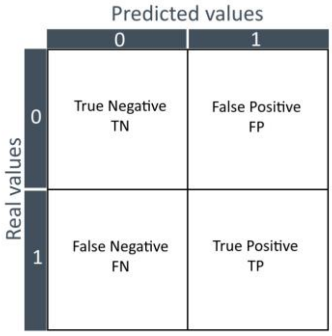
The network was built by using Python 3.8.11 and relies on Tensorflow 2.5.0 library. Graphic processor NVIDIA GEFORCE RTX 3080 (12 GB of memory) was used. Since the whole training dataset was too big to fit in the GPU memory, the code was optimised so that each batch of images could be independently loaded on the GPU after being augmented. This procedure allowed to speed up the training time.

3.3.1. Evaluation metrics

Before proceeding with the description of the network architecture, a remark about the evaluation metrics used from here on is necessary. BACs classification is a binary problem: the network output can be 1, representing the presence of BACs in the mammogram (referred to as BAC+), or 0, indicating the absence of BACs (referred to as BAC-). Classification results can therefore be described by a confusion matrix (Figure 3.3) including:

- True Negatives (TN), mammograms correctly predicted as BAC-;
- True Positives (TP), mammograms correctly predicted as BAC+;
- False Negatives (FN), mammograms incorrectly predicted as BAC-;
- False Positives (FP), mammograms incorrectly predicted as BAC+.

Being the dataset highly unbalanced (BAC+ prevalence per image is about 12% in validation and test set, and 32.7% in the training set, see chapter 4.1.1), the number of TP will always be lower than the one of TN even in case of a perfect prediction. The metrics used to evaluate the network results need to consider this unbalance.



		Predicted values	
		0	1
Real values	0	True Negative TN	False Positive FP
	1	False Negative FN	True Positive TP

Figure 3.3 Structure of a confusion matrix

Accuracy is the most common evaluation metric, and is defined as the number of correct predictions divided by the total number of predictions:

$$Accuracy \ = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

This metrics is misleading when working with an unbalanced dataset: it can be biased by the number of TN, much higher than the one of TP, producing good results even if the minority class (BAC+) prediction is not performing well. To deal with the unbalance, balanced accuracy has been considered, which is defined as:

$$Balanced\ Accuracy \ = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{3.2}$$

Other metrics are particularly suited to tackle data unbalance, such as precision, recall (equivalent to sensitivity) and F1, defined as follows:

$$Precision \ = \frac{TP}{TP + FP} \tag{3.3}$$

$$Recall \ = Sensitivity = \frac{TP}{TP + FN} \tag{3.4}$$

$$F1 \ = \ 2\left(\frac{Precision * Recall}{Precision + Recall}\right) \tag{3.5}$$

Precision represents the percentage of correct positive predictions over the total number of positive predictions. Instead, recall is the percentage of correct positive predictions over the total number of positively labelled images. A tradeoff is often needed between these two measures, and the preference of one or the other is dependent on the application field. F1 represents the harmonic mean of precision and recall, giving equal weights to the two metrics and allowing to consider only one performance metric rather than multiple during comparison of results for different models.

Graphical tools can also be used for performances evaluation; the most common approach is the use of receiver operating characteristic (ROC) curves. They illustrate the network performance when the discrimination threshold of the sigmoidal output neuron is varied, plotting the TP rate (TPR, corresponds to Recall) versus the FP rate (FPR, corresponds to 1-Specificity):

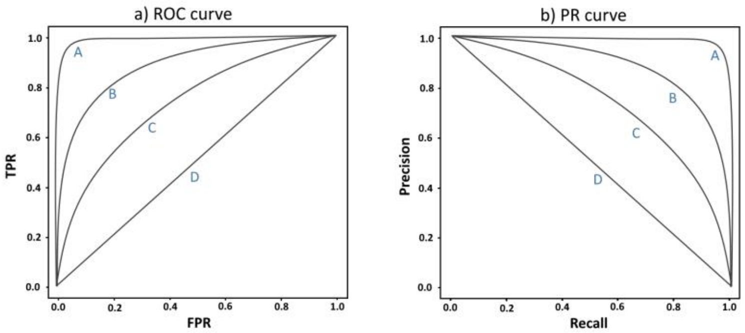
$$TPR = Recall \qquad (3.6)$$

$$Specificity = \frac{TN}{TN + FP} \qquad (3.7)$$

$$FPR \ = \ 1 - Specificity = \frac{FP}{FP + TN} \qquad (3.8)$$

A random classifier performance is plotted as the bisector of the FP rate–TP rate plane. Increases in the classifier's performance are visualized as an increase of the area under the curve (AUC), that would be equal to 1 in case of perfect prediction(Figure 3.4a). ROC AUC is widely used for unbalanced datasets since the effect of the majority class over the AUC result is balanced thanks to the TP rate. A second graphical tool allowing to isolate the results for the minority class is the precision-recall (PR) curve, that evaluates precision and recall while varying the discrimination threshold (Figure 3.4b). This procedure allows to better visualize the trade-off between the two measures. The AUC can be computed for the PR curve as well and, as for the ROC curve, if its value is 1 it represents the perfect network performance.

Overall, balanced accuracy, precision, recall, ROC and PR curves are the metrics considered for the study of classification performances (see chapter 3.4). The tradeoff between precision and recall was evaluated by analysing different discrimination thresholds as shown in chapter 3.5.1.


a) ROC curve

b) PR curve

Figure 3.4 a): ROC curves with increasing AUC; b): PR curves with increasing AUC. In both images, curve D depicts the random prediction (AUC=0.5), curve A the perfect prediction (AUC=1).

3.3.2. Network structure

Ienco et al. CNN is based on VGG16 structure [66], that was designed to process RGB images belonging to 1000 classes of ImageNet dataset. VGG16 has an input shape of (224,224,3) pixels, followed by 13 convolutional layers organised into blocks as depicted in Figure 3.5a. Multiple kernels are used for each layer, allowing the generation of a correspondent number of feature maps. Kernel size is 3x3 pixels, stride and padding are fixed to 1, and ReLU activation function is used for all convolutional layers. After each block of convolutional layers, a pooling layer performs max pooling over a 2 x 2 window, halving the signal's dimensions. The convolutional layers are followed by two fully connected (FC) layers, with ReLU activation function and 4096 neurons. The classification is finally produced by an output layer with SoftMax activation function and 1000 neurons, one per ImageNet class.

This structure was modified to be adapted to BACs classification by Ienco et al (Figure 3.5b). The architecture they proposed is the same used in this thesis, while the hyperparameters of the network were changed according to the tuning strategy (see chapter 3.4).

VGG16 input size was adjusted to one of the preprocessed mammograms (1536x768 pixels); the two fully connected layers were modified, decreasing the number of neurons to 256, setting leaky ReLU as activation (lambda=0.3) and adding a 0.3 dropout rate. A single output neuron characterized by sigmoidal activation was used instead of the SoftMax layer. To allow a binary classification, outputs with values <0.5 were considered as BAC-, while outputs with values ≥0.5 were considered as BAC+. Transfer learning was applied by freezing VGG16 weights for the first 8 convolutional layers, thus retaining the network original ability to detect the basic image characteristics such as borders and their orientation or texture.

The weights of the last 5 convolutional layers, as well as the ones of the FC and output layers, were trained after being initialized with Glorot uniform distribution [102]. A re-weighting method was applied to deal with the unbalance of the dataset, by assigning higher costs to errors made for the minority BAC+ class. Overall, this structure has 13.176.577 trainable and 1.735.488 non-trainable weights (also called network's parameters), for a total of 14.912.065.

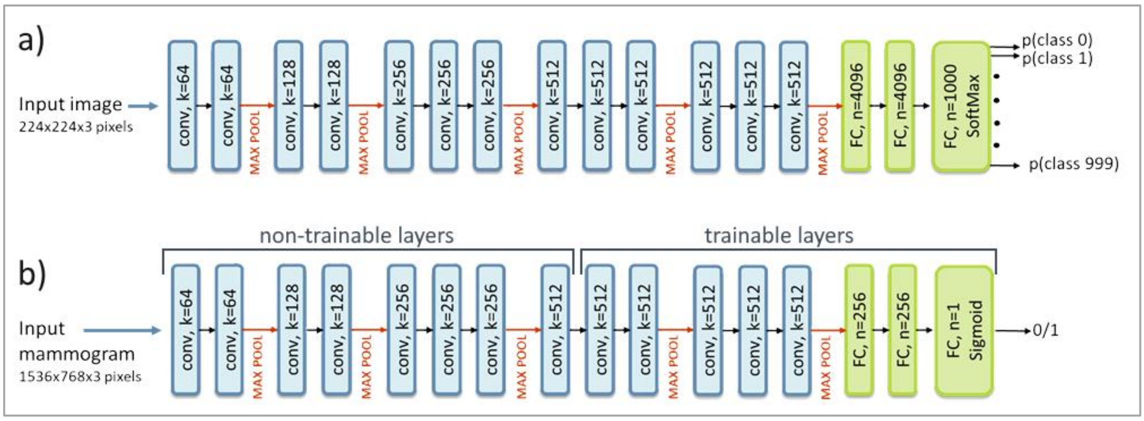


Figure 3.5 a) VGG16 structure; ReLU activation function is used for every layer except for the output layer, where SoftMax is applied; b) Modifications by Ienco et al.; leaky ReLU activation function is used in all layers other than the output layer, where Sigmoid function is used. For both networks, each convolutional layer (conv) has a number k of 3x3 kernels. The max pooling operation (red arrows) halves the images dimensions. The fully connected layers (FC) are composed of n neurons

The network was originally trained for 50 epochs, with a batch size of 8 images. Adam optimizer was used to optimize the binary crossentropy loss function. A learning rate decay was implemented to avoid local minima of the loss function's derivative and to speed up the convergence. Learning rate was computed at each iteration through a Cosine Annealing schedule [103], defined as follows:

$$lr_{eph} = lr_{start} * \frac{cos(\pi * eph/eph_{max}) + 1}{2} \qquad (3.9)$$

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. The value set by Ienco et al. for these parameters are: $lr_{start} = 10^{-5}$ and $eph_{max} = 2 * n°eph = 100$ (where $n°eph$ is the total number of epochs set for training). The network was trained performing 7-fold cross-validation, with a dataset of 992 images. For each cross-validation's fold, 851 images were used for training, 141 for validation. Classification results of the best CNN model produced during cross-validation (from here on called MG-Net) are reported in Table 3.2.

Dataset	Precision	Recall	F1	AUC
Training	0.912	0.881	0.897	0.96
Validation	0.95	0.76	0.84	0.94

Table 3.2 Results of MG-Net as reported by Ienco et al.

3.4. Tuning strategy:

The tuning of the CNN described in chapter 3.5.2 was made by manually searching for the best hyperparameters' values, since both the network and the dataset were too big to allow an automatic grid search on the available hardware. Each network's model was coded with the letter M followed by the model's number; a summary of the networks analysed is reported in Table 3.3.

Model	Initialization	lr_{start}	eph_{max}	n_{eph}	Dropout
M1	**Glorot**	10^-5	100	50	0.3
M2	**MG-Net**	10^-5	100	50	0.3
M3	MG-Net	**10^-4**	200	100	0.3
M4	MG-Net	**10^-5**	200	100	0.3
M5	MG-Net	**10^-6**	200	100	0.3
M6==M5	MG-Net	10^-6	**200**	100	0.3
M7	MG-Net	10^-6	**400**	100	0.3
M8	MG-Net	10^-6	**600**	100	0.3
M9	MG-Net	10^-6	**800**	100	0.3
M10	MG-Net	10^-6	800	**25**	0.3
M11	MG-Net	10^-6	800	**50**	0.3
M12==M9	MG-Net	10^-6	800	**100**	0.3
M13	MG-Net	10^-6	800	**200**	0.3
M14	MG-Net	10^-6	800	**300**	0.3
M15	MG-Net	10^-6	800	25	**0.2**
M16==M10	MG-Net	10^-6	800	25	**0.3**
M17	MG-Net	10^-6	800	25	**0.4**
M18	MG-Net	10^-6	800	25	**0.5**

Table 3.3 . List of studied models. Bold data refer to the parameter under analysis for each model. All networks have been trained with batch size=8, binary crossentropy loss function and Adam optimizer.

Monitored metric

To save the best model produced during training, a metric needs to be chosen for the evaluation of performances over the validation dataset. The metric used by Ienco et al. was the network's loss. In this study, the AUC of PR curve evaluated on the validation dataset was chosen, aiming at maximising BAC+ correct predictions, balancing precision, and recall. A further advantage of using AUC of PR curve is that it is generated by evaluating the results for different thresholds of the sigmoid output (see chapter 3.3.1). This allowed to evaluate the network without fixing the threshold hyperparameter, that was studied in a later step of the hyperparameters analysis.

Network models were saved every time the AUC of the PR curve for the validation set improved, overwriting any older saving, thus reducing the probability of overfitting the training data.

Network initialization

The network is being trained with transfer learning from VGG16. Weights belonging to the first 8 convolutional layers do not need initialization since they are non-trainable, therefore identical to the ones of VGG16 during the whole training. On the other hand, the trainable layers (the last 5 convolutional layers and the fully connected ones) need weights' initialization as a starting point for the training procedure.

As a first attempt, a CNN was trained following Ienco et al.'s method, initializing the trainable weights with Glorot uniform function, that draws samples from a uniform distribution decreasing the probability of vanishing gradient problem (model M1). Moreover, the idea of using MG-Net to perform the initialization of trainable weights was explored in M2. This procedure allows exploiting the knowledge already present in the MG-Net about BACs interpretation, providing the network with a more specialized starting point. All hyperparameters of M1 and M2 were fixed as the ones described by Ienco et al. The model trained with Glorot initialization and the one with MG-Net initialization were compared by analysing the AUC of the PR curve during epochs and the performances on the validation set, allowing to fix the best initialization strategy.

Learning rate

Since learning rate (lr) is considered to be the “most important hyperparameter” [104] in the tuning procedure, it was the first parameter to be explored.

The cosine annealing strategy applied by Ienco et al., also known as stochastic gradient descent with warm restart (SGDR) [103] was maintained, since its efficacy has been proven by literature [105].The initial learning rate $lr_{start}$was explored first, assigning it values of 10^{-n}, with n=[4,5,6] while keeping the eph_{max}fixed at 2*n°eph, to avoid a too small learning rate (models M3 to M5). Values for lr_{start}were chosen according to previous work and based on the typical values used for learning rates in deep neural networks. The number of epochs used for this test was 100, since it allowed to visualize better the network behaviour.

Once fixed lr_{start}, the lr decay rate was explored by changing eph_{max}, assigning values of 200, 400, 600 and 800 (models M6 to M9). Results were evaluated by comparing the metrics described in chapter 3.3.1, computed over the validation set, while assuring to avoid reaching perfect predictions on the training set to avoid overfitting.

Number of epochs

After fixing the initialization and the learning rate decay strategy, the number of training epochs (n_{eph}) was explored; the range considered was between 25 and 300 epochs (model M10 to M14). It must be noticed that n_{eph} is the maximum number of epochs, but the final version of a model may not be the one produced during the last epoch, since the saving of the model is based on AUC of PR curve maximization. Resulting metrics computed over the validation set, along with network overfitting, were evaluated to choose the best number of epochs.

Moreover, the validation set results prior to the application of a classification threshold were examined for models M10 to M14 by histogram visualizations. This allowed to study the potential saturation of the output neuron. Indeed, its activation function is a sigmoid, therefore if its input weights are too high or too low, they might bring the output layer to operate outside the linear sigmoid range, behaving like a binary classifier instead of a continuous one (**Error! Reference source not found.**). This would cause a loss of the ability to discriminate the severity of possible BACs present in the image, and a reduction in the possibility to tune the tradeoff between precision and recall.

A method to avoid saturation is network regularization done by modifying the loss function, as reported in literature[106]. This method was not applied, since it implied a modification of a network that was proven effective on a smaller dataset. On the other hand, it was empirically noted that higher number of epochs led to higher output saturation, therefore to worse results on the validation and test set. This phenomenon was analysed by producing the histograms of output scores assigned to mammograms and considering as an ideal case a uniform distribution between 0 and 1 [106].

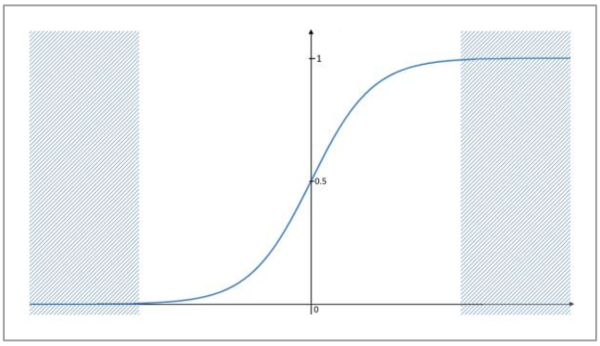


Figure 3.6 Sigmoid function and respective saturations zones (coloured): for any P belonging to these areas, the output can be only a binary classification of 0 or 1.

Dropout value

Dropout was used in the two fully connected (FC) layers of the network; MG-Net was built with a 0.3 dropout rate, meaning that during training, for each weights update, the neurons of the FC layers are turned off (temporarily removed from the network) with a 30% probability, and their weights are not trained. Conversely, when using the model as a predictor, all the network's neurons are active [107], This technique is used to reduce overfitting: the higher the dropout rate, the lower the overfitting probability. However, it must be considered that a high dropout reduces the learning ability of the network to the point where it's not able to acquire new knowledge. To tune the dropout rate, the hyperparameters previously fixed were used, and dropout values between 0.2 and 0.5 (models M15 to M18, Table 4.5) were compared. The choice was based on network results evaluated on validation and training set. Moreover, considering that the behaviour of the CNN with inactive neurons can be studied only during training, the AUC of PR curve was analysed for the training set at each epoch, and the results were compared to confirm the dropout rate choice.

Best performing network

After the hyperparameters analysis, the best performing network was renamed as BAC-Net. BAC-Net was used to perform the testing of results (see chapter 3.5) and for the preliminary study of BACs scoring strategy (see chapter 3.6)

3.5. Evaluation of network performances

After the hyperparameters tuning, the results of the BAC-Net were studied. Firstly, the sigmoid output was analysed without binarizing it with a threshold, to examine the precision-recall tradeoff. This allowed a discussion with the clinicians about which could be the pros and cons of a high precision versus a high recall score, that will be treated in chapter 5. Secondly, since the network results are given image-wise, they were converted into patient-wise classification, to better assess the potential of BACs automatic detection during the production of medical reports. As a last step, several visualization methods were compared and the best one was chosen; an analysis of the resulting heatmaps was performed, paying particular attention to false positives and false negatives results to allow an interpretation of the network behaviour.

3.5.1. Precision-recall tradeoff assessment

The tradeoff between precision and recall was examined on the test set. Firstly, the raw output from the last layer's sigmoidal activation function was computed. A vector of possible classification thresholds was then created with values between 0 and 1 with a step of 0,0016. For each threshold, classification of the test set was performed, allowing to compute precision and recall. The two thresholds maximising precision and recall were extracted and called P-th and R-th respectively. The same procedure was applied to compute the threshold maximising F1, called F1-th.

These evaluations allowed to understand the outputs' behaviour and are valuable especially for clinical use of the network developed. On the other hand, to enable a clearer presentation of the network performances over the test set, and to provide a reference for future works, it was decided to fix an ultimate threshold named τ. Since F1 offers a balance between precision and recall, τ was based on this measure, and calculated as the average of F1-th computed over the test and validation subsets. This procedure was used to avoid over-tuning τ over the test set. The training subset was excluded from the average since BAC+ prevalence is different from the real one in this dataset.

3.5.2. Patients' classification

Patients' classification was performed by following the clinical procedure used to classify a subject during breast arterial calcifications assessment: since BACs detection is used as predictor of risk factor for cardiovascular diseases (CVDs), BACs presence in at least one of the four mammographic views is sufficient to consider the patient as BAC+, and consequently to start an in-depth evaluation of his CVD risk. Practically, after having fixed the final threshold for sigmoidal output binarization, the four binary predictions for each patient were grouped together, and the logic OR between them was computed. Confusion matrix, precision, recall and F1 metrics were assessed, and the results discussed were with the clinicians.

3.5.3. Results visualization

After predicting the class label of mammograms based on BAC-Net, post-hoc explainability methods (see chapter 2.3) were applied and compared. Visual explanation methods were chosen with the expectation that the calcifications' position would be highlighted by them. Visualizations were automatically generated by using the tf-keras-vis library. All the methods proposed in the library were analysed and compared, since no standard for radiologic studies is available in literature.

At first, the strategy introduced by Ienco et al. was tested: saliency maps were generated considering the predicted score of each test image ($S_c(I)$, see chapter 2.3.1) and superimposed to the preprocessed grey-scale mammograms. To obtain better interpretability of the results, the lower gradients of the colour map used were set as transparent. SmoothGrad, GradCAM and GradCAM++ methods were also studied, and the latter was selected as the one producing better results, even when compared with Saliency maps. The GradCAM++ heatmaps of the last convolutional layer were generated by considering the predicted score $S_c(I)$ (see chapter 2.3.1) of images in the test dataset and superimposed to the preprocessed grey-scale mammograms with a transparency of 0.5.

Lastly, GradCAM++ method allowed to study the activations of all convolutional layers, that were computed and displayed in sequence to explore the procedure followed by the network to extract BACs position and compare it to the one used to state their absence.

3.6. BACs severity scoring

The purpose of the work reported in the following chapter is to produce a first assessment of an automatic method to estimate of severity of BACs, after their detection performed by BAC-Net. The aim of such estimation is to evaluate patients' risk of cardiovascular disease (CVD), being it directly correlated with BACs intensity. By thresholding GradCAM++ heatmaps generated by the CNN (see chapter 2.4), it was possible to extract the following scores: area of the calcifications, sum of pixels' intensities, estimation of BACs length. The correlation between these automatic measures and BACs quantification produced by human readers was measured. The latter was taken from the database used for development of the BACs semiquantitative scoring (BAC-SS) [38]. BAC-SS encompasses information about the length of calcified vessels, their number, and the vessels opacification (see chapter 1.4). It was chosen to work only with data regarding the length of BACs, since the opacification had a low weight on the overall score, and the number of calcifications was the measurement showing the lowest inter-reader reproducibility.

3.6.1. Scoring dataset analysis and images preparation

The dataset considered was the one used to produce the BAC-SS score. It is composed of patients positive to BACs and represents a subset of the data used to develop BAC-Net. To reflect the procedure applied by Trimboli et al. [38], only mammograms reporting mediolateral oblique (MLO) views were included in the study, accounting for two images per patient, one for the right view and one for the left. Since a patient is considered positive if at least one of the two MLO views results positive, negative images are present in the dataset as well, despite the inclusion of BAC+ patients only. Measurements of length in millimetres (l_{BAC}) and quartiles-based length score (l_Q) were provided for each image as assessed by two readers. The arithmetic mean between the two readers was computed both for l_{BAC} and l_Q, to mitigate the possible human error and increase the robustness of results. It must be noted that l_Q is defined as follows:

$$l_Q = \begin{cases} 0, & l_{BAC} = 0mm \\ 1, & 0 < l_{BAC} \leq Q_1 \\ 2, & Q_1 < l_{BAC} \leq Q_2 \\ 3, & Q_2 < l_{BAC} \leq Q_3 \\ 4, & Q_3 < l_{BAC} \leq Q_4 \end{cases} \qquad (3.10)$$

Where Q_n with n from 1 to 4 represent the quartiles of BACs length distribution, and l_Q= 0 was assigned to images negative to BACs.

The mammograms belonging to the database were preprocessed following the procedure presented in chapter 3. As already described, to fit the network's input each mammogram was resized maintaining its proportions. Smaller breasts incurred therefore in higher magnification with respect to larger ones; this could represent an error factor during the extraction of automatic scores related to BACs length. In view of the need to rescale the extracted scores, a scaling factor (SF) was computed for each image as follows:

$$SF = \frac{H_{original}}{H_{resized}} \tag{3.11}$$

Where $H_{original}$ is the height of the breast's ROI in the original image and $H_{resized}$ is the height of the breast's ROI after the resizing due to preprocessing. By using SF, any score can be weighted to account for the magnification of the image as follows:

$$score_{scaled} = score * SF \tag{3.12}$$

3.6.2. Network performances on scoring dataset

The database was fed to BAC-Net, and to simulate a real application of the procedure under analysis it was decided to include in the scoring procedure only images classified as BAC+ (both true positive and false positive predictions). Conversely, for negative predictions (true negatives and false negatives), 0 value was assigned to all scores computed, without further processing of the images. The number of BAC+ predictions depends on the classification threshold used on the network sigmoidal output, therefore an evaluation of the tradeoff between precision and recall was performed as described in chapter 3.5.1, generating F1-th, R-th and P-th classification threshold. BAC-Net's results were evaluated for the three thresholds. Severity scores computation was firstly carried on by using P-th as a classification threshold, allowing to include in the positive predictions only images that are classified as BAC+ with high confidence by the network. The procedure was then repeated for F1-th to seek a balance between precision and recall, and for R-th to include the maximum number possible of images predicted as BAC+.

3.6.3. Heatmaps thresholding and scores computation

For each image predicted as BAC+, the GradCAM++ heatmap was thresholded with a binarizing threshold ($T_{heatmap}$). $T_{heatmap}$ values were selected ranging from 0.1 to 0.9 with 0.1 step. Nine binary masks were therefore generated for each image's heatmap (Figure 3.7), with value 1 for over-threshold pixels and 0 for under-threshold ones.

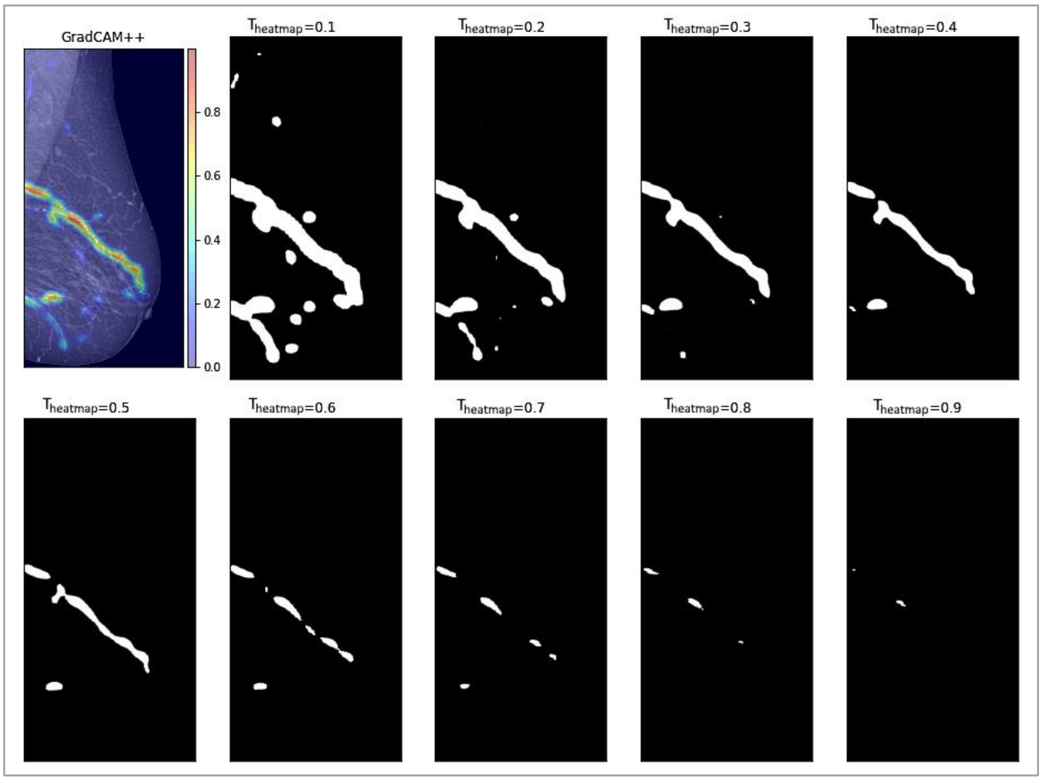


Figure 3.7 Binarization of GradCAM++ heatmap with the application of nine different $T_{heatmap}$. Over-threshold values are considered as 1, under-threshold as 0.

As described in chapter 2.4, pixels labelled as 1 belong to regions where a BAC was detected by the CNN. For this reason, for each $T_{heatmap}$, the area (a) was computed as sum of the over-threshold pixels (Figure 3.8a). Subsequently, the binary mask was multiplied by the original mammogram, allowing to assign a value 0 to all the pixels outside the area of interest. The sum of pixels intensities (i) of the resulting image was computed (Figure 3.8b). Lastly, to extract an estimation of the sum of BACs length (l), skeletonization was applied to the masks. It allowed to iteratively remove from each BAC object its border pixels, on the condition that they didn't break the object's connectivity. Ultimately, a 1-pixel-wide representation was obtained (Figure 3.8c); the number of resulting pixels was computed for all BACs present in the image to extract the predicted length l.

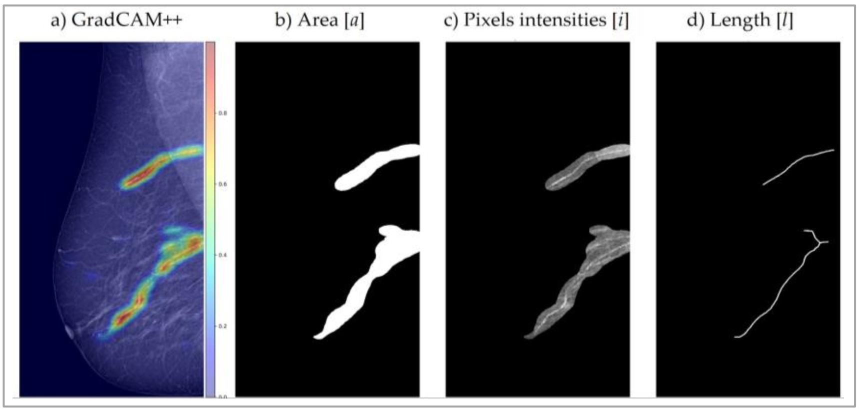


Figure 3.8 a) GradCAM++ heatmap; b) binary mask used to extract the area a; c) original mammogram's pixels belonging to the extracted area, i is the sum of their intensities; d) skeletonization of the extracted area, used to estimate overall BACs length l

Measurements directly extracted from the thresholding of GradCAM++ were further multiplied by SF to account for image scaling, producing the three severity scores ultimately examined: A_{BAC}, I_{BAC} and L_{BAC} respectively.

$$A_{BAC} = a * SF \tag{3.13}$$

$$I_{BAC} = i * SF \tag{3.14}$$

$$L_{BAC} = l * SF \tag{3.15}$$

Moreover, quartiles-based scores were defined by following the same procedure used by clinicians to produce l_Q: the quartiles of A_{BAC}, I_{BAC} and L_{BAC} were computed, and used as thresholds to generate values ranging from 1 to 4; value 0 was assigned to BAC- image. The resulting scores are referred to as A_Q, I_Q and L_Q respectively. It must be noticed that for each image, the scores in Table 3.4. were computed for all the considered $T_{heatmap}$. During the correlation assessment, an analysis of the best heatmap's threshold for each severity score was performed.

score	details	scale
A_{BAC}	Number of pixels over $T_{heatmap}$, scaled by SF	Continuous
I_{BAC}	Sum of intensities f pixels over $T_{heatmap}$, scaled by SF	
L_{BAC}	Number of pixels after skeletonization, scaled by SF	
A_Q	Area score based on A_{BAC} quartiles	0 to 4
I_Q	Pixels' intensity score based on I_{BAC} quartiles	
L_Q	Length score based on L_{BAC} quartiles	

Table 3.4 Summary of the considered severity scores. Note that value of 0 is assigned to all scores when an image is predicted as BAC-

3.6.4. Correlation assessment with gold standards

The assessment of bivariate normality of data was performed through Henze-Zirkler test. This allowed determining whether to use Pearson's or Spearman's correlation coefficient for the computation of the correlation measure [108].

Subsequently, the length of BACs measured in mm (l_{BAC}), considered as gold standard, was compared with A_{BAC}, I_{BAC} and L_{BAC} through linear regression. Regression and correlation measure were computed for each $T_{heatmap}$ and the optimal binarization threshold value was considered as the one maximising correlation. Optimal thresholds for area, intensity and length are indicated respectively as T_{opt-A}, T_{opt-I} and T_{opt-L}.

The quartiles-based length l_Q (gold standard) 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.

4 Results

4.1. Dataset

Mammograms from 1557 female patients were analysed. During each mammography, four images were collected (one per view), for a total of 6228 images. The device for the acquisition of mammograms was GIOTTO 3DL for 80% of patients and GIOTTO TOMO for the remaining 20%. After labelling, 194 patients resulted positive to BACs, representing the 12.46% of the population. Patients' age ranged from 33 to 87 years, and no patient positive to BACs was found with an age lower than 45 years (Figure 4.1). By following the exclusion criterion fixed, 64 patients with age lower than 45 years and negative to BACs were removed from the complete dataset generating the truncated dataset, including 1493 patients. The final BAC+ prevalence was 14.93% (Table 4.1).

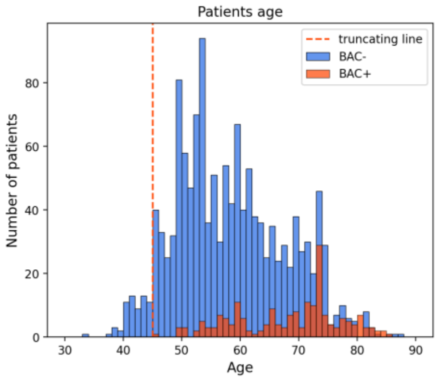


Figure 4.1 Histogram of patient's age. Red bar refers to patients positive to BACs, blue bar to negative patients. The dashed line is at 45 years, all patients below this age were excluded from the dataset. Note that no BAC+ patient is below the truncating line.

	Complete dataset	Truncated dataset
N° of patients	1557	1493
BAC+	194 (12.46%)	194 (14.93%)
BAC-	1363	1299

Table 4.1 Number of patients and BAC+ prevalence for the original and for the truncated datasets

Shapiro-Wilk test was performed on the truncated dataset, resulting in W= 0.96, p-value< 0.01, confirming that patients' ages follow a non-normal distribution. Median age is 59 years, with an interquartile range of 52-68 years.

Quartiles of the BAC+ distribution were computed, and used as age classes named as follows:

- Class 1: corresponding to the first quartile, age 45-60;
- Class 2: corresponding to the second quartile, age 61-70;
- Class 3: corresponding to the third quartile, age 71-73;
- Class 4: corresponding to the fourth quartile, age 74-87.

The truncated dataset was split into these classes to assess the BAC+ prevalence in each of them, confirming that it is correlated with age (Table 4.4).

4.1.1. Training, validation and test sets

Data splitting applied to each age class resulted in three datasets: the training subset with 1042 patients, of which 908 negatives and 134 positives to BACs (12.85% prevalence); the validation subset, containing 222 patients, of which 194 BAC- and 28 BAC+ (12.61%); lastly the test set, with 229 patients, 197 BAC- and 32 BAC+ (13.9%).

The undersampling of training set aimed at reaching around 30% prevalence in each age class. Since Class 3 (71-73 years) and Class 4 (74-87 years) were already characterized by such percentage of BAC+ patients, undersampling was performed only for Class 1 and Class 2. This resulted in randomly removing 474 BAC- patients from Class 1 and 158 BAC- patients from Class 2. The final training dataset is therefore composed of 410 patients, of which 276 BAC- and 134 BAC+ (32.68% BAC+ prevalence). The number of patients and images for each dataset is summarized in Table 4.2 and Table 4.3 respectively.

	Training set	Validation set	Test set
N° of patients	410	222	229
BAC+	134 (32.68%)	28 (12.61%)	32 (13.97%)
BAC-	276	194	197

Table 4.2 Number of patients per subset

	Training set	Validation set	Test set
N° of images	1640	888	916
BAC+	398 (24.27%)	89 (10.03%)	94 (10.26%)
BAC-	1242	799	822

Table 4.3 Number of images per subset. Note that four images per patient are present in the dataset, and that a subject is considered BAC+ if at least one image out of four is labelled as BAC+.

Considering the BAC- undersampling and the patient left out from the study due to inclusion criteria, the original 1557 patients dataset was reduced to 861 patients. For each subset, the histograms of age distribution for BAC+ and BAC- patients are represented in Figure 4.2, while a comparison of the median and BAC+ prevalence of the age classes is reported in Table 4.4. The prevalence of BAC+ for validation and test set was found to be in line with the original dataset prevalence for each quartile.

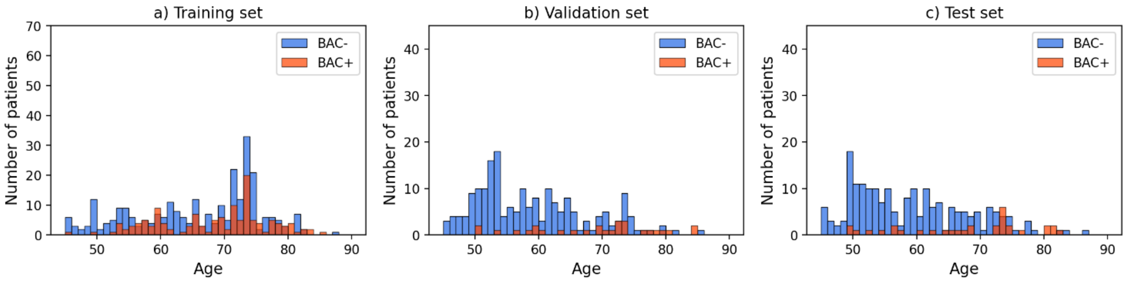


Figure 4.2 Histograms of patient's ages for a) training set, b) validation set, c) test set

	Class 1 (45-60y)		**Class 2** (61-70y)		**Class 3** (71-73y)		**Class 4** (74-87y)	
	Median	BAC+ %	Median	BAC+ %	Median	BAC+ %	Median	BAC+ %
Truncated dataset	53	6.3%	65	11.6%	73	34.3%	77	38.21%
Training	54	30.8%	65	30%	73	34.3%	77	37.64%
Validation	53	6.3%	64	11%	73	33.3%	77.5	38.9%
Test	52	6.9%	65	12.3%	72	34.8%	77	40%

Table 4.4 Median and BAC+ prevalence (BAC+ %) of each age class for the truncated dataset, compared with the training, validation, and test subsets. Note that the training subset has been undersampled aiming at about 30% of BAC+ patients in each class, while the validation and test subsets are in line with the original population

4.2. Tuning of CNN parameters

The following paragraph reports an evaluation of the models listed in Table 3.3 (see chapter 3.4), aimed at determining the best set of hyperparameters amongst the studied combinations. F1, precision and recall metrics evaluated for validation and training set for each model is reported in Table 4.5.

Model	Parameter analysed	Validation set			Training set		
		F1	Precision	Recall	F1	Precision	Recall
M1	Glorot initialization	0,178	0,900	0,090	0,387	1,000	0,323
M2	MG-Net initialization	0,406	0,930	0,259	0,642	0,961	0,482
M3	lr_{start}= 10^{-4}	0,170	1,000	0,090	0,370	1,000	0,230
M4	lr_{start}= 10^{-5}	0,647	0,790	0,540	0,978	1,000	0,956
M5	lr_{start}= 10^{-6}	0,602	0,872	0,460	0,857	0,959	0,775
M6==M5	lr_{max}=200	0,602	0,872	0,460	0,959	0,775	0,990
M7	lr_{max}=400	0,603	0,884	0,457	0,969	0,757	0,990
M8	lr_{max}=600	0,542	0,883	0,391	0,964	0,705	0,980
M9	lr_{max}=800	0,604	0,872	0,462	0,859	0,964	0,775
M10	n_{eph}=25	0,780	0,710	0,870	0,830	0,730	0,960
M11	n_{eph}=50	0,561	0,872	0,414	0,790	0,920	0,698
M12==M9	n_{eph}=100	0,604	0,872	0,462	0,859	0,964	0,775
M13	n_{eph}=200	0,609	0,872	0,468	0,896	0,994	0,816
M14	n_{eph}=300	0,663	0,848	0,544	0,952	0,995	0,914
M15	Dropout=0.2	0,653	0,898	0,513	0,763	0,886	0,669
M16==M10	Dropout=0.3	0,780	0,710	0,870	0,830	0,730	0,960
M17	Dropout=0.4	0,618	0,910	0,468	0,756	0,891	0,656
M18	Dropout=0.5	0,577	0,910	0,421	0,727	0,899	0,609

Table 4.5. Resulting metrics for all the model evaluated. Models are grouped according to the parameter under examination; models highlighted in orange show the final parameter's choice for each group.

Network initialization

M1 model trainable weights were initialized with Glorot uniform function. M1 performances, both for the validation and training set, reported a low precision and a high recall. The initialization of M2 model with MG-Net allowed to increase the precision (Table 4.5). After the first epoch of training, M1 presented an AUC of PR curve of 0.32 over the validation set, while M2 of 0.87 (Figure 4.3). As expected, initializing the trainable layers with MG-Net's weights is a better choice, since this initialization provides a network already able to search for BACs from the first epoch. The training adds to the network's knowledge, allowing to improve the results.MG-Net initialization was therefore considered the best choice and applied to all the models developed later.

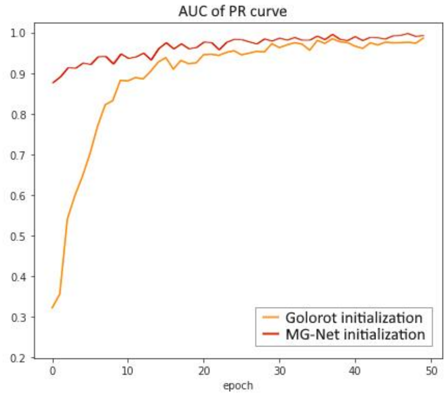


Figure 4.3 Area under the PR curve of the validation set for network initialized with Glorot unform function (model M1) and with MG-Net (model M2)

Learning rate

The initialization of the learning rate at 10^{-4} (model M3) resulted in a completely random behaviour of the network during training, confirming that no maximization of the AUC of PR curve was possible with this learning rate (Figure 4.4).

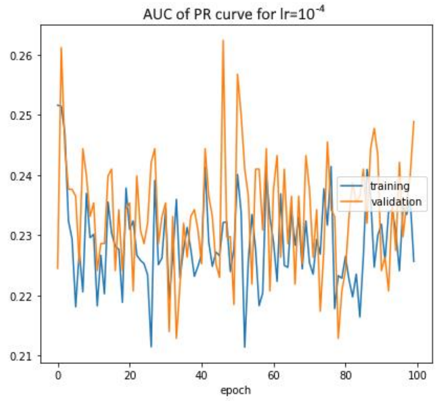


Figure 4.4 Area under PR curve for M3, showing random behaviour during training with learning rate lr= 10^{-4}

The comparison between a learning rate of 10^{-5} (M4) and 10^{-6} (M5) highlighted better results over the validation set for M4. However, the predictions over the training set were almost perfect in M4, and the high difference in results between validation and test set confirmed that the network was overfitting the training set and losing its generalization ability (Table 4.5). Therefore, it was chosen to fix $lr_{start}= 10^{-6}$.

The choice of lr_{max} was done by comparing the results over the validation set for models M6 to M9 (Table 4.5); lr_{max}=800 resulted in better performances.

Number of epochs

Model from M10 to M14 were compared to fix the best number of training epochs. The model giving best results over the validation set was M10; during its 25 epochs, the maximum AUC of PR curve for validation set was found at epoch 23, therefore M10 has been saved at this epoch.

Histograms of the network output before applying the classification threshold allowed to study saturation of the output neuron. It can be noted from Figure 4.5 that when increasing n_{eph} the number of images classified exactly as 0 or as 1 grows, therefore the network behaved approximately a binary classifier. This explains the worsening of results that can be noticed in Table 4.5 from model 10 to model 14.

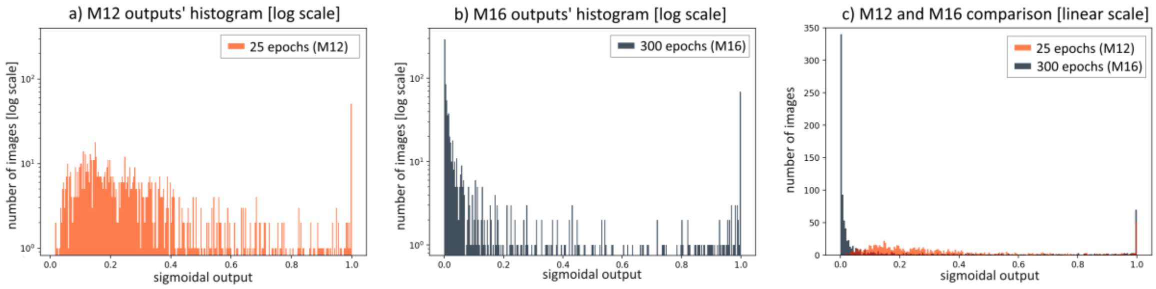


Figure 4.5 Histograms of network sigmoidal output, showing the number of validation images classified at each possible value between 0 and 1. *a*) histogram of the outputs of M10 (n_{eph}=25) in logarithmic scale; b) histogram of the outputs of M14 (n_{eph}=300) in logarithmic scale. c) comparison of M10 and M14 output in linear scale: notice that saturation of M14 is higher than the one of M10

Dropout value

Increasing the dropout value from 0.2 to 0.5 (models from M15 to M18) resulted in a progressive reduction of the learning ability of the network, that was manifested during training. Indeed, in Figure 4.6a it can be observed that, when evaluating the AUC of PR curve over the training set, the network performances improves for lower dropout values. Conversely, as predicted, the performances over the validation set are not directly influenced by the dropout value (Figure 4.6b), since the dropped neurons were turned off during the training but were turned back on when evaluating the validation set. Nonetheless, having an appropriate dropout rate allows to avoid overfitting. A 0.3 rate (M16) confirmed to be the best choice according to the evaluation metrics computed over the validation set (Table 4.5)

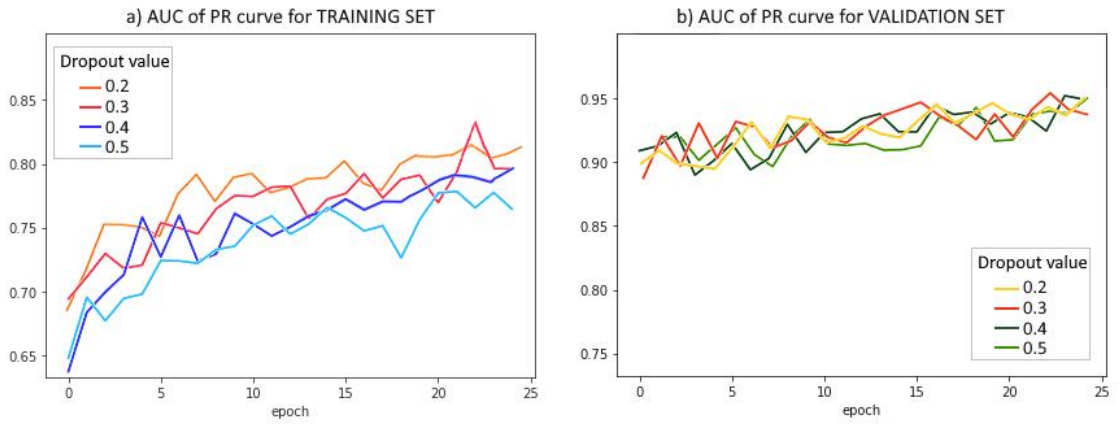


Figure 4.6 a) AUC of PR curve for training set; with increasing dropout value, a reduction in learning ability can be noticed as a decrease in the AUC values. b) AUC of PR curve for validation set; changing the dropout value does not affect the AUC in this case.

Best performing network

According to the previous evaluations, the best performing network was found to be M10, henceforth called BAC-Net. In summary.

BAC-Net is built as follows:

- Network architecture: identical to MG-Net (see chapter 3.3.2);
- Initialization of non-trainable weights (first 8 convolutional layers): VGG16;
- Initialization of trainable weights (last 5 convolutional layers and FC layers): MG-Net;
- Strategy for model's saving: maximization of AUC of PR curve over the validation set;
- Learning rate: set by cosine annealing schedule at each epoch, with lr_{start}= 10^{-6} and eph_{max}= 800;
- Maximum number of training epochs: n_{eph}= 25;
- Dropout rate: 0.3, applied only in the fully connected layers;
- Batch size: 8 images;
- Loss function: binary crossentropy;
- Optimizer: Adam;
- Errors weighting: class 0= 0.65, class 1= 2.14.

4.3. Classification results

4.3.1. Precision-Recall trade-off and image-wise metrics

BAC-Net was fed with the test set. The sigmoidal output was analysed to study the precision-recall trade-off and ultimately extract an optimal discrimination threshold τ, to be used in further work for output's binarization.

The precision and recall metrics resulting from variable thresholds are reported in Figure 4.7. The thresholds maximising precision (P-th), recall (R-th) and F1 (F1-th) on the test are specified in Table 4.6, along with the corresponding metrics.

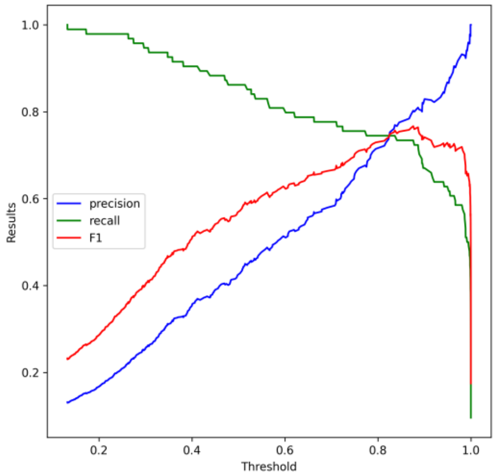


Figure 4.7 Precision (blue curve), recall (green curve) and F1 (red curve) results for a range of classification thresholds from 0.1 to 1.

	Threshold value	Precision	Recall	F1
P-th	0.99	1	0.394	0.565
R-th	0.13	0.131	1	0.232
F1-th	0.88	0.802	0.734	0.767

Table 4.6 Thresholds' evaluation

F1-th was also evaluated for the validation set, resulting in 0.83**.** The ultimate optimal threshold τ was computed as the average of F1-th computed on test and validation set, resulting in τ=0.85.

Results for the classification of the training, validation and test sets by using τ are reported in Table 4.7. For the test set, a confusion matrix of predictions is shown in fig. 4.8; PR and ROC curve are displayed in Figure 4.8.

Dataset	Balanced accuracy	Precision	Recall	F1
Training	0.857	0.963	0.723	0.723
Validation	0.849	0.9	0.707	0.792
Test	0.833	0.831	0.680	0.748

Table 4.7 Resulting metrics (for τ=0.85) evaluated image-wise over train, validation, and test set

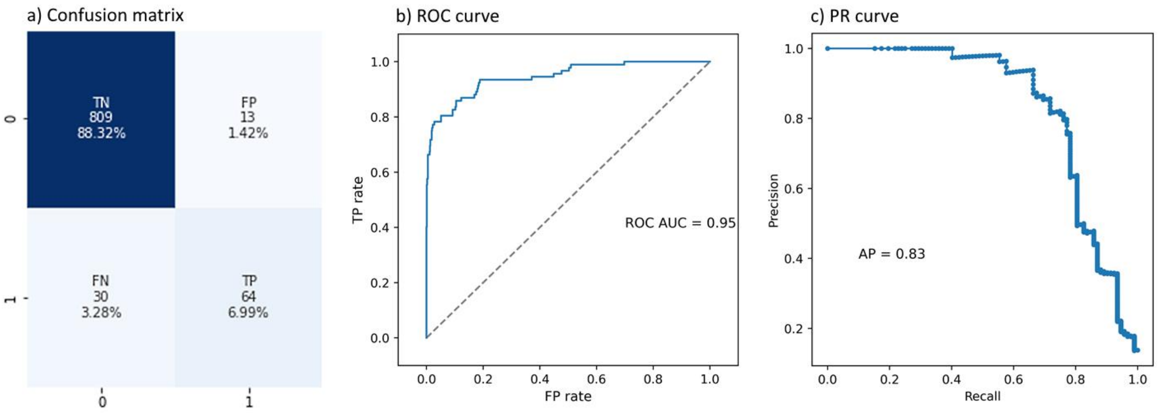


Figure 4.8 Evaluation of network classification over the test set. a) Confusion matrix; b) ROC curve; c) PR curve

4.3.2. Patient-wise metrics

Patient-wise results were computed using the BAC-Net, with the optimal classification threshold τ. They were evaluated on all the subsets, as reported in Table 4.8.

Dataset	Balanced accuracy	Precision	Recall	F1
Train	0.893	0.914	0.873	0.893
Validation	0.866	0.813	0.928	0.866
Test	0.862	0.831	0.680	0.748

Table 4.8 Resulting metrics (for τ=0.85) evaluated patient-wise over train, validation, and test set

4.3.3. Results visualization

A comparison between the four visualization methods is displayed in Figure 4.9. GradCAM++ method was chosen as the one with best precision in BACs localization and lower noise.

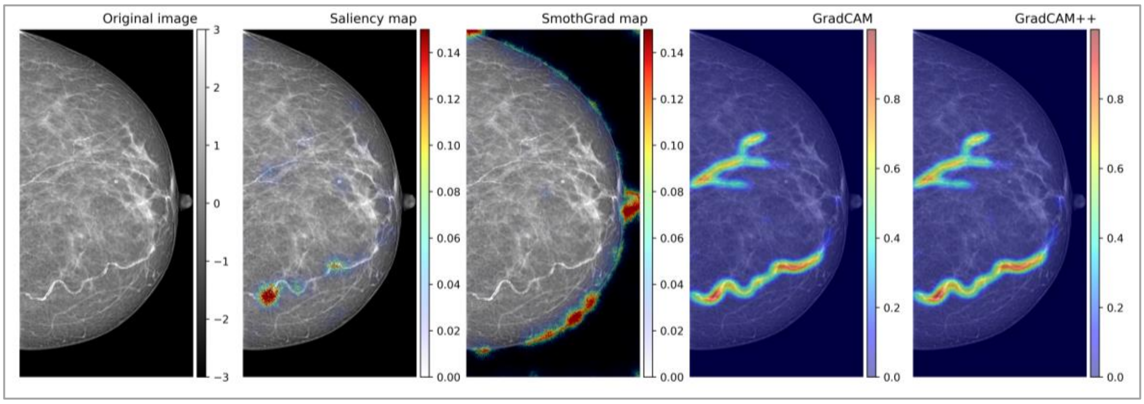


Figure 4.9 Comparison of the analysed visualization methods

Results of GradCAM++ heatmaps are displayed from Figure 4.10 to Figure 4.16, along with the BAC-Net's sigmoidal output and the final classification when considering as classification threshold τ=0.85.

Images from Figure 4.10 to Figure 4.12 illustrate true positive predictions. It can be noticed that in case of severe BACs, the whole calcified artery is highlighted (Figure 4.10). Less intense calcifications can be also detected as shown in Figure 4.11. In case of multiple BACs, the network is able to focus its attention on their different positions in the heatmap. When other types of benign calcifications are present in BAC+ mammograms, the network behaves differently according to their shape and surroundings, as can be seen in Figure 4.12: if the calcification is round and well defined it is not highlighted by the GradCAM++, while less distinct cases might represent a confounding factor. Nonetheless, if true BACs are also detected in the image, the heatmap intensity of BACs area is higher with respect to the ones of other types of misclassified calcifications. Figure 4.13 represents typical cases of false positives, where fibrous tissue or other benign calcifications under dense tissues are mislabelled as BACs, causing the incorrect prediction. GradCAM++ of negative predictions usually highlight the whole breast and do not focus on a specific area (Figure 4.14). In Figure 4.15 four mammograms showing medical implants are reported: pacemakers, cardiac loop recorders, breast implants and radiopaque markers do not interfere with the prediction, as confirmed by the heatmaps. A typical case of false negative is reported in Figure 4.16, highlighting how dense tissue might hinder small BACs detection.

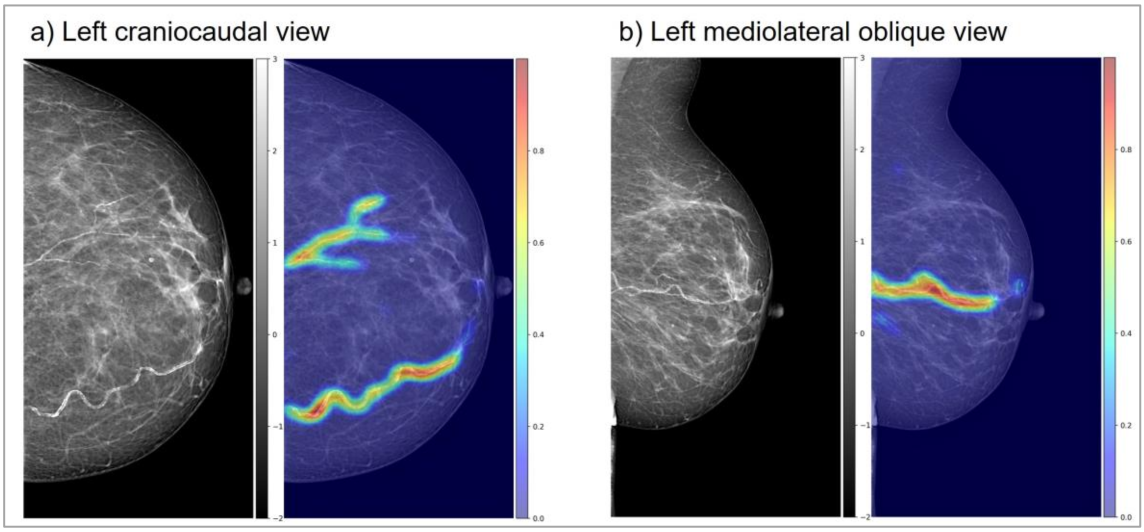


Figure 4.10 True positive predictions of severe BACs: mammograms belonging to a single patient represented in craniocaudal (CC) view (a) and mediolateral oblique (MLO) view (b) of the same breast. The network sigmoidal output for these two mammograms are 1, therefore they are correctly labelled as BAC+ with a high confidence

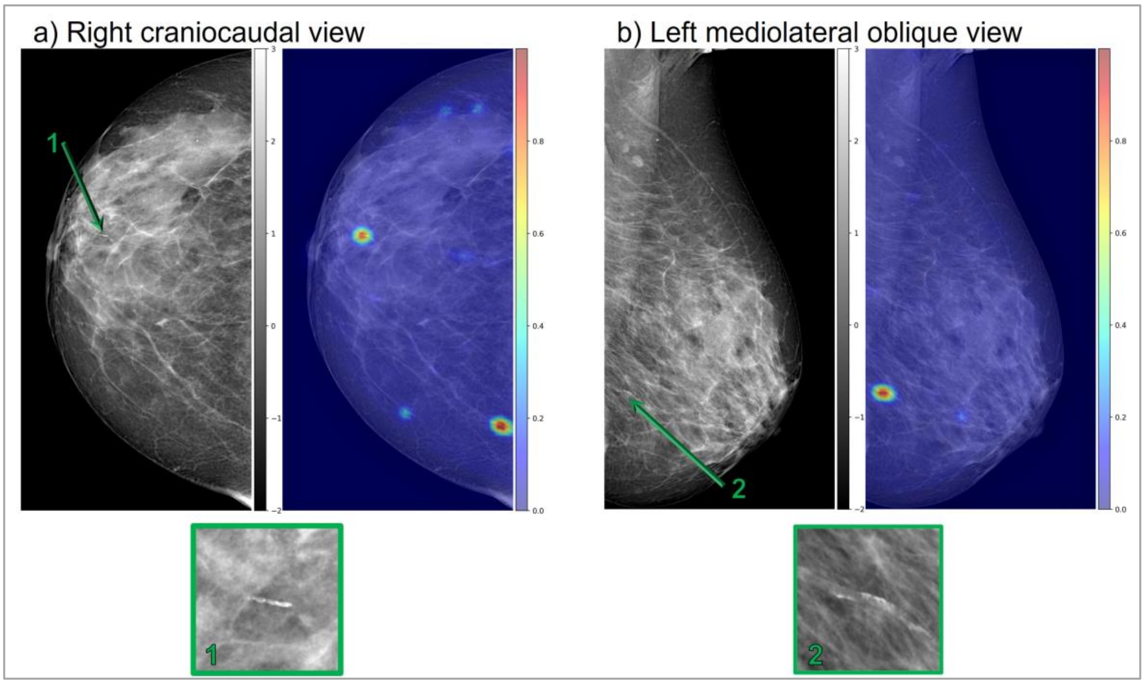


Figure 4.11True positive prediction of small BACs: mammograms belonging to a single patient with dense breast tissue. BACs localization is correct despite tissue density; details of the calcifications can be seen in 1. and 2. The network sigmoidal output is 1 for both mammograms

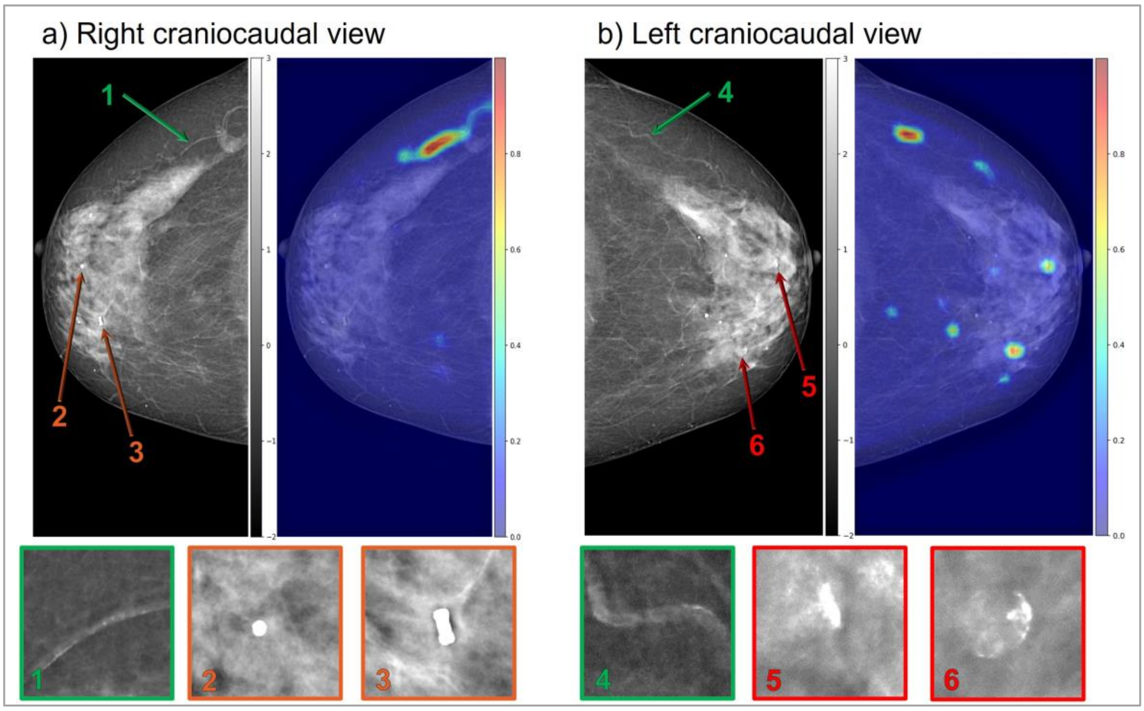


Figure 4.12 True positive prediction of small BACs: mammograms belonging to a single patient. In a) correct detection of BAC position (1) is performed, and distinct microcalcifications (2,3) are not misclassified; network sigmoidal output for this case is 1. In b) BAC position (4) is correctly identified, but microcalcifications under dense tissue are considered as BACs (5,6); network sigmoidal output for this case is 0.87, therefore the image is classified with a lower confidence as BAC+

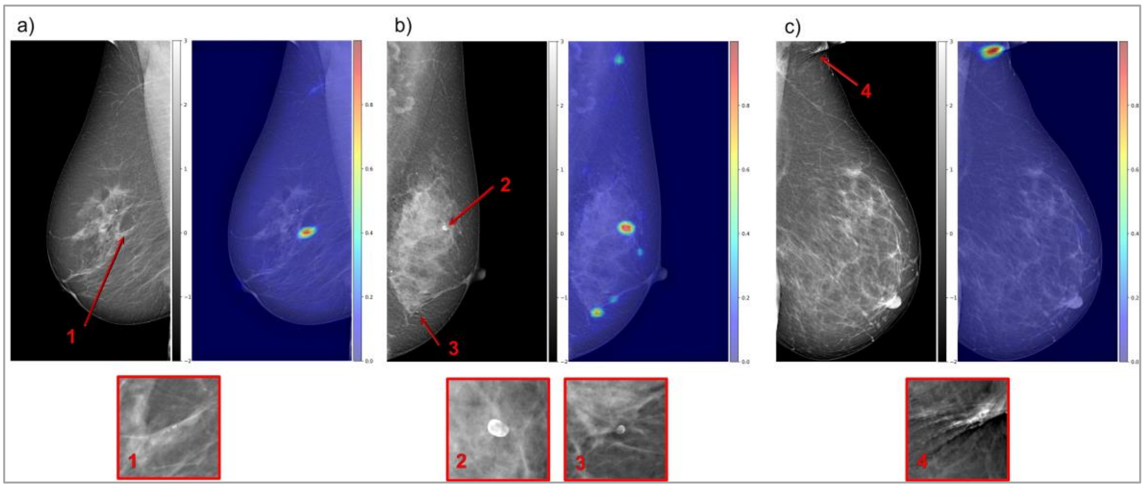


Figure 4.13 False positive cases: a) fibrous tissue (1) misclassified as BAC (network sigmoidal output=0.92). b) Microcalcifications under dense tissue (2,3) misclassified as BACs (network sigmoidal output=0.94). c) Skin calcifications (4) misclassified as BAC (network sigmoidal output=1). Note that in a) and b) the shape of calcifications identified as BACs is linear and therefore misleading for BAC-Net

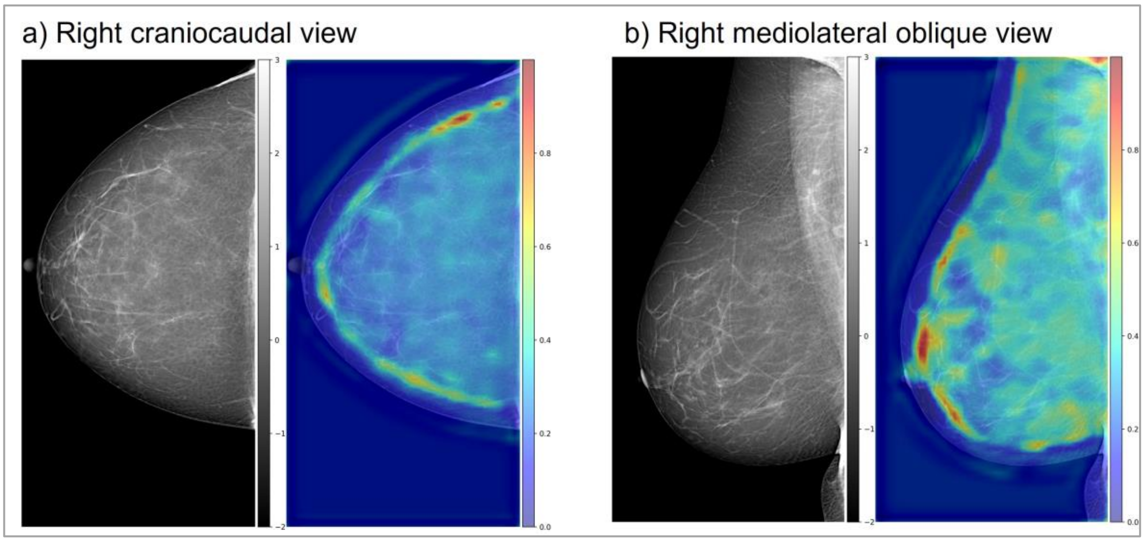

a) Right craniocaudal view
b) Right mediolateral oblique view

Figure 4.14 True negative case, mammograms belonging to a single patient. In both images GradCAM++ highlights the whole breast area, as it is typical in case of true negative predictions. a) network sigmoidal output= 0.07, b)mnetwork sigmoidal output=0.19

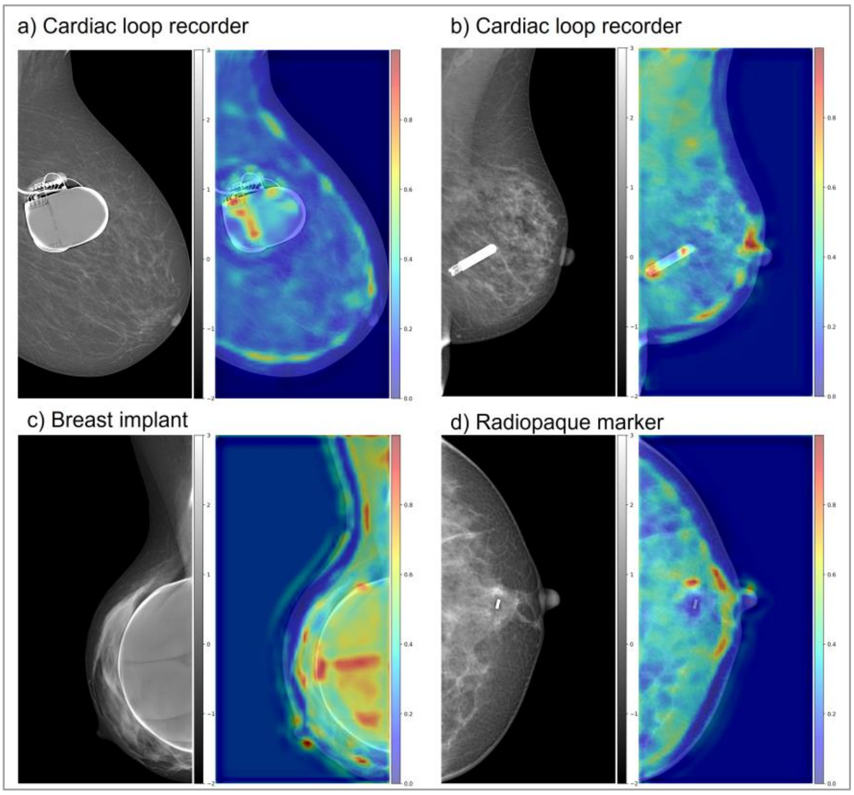


Figure 4.15 True negative cases with medical implants. Note how the implants do not disturb BAC-Net predictions. Network sigmoidal outputs are respectively: a) 0.01, b) 0.06, c) 0.23, d) 0.2

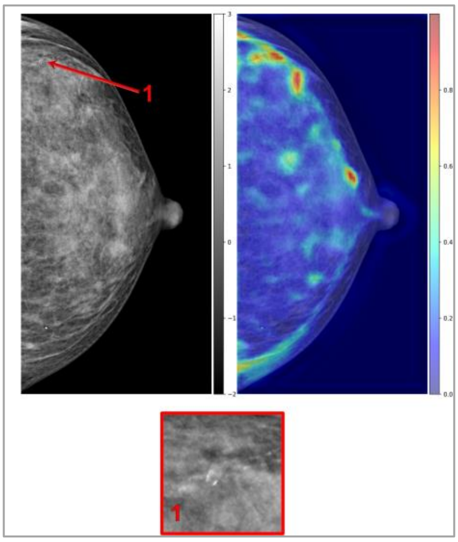


Figure 4.16 False negative prediction: small BAC under dense tissue (1) is not correctly identified by BAC-Net. Network sigmoidal output is 0.55, higher with respect to the typical true negative predictions (Figure 4.14)

Lastly, the GradCAM++ heatmaps were generated for all the convolutional layers of the network for a true positive prediction (Figure 4.17). It was noted that the first eight convolutional layers (not trained with transfer learning) were not distinguishing characteristics of the breast, and as typical of shallow convolutional layers, they focused on the analysis of the objects' shapes and contours. Conversely, in the five deep trainable convolutional layers, the network focused on the detection of the breast's shape and of BACs since deep layers were able to extract high-level features.

When BACs are present, in case of correct image classification, the only area highlighted by GradCAM++ corresponded to the calcification, that was identified not only in the last layer, but also in the previous four layers (Figure 4.18a). For BAC-predictions, the heatmap intensity was high over the whole breast, without any particular area of focus, in the last convolutional layer, while usually only breast contours were considered in shallower layers (Figure 4.18b).

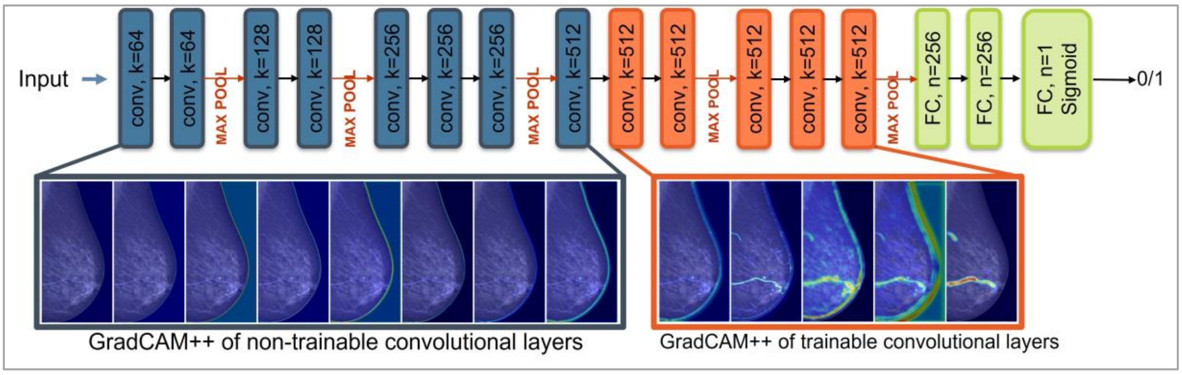


Figure 4.17 Correlation between GradCAM++ visualization and network's training strategy. In blue: non-trainable convolutional layers with VGG16 weights, only able to detect objects contours. In orange: layers trained for BACs detections, showing the ability to highlight calcifications.

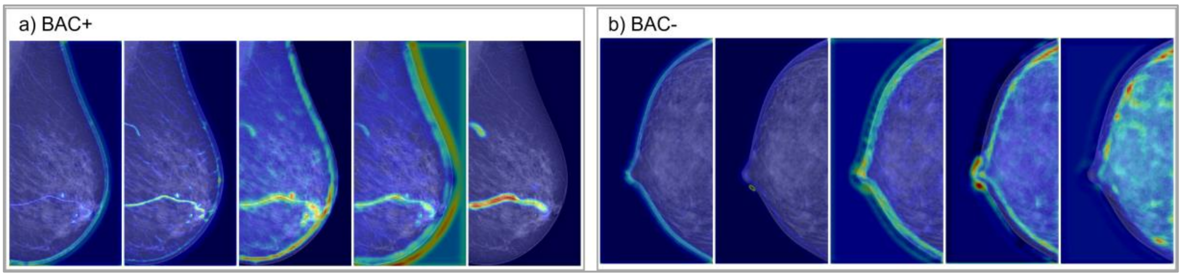


Figure 4.18 Detail of the trainable convolutional layers, for a BAC+ image (a) and a BAC- image (b).

4.4. BACs severity scoring results

4.4.1. Dataset

The dataset used for this preliminary severity scoring study corresponds to the one used by Trimboli et al. [34]. It was originally composed of 57 positive patients, but after double checking the dataset, one patient was excluded from the study since no agreement was found between the clinicians on the actual presence of calcifications on both the right and left views.

Thus 56 patients were analysed, with ages ranging from 49 to 82 years. Since only MLO views were considered, 112 mammograms were used for the study, of which 95 BAC+ and 17 BAC-.

4.4.2. Precision-recall tradeoff and network performances

BAC-Net was fed with the described dataset; the resulting ROC and PR curve are shown in Figure 4.19. P-th, F1-th and R-th were assessed and reported in Table 4.9 along with the classification metrics. Confusion matrices for the three classification thresholds considered are displayed in Figure 4.20.

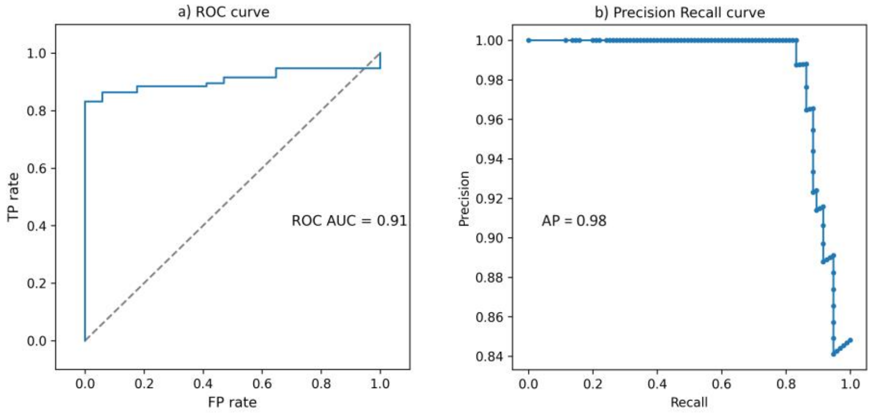


Figure 4.19 a) ROC curve; b) PR curve

	Threshold value	Precision	Recall	F1
P-th	0.7	1	0.832	0.908
F1-th	0.6	0.966	0.884	0.923
R-th	0.1	0.848	1	0.918

Table 4.9 Threshold's evaluation

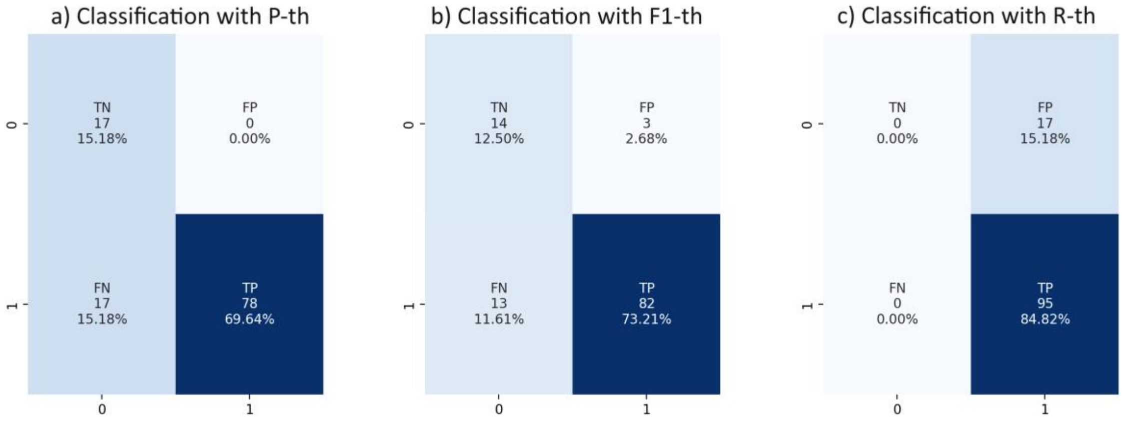


Figure 4.20 Confusion matrices resulting from using different classification thresholds: a) P-th=0.7; b) F1-th=0.6; c) R-th=0.1

4.4.3. Graphical visualization of BAC score computation

The scaling factors computed during image preprocessing ranged from 1.4 to 3.5 (mean value=2.1±0.3). Scores computation was performed through heatmap's thresholding. Figure 4.21 reports examples of GradCAM++ with a thresholding of $T_{heatmap}$=0.2. BACs present in these images are of increasing severity, and this is detected both when computing A_{BAC}, I_{BAC}, L_{BAC} and when assessing quartiles-based scores (A_Q, I_Q ,L_Q).

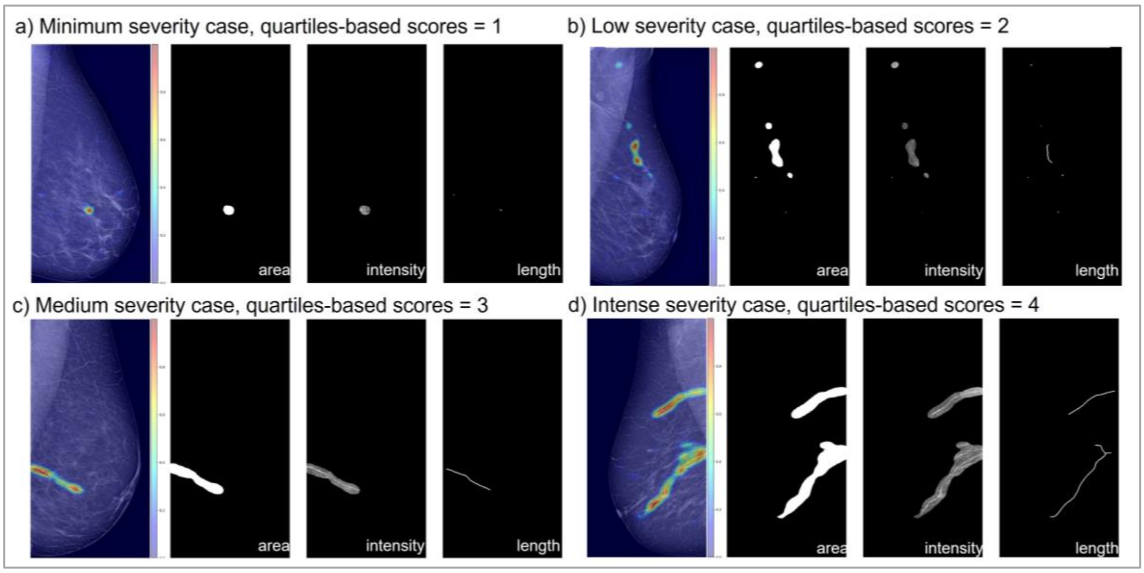

a) Minimum severity case, quartiles-based scores = 1
area
intensity
length
b) Low severity case, quartiles-based scores = 2
area
intensity
length
c) Medium severity case, quartiles-based scores = 3
area
intensity
length
d) Intense severity case, quartiles-based scores = 4
area
intensity
length

Figure 4.21 Examples of GradCAM++ thresholding for increasing BACs intensity from a) to d)

4.4.4. Correlation results

Henze-Zirkler test results proved that the bivariate distributions $p(l_{BAC}, A_{BAC})$, $p(l_{BAC}, I_{BAC})$ and $p(l_{BAC}, L_{BAC})$ were not normal. Spearman's correlation coefficient was therefore used as a correlation measure for all the linear regressions.

P-th was initially used to classify BAC+ images with maximum precision. BAC-Net predicted 78 images as BAC+, 34 images as BAC-. Correlation between A_{BAC}, I_{BAC} and L_{BAC} and l_{BAC} was assessed for all values of the binarization threshold $T_{heatmap}$. The binarization threshold maximising Spearman's correlation coefficient between l_{BAC} and A_{BAC} was found to be $T_{opt\text{-}A}$= 0.2. The same value resulted to be the optimal binarization threshold also for I_{BAC}, so that $T_{opt\text{-}I}$= 0.2. Lastly, optimization of the $T_{heatmap}$ for L_{BAC} resulted in $T_{opt\text{-}L}$= 0.3. These three optimal thresholds were also the one minimizing p-value for Spearman's coefficient for the respective correlations, as reported in Figure 4.22.

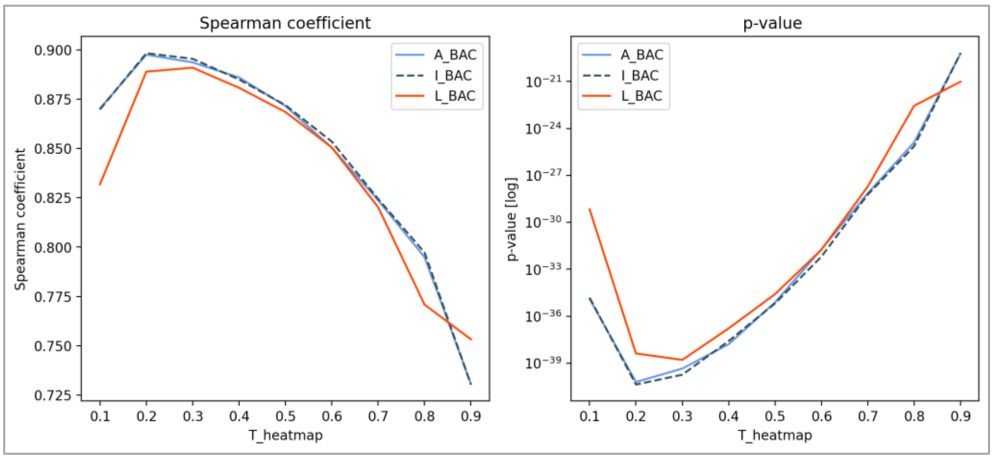
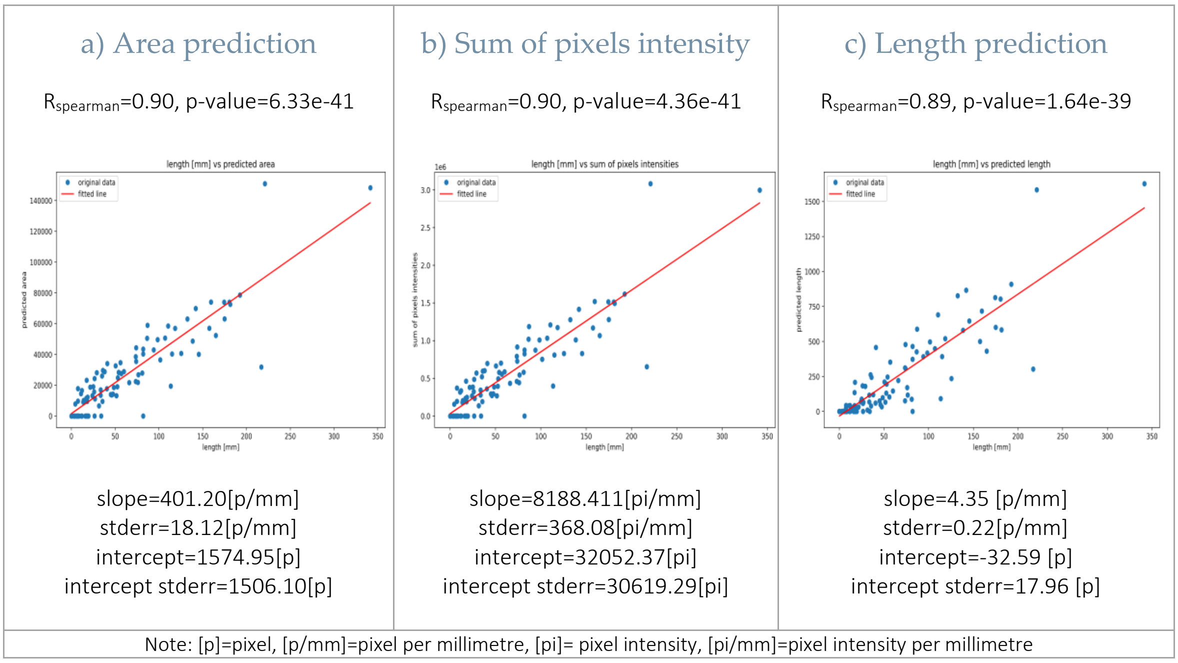


Figure 4.22 Evaluation of Spearman's coefficient and p-value for every $T_{heatmap}$ value. Notice the similar behaviour of predicted area (A_{BAC}) and pixels intensities sum (I_{BAC}) score, that results in the same optimal threshold $T_{opt\text{-}A}$= $T_{opt\text{-}I}$= 0.2. Predicted length (L_{BAC}) behaves differently, with $T_{opt\text{-}L}$=0.3

Linear regression results and correlation metrics obtained by using P-th as a classification threshold are reported in Figure 4.23. Comparison between the quartile-based length l_Q and the scores A_Q, I_Q and L_Q is displayed in Figure 4.24.

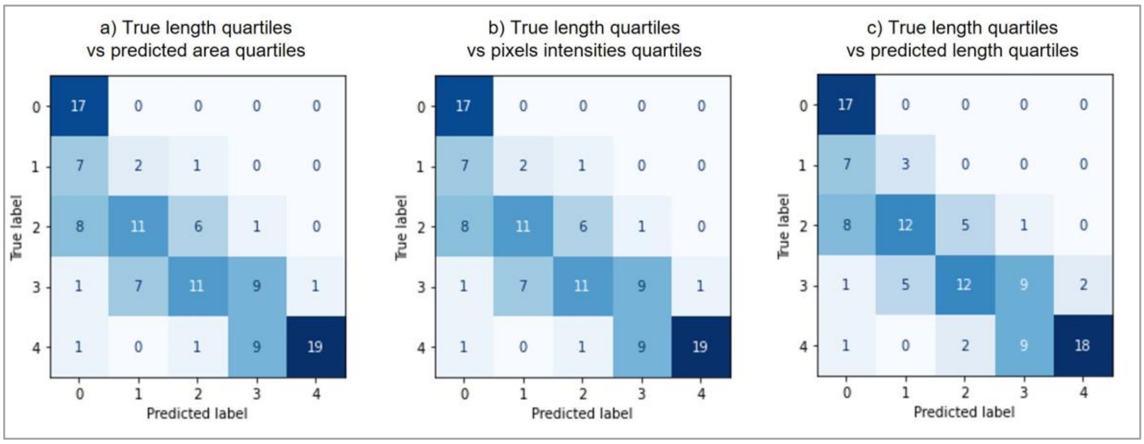


a) Area prediction	b) Sum of pixels intensity	c) Length prediction
R$_{spearman}$=0.90, p-value=6.33e-41	R$_{spearman}$=0.90, p-value=4.36e-41	R$_{spearman}$=0.89, p-value=1.64e-39

slope=401.20[p/mm]
stderr=18.12[p/mm]
intercept=1574.95[p]
intercept stderr=1506.10[p]
	slope=8188.411[pi/mm]
	stderr=368.08[pi/mm]
	intercept=32052.37[pi]
	intercept stderr=30619.29[pi]
		slope=4.35 [p/mm]
		stderr=0.22[p/mm]
		intercept=-32.59 [p]
		intercept stderr=17.96 [p]

Note: [p]=pixel, [p/mm]=pixel per millimetre, [pi]= pixel intensity, [pi/mm]=pixel intensity per millimetre

Figure 4.23 Results for P-th classification threshold. a) Linear regression between l_{BAC} and A_{BAC} (T$_{opt\text{-}A}$=0.2); b) Linear regression between l_{BAC} and I_{BAC} (T$_{opt\text{-}I}$=0.2); c) Linear regression between l_{BAC} and L_{BAC} (T$_{opt\text{-}L}$=0.3)



a) True length quartiles vs predicted area quartiles

True label \ Predicted label	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

b) True length quartiles vs pixels intensities quartiles

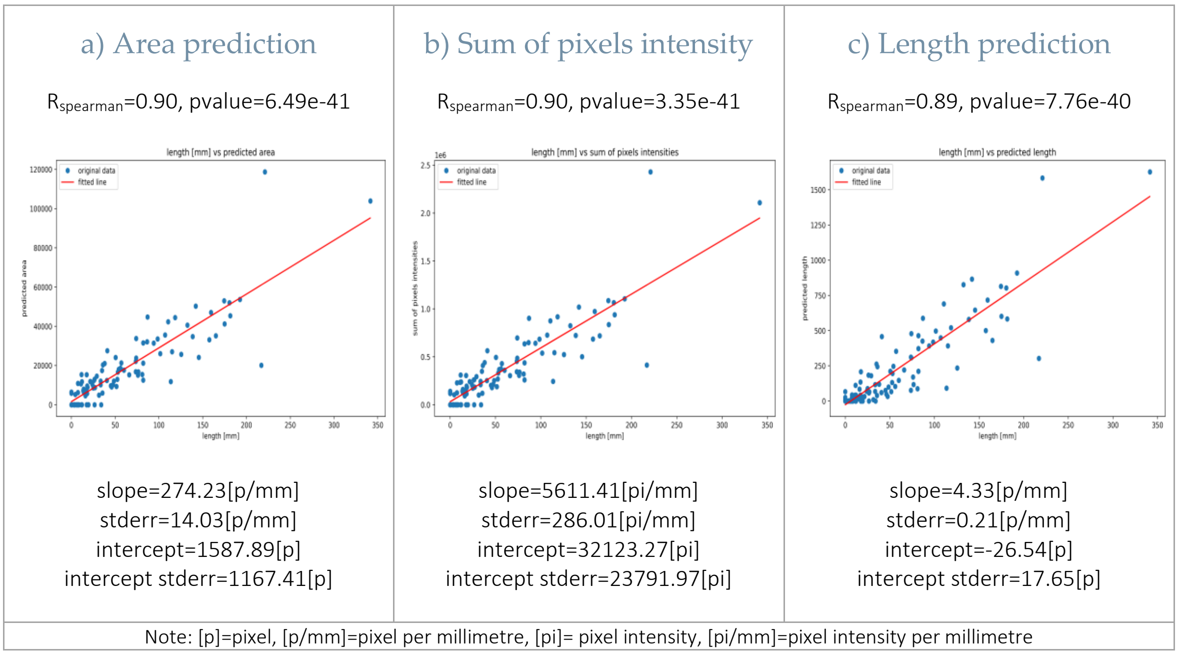
True label \ Predicted label	0	1	2	3	4
0	17	0	0	0	0
1	7	2	1	0	0
2	8	11	6	1	0
3	1	7	11	9	1
4	1	0	1	9	19

c) True length quartiles vs predicted length quartiles

True label \ Predicted label	0	1	2	3	4
0	17	0	0	0	0
1	7	3	0	0	0
2	8	12	5	1	0
3	1	5	12	9	2
4	1	0	2	9	18

Figure 4.24 Results for P-th classification threshold. a) Confusion matrix of l_Q and A_Q (T$_{opt\text{-}A}$=0.2); b) confusion matrix of l_Q and I_Q(T$_{opt\text{-}I}$=0.2); c) confusion matrix of and l_Q and L_Q (T$_{opt\text{-}L}$=0.3)

Secondly, classification of mammograms based on F-th was performed, maximising F1 measure. 85 images were predicted as BAC+, 88 as BAC-. The three optimal binarization thresholds computed were identical, corresponding to $T_{opt\text{-}A}= T_{opt\text{-}I}= T_{opt\text{-}L}= 0.3$. Linear regression results and correlation metrics obtained by using F1-th as a classification threshold are reported in Figure 4.25. Quartiles-based scores correlation was assessed by confusion matrix, as reported in Figure 4.26.



a) Area prediction

$R_{spearman}$=0.90, pvalue=6.49e-41

slope=274.23[p/mm]
stderr=14.03[p/mm]
intercept=1587.89[p]
intercept stderr=1167.41[p]

b) Sum of pixels intensity

$R_{spearman}$=0.90, pvalue=3.35e-41

slope=5611.41[pi/mm]
stderr=286.01[pi/mm]
intercept=32123.27[pi]
intercept stderr=23791.97[pi]

c) Length prediction

$R_{spearman}$=0.89, pvalue=7.76e-40

slope=4.33[p/mm]
stderr=0.21[p/mm]
intercept=-26.54[p]
intercept stderr=17.65[p]

Note: [p]=pixel, [p/mm]=pixel per millimetre, [pi]= pixel intensity, [pi/mm]=pixel intensity per millimetre

Figure 4.25 Results for F1-th classification threshold. a) Linear regression between l_{BAC} and A_{BAC} ($T_{opt\text{-}A}$=0.3); b) Linear regression between l_{BAC} and I_{BAC} ($T_{opt\text{-}I}$=0.3); c) Linear regression between l_{BAC} and L_{BAC} ($T_{opt\text{-}L}$=0.3)

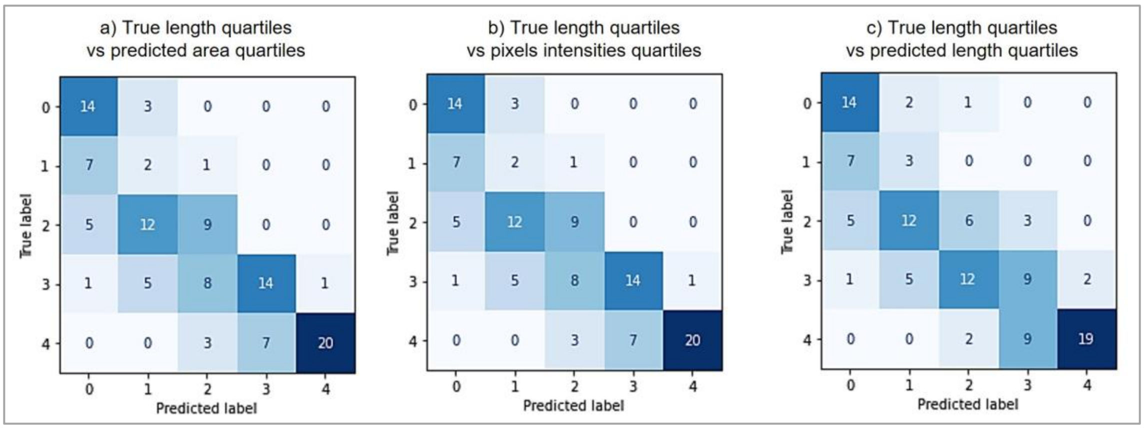
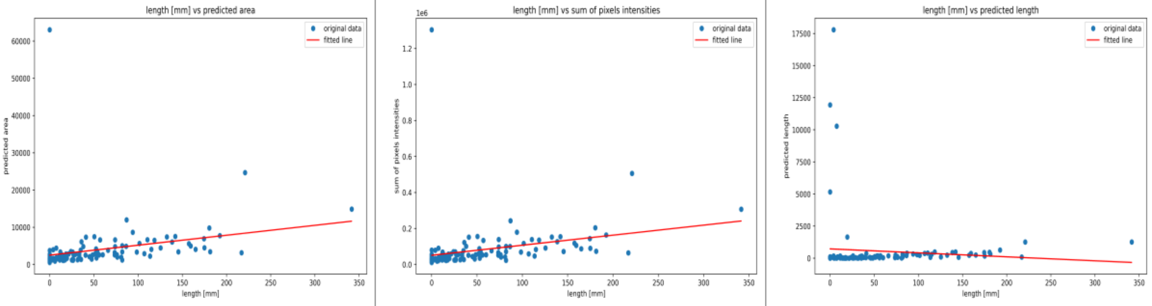


Figure 4.26 Results for F1-th classification threshold. a) Confusion matrix of l_Q and A_Q ($T_{opt\text{-}A}$=0.3); b) confusion matrix of l_Q and I_Q($T_{opt\text{-}I}$=0.3); c) confusion matrix of and l_Q and L_Q ($T_{opt\text{-}L}$=0.3)

Lastly, images were classified with R-th, maximising recall. All 112 images were predicted as BAC+. Optimal binarization thresholds were computed as T$_{opt-A}$= 0.7, T$_{opt-I}$= 0.7, T$_{opt-L}$= 0.5. Linear regression results and correlation metrics obtained by using F1-th as a classification threshold are reported in Figure 4.27. Correlation of quartiles-based scores is displayed in Figure 4.28.

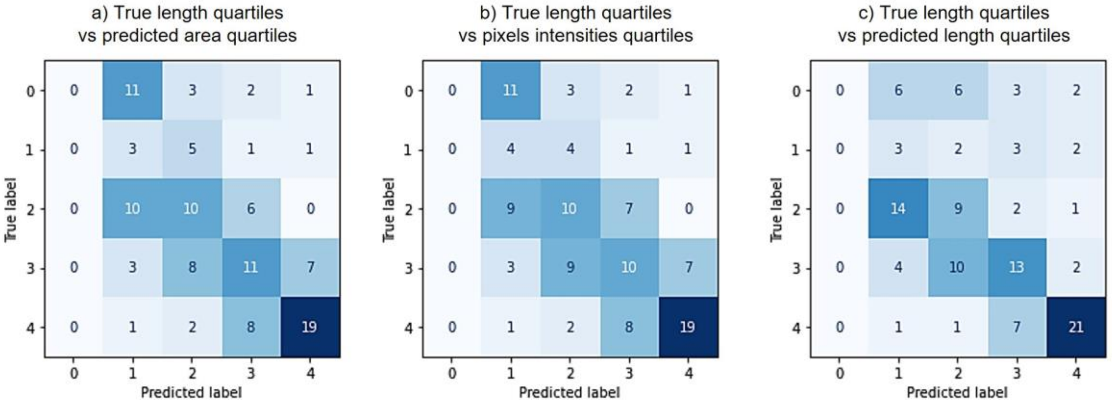
a) Area prediction	b) Sum of pixels intensity	c) Length prediction
R$_{spearman}$=0.67 pvalue=8.83e-16	R$_{spearman}$=0.67 pvalue=8.08e-16	R$_{spearman}$ =0.54 pvalue=1.05e-09



slope=26.71[p/mm]
stderr=9.58[p/mm]
intercept=2514.29[p]
intercept stderr=796.55[p/mm]
	slope=558.14[pi/mm]
	stderr=197.96[pi/mm]
	intercept=51188.28[pi]
	intercept stderr=16467.63[pi]
		slope=-3.09 [p/mm]
		stderr=3.46[p/mm]
		intercept=725.09[p]
		intercept stderr=287.81[p/mm]

Note: [p]=pixel, [p/mm]=pixel per millimetre, [pi]= pixel intensity, [pi/mm]=pixel intensity per millimetre

Figure 4.27 Results for R-th classification threshold. a) Linear regression between l_{BAC} and A_{BAC} (T$_{opt-A}$=0.7); b) Linear regression between l_{BAC} and I_{BAC} (T$_{opt-I}$=0.7); c) Linear regression between l_{BAC} and L_{BAC} (T$_{opt-L}$=0.5)



a) True length quartiles vs predicted area quartiles

	0	1	2	3	4
0	0	11	3	2	1
1	0	3	5	1	1
2	0	10	10	6	0
3	0	3	8	11	7
4	0	1	2	8	19

b) True length quartiles vs pixels intensities quartiles

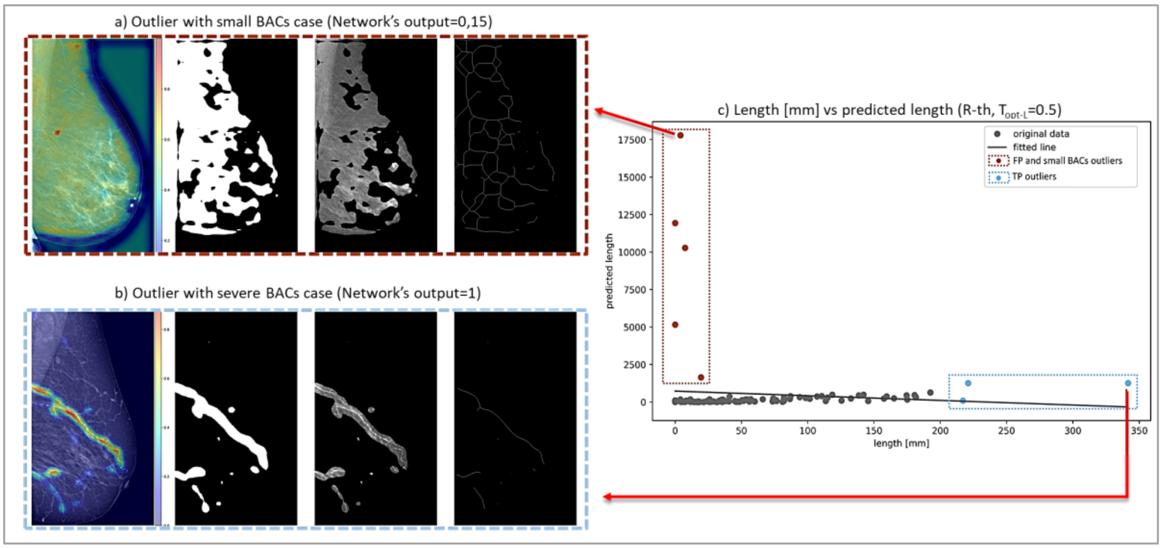
	0	1	2	3	4
0	0	11	3	2	1
1	0	4	4	1	1
2	0	9	10	7	0
3	0	3	9	10	7
4	0	1	2	8	19

c) True length quartiles vs predicted length quartiles

	0	1	2	3	4
0	0	6	6	3	2
1	0	3	2	3	2
2	0	14	9	2	1
3	0	4	10	13	2
4	0	1	1	7	21

Figure 4.28 Results for R-th classification threshold. a) Confusion matrix of l_Q and A_Q (T$_{opt-A}$=0.7); b) confusion matrix of l_Q and I_Q(T$_{opt-I}$=0.7); c) confusion matrix of and l_Q and L_Q (T$_{opt-L}$=0.5)

In each linear correlation outlined above, it is possible to notice three outliers on the right side of the image, corresponding to three true positive predictions with high BAC severity; the heatmap of these cases correctly predicts the position of calcifications, but the resulting scores are overestimated, most likely due to noise in the heatmap that generates small dots during thresholding that are considered in scores computation (Figure 4.29b). For score computation performed with R-th=0.1, it is also possible to notice outliers on the left side of the image (Figure 4.29a); these are either false positive predictions or cases of low BACs length and correspond to network's sigmoid outputs between 0.1 and 0.22., therefore were predicted as positive with low confidence. Heatmap thresholding of this images gives high scores results since their GradCAM++ are not highlighting only BACs regions but the entire breast, as is typical of true negative heatmaps (see chapter 4.3.3), therefore the scores computed are biased (Figure 4.14).


a) Outlier with small BACs case (Network's output=0,15)

b) Outlier with severe BACs case (Network's output=1)

c) Length [mm] vs predicted length (R-th, T_{opt-L}=0.5)

Figure 4.29 a) Worst case of outlier with small BAC; b) Worst outlier with significative BACs Network's output for this case is 1. c) Highlighting of outlier cases on the regression plot for l_{BAC}-L_{BAC} (R-th classification, T_{opt-L}=0.5). Cases on bottom right are common to all linear regressions performed; cases on top left are not present for P-th and F1-th regressions.

Amongst the three continuous scores proposed, the best performing one on this dataset was A_{BAC} computed with P-th and T_{opt-A}=0.2. A comparison of linear regressions of A_{BAC} computed for the three combinations of thresholds considered is displayed in Figure 4.30. For the ordinal scores, A_Q and I_Q had almost identical performances, and were better BACs severity predictors with respect to L_Q. When considering A_Q, a reduction of classification threshold from 0.7 (P-th) to 0.6 (F1-th) improved the results by increasing accuracy from 0.47 to 0.53, while classification with threshold 0.1 (R-th) resulted in an accuracy of 0.36 (Figure 4.31).

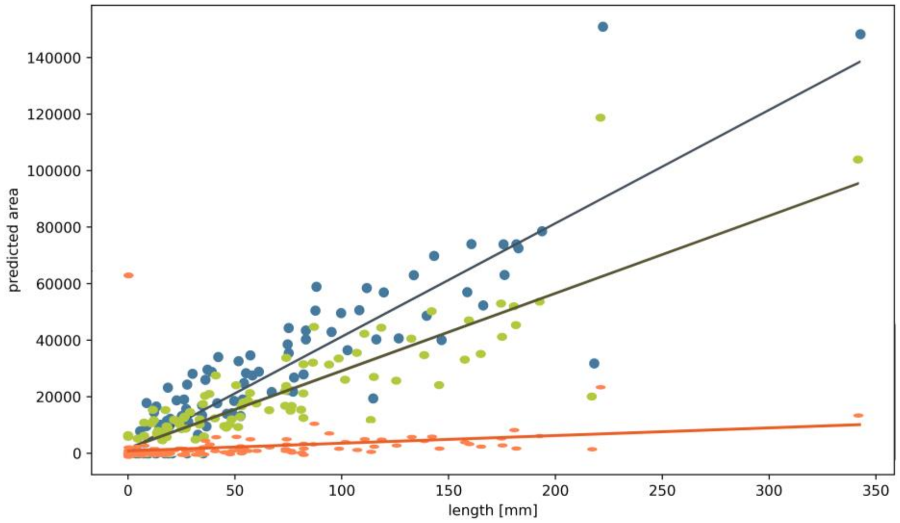
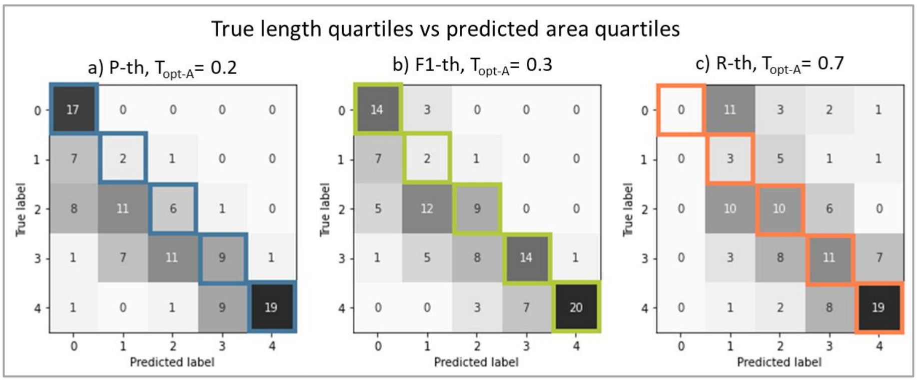


Figure 4.30 Comparison of linear regression results for l_{BAC}-A_{BAC}computed with the three combinations of thresholds considered.



True label \ Predicted label (a)	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

True label \ Predicted label (b)	0	1	2	3	4
0	14	3	0	0	0
1	7	2	1	0	0
2	5	12	9	0	0
3	1	5	8	14	1
4	0	0	3	7	20

True label \ Predicted label (c)	0	1	2	3	4
0	0	11	3	2	1
1	0	3	5	1	1
2	0	10	10	6	0
3	0	3	8	11	7
4	0	1	2	8	19

Figure 4.31 Comparison of confusion matrices for l_Q-A_Q for the three combinations of thresholds considered, highlighting correct predictions: a) P-th, T$_{opt-A}$= 0.2, resulting in accuracy=0.47; b) F1-th, T$_{opt-A}$= 0.3, accuracy=0.53; c) R-th, T$_{opt-A}$= 0.7, accuracy=0.36.

5 Discussion and future developments

In this thesis, an automatic method for detection and quantification of breast arterial calcifications has been investigated and validated. BACs are frequent findings in mammograms and are indicated as a woman-specific cardiovascular risk factor: their intensity is directly correlated to the severity of CVDs risk. Cardiovascular diseases are the leading death cause in the world, and great effort is being made to reduce their occurrences. Despite that, women's CVD risk is often underestimated and the rate of decline of deaths for CVDs is lower for women than for men. The introduction of sex-specific risk factors such as BACs in clinical practice is therefore a priority.

Mammography screening is already performed on the majority of European and US female population above 50 years, hence women screened for breast cancer are the same that would benefit from BACs screening: no further radiation exposure for the patients nor sanitary expenses would be required. However, BACs presence is not yet included by most clinicians in mammography reports, therefore its predictive ability for CVD risk is poorly exploited, so far. The procedure proposed in this work could play a role in the solution of this problem, by automatically evidencing the presence and intensity of BACs to the radiologists, thus reducing the workload required by this additional inspection.

The development of this thesis has been made possible with the collaboration of the IRCCS Policlinico San Donato radiology team, that provided an annotated dataset of 6228 mammographic images (1557 patients) labelled as BAC+ or BAC-, i.e. with or without BACs, respectively. Is worth remarking that such huge annotation work is a necessary to the subsequent training, validation, and testing of AI approaches, such as the applied CNN. Compared to the previous thesis work (Ienco et al.) the annotated data-base grew about two fold, which permitted the present improvement of the study. BAC+ patients in the actual dataset (12.46%) reflected BACs prevalence amongst women reported in literature (12.7%) [26]. The lowest age for a BAC+ patients was found to be 45 years. Therefore, along with the clinicians, it was decided to fix an exclusion criterion for patients younger than 45 years, that will be used also for future studies.

The problem of age distribution amongst training, validation and test subsets has not been investigated by other studies applying neural networks to BACs detection [59], [61], [62]. These studies perform splitting and undersampling by random data extraction on the whole data sets, overlooking the issue of uneven BAC+ prevalence, strongly increasing with age. Conversely, the dataset splitting procedure here proposed was based on the extraction from four age classes independently, thus maintaining BAC+ original age distribution amongst subsets. Validation and test sets need to be representative of the real population that might be subject to BACs screening, therefore their age distributions must reflect the one of the original dataset; so, no BAC- undersampling was performed on the validation and test sets. Still, the segmentation into 4 age classes was applied in the splitting procedure, which had a significant positive impact on the whole procedure. Indeed, the problem of feature extraction permitting to detect BACs within the breast background does change significantly with age since the breast density significantly decreases. We might say that we treated our CNN as an expert radiologist would challenge a young one to be trained showing him/her mammographies from different age classes.

The architecture of the CNN used was firstly developed by Ienco et al. [1] for BACs classification (MG-Net) and relied on transfer learning from VGG16 network. Network's hyperparameters were not previously explored, and lack of data did not allow for testing of MG-Net's performances on an independent dataset. As a first step for the development of the new BAC-Net, hyperparameters tuning was performed. Learning rate decay and dropout were fixed by maximising results over the validation set. Tuning the number of training epochs was found to be particularly challenging, since it was noted that network performances on the validation set were worsening, especially in sensitivity, when increasing the number of epochs. On the other hand, the network wasn't showing signs of overfitting in the loss curve. An analysis of the sigmoidal output before the application of classification's threshold allowed to hypothesize that this behaviour was due to saturation of the output sigmoid neuron, that led it to behave like a binary classifier instead of operating in its linear region. Considering that BAC-Net trainable weights were initialized with MG-Net, results were already promising with a low number of epochs, therefore no measure has been taken to counterbalance the saturation. However, future development of BAC-Net might explore regularization methods such as L2 or modify errors weighting to overcome this issue and improve upon classification results.

Testing of BAC-Net was performed on the independent test dataset. Since the raw output of BAC-Net has a sigmoidal distribution, a classification threshold must be fixed to extract a binary label indicating presence or absence of BACs. While choosing this threshold, a tradeoff between precision and recall is necessary, but the decision of which measure to favour highly depends on the scope of the prediction. If the network is to be used for screening purposes, recall must be privileged, therefore a low classification threshold must be used to increase the number of patients predicted as BAC+, avoiding the exclusion of positive patients from further analysis. On the other hand, if BACs classification is being made for research purposes, such as the extraction of positive images to be used for the scoring procedure, the threshold must be high, favouring precision, in order to consider as BAC+ only images where the network finds BACs with a good confidence. Results of different classification thresholds were reported in this work, and it was decided to leave the final decision to future users of the network to allow for adaptability of the system to the clinicians' aims. For clarity in results presentation a threshold maximising F1 measure over validation and test set was fixed as τ=0.85, balancing precision and recall. This resulted in a prediction over the test set with precision=0.831, recall=0.68, F1=0.748 and ROC AUC=0.95. The difference between current results and the best performing network reported by Ienco et al. is faint, but when considering MG-Net average results over the 7-fold cross validation, an improvement can be noticed (average precision=0.864±0.040, average recall=0.667±0.132, average F1=0.744±0.094, average ROC AUC=0.86 ± 0.07). Moreover, MG-Net was evaluated on validation datasets of 141 images each, not independent from the training procedure, while BAC-Net results refer to an independent test set of 916 images.

As a further step in BAC-Net results analysis, an exploration of state-of-the-art visual explainability methods was performed. Saliency maps were previously used for this scope by Ienco et al. [1]. They correctly identified the position of calcifications but resulted noisy and did not follow the calcified vessels. On the other hand, SmoothGrad maps often displayed high values for pixels not related to BACs. Conversely, GradCAM and GradCAM++ heatmaps were similarly able to locate and highlight the whole calcified vase as shown in Figure 4.9. Considering the lower noise present in the GradCAM++ method with respect to GradCAM [87], GradCAM++ heatmaps were considered as the ones with best performance.

The generation of heatmaps with this technique is fast enough to be used in real applications and allows not only to analyse the image areas that are considered by the network as prevalent in reaching a decision, but also to study the behaviour of each convolutional layer. Moreover, use of GradCAM++ for radiologic images has been already reported in literature, e.g., for detection of breast cancer [88] and Covid-19 [89].

GradCAM++ showed a good precision in highlighting the position of BACs, even though the area considered was always wider than the calcified arteries. This could be explained by the fact that the network bases its decision not only on pixels belonging to the calcification, but also on their contrast with respect to other breast tissues. Indeed, BAC pixels have higher intensity with respect to soft tissues, since they are more radiopaque. Shape also plays a role in BAC classification: linear calcium deposits of different origins or fibrous tissues might represent a confounding factor for network prediction, generating false positive results. Conversely, round calcifications and microcalcifications usually do not disturb BACs detection, unless they are superimposed to dense tissue that makes their boundaries not well defined, sometimes resembling a line. Interestingly, heatmaps of the last convolutional layer (Figure 5.1f) demonstrated that well-distinguished round calcifications (that are not BACs) are not considered for the final prediction, while are analysed in previous convolutional layers (Figure 5.1b-e) presumably for their pixel intensities.

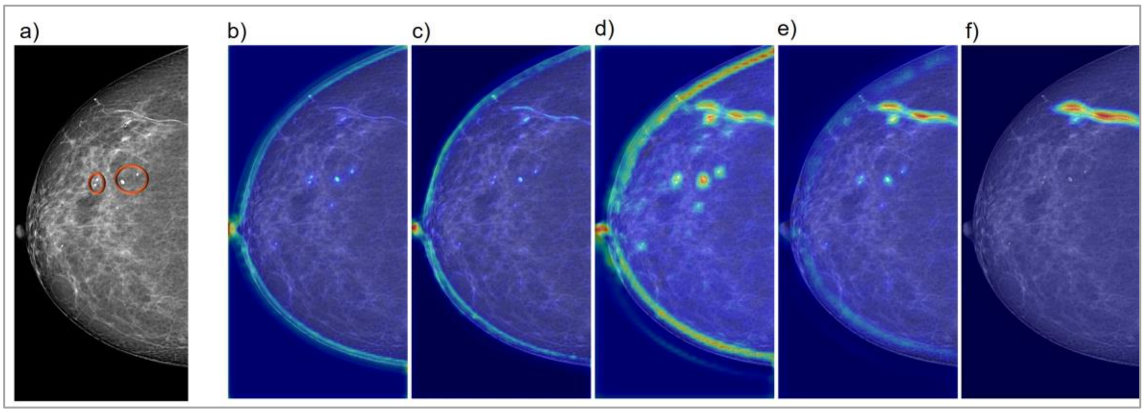


Figure 5.1 a) BAC+ image with round well-distinguished microcalcifications. b)-f) heatmaps of the last five convolutional layers: the microcalcifications are highlighted by BAC-Net until the penultimate layer, presumably for their pixel intensity, but do not have any influence on the last convolutional layer (where only the real BAC is highlighted) since their shape is clearly distinguishable from BACs linear shape.

The analysis of false positive predictions performed together with a radiologist allowed to understand that for most false positive cases, the network's detection of BACs position on the heatmap is considered wrong from the clinician without any doubt (see chapter 4.3.3, Figure 4.13). On the other hand, some images labelled as BAC- by human readers, and predicted as BAC+ by BAC-Net, were indeed reviewed as positive to BACs presence when analysing the heatmap (Figure 5.2). These two cases are suggesting that, in clinical application, the visualization of results could play an important role in the mitigation of both human and network errors. Moreover, the possibility to study BAC-Net results with graphical representations encouraged a discussion between engineers and radiologists that would have never been possible by only considering quantitative results of network performances; this allowed both groups to improve their understanding and confidence in the CNN predictions, and to better contribute to the development of the proposed BACs scoring procedure.

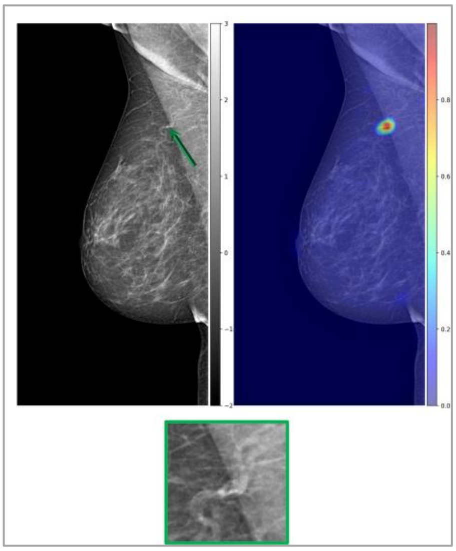


Figure 5.2 Case labelled as BAC- by manual reader, as BAC+ by the network (false positive). After heatmap analysis was considered as BAC+ by the radiologist.

The radiology team, collaborating with this study, had previously developed a semiquantitative score (BAC-SS), based on BACs number Nv, vessels opacification Ov and an ordinal BACs length score l_Q. Data for Nv and Ov can be quickly retrieved from the mammogram. On the other hand, l_Q computation requires to extract the length of BACs through manual segmentation, a high precision and time-consuming procedure. The obtained length is further compared with quartiles of BACs lengths computed for a wide population, and the final length score considered is the one corresponding to the quartile to which the calcifications belong, so that l_Q ranges from 1 to 4.

The idea for an automation of the scoring process originates from the need to speed up the described procedure. For this preliminary study a subset of 56 BAC+ patients (scoring dataset), whose mammograms were already classified with BAC-SS, was used for scores extraction and correlation tests. BAC-Net sigmoidal outputs were computed for images belonging to the scoring dataset, and again an analysis of the precision-recall tradeoff was performed. It was decided to compute correlation results considering three possible classification thresholds: the one maximising precision (P-th), that was discussed as the primary case since it included in the positive predictions only true positive images classified with high confidence by

BAC-Net, and the ones maximising F1 (F1-th) and recall (R-th). As a ground truth both continuous BACs length in millimetres (l_{BAC}) and quartiles-based semiquantitative length score (l_Q) were considered. The scores examined as possible candidates for a definitive BACs numerical score (S_{BAC}) were the heatmap's area with intensity above threshold $T_{heatmap}$ (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}). For the definition of a semiquantitative BACs score (S_Q), three quartiles-based scores were derived from the continuous ones previously defined: A_Q, I_Q, L_Q .

Linear regression was firstly performed for data classified with P-th, and showed high correlation coefficient for all three proposed continuous scores: l_{BAC}-A_{BAC} correlation resulted in $R_{Spearman}$=0.90, p-value=6.33e^{-41} ; l_{BAC}-I_{BAC} resulted in $R_{Spearman}$=0.90, p-value=4.36e^{-41} ; l_{BAC}-L_{BAC} resulted in $R_{Spearman}$=0.89, p-value=1.64e^{-39}. Lowering the classification threshold to F1-th and R-th produced worse results due to the increase of false positive predictions, that represented outliers in the regression's scatterplot. Interestingly, for each classification threshold considered, linear regression results between l_{BAC} and A_{BAC}, I_{BAC}, L_{BAC} were similar; a possible reason is that the three scores might be proportional to each other, since I_{BAC} could be considered as product of A_{BAC} and mean pixels intensity, while L_{BAC} as ratio between A_{BAC} and mean width of the objects considered as BACs. Ultimately, in this analysis, A_{BAC} has shown the best correlation results, but further testing with a higher number of images needs to be performed in order to choose a definitive continuous score S_{BAC}.

The correlation between l_Q and the quartiles-based scores A_Q, I_Q, L_Q was analysed by confusion matrices. A_Q and I_Q, had almost identical performances, and were better BACs severity predictors with respect to L_Q; these behaviours well reflect the correlations of continuous scores with length in millimetres. A comparison of A_Q results for P-th, F1-th and R-th can be used to illustrate the predictive abilities of quartiles-based scores for variable classification threshold. Considering results produced using P-th, the predictions resulted either coincident with the real score or lower (the confusion matrix was lower-triangular, accuracy=0.47); the diagonality of confusion matrix was improved with F1-th (accuracy=0.53), while R-th produced the worst confusion matrix (accuracy=0.36) due to high number of false positives. This suggests that using lower classification thresholds improves results, as long as the number of false positive is not too high; this is caused by a shift of the quartiles considered for the production of A_Q, I_Q and L_Q toward lower values, due to inclusion of mammograms with smaller BACs (that are classified by BAC-Net with a lower sigmoid output). This behaviour demonstrates the future necessity of a careful study of the proper classification threshold to obtain optimal results, together with the need for testing on a wider dataset to fix the best S_Q.

The use of thresholded heatmaps for unsupervised segmentation applied to medical images has been poorly explored in literature, and the possibility to use this method for BACs quantification has not been previously investigated. The majority of literature studies use heatmaps for the extraction of region of interest instead of dealing with a precise segmentation: an example of this application can be found in research by Guan et al. [93], applied on chest X-rays (Figure 5.3c). Segmentation for skin cancer starting from VGG-16 network and using GradCAM was previously explored [91] following a procedure very similar to the one here applied. However, skin cancer presents a different shape with respect to BACs, and its analysis is based on skin digital pictures with high contrast of the detected shape, so that GradCAM appears to highlight with more precision the cancer area (Figure 5.3b) when compared to BACs detection (Figure 5.3a).

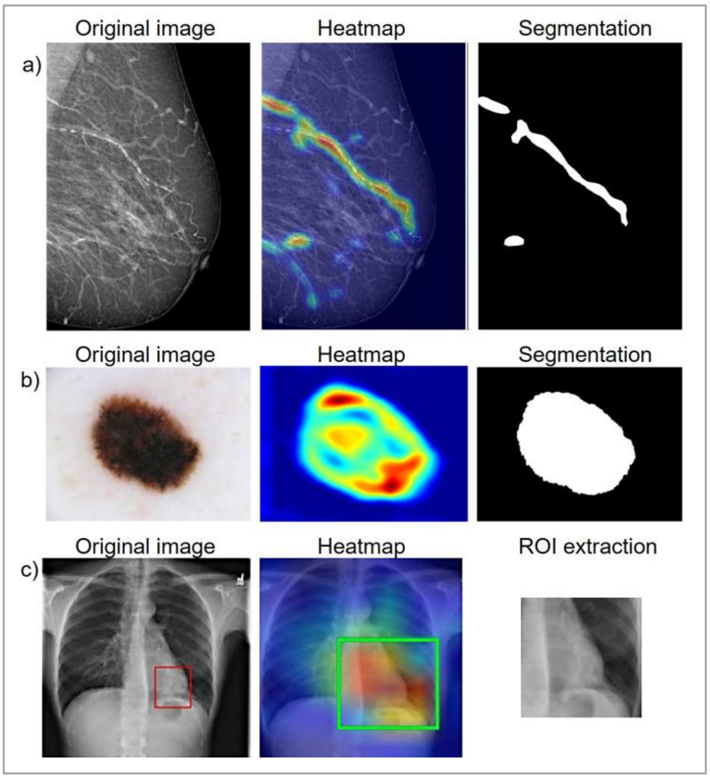


Figure 5.3 a) BACs unsupervised segmentation proposed in the present work; b) skin cancer unsupervised segmentation [91]; c) unsupervised ROI extraction of chest diseases from chest X-ray [89]

Overall, the final workflow proposed in this thesis is illustrated in Figure 5.4. Images of the four mammographic views for each patient are automatically preprocessed, and each mammogram is fed to BAC-Net. The network assesses the presence (BAC+ image) or absence (BAC- image) of BACs in the four views; the patient is considered as positive to BAC if at least one image was detected as BAC+. Secondly, for the scoring procedure, scores S_{BAC} = 0 and S_Q= 0 are assigned to all images labelled as BAC-. Conversely, heatmaps of mammograms labelled as BAC+ are processed to extract a raw continuous score S_{BAC} related to BACs extent, that could be either area, sum of pixels' intensities or length. Score S_Q from 1 to 4 is then computed evaluating the correspondence of the raw score to one of its four quartiles assessed on a wide population. Finally, the heatmaps of positive views highlighting areas where BACs were detected are presented to the radiologist, along with indications about the BAC-Net classification, the raw scores and the quartiles-based scores computed. This would allow a complete patient's assessment, without requiring any manual work for BACs detection or segmentation, leaving to the radiologist only the final decision about the need of further investigation on patient's CVD risk.

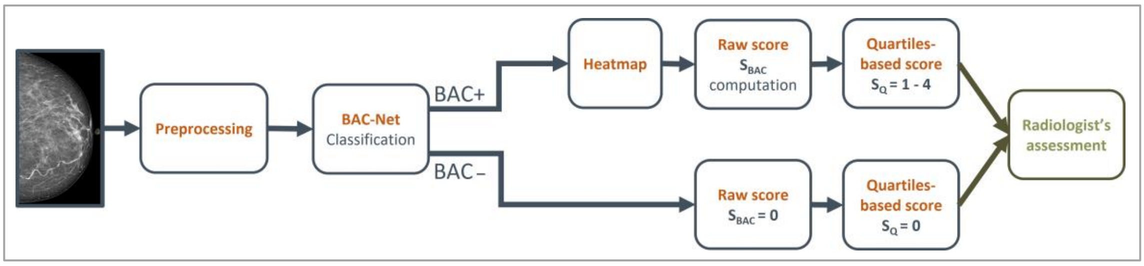


Figure 5.4 Workflow of the proposed system for automatic BACs detection and quantification. Text in orange highlights the steps developed in the present work: preprocessing, BAC-Net classification, heatmaps visualization and scoring process. The only non-automated step is the radiologist's final assessment on each image (highlighted in green), performed with the aid of heatmaps and BACs severity scores.

The final aim of the proposed workflow is BACs quantification based on BAC-Net CNN; literature reports a single case of CNN used for this aim [59], that takes in input single pixels and a patch of their surroundings, and returns as output a prediction of the probability of each pixel to belong to a BAC. Moreover, two architectures based on U-Net for automatic BACs segmentation have been proposed (named SCU-Net and DU-Net) [61], [62]. Both the CNN and U-Nets require manually segmented images as gold standard, which are time-consuming to produce. Indeed, the output of these networks is evaluated by measuring the number of pixels correctly detected as BACs when compared to ground-truth segmentation. This results in a precise extraction of BACs positions and dimensions BACs as shown in Figure 5.5. It is foreseen that such a clean visual enhancement procedure can concretely help clinicians in BAC diagnosis supported by a semiquantitative score. Indeed, at clinical level there is no need for exact BACs measurement: state-of-the-art BACs manual scoring methods are all based on semiquantitative scales (Table 1.1), that according to the radiology team contributing to this study provide enough information to evaluate CVDs risk.

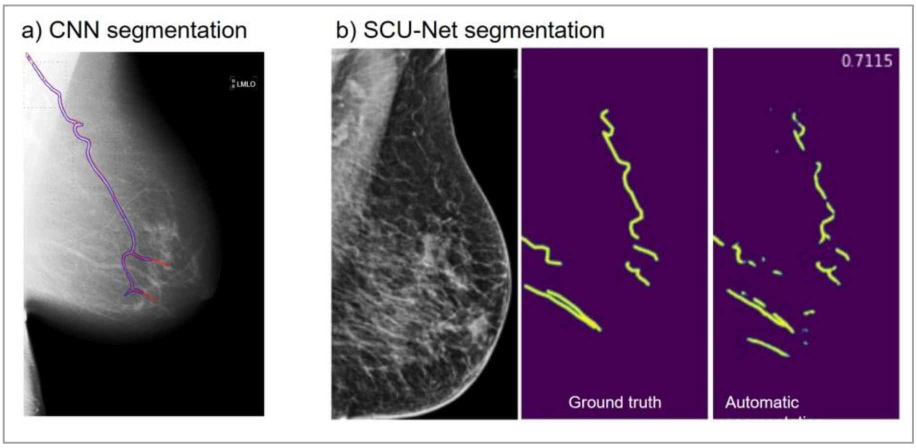


Figure 5.5 a) Segmentation performed through CNN: red lines represent ground truth segmentation, blue lines automatic segmentation [59], [62]; b) Segmentation of a mammogram (left) performed through SCU-Net (right) compared with ground truth (middle) [62]

The method here proposed shows low precision in segmenting BACs contours but allows to extract BACs scores from GradCAM++ heatmaps BAC-Net without the need of a ground truth about BACs segmentation during training. Indeed, BAC-Net is trained by providing mammograms with image-level annotations (presence/absence of BACs), that can be more easily produced: BAC-Net was trained with 1640 images, while the state-of-the-art CNN was trained with 420 images, SCU-Net with 527 images and DU-Net with 689 images. Moreover, according to Trimboli et al. [38], intra-reader agreement on BAC presence/absence is 99%, and inter-reader agreement is 98%; on the other hand, length of calcified vessels shows 87% intra-reader and 82% inter-reader agreement. This suggests that annotated data for CNN training based on BAC presence/absence only do contain a lower number of human annotation errors, compared to segmentation annotations, thus lowering gold-standard related biases.

A first limitation of this thesis is related to BAC-Net development: hyperparameters such as batch size, loss function, optimizer and errors weighting were not analysed, and regularization techniques e.g., L1 or L2 were not explored. Thus, a possible near-future development of this work should consider these factors to possibly improve network results and reduce neurons saturation. Moreover, increase in number of data used for training and testing could be beneficial, especially if mammograms were provided by different centres, including in the database images acquired from multiple devices. Continual learning [109] might also be envisioned: new data produced every day in the radiology departments could be quickly annotated as BAC+ or BAC- and provided to BAC-Net, increasing the variability of cases included in the training set, ultimately improving network performances. A second main limitation is related to the scoring procedure, that has been tested on a small dataset, and still needs to be improved. Further testing with a higher number of data will allow to fix S_{BAC}, S_Q, and efficient thresholds for classification and heatmaps binarization. Moreover, a choice between continuous $S_{BAC}$and ordinal S_Q might also be necessary for future developments. Lastly, it's clear that in order to create a software useful in clinical practice, the workflow proposed has to be unified and a graphic user interface needs to be created to allow clinicians to visualize BACs heatmaps and scoring without the engineers' assistance, that for now is required to properly run the algorithm.

It can be envisioned that, after its finalisation, the procedure here proposed might contribute not only to BACs screening in population, but also to research. Indeed, automatization of the scoring procedure allows to extract BACs severity for a higher number of cases and in a fraction of time with respect to manual scoring. With these data, the correlation of BACs automatic score with CVDs could be evaluated to ensure its efficacy as a cardiovascular risk factor. Moreover, a correlation test with coronary arteries calcifications (CAC) score would be possible. Lastly, wider applications in clinical research might be explored related to comorbidities typical in the aged population. E.g., contributing to the investigation of correlation between BAC and white matter hyperintensities [110], bone density [111] or chronic kidney disease [112].

Bibliography

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

[2] P. Hogg, J. Kelly, and C. Mercer, "Digital Mammography: A Holistic Approach," *Springer*, 2015.

[3] F. L. J. Visseren *et al.*, "2021 ESC Guidelines on cardiovascular disease prevention in clinical practice," *European Heart Journal*, vol. 42, no. 34, pp. 3227–3337, Sep. 2021, doi: 10.1093/eurheartj/ehab484.

[4] M. J. Yaffe, "Basic Physics of Digital Mammography," *Springer*, 2010.

[5] R. M. Rangayyan, T. M. Nguyen, F. J. Ayres, and A. K. Nandi, "Effect of pixel resolution on texture features of breast masses in mammograms," *Journal of Digital Imaging*, vol. 23, no. 5, pp. 547–553, Oct. 2010, doi: 10.1007/s10278-009-9238-0.

[6] J. Lian and K. Li, "A Review of Breast Density Implications and Breast Cancer Screening," *Clinical Breast Cancer*, vol. 20, no. 4. Elsevier Inc., pp. 283–290, Aug. 01, 2020. doi: 10.1016/j.clbc.2020.03.004.

[7] J.-Z. Cheng, E. B. Cole, E. D. Pisano, and D. Shen, "Detection of Arterial Calcification in Mammograms by Random Walks," 2009.

[8] P. L. Arancibia Hernández, T. Taub Estrada, A. López Pizarro, M. L. Díaz Cisternas, and C. Sáez Tapia, "Breast calcifications: Description and classification according to BI-RADS 5th edition," *Revista Chilena de Radiologia*, vol. 22, no. 2. Sociedad Chilena de Radiologia, pp. 80–91, Jun. 01, 2016. doi: 10.1016/j.rchira.2016.06.004.

[9] World Health Organization, "Global health estimates: Leading causes of death," 2019. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death (accessed Oct. 19, 2021).

[10] M. Jakab and World Health Organization. Regional Office for Europe, *Health systems respond to noncommunicable diseases : time for ambition*. World Health Organization, Regional Office for Europe, 2018.

[11] K. A. Wilmot, M. O'Flaherty, S. Capewell, E. S. Ford, and V. Vaccarino, "Coronary heart disease mortality declines in the United States from 1979 through 2011: Evidence for stagnation in young adults, especially women," *Circulation*, vol. 132, no. 11, pp. 997–1002, Sep. 2015, doi: 10.1161/CIRCULATIONAHA.115.015293.

[12] D. C. Goff *et al.*, "2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines," *Circulation*, vol. 129, no. 25 SUPPL. 1. Lippincott Williams and Wilkins, Jun. 24, 2014. doi: 10.1161/01.cir.0000437741.48606.98.

[13] R. B. D'Agostino *et al.*, "General cardiovascular risk profile for use in primary care: The Framingham heart study," *Circulation*, vol. 117, no. 6, pp. 743–753, Feb. 2008, doi: 10.1161/CIRCULATIONAHA.107.699579.

[14] P. Ridker, J. Buring, N. Rifai, and N. Cook, "Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women," *JAMA : the journal of the American Medical Association*, vol. 297, pp. 611–619, Feb. 2007, doi: 10.1001/jama.297.6.611.

[15] A. P. DeFilippis *et al.*, "An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort," *Annals of Internal Medicine*, vol. 162, no. 4, pp. 266–275, Feb. 2015, doi: 10.7326/M14-1281.

[16] E. F. van Bussel *et al.*, "Predictive value of traditional risk factors for cardiovascular disease in older people: A systematic review," *Preventive Medicine*, vol. 132. Academic Press Inc., Mar. 01, 2020. doi: 10.1016/j.ypmed.2020.105986.

[17] M. Garcia, S. L. Mulvagh, C. N. B. Merz, J. E. Buring, and J. A. E. Manson, "Cardiovascular disease in women: Clinical perspectives," *Circulation Research*, vol. 118, no. 8. Lippincott Williams and Wilkins, pp. 1273–1293, Apr. 15, 2016. doi: 10.1161/CIRCRESAHA.116.307547.

[18] A. Abuful, Y. Gidron, and Y. Henkin, "Physicians' Attitudes toward Preventive Therapy for Coronary Artery Disease: Is There a Gender Bias?," 2005. doi: 10.1002/clc.4960280809.

[19] K. K. Hyun *et al.*, "Gender inequalities in cardiovascular risk factor assessment and management in primary healthcare," 2017, doi: 10.1136/heartjnl.

[20] H. Tabenkin, C. B. Eaton, M. B. Roberts, D. R. Parker, J. H. McMurray, and J. Borkan, "Differences in cardiovascular disease risk factor management in primary care by sex of physician and patient," *Annals of Family Medicine*, vol. 8, no. 1, pp. 25–32, 2010, doi: 10.1370/afm.1071.

[21] D. K. Arnett *et al.*, "2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines," *J Am Coll Cardiol*, vol. 74, no. 10, pp. e177–e232, Sep. 2019, doi: 10.1016/j.jacc.2019.03.010.

[22] P. Greenland, M. J. Blaha, M. J. Budoff, R. Erbel, and K. E. Watson, "Coronary Calcium Score and Cardiovascular Risk," *Journal of the American College of Cardiology*, vol. 72, no. 4. Elsevier USA, pp. 434–447, Jul. 24, 2018. doi: 10.1016/j.jacc.2018.05.027.

[23] J. Vilar-Palop, J. Vilar, I. Hernández-Aguado, I. González-Alvarez, and B. Lumbreras, "Updated effective doses in radiology," *Journal of Radiological Protection*, vol. 36, no. 4, pp. 975–990, Dec. 2016, doi: 10.1088/0952-4746/36/4/975.

[24] S. M. Grundy *et al.*, "2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines," *J Am Coll Cardiol*, vol. 73, no. 24, pp. 3168–3209, Jun. 2019, doi: 10.1016/j.jacc.2018.11.002.

[25] P. Salvi, "Arterial Stiffness in Chronic Kidney Disease," in *Pulse Waves*, Springer International Publishing, 2017, pp. 199–206. doi: 10.1007/978-3-319-40501-8_7.

[26] E. J. E. Hendriks, P. A. de Jong, Y. van der Graaf, W. P. T. M. Mali, Y. T. van der Schouw, and J. W. J. Beulens, "Breast arterial calcifications: A systematic review and meta-analysis of their determinants and their association with cardiovascular events," *Atherosclerosis*, vol. 239, no. 1. Elsevier Ireland Ltd, pp. 11–20, Mar. 01, 2015. doi: 10.1016/j.atherosclerosis.2014.12.035.

[27] N. Loberant, V. Salamon, N. Carmi, and A. Chernihovsky, "Prevalence and degree of breast arterial calcifications on mammography: A cross-sectional analysis," *Journal of Clinical Imaging Science*, vol. 3, no. 1, 2013, doi: 10.4103/2156-7514.119013.

[28] Q. M. Bui and L. B. Daniels, "A Review of the Role of Breast Arterial Calcification for Cardiovascular Risk Stratification in Women," *Circulation*, vol. 139, no. 8. Lippincott Williams and Wilkins, pp. 1094–1101, Feb. 19, 2019. doi: 10.1161/CIRCULATIONAHA.118.038092.

[29] S. C. Lee, M. Phillips, J. Bellinge, J. Stone, E. Wylie, and C. Schultz, "Is breast arterial calcification associated with coronary artery disease?—A systematic review and meta-analysis," *PLoS ONE*, vol. 15, no. 7 July. Public Library of Science, Jul. 01, 2020. doi: 10.1371/journal.pone.0236598.

[30] L. Margolies *et al.*, "Digital Mammography and Screening for Coronary Artery Disease," *JACC: Cardiovascular Imaging*, vol. 9, no. 4, pp. 350–360, Apr. 2016, doi: 10.1016/j.jcmg.2015.10.022.

[31] Eurostat, "Healthcare activities statistics-preventive services Statistics Explained Breast cancer screening," 2020. [Online]. Available: https://ec.europa.eu/eurostat/statisticsexplained/

[32] National Center for Health Statistics, "Health, United States 2019," 2019, [Online]. Available: https://www.cdc.gov/nchs/hus/contents2019.htm#Table-033

[33] J. L. Rodgers *et al.*, "Cardiovascular Risks Associated with Gender and Aging," *Journal of Cardiovascular Development and Disease*, vol. 6, no. 2, p. 19, Apr. 2019, doi: 10.3390/jcdd6020019.

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

[35] L. Mostafavi *et al.*, "Prevalence of coronary artery disease evaluated by coronary CT angiography in women with mammographically detected breast arterial calcifications," *PLoS ONE*, vol. 10, no. 4, Apr. 2015, doi: 10.1371/journal.pone.0122289.

[36] D. Ružičić *et al.*, "The correlation of SYNTAX score by coronary angiography with breast arterial calcification by digital mammography," *Clinical Radiology*, vol. 73, no. 5, pp. 454–459, May 2018, doi: 10.1016/j.crad.2017.12.002.

[37] S. Molloi, T. Mehraien, C. Iribarren, C. Smith, J. L. Ducote, and S. A. Feig, "Reproducibility of Breast Arterial Calcium Mass Quantification Using Digital Mammography," *Academic Radiology*, vol. 16, no. 3, pp. 275–282, Mar. 2009, doi: 10.1016/j.acra.2008.08.011.

[38] R. M. Trimboli *et al.*, "Semiquantitative score of breast arterial calcifications on mammography (BAC-SS): intra- and inter-reader reproducibility," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 5, pp. 2019–2027, May 2021, doi: 10.21037/qims-20-560.

[39] S. Molloi, T. Xu, J. Ducote, and C. Iribarren, "Quantification of breast arterial calcification using full field digital mammography," *Medical Physics*, vol. 35, no. 4, pp. 1428–1439, 2008, doi: 10.1118/1.2868756.

[40] G. Sianos *et al.*, “The SYNTAX Score: An angiographic tool grading the complexity of coronary artery disease,” *EuroIntervention*, vol. 1, pp. 219–227, Nov. 2005.

[41] “Scopus - Analyze search results | Signed in.” https://www.scopus.com/term/analyzer.uri?sid=ad7cfd8cdf441356741ab30878c2c99a&origin=resultslist&src=s&s=TITLE-ABS-KEY%28%28%27artificial+intelligence%27+OR+%27artificial+intelligence%27+OR+%27machine+learning%27+OR+%27machine+learning%27+OR+%27deep+learning%27+OR+%27deep+learning%27%29+AND+%28%27radiology%27+OR+%27radiology%27+OR+%27diagnostic+imaging%27+OR+%27diagnostic+imaging%27%29%29&sort=plf-f&sdt=a&sot=a&sl=234&count=3513&analyzeResults=Analyze+results&txGid=adbb7f16f576b7d857533a55d54600cc (accessed Oct. 11, 2021).

[42] Philip Ward, Erik L. Ridley, Wayne Forrest, and Brian Casey, “Top 5 trends from ECR 2019 in Vienna,” 2019. https://www.auntminnieeurope.com/index.aspx?sec=rca&sub=ecr_2019&pag=dis&ItemID=617117 (accessed Oct. 11, 2021).

[43] K. G. van Leeuwen, S. Schalekamp, M. J. C. M. Rutten, B. van Ginneken, and M. de Rooij, “Artificial intelligence in radiology: 100 commercially available products and their scientific evidence,” 2021, doi: 10.1007/s00330-021-07892-z/Published.

[44] Russel Stuart and Norvig Peter, “Artificial Intelligence A Modern Approach, 3rd Edition,” 2010.

[45] I. Castiglioni *et al.*, “AI applications to medical images: From machine learning to deep learning,” *Physica Medica*, vol. 83. Associazione Italiana di Fisica Medica, pp. 9–24, Mar. 01, 2021. doi: 10.1016/j.ejmp.2021.02.006.

[46] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, May 27, 2015. doi: 10.1038/nature14539.

[47] R. J. McDonald *et al.*, “The Effects of Changes in Utilization and Technological Advancements ofCross-Sectional Imaging onRadiologist Workload,” *Academic Radiology*, vol. 22, no. 9, pp. 1191–1198, Sep. 2015, doi: 10.1016/j.acra.2015.05.007.

[48] S. Jha and E. J. Topol, “Adapting to artificial intelligence: Radiologists and pathologists as information specialists,” *JAMA - Journal of the American Medical Association*, vol. 316, no. 22. American Medical Association, pp. 2353–2354, Dec. 13, 2016. doi: 10.1001/jama.2016.17438.

[49] F. Pesapane, M. Codari, and F. Sardanelli, “Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine,” *European Radiology Experimental*, vol. 2, no. 1. Springer, Dec. 01, 2018. doi: 10.1186/s41747-018-0061-6.

[50] P. Lambin *et al.*, “Radiomics: The bridge between medical imaging and personalized medicine,” *Nature Reviews Clinical Oncology*, vol. 14, no. 12. Nature Publishing Group, pp. 749–762, Dec. 01, 2017. doi: 10.1038/nrclinonc.2017.141.

[51] M. R. Chetan and F. v. Gleeson, “Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives,” *European Radiology*, vol. 31, no. 2. Springer Science and Business Media Deutschland GmbH, pp. 1049–1058, Feb. 01, 2021. doi: 10.1007/s00330-020-07141-9.

[52] A. S. Tagliafico, M. Piana, D. Schenone, R. Lai, A. M. Massone, and N. Houssami, “Overview of radiomics in breast cancer diagnosis and prognostication,” *Breast*, vol. 49, pp. 74–80, Feb. 2020, doi: 10.1016/j.breast.2019.10.018.

[53] D. Dey and F. Commandeur, “Radiomics to Identify High-Risk Atherosclerotic Plaque from Computed Tomography: The Power of Quantification,” *Circulation: Cardiovascular Imaging*, vol. 10, no. 12. Lippincott Williams and Wilkins, Dec. 01, 2017. doi: 10.1161/CIRCIMAGING.117.007254.

[54] M. Kolossváry, M. Kellermayer, B. Merkely, and P. Maurovich-Horvat, “Cardiac Computed Tomography Radiomics,” in *Journal of Thoracic Imaging*, 2018, vol. 33, no. 1, pp. 26–34. doi: 10.1097/RTI.0000000000000268.

[55] P. Omoumi *et al.*, “To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines),” *Eur Radiol*, vol. 31, no. 6, pp. 3786–3796, Jun. 2021, doi: 10.1007/s00330-020-07684-x.

[56] D. B. Larson, H. Harvey, D. L. Rubin, N. Irani, J. R. Tse, and C. P. Langlotz, “Regulatory Frameworks for Development and Evaluation of Artificial Intelligence–Based Diagnostic Imaging Algorithms: Summary and Recommendations,” *Journal of the American College of Radiology*, vol. 18, no. 3, pp. 413–424, Mar. 2021, doi: 10.1016/j.jacr.2020.09.060.

[57] Frangi Alejandro F., W. J. Niessen, Vincken Koen L., and Viergever Max A., “Multiscale vessel enhancement filtering,” in *Medical Image Computing and Computer-Assisted Intervention — MICCAI’98*, 1998, pp. 130–137.

[58] J.-Z. Cheng, C.-M. Chen, and D. Shen, “Identification of breast vascular calcium deposition in digital mammography by linear structure analysis,” in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, 2012, pp. 126–129. doi: 10.1109/ISBI.2012.6235500.

[59] J. Wang *et al.*, “Detecting Cardiovascular Disease from Mammograms with Deep Learning,” *IEEE Transactions on Medical Imaging*, vol. 36, no. 5, pp. 1172–1181, May 2017, doi: 10.1109/TMI.2017.2655486.

[60] Md. Z. Alom, M. Hasan, C. Yakopcic, T. Taha, and V. Asari, “Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation,” Nov. 2018.

[61] M. Alghamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, “DU-Net: Convolutional Network for the Detection of Arterial Calcifications in Mammograms,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3240–3249, Oct. 2020, doi: 10.1109/TMI.2020.2989737.

[62] X. Guo *et al.*, “SCU-Net: A deep learning method for segmentation and quantification of breast arterial calcifications on mammograms,” *Medical Physics*, vol. 48, no. 10, pp. 5851–5861, 2021, doi: https://doi.org/10.1002/mp.15017.

[63] W. S. Mcculloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” 1943.

[64] A. Aggarwal, M. Mittal, and G. Battineni, “Generative adversarial network: An overview of theory and applications,” *International Journal of Information Management*, p. 100004, Jan. 2021, doi: 10.1016/j.jjimei.2020.100004.

[65] A. C. Tsoi and A. D. Back, “Discrete time recurrent neural network architectures: A unifying review,” *Neurocomputing*, vol. 15, pp. 183–223, 1997.

[66] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[67] D. Purves, “Neuroscience, 3rd Edition,” 2004.

[68] V. Nair and G. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair,” in *Proceedings of ICML*, Nov. 2010, vol. 27, pp. 807–814.

[69] A. L. Maas, “Rectifier Nonlinearities Improve Neural Network Acoustic Models,” 2013.

[70] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical Evaluation of Rectified Activations in Convolutional Network,” May 2015, [Online]. Available: http://arxiv.org/abs/1505.00853

[71] A. E. Guissous, “Skin Lesion Classification Using Deep Neural Network.” Feb. 2019.

[72] E. A. Badr, C. Joun, and G. E. Nasr, “Cross Entropy Error Function in Neural Networks: Forecasting Gasoline Demand,” 2002. [Online]. Available: www.aaai.org

[73] S. Ruder, “An overview of gradient descent optimization algorithms,” Sep. 2016, [Online]. Available: http://arxiv.org/abs/1609.04747

[74] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.6980

[75] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10. pp. 1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

[76] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks.” [Online]. Available: http://code.google.com/p/cuda-convnet/

[77] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, 2012, vol. 25. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[78] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *CoRR*, vol. abs/1512.03385, 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[79] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A Survey on Deep Transfer Learning,” in *Artificial Neural Networks and Machine Learning – ICANN 2018*, 2018, pp. 270–279.

[80] N. Tajbakhsh *et al.*, “Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, doi: 10.1109/TMI.2016.2535302.

[81] A. Khamparia *et al.*, “Diagnosis of breast cancer based on modern mammography using hybrid transfer learning,” *Multidimensional Systems and Signal Processing*, vol. 32, no. 2, pp. 747–765, Apr. 2021, doi: 10.1007/s11045-020-00756-7.

[82] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, “Explainable artificial intelligence: an analytical review,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, Sep. 2021, doi: 10.1002/widm.1424.

[83] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, “What do we need to build explainable AI systems for the medical domain?,” Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.09923

[84] P. Angelov and E. Soares, “Towards explainable deep neural networks (xDNN),” *Neural Networks*, vol. 130, pp. 185–194, Oct. 2020, doi: 10.1016/j.neunet.2020.07.010.

[85] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,” Dec. 2013, [Online]. Available: http://arxiv.org/abs/1312.6034

[86] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” Oct. 2016, doi: 10.1007/s11263-019-01228-7.

[87] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, May 2018, vol. 2018-January, pp. 839–847. doi: 10.1109/WACV.2018.00097.

[88] Y. J. Suh, J. Jung, and B. J. Cho, “Automated breast cancer detection in digital mammograms of various densities via deep learning,” *Journal of Personalized Medicine*, vol. 10, no. 4, pp. 1–11, Nov. 2020, doi: 10.3390/jpm10040211.

[89] T. C. Lin and H. C. Lee, “Covid-19 chest radiography images analysis based on integration of image preprocess, guided grad-CAM, machine learning and risk management,” in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, Aug. 2020, pp. 281–288. doi: 10.1145/3418094.3418096.

[90] M. Hinata and T. Ushiku, “Detecting immunotherapy-sensitive subtype in gastric cancer using histologic image-based deep learning.,” *Sci Rep*, vol. 11, no. 1, p. 22636, Nov. 2021, doi: 10.1038/s41598-021-02168-4.

[91] F. Nunnari, M. A. Kadir, and D. Sonntag, “On the Overlap Between Grad-CAM Saliency Maps and Explainable Visual Features in Skin Cancer Images,” in *Machine Learning and Knowledge Extraction*, 2021, pp. 241–253.

[92] M. Jahanifar, N. Z. Tajeddin, N. A. Koohbanani, A. Gooya, and N. M. Rajpoot, “Segmentation of Skin Lesions and their Attributes Using Multi-Scale Convolutional Neural Networks and Domain Specific Augmentations,” *arXiv: Computer Vision and Pattern Recognition*, 2018.

[93] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, “Thorax disease classification with attention guided convolutional neural network,” *Pattern Recognition Letters*, vol. 131, pp. 38–45, Mar. 2020, doi: 10.1016/j.patrec.2019.11.040.

[94] Y. and L. L. and L. Z. and B. M. and S. R. M. Wang Xiaosong and Peng, “ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases,” in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, X. and C. G. and Y. L. Lu Le and Wang, Ed. Cham: Springer International Publishing, 2019, pp. 369–392. doi: 10.1007/978-3-030-13969-8_18.

[95] Z. Li *et al.*, “Thoracic Disease Identification and Localization with Limited Supervision,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8290–8299. doi: 10.1109/CVPR.2018.00865.

[96] N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979, doi: 10.1109/TSMC.1979.4310076.

[97] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[98] H. Hernandez, “Testing for Normality: What is the Best Method?,” 2021. [Online]. Available: www.forschem.org

[99] M. Farzan, M. Ezati Asar, and M. Hosseini, “Common Statistical Mistakes in Descriptive Statistics Reports of Normal and Non-Normal Variables in Biomedical Sciences Research,” *Iranian Journal of Public Health*, vol. 44, pp. 1557–1558, Jan. 2015.

[100] M. M. Rahman and D. N. Davis, “Addressing the Class Imbalance Problem in Medical Datasets,” *International Journal of Machine Learning and Computing*, pp. 224–228, 2013, doi: 10.7763/ijmlc.2013.v3.307.

[101] C. V. K. Veni and T. S. Rani, “Quartiles based UnderSampling(QUS): A Simple and Novel Method to increase the Classification rate of positives in Imbalanced Datasets,” in *2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR)*, 2017, pp. 1–6. doi: 10.1109/ICAPR.2017.8593202.

[102] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” *Journal of Machine Learning Research - Proceedings Track*, vol. 9, pp. 249–256, Jan. 2010.

[103] I. Loshchilov and F. Hutter, “SGDR: Stochastic Gradient Descent with Warm Restarts,” Aug. 2016, [Online]. Available: http://arxiv.org/abs/1608.03983

[104] G. Montavon, G. Orr, and K.-R. Müller, *Neural Networks: Tricks of the Trade: Second Edition*. 2012. doi: 10.1007/978-3-642-35289-8.

[105] K. You, M. Long, J. Wang, and M. I. Jordan, “How Does Learning Rate Decay Help Modern Neural Networks?,” Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.01878

[106] L. Garrido, S. Gòmez, V. Gaitàn, and M. Serra-Ricart, “A regularization term to avoid the saturation of the sigmoids in multilayer neural networks,” *International Journal of Neural Systems*, vol. 07, no. 03, pp. 257–262, Jul. 1996, doi: 10.1142/S0129065796000233.

[107] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” 2014.

[108] L. Myers and M. J. Sirois, “Spearman Correlation Coefficients, Differences between,” in *Encyclopedia of Statistical Sciences*, John Wiley & Sons, Ltd, 2006. doi: https://doi.org/10.1002/0471667196.ess5050.pub2.

[109] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu, “Embracing Change: Continual Learning in Deep Neural Networks,” *Trends in Cognitive Sciences*, vol. 24, no. 12, pp. 1028–1040, Dec. 2020, doi: 10.1016/J.TICS.2020.09.004.

[110] F. Moroni, E. Ammirati, M. A. Rocca, M. Filippi, M. Magnoni, and P. G. Camici, “Cardiovascular disease and brain health: Focus on white matter hyperintensities,” *Int J Cardiol Heart Vasc*, vol. 19, pp. 63–69, May 2018, doi: 10.1016/j.ijcha.2018.04.006.

[111] J. Reddy, J. P. Bilezikian, S. J. Smith, and L. Mosca, "Reduced bone mineral density is associated with breast arterial calcification," *J Clin Endocrinol Metab*, vol. 93, no. 1, pp. 208–211, Jan. 2008, doi: 10.1210/jc.2007-0693.

[112] V. Duhn, E. T. D'Orsi, S. Johnson, C. J. D'Orsi, A. L. Adams, and W. C. O'Neill, "Breast arterial calcification: a marker of medial vascular calcification in chronic kidney disease," *Clin J Am Soc Nephrol*, vol. 6, no. 2, pp. 377–382, Feb. 2011, doi: 10.2215/CJN.07190810.

List of Figures

List if Tables

Acknowledgments

Un sentito ringraziamento a tutte le persone che mi hanno permesso di arrivare fin qui e di portare a termine questo percorso.

Grazie al Professor Baselli e a Nazanin, per la grande disponibilità nell'ascoltare le mie idee e i miei dubbi e per i preziosi consigli senza i quali non avrei potuto portare a conclusione questo progetto.

Ringrazio il Professor Sardanelli, Davide, e tutto il gruppo di radiologia dell'ospedale San Donato per avermi permesso di lavorare in un ambiente così stimolante, per la professionalità, e per la fiducia che mi è stata dimostrata.

Un grazie sentito ai miei genitori, a mio fratello, ai nonni e a tutta la mia famiglia per avermi sempre sostenuta in ogni modo possibile e per avere creduto in me durante questi anni.

Infine, ringrazio Federico ed i miei amici, che mi hanno supportata in questo percorso. Grazie per avermi aiutata non solo a tenere duro nei momenti difficili, ma anche a prendermi del tempo per rilassarmi quando necessario.