UNIVERSITÀ
DEGLI STUDI
DI MILANO

Department of Biomedical Sciences for Health
Doctoral Program in Translational Medicine
R46–XXXVI Cycle

Artificial intelligence for detection and quantification of breast arterial calcifications on mammograms as a biomarker of cardiovascular disease

PhD Thesis of
Nazanin MOBINI
Matr. n. **R12981**

Supervisor: Prof. **Francesco SARDANELLI**
Co-supervisor: Prof. **Giuseppe BASELLI**
PhD Coordinator: Prof. **Chiarella SFORZA**

Academic Year 2022-2023

Table of Contents

Abbreviations

AI, artificial intelligence

ANN, artificial neural network

ASCVD, atherosclerotic cardiovascular disease

AUC-PR, area under the precision-recall curve

AUC-ROC, area under the receiver operating characteristic curve

BAC, breast arterial calcifications

BCE, binary cross-entropy loss

CAC, coronary artery calcium

CC, cranio-caudal

CI, confidence interval

CNN, convolutional neural network

CVD, cardiovascular diseases

DBT, digital breast tomosynthesis

DenseNet, dense convolutional network

DL, deep learning

DT, decision tree

FC, fully connected

FFDM, full-field digital mammogram

FN, false negative

FP, false positive

FROC, free-response receiver operating characteristic

GAN, generative adversarial network

GPU, graphic processing unit

Grad-CAM++, generalized gradient-weighted class activation mapping

IoU, intersection over union

IQR, interquartile range

ML, machine learning

MLO, medio-lateral oblique

PCA, principal component analysis

ReLU, rectified linear unit

ResNet, residual network

SD, standard deviation

SVM, support vector machine

TL, transfer learning

TN, true negative

TP, true positive

TPR, true positive rate

U-Net, u-shaped encoder-decoder network

VGG, visual geometry group

VRAM, video random access memory

Xception, extreme inception

Summary

Cardiovascular diseases (CVD) represent the leading cause of morbidity and mortality worldwide, imposing a significant healthcare and economic burden. Current risk stratification scores as the main approach for assessing cardiovascular health often underestimate the risk in female population, leading to missed opportunities for primary prevention, early diagnosis, appropriate treatment, and ultimately contributing to the elevated cardiovascular disease burden. This gender-based disparity has prompted the development of innovative sex-specific predictors that could improve women's CVD risk stratification. In this light, breast arterial calcifications (BAC) have gained traction as one of the most promising women-specific biomarkers: BAC are localized Mönckeberg sclerosis expression involving within the tunica media of breast arteries and detectable as parallel line opacities on about 13% of routine mammograms. They have been shown to be associated with an elevated hazard of cardiovascular adverse events, more accurate than other traditional risk factors in asymptomatic middle-aged women, and also independent of them, indicating the different pathogenesis of BAC from that of atherosclerotic plaques. Considering the widespread diffusion of mammography breast cancer screening programs, systematic BAC assessment could offer a cost-effective cardiovascular risk stratification in women without additional examinations. However, their assessment is a challenging and time-consuming manual task, vulnerable to intra- and inter-observer variability; also, the considerable diversity of BAC's appearance and the lack of a standard reporting guideline or a reliable quick quantification method have limited their adoption as a robust imaging biomarker in clinical practices. Automated methods using artificial intelligence (AI) and deep learning (DL) algorithms hold promise in addressing the limitations, improving diagnostic reproducibility, reducing radiologists' post-processing workload, and facilitating broader utilization of BAC to improve cardiovascular risk

stratification in women and promote awareness of their cardiovascular health, leveraging the large-scale mammographic screening programs. Accordingly, this thesis will present an overview of the current state of knowledge on the automatic BAC assessment using AI-based algorithms (section I), propose a novel DL-based approach for detection and estimation of BAC burden (section II), and subsequently explore the method by a comparative analysis with other established CNN architectures (section III).

Section I:

Introduction to BAC as a biomarker of cardiovascular disease and applications of AI in automated detection

Based on:

- N Mobini, D Capra, G Baselli, and F Sardanelli. Role of deep learning in detecting breast arterial calcifications: a narrative review. *In Submission,*
- V Magni, D Capra D, A Cozzi, CB Monti, N Mobini, A Colarieti, and F Sardanelli. Mammography biomarkers of cardiovascular and musculoskeletal health: A review. *Maturitas* (2023)

  DOI: 10.1016/j.maturitas.2022.10.001

Breast arterial calcifications and cardiovascular risk

Cardiovascular diseases (CVD) are the leading cause of morbidity and mortality worldwide, imposing a substantial healthcare and economic burden [1], [2]. Despite the common belief that mostly regarded CVD as a male affliction [3], [4], 45% of all women's deaths in Europe are attributed to heart diseases, whereas men have a relatively lower CVD death rate of 39% [2]. Indeed, even though oestrogen has a protective role against CVD during the fertile age [5], this protection tends to vanish during the menopause transition, thus contributing to increase CVD risk, together with other adverse physiological and metabolic changes occurring in this period, such as alterations in body composition, lipid profile, and vascular function [6]. Furthermore, female-specific risk factors strictly related to reproductive life (such as preterm delivery, hypertensive pregnancy disorders, and gestational diabetes mellitus) might contribute to the worsening of CVD risk profiles, especially in young women.

Even though the awareness about CVD in women has increased during the past decades with a corresponding decline in female CVD mortality (in Europe, from 374 to 209 deaths per 100000 in the period between 1985 and 2014) [7], both women and primary care physicians still have a tendency to underestimate this risk of developing CVD, increasing the disparity between men and women in the prevention, diagnosis, and treatment of CVD [8], [9]. Even the most updated prediction models used to estimate the risk of fatal and nonfatal CVD apply age- and sex-specific multipliers without including risk factors specific to the female sex, further limiting the development of sex-specific strategies for the primary prevention of CVD [9].

In Europe, breast cancer awareness campaigns have been crucial to highlight the importance and efficacy of early diagnosis through mammographic screening [10], achieving satisfactory attendance rates in the majority of organized screening programs [11]. However, alongside the early detection of breast cancer, mammography has been reported to be useful for the identification of ancillary features unrelated to oncological disease, such as mammographic breast arterial calcifications (BAC), which have been recognized as important biomarkers of cardiometabolic risk [12]. This progressive awareness is paving the way towards an extension of the preventive role of mammography beyond breast cancer screening, acknowledging its potential to offer an insight into women's cardiometabolic health.

In the context of breast cancer screening mammography reading, calcifications are classified either as typically benign or of suspicious morphology: the former are discarded, while the latter prompt second-level investigations. However, some calcifications (which are considered as surely benign) carry information about women's cardiovascular health. Specifically, BAC are a local expression of Mönckeberg sclerosis appear as parallel or tubular opacities associated with blood vessels and evolve within the tunica media and the internal elastic lamina of large and medium-sized arteries [13], [14], [15], which have been associated with cardiovascular risk for more than two decades [16]. Mönckeberg sclerosis is a histopathologic entity distinct from atherosclerosis involving coronary arteries, related to a pro-osteogenic environment, with the deposition of hydroxyapatite crystals in conditions of altered mineral metabolism [17] while atheromatic plaques are characterized by macrophagic activation and cholesterol deposition. In fact, no signs of inflammation were found in BAC plaques by histologic studies [15], [18]. It is supposed that calcified vessels become stiffer, leading to increased pulse pressure that could lead to CVD [19]. Indeed, postmenopausal women with BAC included in a substudy of the MINERVA (multiethnic study of breast arterial calcium gradation and CVD) cohort [20], had an odds ratio of 1.36 (95% CI 1.01–1.87, $p = 0.04$) for having an ankle-brachial index < 0.90, a marker of peripheral artery disease [21]. The authors however did not observe any significant association between BAC severity and peripheral artery disease, perhaps because of the relatively small size of the BAC positive group. Conversely, such quantitative association was reported in a previous case-cohort study by Hendriks et al. [22], who reported a hazard ratio of 2.93 (95% CI 1.05–8.16) for peripheral artery disease compared to women without BAC.

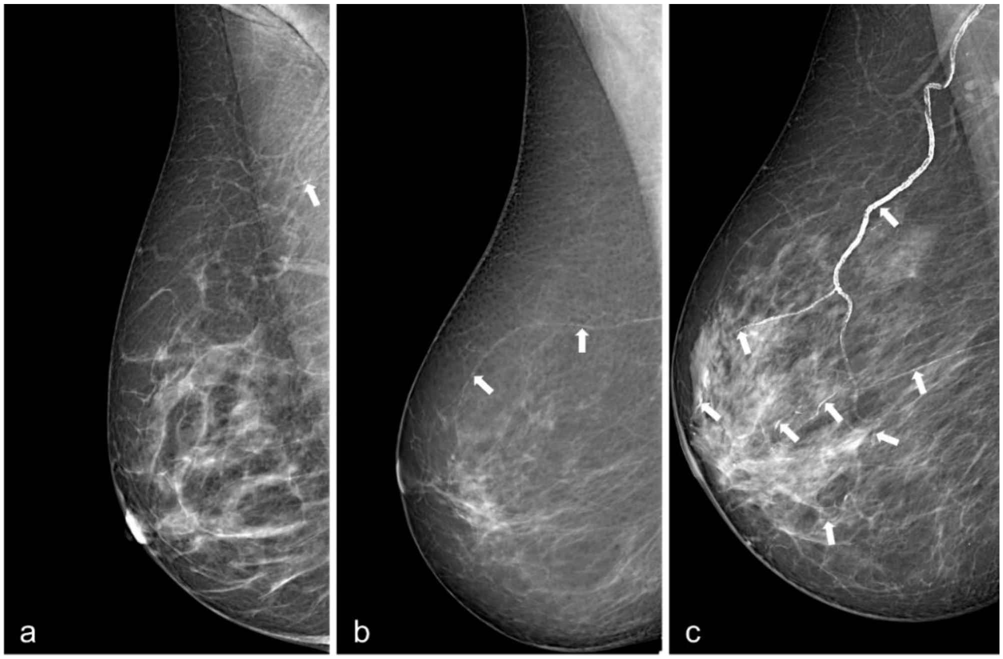


Figure 1 Examples of breast arterial calcifications on screening mammograms (white arrows). **a** Low, **b** mild, and **c** severe burden of BAC [23]

BAC are a relatively common incidental finding, observed in around 13% of mammograms (**Figure 1**), their most important predictors being increasing age, diabetes, and parity [24]. Furthermore, hormonal levels seem to impact BAC, as BAC prevalence rises after menopause [24], while it is reduced by 50% in women aged over 65 years under hormonal replacement therapy [25]. BAC are associated with hypertension (pooled odds ratio 1.80, 95% CI 1.47–2.21), but not with other established cardiovascular risk factors, such as hypercholesterolemia (odds ratio 1.31, 95% CI 0.97–1.77), and present a negative association with smoking habit (odds ratio 0.54, 95% CI 0.42–0.70) [26], which again underlines the distinct pathologic pathway that leads to BAC pathogenesis.

Nevertheless, BAC presence is significantly associated with coronary artery disease (odds ratio 2.61, 95% CI 2.12–3.21), and women with a moderate to severe BAC load have a 2.95 odds ratio (95% CI 1.49–5.84) for coronary artery disease. A retrospective study published by Margolies et al. [27] in 2016 found a strong, quantitative association between BAC and coronary artery disease, with the incidence of higher BAC scores increasing accordingly to coronary artery calcium (CAC) score measured at coronary computed tomography. Furthermore, BAC scores from 4 to 12 (representing a marked BAC burden) had an adjusted odds ratio of 3.2 (95% CI 1.8–5.9) for the presence of coronary artery calcium. Moreover, a BAC score > 0 showed an equivalent area under the receiving operator curve to that of Framingham risk score for the detection of CAC. A subsequent retrospective cohort study by Yoon et al. [28] confirmed the association between BAC presence and BAC score to subclinical coronary artery calcium, with adjusted odds ratios of 2.87 (95% CI 1.67–4.93) and 1.20 (95% CI 1.10–1.31) respectively. They also confirmed the prognostic value of BAC assessment, showing net reclassification improvements after adding BAC presence to the 10-years atherosclerotic cardiovascular disease−ASCVD risk score calibrated for the Korean population, with a net reclassification index of 0.052, and significant, albeit small improvement of the AUC from 0.66 to 0.68 (p = 0.010) for the presence of coronary arteries plaques. The recently published results from the MINERVA cohort study [29], conducted on women aged between 60 and 79, reported that women with BAC have a 1.51 (95% CI 1.08–2.11) increased hazard of hard atherosclerotic CVD events (acute myocardial infarction, ischemic stroke, CVD death), and a 1.23 (95% CI 1.00–1.52) increased hazard of global CVD events. Iribarren and colleagues also evaluated the performances of the American College of Cardiology/American Heart Association Pooled Cohort Equations for atherosclerotic CVD risk assessment combined with the presence of BAC, significantly improving its performances, with a net reclassification index of 0.11. A previous prospective cohort study [30]on 1454 women with a 5-year follow up reached similar conclusions, reporting a significantly higher likelihood of developing coronary artery disease for women with BAC than those without (6.3% vs 2.3%, p = 0.003) (**Figure 2**).

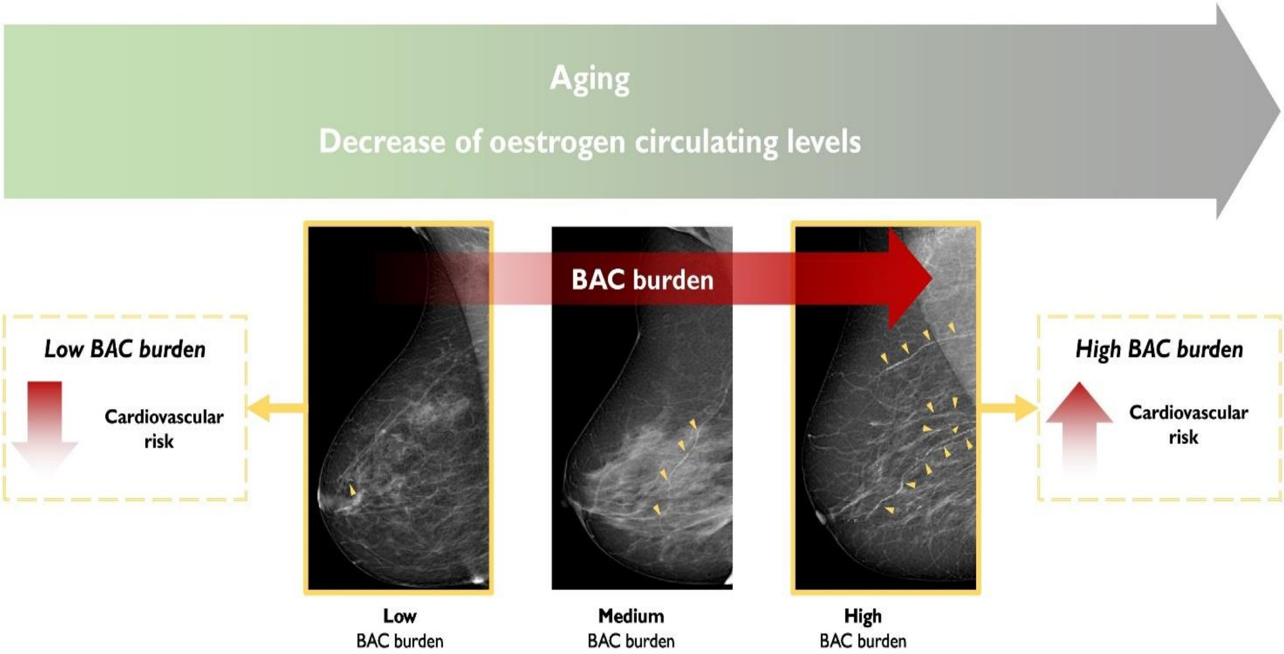


Figure 2 An overview of how modifications in breast arterial calcifications burden influence cardiovascular risk. Yellow arrowheads in mammographic images indicate BAC [31]

The aforementioned results and the widespread diffusion of screening mammography [10], [11], advocate for the inclusion of BAC in cardiovascular risk scores [32], particularly for postmenopausal women. Nonetheless, despite the fact that an overwhelming majority of women would prefer to be informed about their BAC status [33], BAC are currently not integrated in CVD risk prevention strategies and even BAC reporting in mammography interpretation is still scarce [34]: although over 80% of European breast radiologists declared they are aware of the association between BAC and CV risk, less than 65% of them report on BAC. Moreover, most of the radiologists who report BAC merely describe them as present, while just over 25% of radiologists use an ordinal visual scale for BAC evaluation and only one radiologist uses a quantitative assessment. Indeed, there are a few issues hindering a more widespread adoption of BAC detection and reporting in routine clinical practice.

One pivotal obstacle in BAC assessment lies in the time needed to evaluate them. Indeed, if spotting BAC presence may be considered relatively immediate (excepting the case of small, tiny calcifications not definable as surely being BAC), measuring their extension may be a painstaking process. In fact, quantification methods based on manual measurements may take up to 3 minutes per mammogram [35], which would put further strain on radiologists, especially in the case of screening reading. In addition, methods based on subjective visual assessment may not ensure optimal reproducibility.

Several quantification methods have been proposed over the years, from 4-points Likert scales [36] and 12-points semiquantitative scales [37], [35] to quantitative scores that evaluate the calcium mass performing a densitometry using carefully calibrated mammography systems, as in the MINERVA study [38]. However, the necessity of calibrating mammography systems clashes with the potential immediate application of BAC evaluation in the context of the available mammography systems already employed for routine breast cancer screening and clinical assessment.

Artificial Intelligence for BAC assessment

The term Artificial Intelligence (AI) was coined in the 1950s and refers to a broad field of computational science (**Figure 3**) focused on developing automatic systems capable of imitating humanlike intelligence and behaviours such as learning, problem-solving, and natural language understanding to perform various tasks [39], [40], [41], [42]. Machine learning (ML), the key subfield of AI, involves algorithms and statistical models that can learn from relevant extracted features, identify patterns, and make decisions based on that learning without external reprogramming [43], [44]. The main categories in ML techniques include supervised learning (training on labelled data), unsupervised learning (finding patterns in unlabelled data), semi-supervised learning (dealing with semi-labelled data to minimize annotation requirements), and reinforcement learning (learning from interactions with an environment) [44], [45]. Decision Trees (DT) [46], Support Vector Machine (SVM) [47], and Principal Component Analysis (PCA) [48] are some representatives of classic ML models.

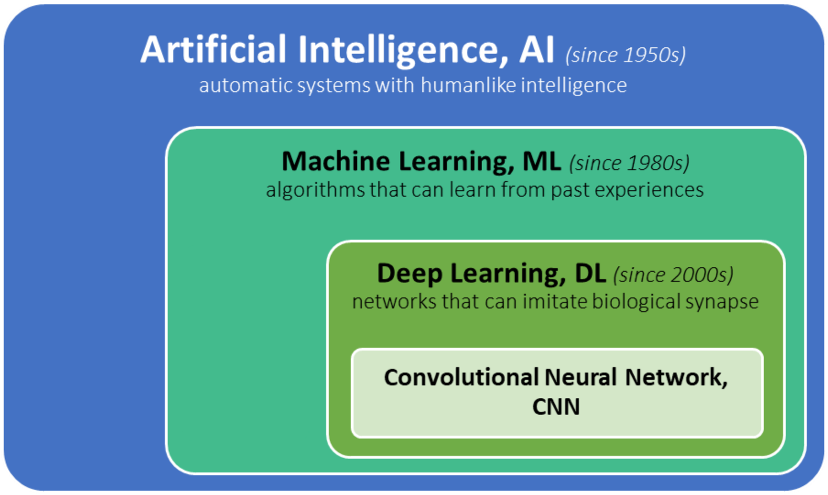

Artificial Intelligence, AI (since 1950s)
automatic systems with humanlike intelligence
Machine Learning, ML (since 1980s)
algorithms that can learn from past experiences
Deep Learning, DL (since 2000s)
networks that can imitate biological synapse
Convolutional Neural Network, CNN

Figure 3 The broad AI family. Since 2010s and the evolution of technologies, CNNs have become the leading architecture for most image classification tasks

Moving forward to the early 2000s, the evolution of technologies, growth of big data, and availability of open-source algorithms have shifted AI from research to practical applications [49], [50], where a specialized subfield of ML called Deep learning (DL) has outperformed its predecessors [51], [52]. These networks, also known as Artificial Neural Networks (ANNs), inspired by the hierarchical structure of the biological synapse system are composed of multi-layer interconnected artificial neurons organized into input, hidden, and output layers to directly process raw form of data and generate predictions or results [45], [53]. Unlike traditional ML, DL techniques minimize human intervention by embedding feature extraction steps into the network architecture through adjustable model parameters (**Figure 4**), and therefore, have gained prominence for robust handling of complex tasks requiring sophisticated pattern recognition and representation learning [52]. The performance of DL algorithms improves as dataset size rises; though there is no rule to determine the exact size of dataset required, training sets must be sufficiently broad and diverse to incorporate the wide-ranging features of the classes being classified.

Convolutional neural networks (CNNs) are the most popular of DL algorithms, which have been specially modified for processing structured grid data and applied with great success to the detection, segmentation, and classification of objects in images [54], [55]. AlexNet [56], VGG [57], and U-Net [58] are well-known examples of such architectures, which could surpass human expert performance in some cases. However, a successful CNN implementation requires a substantial volume of labelled training data, posing a significant barrier in real-world applications [55], particularly in fields like medicine where annotations demand professional expertise and instances of diseases are scarce [59], [60]. Transfer learning (TL) is an appealing solution, where knowledge and feature representations acquired from processing a large-scale annotated dataset like ImageNet can be leveraged to address another problem with fewer input samples, instead of training from scratch [61], [62]. The earlier layers of a CNN model capture simple features such as edges or contours generic to all kinds of images, while the deeper layers extract high-level details specific to the tasks in hand, and therefore, can be transferred through fine-tuning across domains and adapted to new tasks.

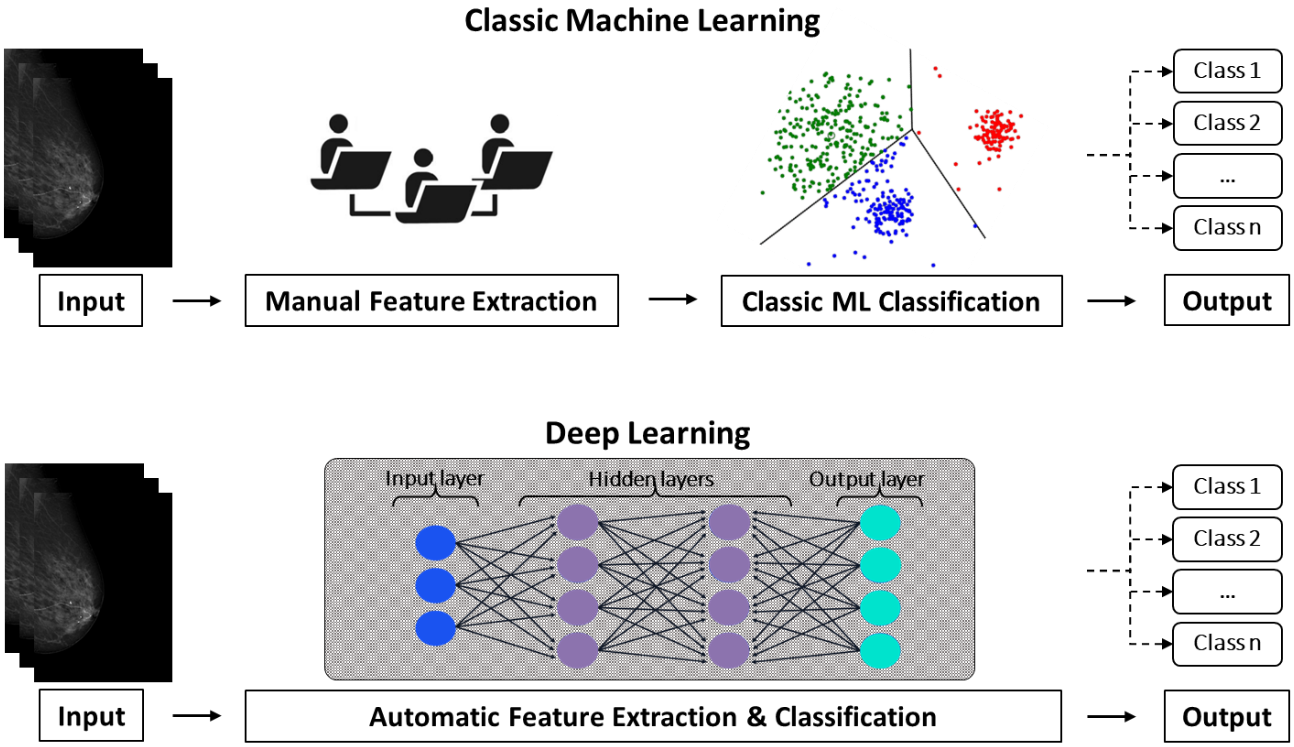


Figure 4 Differences between machine learning and deep learning in image classification tasks: While traditional ML algorithms mainly rely on manually engineered features, DL-based methods integrate feature extraction directly into the network architecture, thus minimizing the manual intervention by human expert

Given the growing popularity of DL in radiology following the dynamic development of AI systems and hardware technologies [60], [63], researchers show interest in using neural networks for BAC assessment as a potential risk biomarker for CVD in women. AI could be effective in addressing the intrinsic challenges associated with manual BAC evaluation ensuring a processing time compatible with everyday practice, helping reduce the ever-increasing radiologic workload, and minimising operator-dependency. Considering the intricate topology of BAC appearance over a broad spectrum of signal intensities on mammographic images, various strategies have been presented in the literature, developing either semiautomatic or automatic AI-based tools. **Table 1** presents an overview of the key published studies, detailing the datasets and the CNN architectures employed.

The group of Wang et al. in 2017 [64] was the pioneer to investigate the viability of automating accurate BAC detection with DL systems, developing a customized 12-layer deep architecture that involved stacks of ten convolutional blocks, two fully connected (FC) layers, and a SoftMax activation function. They formulated the problem as a binary classification task and used case-based 3-fold cross validation to implement a pixel-wise patch-based procedure to predict the probability of each central pixel belonging to the BAC class. The performance was evaluated using free-response receiver operating characteristic (FROC) analysis and quantitative calcium mass study on a total dataset of 210 four-view standard mammography exams with 146 BAC and 64 non-BAC cases. Their findings demonstrated an overall true positive rate (TPR) of 60% comparable to that of the human reader and a strong correlation between the automatically estimated calcium mass and that of the ground truth annotations (coefficient of determination of 0.96). Investigation into the interchangeability of mammographic views revealed that integrating samples from both projections (i.e., cranio-caudal (CC) and medio-lateral oblique (MLO)) can effectively improve the model's performance.

Conversely, Wang et al. 2019 [65] did not observe any advantage of deep networks over classical image processing approaches in detecting or segmenting calcified vessels. They tested the efficacy of three available networks, including YOLO, U-Net, and DeepLabv3+, using a rather small dataset of 135 BAC images sourced from both standard full-field digital mammographic (FFDM) or digital breast tomosynthesis (DBT) data. Evaluation metrics included intersection over union (IoU) and a new measure analysing the proportion of acceptable small-object segmentation over the validation subset (20% of the total dataset, i.e., 27 images), to consider the complex topology and the small dimension of BAC. However, all models exhibited poor performance in the experiment, failing to detect BAC from the digital images. In comparison, a simple Hessian-based multiscale filter paired with a self-adaptive thresholding technique yielded the maximum validation accuracy (small-object detection score of 0.78), with the minimum configuration complexity and computation cost. Considering the imbalanced distribution of calcified pixels across images, they concluded that an appropriate selection of training parameters, such as loss functions, would significantly impact the results.

In 2020, Alghamdi et al. [66] explored the applicability of the U-Net alterations for a same purpose, which had shown remarkable results in various medical segmentation tasks [58]. They extended the conventional structure by integrating both the contracting and expanding paths with summation long-connections and dense blocks of DenseNet [67], to avoid learning redundant features, preserve local details, and reduce computational costs. The problem was framed as a semantic segmentation and studied on 816 digital mammograms with an equal number of BAC and non-BAC images, which were collected from a publicly available breast cancer screening database [68] and meticulously annotated by expert radiologists to provide ground-truth BAC information. The experimental results from training the proposed DU-Net using 5-fold cross validation demonstrated an impressive overall F1 score of 0.92 and a Jaccard index of 0.85, outperforming the expert readers. Further quantitative evaluation showed DU-Net's efficacy in the BAC detection task, achieving an accuracy of 0.91 with an evaluation time of 10.8 seconds per epoch—faster than the preliminary CNN model [64], which attained an accuracy of 0.62 on the current annotated dataset.

Similarly, Guo et al. 2021 [69] presented the Simple Context U-Net (SCU-Net), a lightweight design to enhance the efficacy of fine vessel segmentation and calcification quantification. The network exploited the advantage of both dilated convolution operations and skip connections to learn multilevel contextual features, thereby improving the predictive accuracy of the typical U-Net while maintaining an order of magnitude less trainable parameters. For model training and validation on a dataset of 661 FFDM from 216 participants, each mammographic image was trimmed into fixed-size patches of 512×512 with pixels of overlap and then corresponding patches were concatenated together to generate whole-image final predictions. Extensive quantitative and qualitative results displayed comparable or superior performance of the SCU-Net, as compared to a series of semantic segmentation models including DeepLabv3 and U-Net, achieving 0.99 accuracy, 0.72 F1 score, and 0.58 Jaccard index value on the validation scans, capable of correctly disregarding benign ductal calcifications unrelated to BAC. The severity estimation of BAC within the segmented mask by the model was strongly correlated with calcified volume (R2-correlation = 0.84) and calcium mass (R2-correlation = 0.87) in a cohort of 10 subjects who had previous breast CT examinations. Notably, the SCU-Net's automated tracking of calcification progression in a longitudinal study of 26 patients with almost 10 years of retrospective mammograms revealed a gradual increase in BAC burden over time.

A subsequent study by Alamir et al. 2023 [70] recommended adopting the generative adversarial network (GAN), consisting of a generator followed by a discriminator interdependent grid, to extract binary masks and classify them as either BAC or non-BAC. They integrated a multiscale difference-of-Gaussian (DoG) pyramid into the contracting path of the U-Net as the generator for segmenting input mammographic images, to enhance the salient features of calcified vessels while reducing high-level details of others or noise, thereby improving feature extraction capability. To develop and assess the proposed DoG-GAN model, researchers prepared a BAC-enriched dataset comprising 750 synthetic 2D images from DBT examinations, offering enhanced tissue and lesions visualization compared to a standard digital mammography. The network's ability to detect the calcified patterns from the processed image of the DoG containing the edge information, exceeded the traditional U-Net trained on raw mammograms themselves, leading to an overall AUC-ROC value of 0.99 and sensitivity of 0.75, within an evaluation time of 14 seconds per epoch.

Wang et al. 2023 [71] further analyzed the effect of training factors, including input size, pre-processing techniques, loss functions, deep network characteristics, and annotation quality, on the automatic BAC segmentation performance. The study involved 6573 raw tomosynthesis central projections with a BAC positive rate of 95%, annotated for training and evaluating deepLavV3+ and U-Net models with various learning strategies. While higher image resolution, proper contrast adjustments, and deeper complex architectures can typically contribute to optimal classification results, they found labelling quality to be a pivotal determinant of the calcification segmentation performance, specifically when annotations were made by non-expert readers. A solid annotation alone delivered the most significant improvement in network outcomes compared to other learning settings investigated; however, achieving pixel-perfect ground-truth markings remain challenging even for experienced radiologists. Additionally, recognizing the clinical practice of measuring BAC by its length rather than area size, they advocated for length-based quantification to better capture the linear trajectory of BAC.

Table 1 Summary table of the state-of-the-art research focused on AI-driven automatic BAC assessment on mammograms

Reference	Dataset	Network	Validation method	Outcome
Wang et al. (2017) [64]	Private FFDM dataset, including 840 images (506 BAC) with pixel-wise ground truth annotations	Customized 12-layer CNN, trained with patches of 95×95	3-fold cross validation (case-wise data splitting)	Pixel-wise detection and quantification of BAC burden
Wang et al. (2019) [65]	Private FFDM and DBT (central slices) dataset, including 135 images (135 BAC) with pixel-wise ground truth annotations and the bounding box of the significant calcified regions	YOLO, U-Net, and DeepLabv3+	80% training, 20% validation (image-wise data splitting)	Failed

Alghamdi et al. (2020) [66]	Public FFDM dataset [68], including 826 images (413 BAC) with pixel-wise ground truth annotations	DU-Net	5-fold cross validation (image-wise data splitting)	Pixel-wise BAC detection
Guo et al. (2021) [69]	Private FFDM dataset, including 661 images from 216 cases with pixel-wise ground truth annotations	SCU-Net, trained with BAC patches of 512×512	80% training, 20% validation (image-wise data splitting)	Pixel-wise detection and quantification of BAC burden
Alamir et al. (2023) [70]	Private DBT (synthetic 2D view) dataset, including 750 images (600 BAC) with pixel-wise ground truth annotations	DoG-Gan	80% training, 20% validation (image-wise data splitting)	Pixel-wise BAC detection
Wang et al. (2023) [71]	Private DBT (central slices) dataset, including 6573 images (95% BAC) with pixel-wise ground truth annotations	U-Net and DeepLabv3	80% training, 10% validation, 10% testing (image-wise data splitting)	Pixel-wise BAC detection

Conclusions

In the framework of ongoing efforts aiming to reduce gender-based disparities in cardiac health assessment, BAC have emerged as a beneficial and cost-effective biomarker that can be easily obtained from the already established mammographic screening practices to improve women's CVD risk stratification. AI-based tools can play a significant role in detecting and quantifying BAC reproducibly, without increasing the radiologists' workload. Nevertheless, despite all promising results from previous studies, fully automated BAC quantification remains an open challenge, as BAC present with a complex topology, strongly influencing their appearance on different mammographic views, with a large spectrum of extent and x-ray attenuations. The patch-based training models prove time-consuming and impractical for clinic [64], some models may mis-detect non-continuous BAC structures [66], [70], and improvements could be artificially inflated when BAC samples deviate from real-world prevalence, when the dataset is not sufficient to form a fully unknown test set, or when bias is introduced through image-wise data splitting [69], [70], [71]. Foremost, the available approaches still rely on manual pixel-wise BAC segmentation to train the models, thus being vulnerable to inter-reader variability, as highlighted in a paper by Trimboli et al. [72] in which the authors developed a semi-automatic tool for BAC quantification. In the future, weakly supervised approaches may overcome the need for images annotated on a pixel basis as ground truth, further reducing operator-dependency and facilitating their translation from research into clinical practice.

Section II:

Development of an innovative deep learning approach for detection and quantification of BAC

Based on:

- N Mobini, M Codari, F Riva, MG Ienco, D Capra, A Cozzi, S Carriero, D Spinelli, RM Trimboli, G Baselli, and F Sardanelli. Detection and quantification of breast arterial calcifications on mammograms: a deep learning approach. *European Radiology* (2023) DOI: 10.1007/s00330-023-09668-z

Background

Cardiovascular diseases (CVD) are the leading cause of death in the female population [72]. Although it is commonly assumed that males have a greater mortality rate from CVD [3], almost as many women as men die from heart disease yearly. Traditional approaches for cardiovascular risk assessment perform worse in women [9], [73], as up to 20% of women's cardiovascular adverse events occur in the absence of traditional risk factors [74], and women are less likely to be prescribed CVD prevention therapy in primary care settings [75]. Hence, innovative imaging biomarkers that could improve cardiovascular risk stratification in women have been proposed over the last two decades [31].

In particular, breast arterial calcifications (BAC) have been suggested as a sex-specific predictor of cardiovascular risk [13], [29], [30], [76], [77], [78], [79]. BAC are a common incidental finding on mammograms, where they appear as parallel linear opacities within vessel walls (illustrated in **Figure 1**) [13], [76]. Their approximate prevalence, although in a wide range, has been estimated around 13% [27], [29], [35], [78], [79]. BAC presence has been associated with a 1.23 increased risk of CVD in postmenopausal women [29] and has higher diagnostic accuracy than other traditional cardiovascular risk factors in asymptomatic middle-aged women, especially under 60 years of age [30], [78], [79].

Considering the widespread diffusion of screening mammography [10], [11], systematic BAC assessment could provide a low-cost cardiovascular risk stratification in women without any additional tests. Although most radiologists are aware of the link between BAC and CVD, BAC reporting in routine mammography interpretation is scarce [34], being further prevented by the lack of standard BAC reporting guidelines and of reliable and quick methods for BAC detection and quantification [35]. As BAC vary considerably in size, length, and density, several methods for BAC burden estimation have been proposed, either with manual semiquantitative scoring [27], [35] or with quantitative scoring based on automated segmentation by artificial neural networks [64], [69]. Despite promising results, these supervised algorithms still required time-consuming manual pixel-wise annotations in a large number of images for the training process. Conversely, deep learning (DL) algorithms and convolutional neural networks (CNN) trained by a simple dichotomic supervision to detection can provide higher robustness and lesser human image postprocessing workload [53], [80]. BAC positive (BAC+) and BAC negative (BAC-) annotation can be adopted in place of a full manual segmentation of BAC and throughout the work we name the former "weak supervision" as opposed to the latter.

The objective of our study was to develop a weakly supervised deep CNN that can distinguish mammograms with and without BAC. Additionally, we aimed to obtain an estimate of the BAC burden as a by-product of our detection algorithm. To achieve this, we formulated the problem as a binary classification task and used an AI explainability algorithm to identify the approximate location of BAC, without relying on ground truth segmentation.

Methods

Patient enrolment and data collection

This retrospective study was approved by the local Ethics Committee (protocol code SenoRetro, approved on November 9, 2017, amended on May 12, 2021) and the need for informed consent was waived. We included a series of consecutive patients aged ≥45 years, who were referred to the IRCCS Policlinico San Donato between January and March 2018 to undergo spontaneous or organized population-based screening mammography.

All included examinations were bilateral mammograms with cranio-caudal (CC) and medio-lateral oblique (MLO) projections, acquired using full-field digital systems (Giotto IMAGE 3DL or Giotto TOMO series, IMS). Three readers (R.M.T., D.S., and S.C. with 10, 3, and 2 years of experience in breast imaging, respectively) reviewed the included mammograms to perform a patient-based classification as BAC+ or BAC-. BAC+ patients had at least one BAC detectable on a mammographic view, whereas all other patients were considered BAC-. A fourth reader (D.C. with 3 years of experience in breast imaging) then labelled each mammographic view of BAC+ patients as BAC+ or BAC-. All the labels were encoded in a database and served as the ground truth during training and testing of the BAC detection model.

Clinical dataset preparation and pre-processing

To preserve the age distribution of the positives, BAC+ data was divided into four age classes using our population's age quartiles as thresholds: first class, 45 years–Q1; second class, Q1–Q2; third class, Q2–Q3; fourth class, Q3–maximum age of the participants (see

Results for details). Then, we performed a stratified split of the BAC+ dataset into three subsets within the classes to preserve the BAC+ age distribution: 70% of the random shuffled positive cases entered the training subset, 15% entered the validation subset to tune model hyperparameters based on the highest precision-recall curve (AUC-PR), and the remaining 15% entered the test subset to evaluate the performance of the final optimized CNN. Subsequently, the whole BAC-dataset was randomly partitioned into training, validation, and test sets containing 70%, 15%, and 15% of the negative cases, respectively. The relevant BAC+ and BAC- splits were then consolidated to complete the three subsets. In other words, we combined the training divisions of BAC+ and BAC- to form the final training set, the validation divisions of BAC+ and BAC- to create the final validation set, and likewise the test divisions of BAC+ and BAC- to form the final test set. To account for class imbalance during model training [81], [82], the majority class (BAC-) in the training subset was randomly under-sampled to reach a BAC+ prevalence of 30% at patient-level. The validation and test sets remained intact to mirror the real BAC prevalence. To eliminate any bias that may happen by allocating different views of a single case into different subsets, data splitting at patient-level preserved all the mammogram views of each case the same subset.

The dataset consisted of images with various matrix sizes up to 3584 × 2816, depending on the compacting plates used during acquisition. Therefore, a pre-processing step was required to exclude non-tissue areas and normalize the signal intensities. Using histogram analysis following the Otsu's method, we successfully extracted the tissue regions from the dark background pixels [83], [84]. After defining the smallest rectangular area surrounding the breast tissue, the cropped image was scaled to a fixed-size 1536 × 768 matrix that would define the size of the input layer of the CNN (**Figure 5**). Then, pixels belonging to the breast region were normalized to zero mean and unit variance to improve the convergence of training, thus accounting for the high variability of mammogram pixel intensities caused by acquisition and biological factors like technical differences between mammography units and tissue density, as follows:

$$x_i' = \frac{x_i - \mu(x)}{\sigma(x)}$$

where x_i' represents the normalized intensity of the ith pixel, μ is the mean, and σ is the standard deviation of the pixel values in the image.

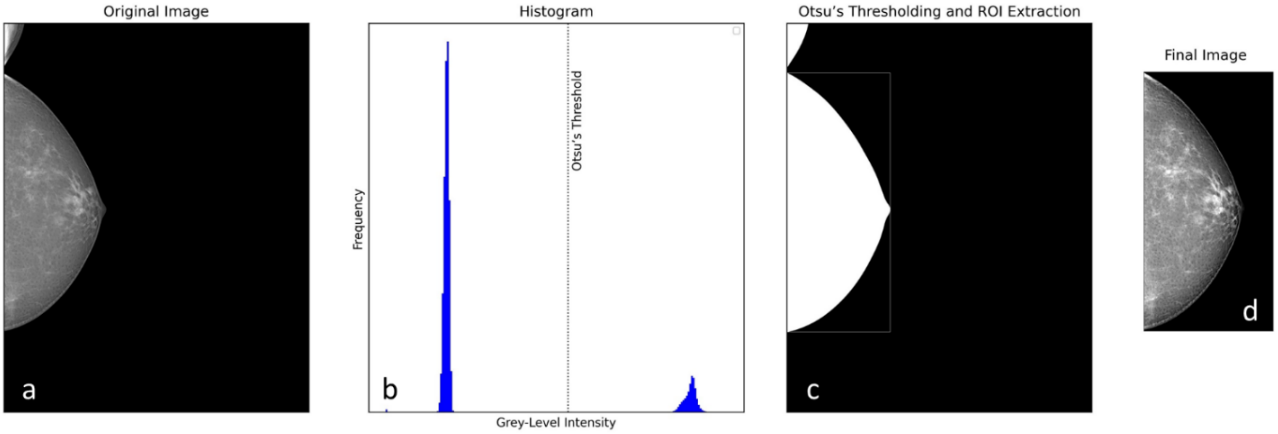


Figure 5 Overall workflow of the pre-processing step to exclude non-tissue areas. **a** a sample mammogram with a matrix size of 3580 × 2812, **b** the bimodal histogram (low intensities referring to the background, high intensities referring to the tissue) and the calculated Otsu's threshold, **c** the smallest rectangular contour surrounding the largest over-threshold area as the breast tissue, and **d** final cropped and padded image to a fixed-size matrix of 1536 × 768

Neural network architecture and implementation

We implemented a BAC detection model using a deep transfer learning strategy [61] based on the 16-layer pre-trained Visual Geometry Group (VGG16) image classification model with modifiable connection weights [57]. We replaced the last dense layer with two fully connected layers (256 channels each) including leaky rectified linear unit activation functions ($\alpha = 0.3$) trained from scratch, and a sigmoid activation function as final output layer, as appropriate for our binary classification problem (presence or absence of BAC). Next, we optimized the number of the initial convolutional layers to be fixed as "non-trainable layers" and of the later ones to be fine-tuned on the new binary classification. This was done by trial and error, each time training the modified CNN and assessing its performance on the validation set. The best-performing structure was found to be that with five fine-tuning layers. **Figure 6** summarizes the complete architecture of the proposed CNN. VGG16 input structure constrained a fixed dimension of red-green-blue colour coding (**Figure 6a**); hence, grey-level mammograms were resampled to fixed-size 1536 × 768 images and input three times in parallel (**Figure 6b**). Our model elaborated each mammographic view independently.

We applied online data augmentation during training, including random rotations, width/height shift, horizontal/vertical flip, and zoom, as well as random Gaussian and salt-pepper noise addition to learn more robust features. During training, the Adam optimizer [85] was applied to minimise the binary cross-entropy loss (BCE) function. In addition to the data-level solution using random under-sampling, a class-balanced re-weighting strategy (weighted BCE) was also utilized to deal with the imbalanced dataset at algorithmic-level which automatically altered the loss inversely proportional to the class frequency, thereby assigning a higher weight to the minority

BAC+ class in the loss function (the number of positive and negative incidences are summarized in **Table 3**), helping to balance the impact of classes on the model's training process.

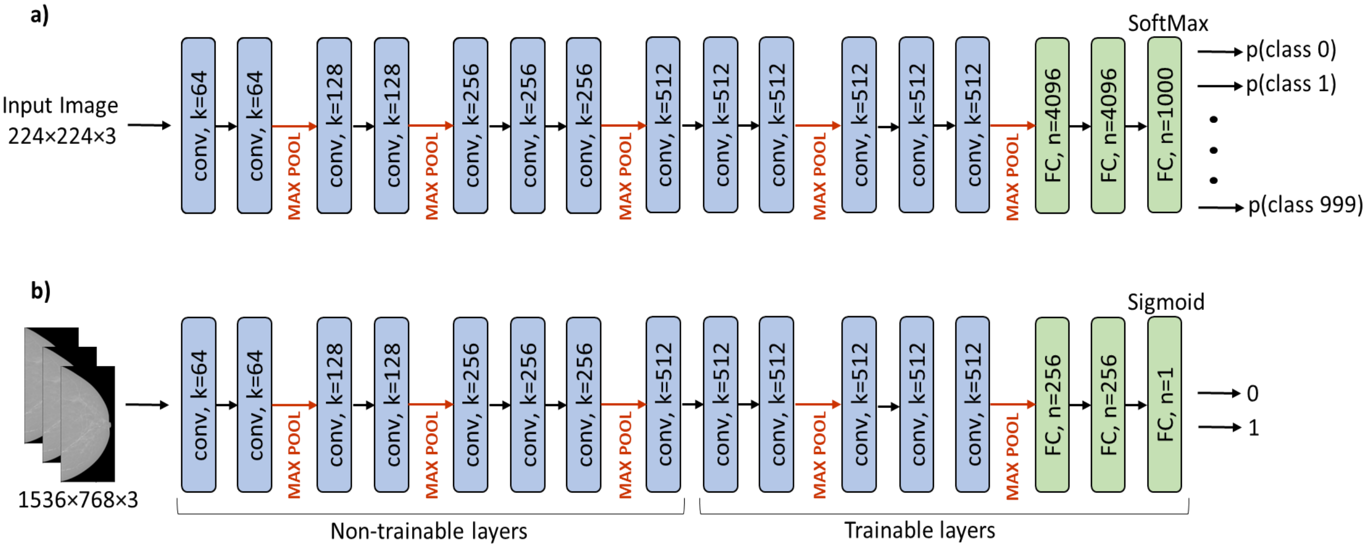


Figure 6 General VGG16 architecture consisting of 13 convolutional layers (kernel 3 × 3, depth k), 5 pooling layers (non-trainable), and 2 fully connected (FC, n: number of neurons) layers followed by a Softmax activation function to solve the multiclass classification problem (**a**), and the final CNN for automated binary BAC detection where the "non-trainable layers" exploited VGG16 transfer learning (**b**). Rectified linear unit (ReLU) activation functions (in model a) and leaky ReLUs (in model b) following each convolutional kernel are not shown

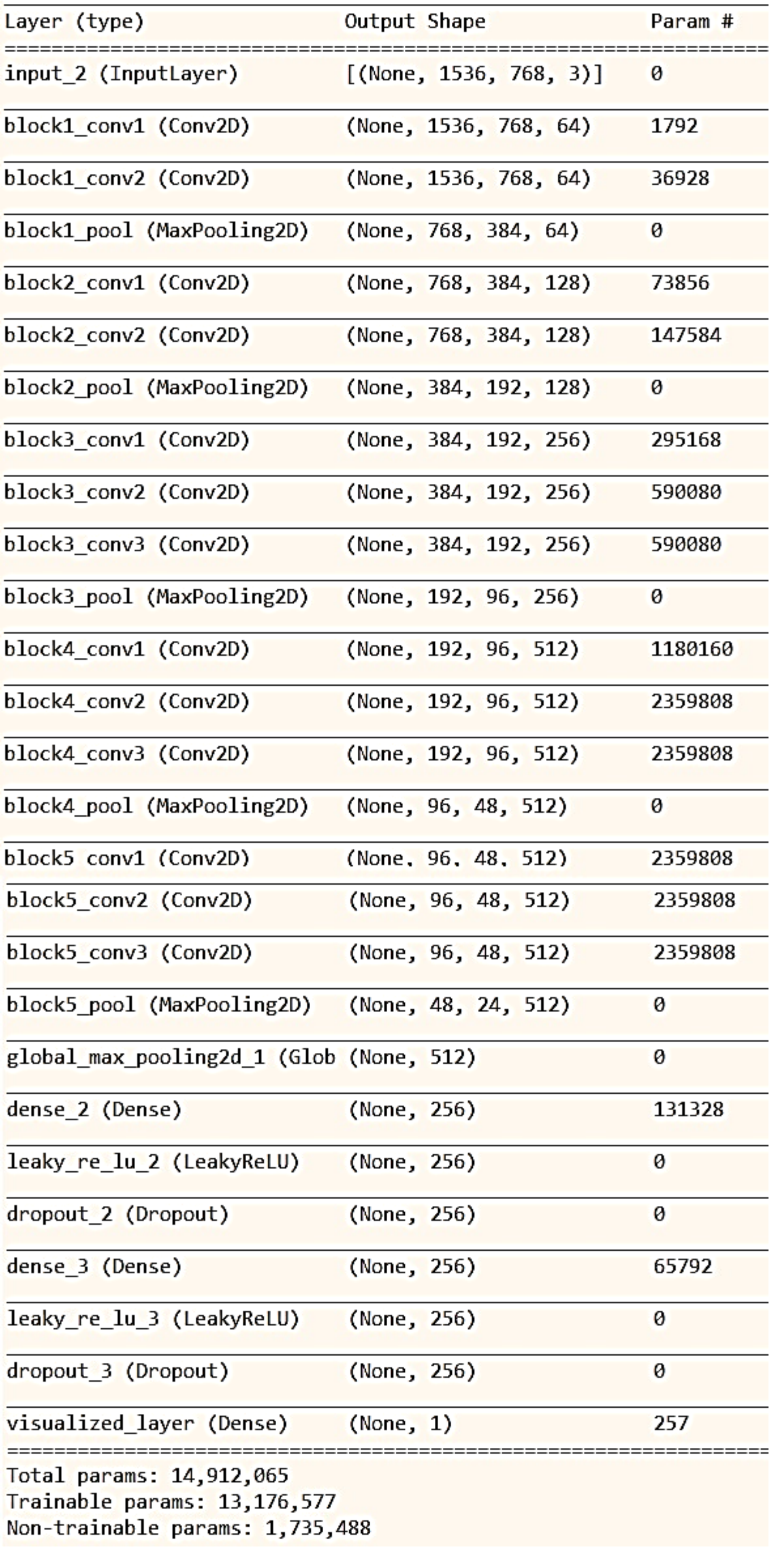
Learning rate was initially set to 10^{-6} and adjusted over the epochs using cosine annealing scheduler as follows:

$$lr_x = lr\left({}^{\cos\left({}^{x\pi}/_{100}\right) + 1}/_{2}\right)$$

where lr is the starting rate, x is the epoch number, and lr_x is the learning rate at epoch x [86]. Due to the highly imbalanced dataset, the area under the PR curve was monitored and the parameters related to the maximum quantity provided the best model configuration at the end of each epoch. The number of epochs and batch size were empirically selected as 25 and 8 images, respectively, to optimize model performance while ensuring compatibility with the available hardware resources. Dropout regularisation was set to 0.3 for each dense layer. The proposed model summary, developed using Python V3.8.11 on a system with NVIDIA GeForce RTX 3080 and 10GB vRAM, is represented in **Figure 7**.

Finally, visual explanations of the proposed CNN were generated using the generalized gradient-weighted class activation mapping (Grad-CAM++) method after the deepest convolutional layer [87], [88], providing heat-maps highlighting the pixels that were significant for predictions. Simple binarization thresholding of the heatmaps in positive predictions enabled us to delineate an estimated BAC region from the total tissue.

The time required for automatic mammogram classification and generation of Grad-CAM++ heatmaps was recorded and reported as average image elaboration time.



```
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         [(None, 1536, 768, 3)]    0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 1536, 768, 64)     1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 1536, 768, 64)     36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 768, 384, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 768, 384, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 768, 384, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 384, 192, 128)     0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 384, 192, 256)     295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 384, 192, 256)     590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 384, 192, 256)     590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 192, 96, 256)      0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 192, 96, 512)      1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 192, 96, 512)      2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 192, 96, 512)      2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 96, 48, 512)       0
_________________________________________________________________
block5 conv1 (Conv2D)        (None. 96. 48. 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 96, 48, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 96, 48, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 48, 24, 512)       0
_________________________________________________________________
global_max_pooling2d_1 (Glob (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 256)               131328
_________________________________________________________________
leaky_re_lu_2 (LeakyReLU)    (None, 256)               0
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_3 (Dense)              (None, 256)               65792
_________________________________________________________________
leaky_re_lu_3 (LeakyReLU)    (None, 256)               0
_________________________________________________________________
dropout_3 (Dropout)          (None, 256)               0
_________________________________________________________________
visualized_layer (Dense)     (None, 1)                 257
=================================================================
Total params: 14,912,065
Trainable params: 13,176,577
Non-trainable params: 1,735,488
```

Figure 7 Summary of the model implemented in Python

Quantification

We assessed the correlation of the estimated BAC region delineated on the Grad-CAM++ in a subset of MLO views with manual measurements of calcified segments length obtained from a previously published study (**Figure 8**) [35]. The BAC length was calculated as follows:

$$BAC = P \sum_{i=1}^{n} 1_{G(i)>Th}$$

where P is the pixel size, n the total number of pixels in the image, and $G(i)$ the Grad-CAM++ heatmap value at pixel ith. Th represents the best binarization threshold, which was set to 0.3 by trial and error.

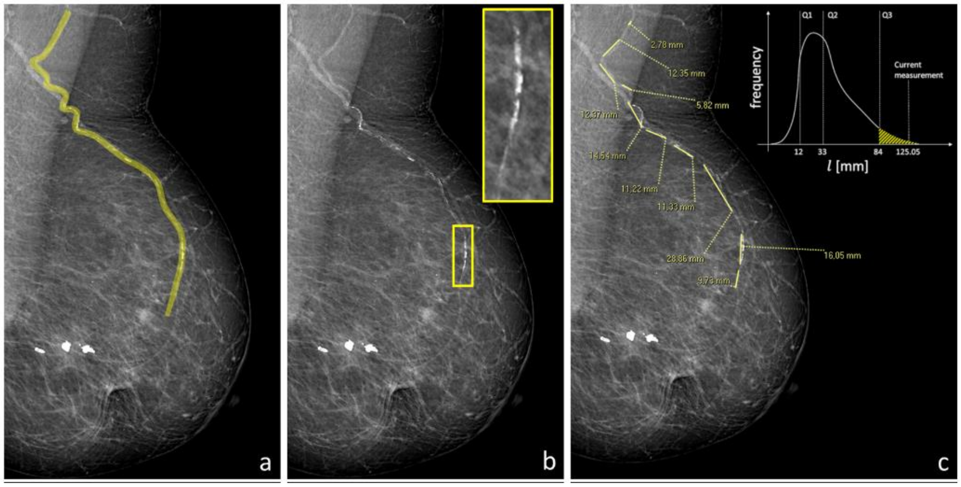


Figure 8 An example of manual BAC-length measurement (adapted from Trimboli et al. 2021 [35]) showing: **a** a single involved vessel, **b** opacification of the vessel from side to side, and **c** the resulting calcified segments of 125.05 mm

Statistical analysis

The Kolmogorov–Smirnov test was used to assess the normality of the continuous variables' distributions; normal variables were reported as mean ± standard deviation (SD), whereas non-normal variables were reported as median and interquartile range (IQR). The Mann-Whitney *U* test was performed to compare the age distributions in the BAC+ and BAC- groups; *p* values less than 0.05 were considered statistically significant [89].

The overall diagnostic performance of the proposed CNN model was evaluated against the ground truth labels provided by the readers, using the following metrics: accuracy, balanced accuracy, precision, recall (sensitivity), F1 score, and area under the receiver operating characteristic curve (AUC-ROC).

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

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

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

$$Recall = \frac{TP}{TP + FN}$$

$$F_1\ score = 2\ \frac{precision \times recall}{precision + recall} = \frac{2\,TP}{2\,TP + FP + FN}$$

where *TP*, *TN*, *FP*, *FN* denote the number of true positive, true negative, false positive, and false negative detections, respectively. Correlations were appraised by Pearson r or Spearman ρ as appropriate, and the resulting coefficients were interpreted according to Evans [90].

Results

A total of 1557 patients underwent screening mammography at our institute between January and March 2018. After excluding patients younger than 45 years of age, 1493 women with a median age of 59 years (IQR 52–68) were finally included, for a total of 5972 mammographic views. BAC were present in 194 of 1493 women (13.0%) and 581 of 5972 views (9.7%), respectively (**Table 2**). Prevalence of BAC increased with age, from 6.3% in the first age class (45–60 years), to 11.6% in the second age class (61–70 years), 34.3% in the third age class (71–73 years), and 38.2% in the fourth age class (74–87 years). The 194 BAC+ women had a significantly higher median age (70.5 years, IQR 60–73) than the 1299 BAC- women (median 57 years, IQR 52–65, $p < 0.001$) (**Figure 9**).

Table 2 Study population

	Total	**BAC Positive**		**BAC Negative**	
	Population	**Patient-level**	**Image-level**	**Patient-level**	**Image-level**
Frequency	1493	194	581	1299	5391
Prevalence	100%	13.0%	9.7%	87.0%	90.3%
Age [IQR]	59 [52–68]	70.5 [60–73]		57 [52–65]	

IQR Interquartile range

Table 3 reports training, validation, and test sets composition. Following data partitioning, 1042 women (4168 mammograms) were assigned for training, 222 (888 mammograms) for validating, and 229 (916 mammograms) for testing, each containing 398, 89, and 94 BAC+ views, respectively. To reduce class imbalance during model training we artificially increased the prevalence of BAC+ patients to around 30% in the training set by randomly under sampling BAC- mammograms from those assigned to the training dataset, reaching 1640 images from 410 women. Eventually, image-level BAC prevalence was lower, given that not all mammographic views of BAC+ patients showed BAC. BAC prevalence in validation and test sets was left unchanged (**Figure 10**).

Table 3 Training, validation, and test set composition

	Training	**Validation**	**Testing**
BAC+ (n[%])	398 (24.27)	89 (10.02)	94 (10.26)
BAC- (n[%])	1242 (75.73)	799 (89.98)	822 (89.74)
Total images	1640	888	916

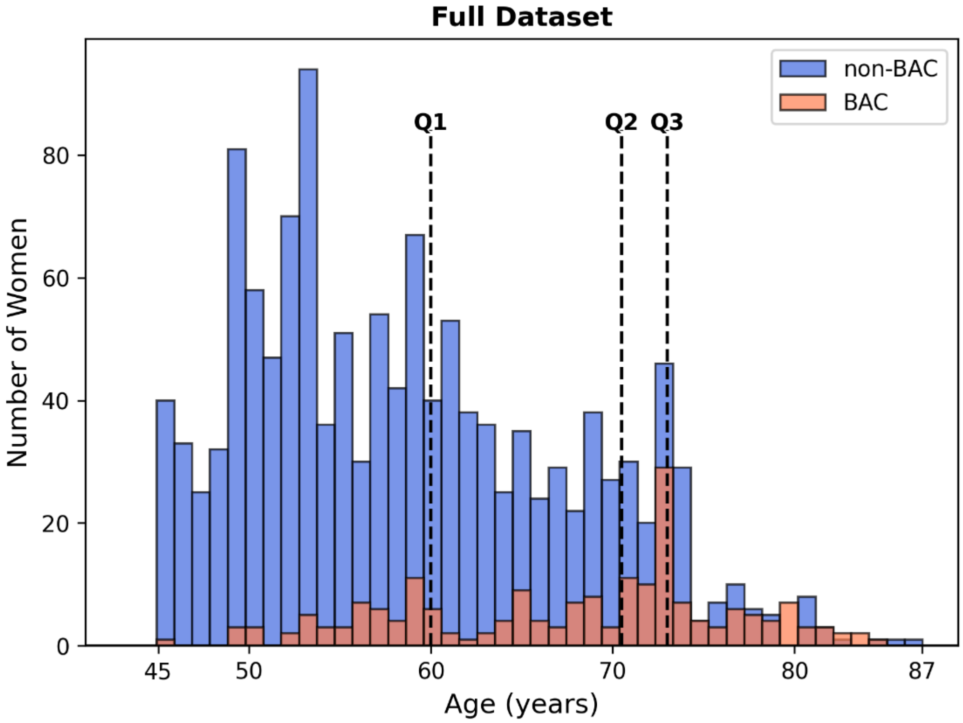


Figure 9 Age distribution of the study population (blue: non-BAC, orange: BAC) and the BAC's quartiles defining the age classes. First class, Minimum age (45 years)–Q1; second class, Q1–Q2; third class, Q2–Q3; fourth class, Q3–maximum age of the participants (87 years)

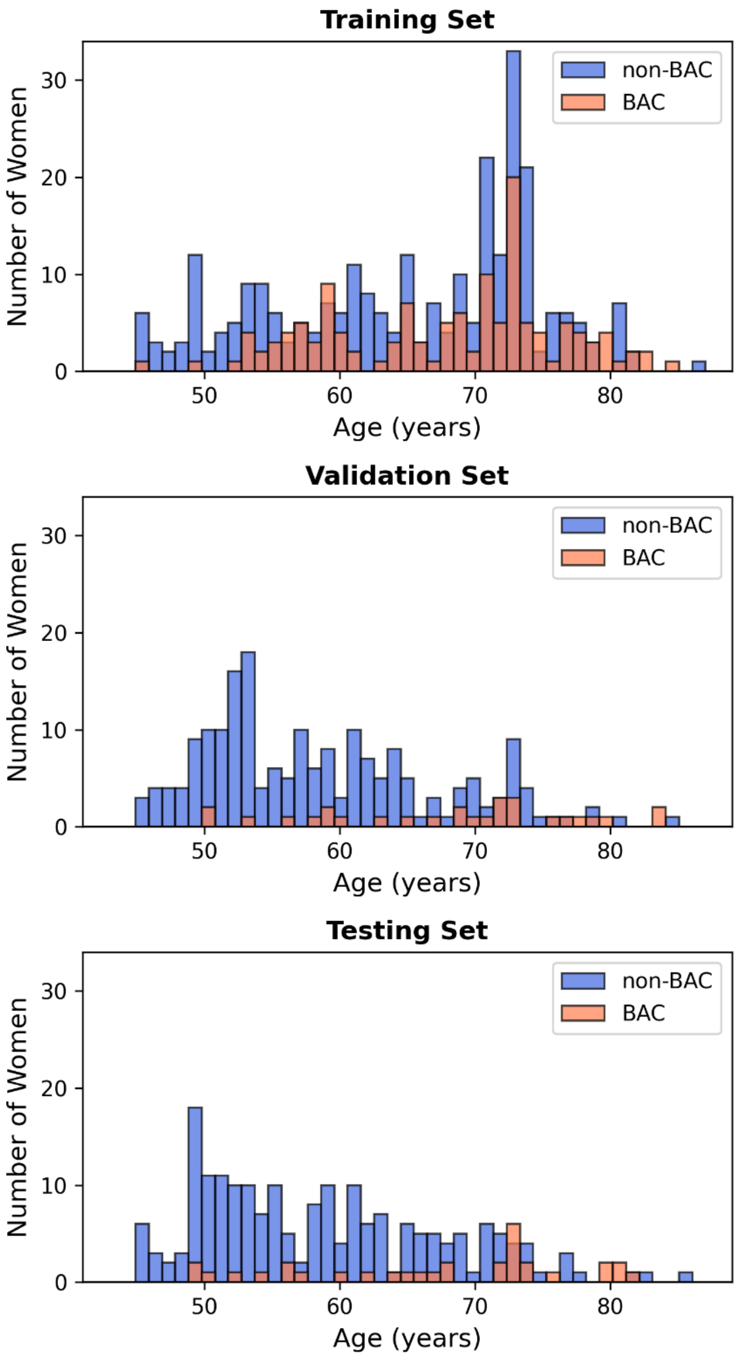


Figure 10 Age distribution of the subsets (blue: non-BAC, orange: BAC). The training set was under-sampled to a BAC prevalence of 30% to address the imbalanced dataset bias, while the validation and testing sets remained intact to reflect the real-world prevalence

Table 4 represents the overall corresponding image-level performances of the proposed CNN model in detecting the presence or absence of BAC in the subsets. Training was performed at image-level and optimised based on the highest AUC-PR. In the independent test set, the best-trained CNN achieved a 0.95 accuracy, a 0.76 F1 score, and a 0.94 AUC-ROC, highlighting good overall performances in BAC detection. The training phase loss curves and the resulting ROC and PR plots are presented in **Figure 11** and **Figure 12**, respectively.

Table 4 Diagnostic performance of the model in detecting BAC on mammograms

	TN	**TP**	**FN**	**FP**	**Accuracy**	**Balanced-Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC-ROC**	**AUC-PR**
Training	1222	312	86	20	0.93	0.88	0.94	0.78	0.85	0.96	0.93
Validation	787	64	25	12	0.96	0.85	0.84	0.72	0.78	0.95	0.86
Test	803	69	25	19	0.95	0.86	0.78	0.73	0.76	0.94	0.81

TN true negative, *TP* true positive, *FN* false negative, *FP* false positive, *AUC-ROC* area under the receiver operating characteristic curve, *AUC-PR* area under the prediction-recall curve

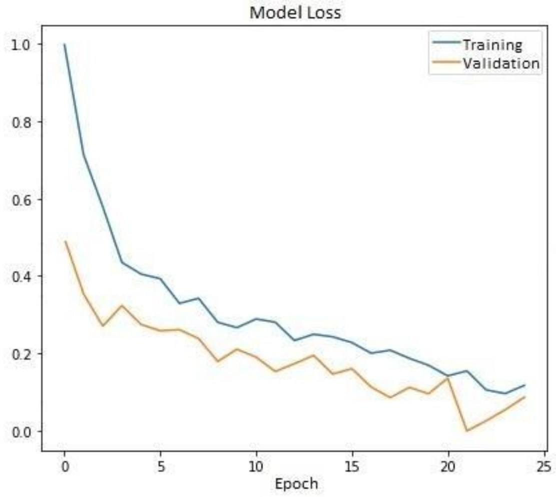


Figure 11 The model loss curves on the training (blue) and validation (orange) datasets, over the epoch numbers

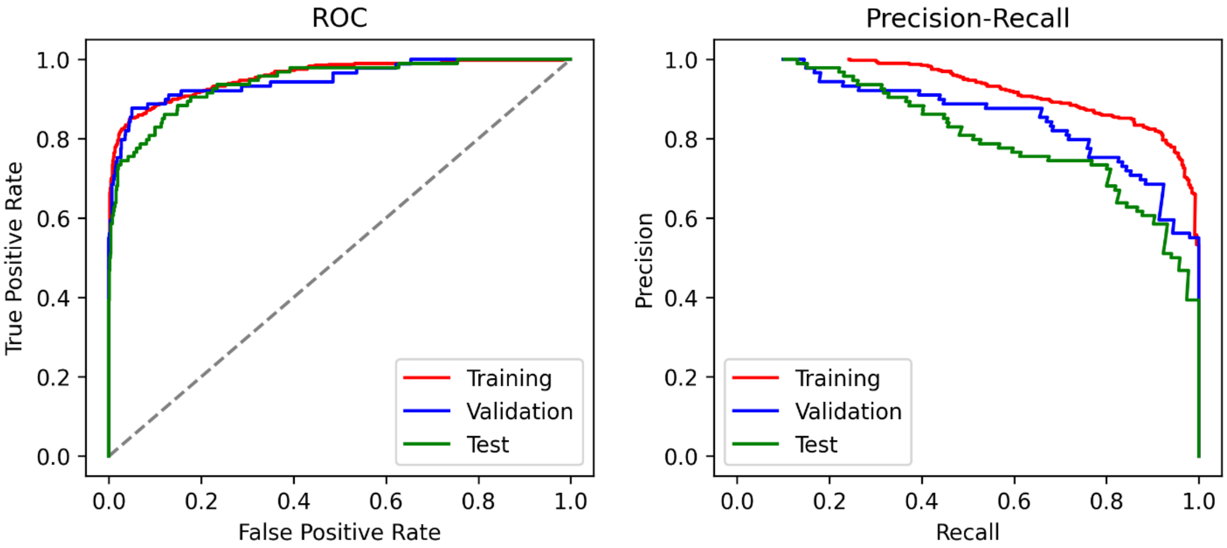


Figure 12 ROC and PR plots of training (red line), validation (blue line), and test (green line) subsets

The Grad-CAM++ heatmaps across the convolutional layers of the fine-tuned VGG16 offer a visual explanation of the decision-making process by highlighting the significant regions for predictions, thereby opening the black box of the presented deep model. As illustrated in **Figure 13**, with a confirmed BAC mammographic input, the heatmaps at the initial layers show diffuse and broad activations, primarily capturing basic edge and contour information. However, as we move deeper into the network and into the trainable layers, the heatmaps become progressively focused, recognizing more specific features, with the final layer showing the highest concentration of activation over the BAC area.

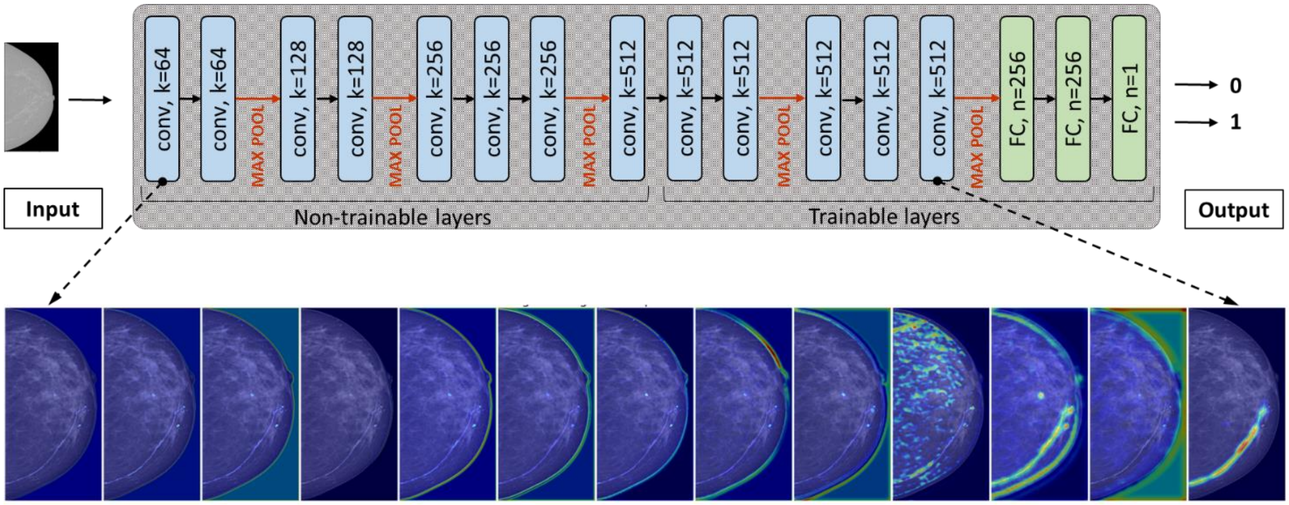


Figure 13 Visual explanation of each convolutional layer to open the model's black box. Generally, the earlier blocks detect generic information such as edges and contours, while the deeper blocks capture high-level task-specific features

Accordingly, **Figure 14** shows the performance of our CNN model through Grad-CAM++ heatmaps generated from the deepest convolutional layer. In true positive detections, BAC are accurately localized also when multiple incidences of BAC are present in the same view (**Figure 14a, a'**). Furthermore, our CNN demonstrated to be capable of detecting even small or non-continuous BAC occurrences (**Figure 14b, c**). Conversely, Grad-CAM++ heatmaps of true negative predictions emphasize BAC-like structures in the whole breast without reaching the threshold for BAC+ classification (**Figure 15**) and without being confounded by typically benign rounded micro calcifications or medical implants such as loop recorders and biopsy markers. Examples of wrong detection are reported in **Figure 16**. The average image elaboration time, including automatic BAC detection and Grad-CAM++ generation, was 0.80 ± 0.07 s.

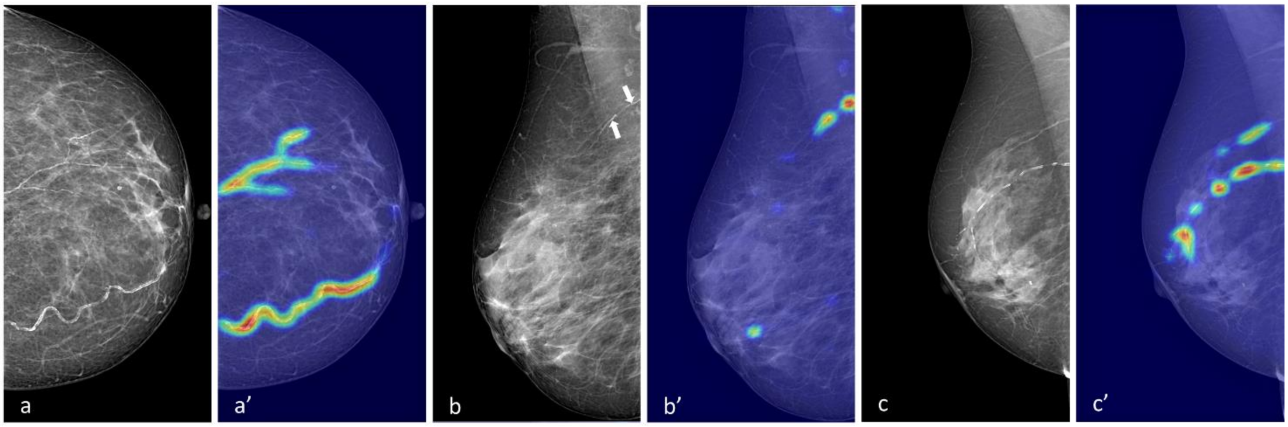


Figure 14 Grad-CAM++ heatmaps of the automatic detection results by the proposed model. Examples of true-positive cases with **a, a′** a high burden of BAC in multiple vessels, **b, b** small BAC (arrows), and **c, c′** non-continuous BAC

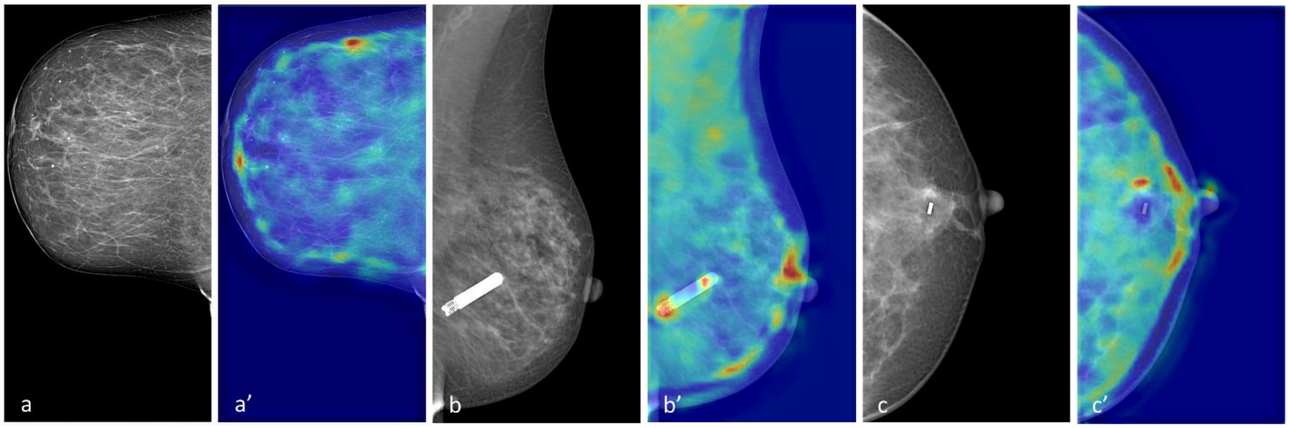


Figure 15 True-negative cases with confounding factors such as: **a, a′** various benign calcifications unrelated to BAC, **b, b′** an implantable loop recorder, and **c, c′** a radiopaque biopsy marker. None of the structures coloured on the heatmaps reach the threshold for being finally detected as BAC and are correctly ignored by the model

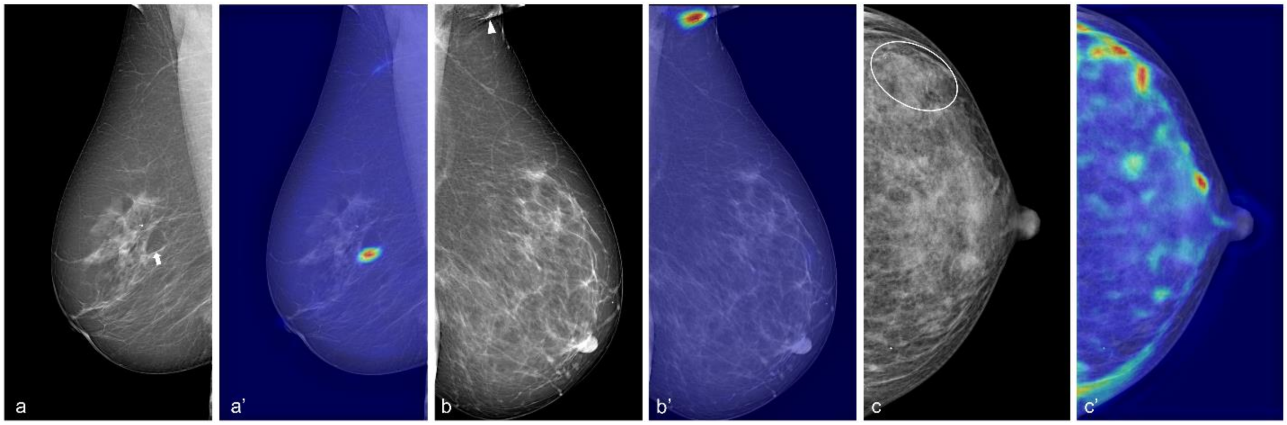


Figure 16 Examples of misclassification. **a, a′** False-positive case with small calcifications within a Cooper's ligament mistaken as BAC (arrow), **b, b′** false-positive case with skinfold including cutaneous calcifications mislabelled as BAC (arrowhead), **c, c′** false-negative case with small BAC concealed under dense tissue (circle)

A preliminary quantitative evaluation was performed on a subgroup of 57 patients with previous manual BAC length measurements. One patient had a discordant assessment of her BAC status between assigned label and BAC length measurement and was hence discarded. The analysis was therefore performed on MLO views of 56 BAC+ women aged 49–82 years. In total, 112 MLO views were analysed, and presence of BAC was reported in 95 of them. Automatic BAC burden estimation was performed by Grad-CAM++ heatmaps thresholding as depicted in **Figure 17a**. The automatically detected BAC region showed a strong correlation with the manually measured length (Spearman ρ = 0.88, $p < 0.001$) (**Figure 17b**).

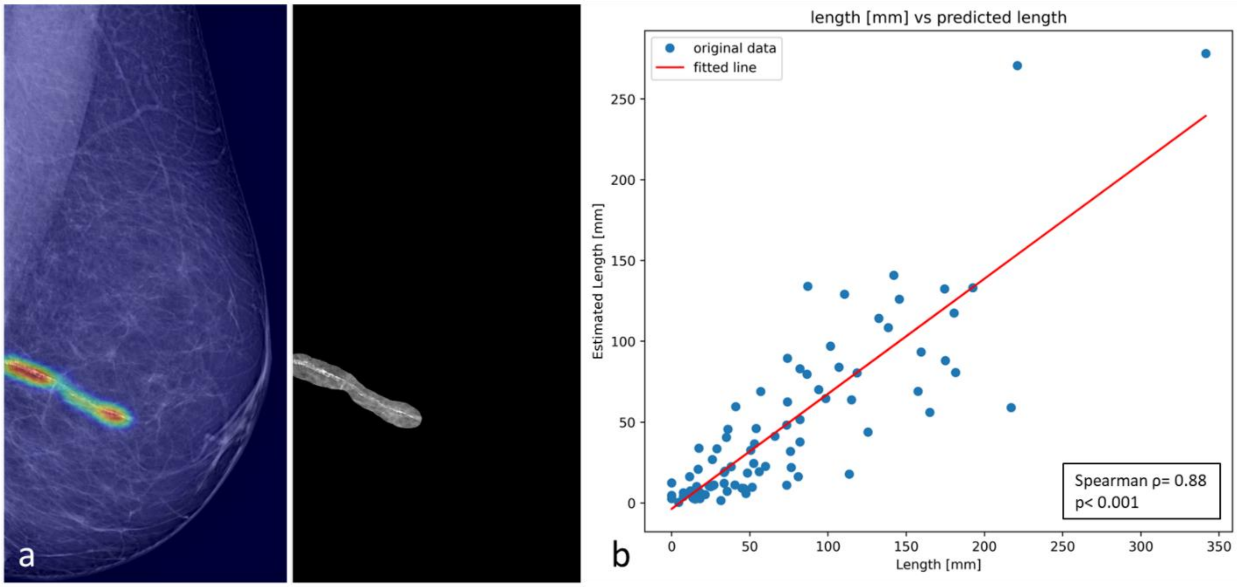


Figure 17 a Automatic segmentation of a BAC by thresholding the Grad-CAM++ heatmap of a mammogram with moderate burden of BAC (length 41 mm). **b** Scatterplot of the estimated BAC (y-axis) compared to the manually measured length (x-axis) for all 56 women in the subgroup (112 views)

Discussion and Conclusions

We implemented a CNN for the automatic detection of BAC on mammograms. Our model showed good performances in BAC detection, with an AUC-ROC of 0.95 in the test set, and it proved capable of estimating BAC region with a correlation of 0.88 with manual measurements. The application time of our model was less than a second for each image, a time suitable for a swift integration in everyday clinical practice.

In the framework of the research effort aiming to reduce the gender gap in CVD prevention and cardiovascular risk assessment [91], BAC stand out as a beneficial and low-cost biomarker of cardiovascular risk that can be easily obtained from the already established screening practice [12]. Nonetheless, BAC presence is seldom reported during mammography interpretation [34]: this can be ascribed both to the primary focus on cancer detection that clinicians keep in the context of mammographic screening and to the lack of fast, automated, and reliable tools for BAC detection and quantification. Therefore, automatic tools for BAC detection and quantification could overcome this issue without increasing the radiologists' workload.

A previous experience in BAC semiautomatic detection and quantification demonstrated that human detection is the main source of variability in developing an automated tool [92]. Therefore, we chose to address the classification problem by training a weakly supervised CNN, which may allow to partially overcome the intra and inter-reader variability. Our CNN was trained using image level labels in order to obtain as by-product then pixel-wise detection of BAC on mammograms. This strategy allowed us to reach high performances with an accuracy of 0.95, a recall (*i.e.*, sensitivity) of 0.73, a precision (*i.e.*, positive predictive value) of 0.78, and an AUC-ROC of 0.94 in the independent test set, which consisted of 916 images. Furthermore, our model proved to be capable of estimating BAC length with a strong correlation (ρ = 0.88) with manual annotation in a subset of 56 positive cases.

Our performances are similar to those reported by previous studies: Khan and Masala [93] recently published a study on BAC detection using transfer learning, comparing the results obtained from different deep learning architectures trained on a small population of just 104 mammograms from 26 patients. They reported an accuracy of 0.96 of VGG19, marginally lower than that yielded by deeper CNNs such as ResNet50 or DenseNet-121, which shown an accuracy of 0.97 and 0.98 respectively. In 2017, Wang et al. [64] developed a CNN for BAC detection using the mammograms of 210 women, 146 BAC+ and 64 BAC-, demonstrating a detection rate comparable to that of human readers, and a very strong correlation between the automatically estimated BAC area and the ground truth (Pearson coefficient 0.94). In 2021, Guo et al. [69] trained a Simple Context U-Net capable of segmenting BAC with a R^2 correlation > 0.95 with ground truth. The estimated area using this model were strongly correlated with calcification volume ($R^2 = 0.84$) and calcification mass ($R^2 = 0.87$) on breast computed tomography. However, some notable advantages of our model over these previously developed tools are worth noting. First, we did not input any information regarding BAC quantity for CNN training, whereas Guo and Wang's works relied on manual, pixel-by-pixel BAC annotations as ground truth [64], [69]. Our weakly supervised approach yielded a twofold benefit: a considerable facilitation in the dataset formation (as our readers only had to classify each image either as BAC+ or BAC-) and a sizable computational efficiency, given that we obtained good estimations of BAC burden as a by-product of BAC detection using a relatively simple CNN, with fast processing times (around 1 s for each image). Furthermore, differently from previous works, we tested our model on an independent test set which reflected real world BAC prevalence (around 12%), whereas the datasets employed in other works [64], [69] included a majority of BAC+ patients, which might have led to model overfitting [94]. Instead, we chose to artificially augment BAC prevalence to 30% only in the training set, in order to select the best performing hyperparameters for BAC detection, and then reverted to a 12% prevalence for validation and testing. Therefore, as we already tested the CNN on a realistic and imbalanced set, we hypothesise that our model's performances will be stable and robust in the upcoming external validation, where BAC are the minority class.

A visual examination of the wrong predictions by our model showed that the majority of false positives were due to small calcifications that mimicked BAC usual appearance, *i.e.*, lined-up, punctuated calcifications often within linear formations such as skin folds or Cooper's ligaments (**Figure 16a,b**). Conversely, false negatives occurred in situations where BAC detection could be difficult also for trained human readers, such as BAC in dense breasts (**Figure 16c**) or very faint BAC. Of note, the latter could perhaps be of lower clinical value for CVD risk prediction.

Our work presents some limitations. First, the model was trained and tested on a consecutive series of women from a single institution studied using two mammographic units from a single manufacturer. Even though our dataset consisted of over 1400 patients and we allotted 15% of the dataset for independent testing, an external validation of our model on different machines is warranted. Second, the correlation coefficient of BAC burden estimation with manual measurement in our work (0.88) was marginally lower than those reported in previous studies (0.95 [69] and 0.94 [64]). However, we must note that differently from previous studies we did not train our model using manual segmentations as ground truth, and that extremely precise BAC segmentation may not be necessary from a clinical point of view. Indeed, according to the most recent meta-analysis on the association between BAC and CVD [26], only moderate and severe BAC (*i.e.*, extensive calcifications on one or more vessels, clouding vessels' lumen and involving notable portions of their length – see **Figure 6a**) were associated with coronary artery disease. Therefore, our model would still allow to identify women at higher CVD risk, albeit with a less precise BAC segmentation. Third, we performed a stratified split of BAC+ cases into training, validation and test sets to preserve the BAC age distribution and avoid any age-related potential bias. However, this procedure might have introduced some degree of sampling bias, considering the age constrains in the randomization. Finally, we did not perform any experimental comparison between the performances of our model and that obtainable with other available CNN architectures, such as ResNet 50 or DenseNet. However, such comparison was beyond the aims of the present work.

In conclusion, we developed a CNN that can detect BAC with good performance (AUC-ROC of 0.94 in the test set) and can also output a segmentation of BAC with a very strong correlation with manual measurements ($\rho = 0.88$). The integration of our model to clinical practice could improve BAC reporting without increasing clinical workload, potentially facilitating large scale studies on the impact of BAC use as a biomarker to consistently guide cardiovascular risk assessment and management, ultimately contributing to raise awareness on women cardiovascular health in the context of mammographic screening practice.

Section III:

Comparative study of CNN architectures for detection and quantification of BAC

Based on:

- N Mobini, D Capra, A Colarieti, M Zanardo, G Baselli, and F Sardanelli. Deep transfer learning for detection of breast arterial calcifications on mammograms: a comparative study. *European Experimental Radiology* (2024)

  DOI: 10.1186/s41747-024-00478-6

Background

Cardiovascular diseases (CVD) are the primary cause of mortality and morbidity in women worldwide [1], [2]. Traditional risk scores such as the Framingham score often underestimate the risk in women, leading to missed opportunities for early diagnosis and appropriate primary prevention [29], [76], [77], [79]. Over the past decades, breast arterial calcifications (BAC) have been advocated as a promising sex-specific biomarker of CVD to improve women's cardiovascular stratification [12], [31], [78], [95]. BAC are medial calcium depositions detectable as parallel line opacities on about 13% of routine mammograms [14], [24] and have been shown to be associated with an elevated hazard of CVD, independent of most conventional risk factors such as smoking [22], [96], [97]. With the increasing use of mammography for breast cancer screening, BAC present an opportunity for CVD risk stratification in asymptomatic women [78], [98]. Nevertheless, their assessment is a time-consuming manual task, vulnerable to intra- and inter-observer variability [35], [92]; also, the considerable diversity of BAC's appearance and the lack of a standard reporting guideline limited their adoption as a robust imaging biomarker in clinical practice [13], [99].

Automated methods using artificial intelligence (AI) have been recommended in the literature to overcome the intrinsic limitations of BAC detection [64], [66], [69]. The potential capability of deep learning (DL)-based approaches in extracting complex topologies of large datasets, could improve the reproducibility of diagnosis while reducing radiologists' post-processing workload. A twelve-layer deep convolutional neural network (CNN) was the first DL model developed for pixel-wise patch-based BAC detection and exhibited comparable overall performances to a human expert considering the free-response receiver operating characteristic (FROC) analysis [64] analysis [64]. In subsequent studies, modified versions of U-Net were explored for the similar purpose of segmenting calcified vessels and achieved higher levels of accuracy [66], [69]. However, training supervised learning models requires large-scale images with manual segmentation-level annotations, therefore still exposing the models to biases related to the inherent variability of human assessment. Nonetheless, techniques such as transfer learning from a pretrained CNN are well recognized to mitigate this issue [61], [62].

In a recent study [23] addressing automatic BAC detection and quantification, we proposed a novel transfer learning-based weakly supervised framework that effectively reduced operator dependency. By formulating the problem as a simple dichotomous classification task that only requires image-level annotations, i.e. BAC or non-BAC labels instead of time-consuming pixel-by-pixel ground truth, the approach allowed estimation of calcified regions through weak supervision. Further improvements were achieved by fine-tuning a pre-trained VGG16 classification model on challenging open-source datasets, allowing the transfer of previously acquired knowledge for solving the specific BAC classification problem, without starting from scratch. Despite the study demonstrated promising results in BAC recognition, it primarily focused on optimizing VGG16 architecture, leaving the exploration of the optimal models among the state-of-the-art deep CNN networks as an open challenge subject to further research.

In this article, we compare the performance of different neural network architectures using a deep transfer learning strategy and aim to find the best models for the binary classification task of discriminating mammograms with and without BAC. The findings would assist researchers in selecting exemplary networks for detecting BAC and developing efficient tools for early CVD risk stratification, with the potential for widespread integration into clinical practices.

Material and Methods

Dataset description

The dataset as well as the preparation process were similar to our previously published study ([23]. In summary, this retrospective single-center study included 1493 screening mammography exams acquired using full-field digital IMS systems (Giotto IMAGE 3D or Giotto TOMO series). Each examination consisted of bilateral craniocaudal (CC) and mediolateral oblique (MLO) view images of both breasts, which were reviewed by four expert readers and labelled as either BAC or non-BAC. These annotated labels were encoded as the ground truth for model training, hyperparameter tuning, and performance evaluation. As fully discussed earlier (also in **Section II**) [23], BAC incidence was found to be positively associated with women's age [35] and therefore, a specific strategy was conducted to split the data while preserving BAC age distribution; 70% of the exams were allocated to the training subset, 15% to the validation subset, and the remaining 15% to the testing subset. The training images were further randomly under-sampled reaching a BAC prevalence of 30%, to alleviate the classification bias toward the majority class of our imbalanced dataset [81], [82]. The validation and testing subsets were instead fully preserved to ensure an accurate representation of the real-world BAC prevalence. Similarly (as illustrated in **Figure 5**), the data preprocessing step involved extracting the breast regions from the dark background pixels by defining the smallest rectangular area surrounding the breast and rescaling the cropped images to a common fixed-size dimension of 1536 × 768 pixels accepted by all the networks. Histogram analysis and Otsu's thresholding method were used to separate the image pixels into tissue and background [83], [84]. Next, over-threshold pixel values corresponding to the breast region were normalized to reduce the intensity variation of mammographic images caused by technical or biological reasons, thus enhancing the convergence of training.

Training setting

Throughout the experiment, we used a total of eleven deep neural networks, namely Xception [100], VGG16, VGG19 [57], ResNet50V2, ResNet101V2, ResNet152V2 [101], MobileNet [102], MobileNetV2 [103], DenseNet121, DenseNet169, and DenseNet201 [67]. The models were previously pretrained on the ImageNet dataset, comprising more than 14 million annotated color images from 1000 categories [104], and were publicly available through Keras Applications. Then, we implemented a uniform transfer learning strategy and a harmonized set of hyperparameters across all the networks to directly compare the performance of the various architectures, regardless of specific optimization. Since the source and our target datasets were from disparate domains, the classification layer of each was replaced with two randomly initialized fully connected layers followed by a sigmoid activation function in the output layer, as appropriate for the binary BAC classification task. For transferring knowledge, all layers in the convolutional base except the last were kept frozen with initial pre-trained weights, while the rest of the deeper layers and the new classification top were fine-tuned on the mammographic dataset specifically, as illustrated in **Figure 18**.

The training and evaluations were implemented using Keras and TensorFlow2 framework of Python V3.8, on a system equipped with Intel Core i7-10700KF CPU, NVIDIA GeForce RTX 3080 card, and 10GB video memory (vRAM). Each network was retrained over 100 epochs, with a batch size of eight images limited by the available graphic processing unit (GPU) memory. The Adam optimizer with an initial learning rate of 10^{-3} decayed by a cosine annealing scheduler was exploited to minimize the binary cross-entropy loss [85], [86]. Furthermore, augmentation techniques including random rotation, shifting, flipping, and zooming were applied online to the training data to avoid overfitting and improve robustness of the classifications [105], [106]. Model checkpoint executed on the validation subset while tuning the hyperparameters and the best-performing configuration was saved at the end of each training.

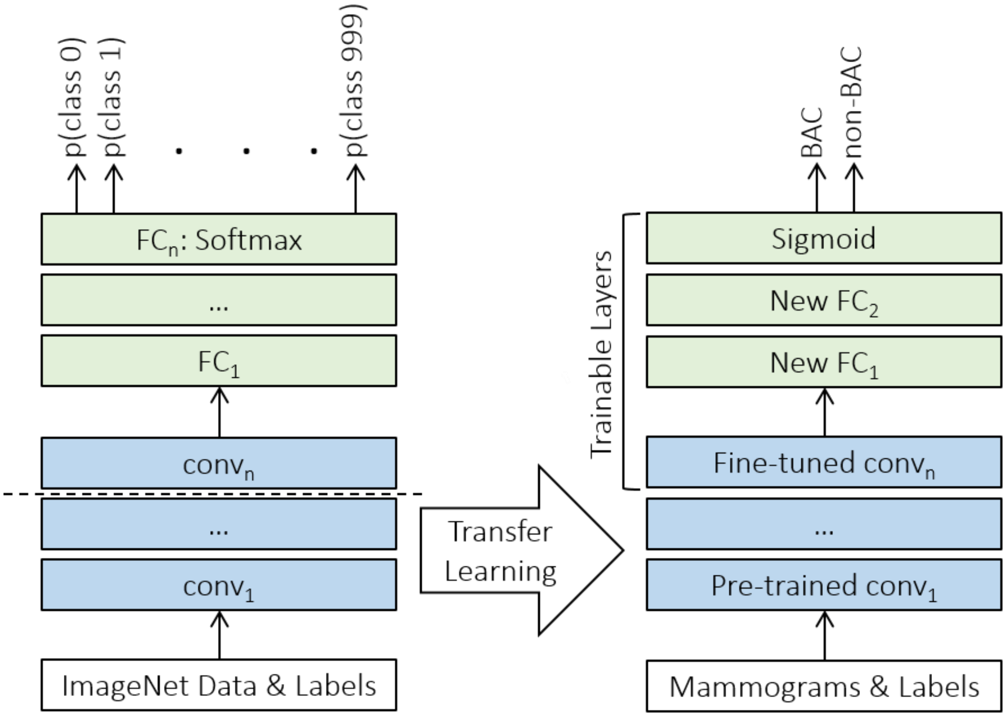


Figure 18 The transfer learning strategy using fine-tuning. *FC,* fully connected

Performance evaluation

The Kolmogorov–Smirnov test was used to evaluate the normality. The continuous variables were presented by mean ± standard deviation (SD) or median and interquartile range (IQR) according to their distribution. Further, the Mann–Whitney *U* test was adopted to evaluate the age distribution disparities between the BAC and non-BAC groups, where a *p*-value less than 0.05 was considered statistically significant [89].

The overall diagnostic performance of the models against the ground truth labels was evaluated using the receiver operating characteristic curve (ROC) and area under the curve (AUC), independent of classification thresholds. Then, the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values were calculated at an optimal cutoff point, corresponding to the maximum F_1 score achieved by each network on the validation dataset. The F_1 score is a harmonic mean of precision and recall metrics that sought to balance the concerns of both classes in our binary classification problem:

$$F_1\ score = \frac{2\,precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Furthermore, we conducted a qualitative evaluation of the models' detection and localization abilities using the generalized gradient-weighted class activation mapping (Grad-CAM++) method, which can provide promising reader-interpretable visual explanation of the CNN models in the presence of multiple object instances within a single image, compared to the state-of-the-art [87], [88]. The technique exploited the last convolutional layer's rich semantic and spatial information to generate a heatmap that highlighted the most informative pixels contributing to the decision-making process of the network [88], [107]. To rank these visual explanations in a somewhat quantitative manner, we assessed the spearman correlation coefficient of the estimated calcified region delineated through thresholding of the heatmaps [23], against the corresponding manual measurements of BAC lengths previously measured in a subgroup of BAC exams with MLO views [35].

Results

The ground truth annotation indicated the presence of BAC in 194/1493 women (13.0%) and 581/5972 images (9.7%). The participants' median age was 59 years (interquartile range (IQR) 52−68), where women with BAC had a significantly higher median age of 70.5 years (IQR 60–73) compared to non-BAC women (median age 57, IQR 52–65, $p < 0.001$). Following data partitioning, 410 women were assigned for training (1640 views, including 398 BAC), 222 for validating (888 views, including 89 BAC), and 229 for testing (916 views, including 94 BAC). The training subset BAC prevalence was artificially increased by random under-sampling to address the class imbalance bias. **Table 3** presents the final composition of the subsets. The patient-level data splitting prevented biases that could arise from allocating different views of an individual to different subsets.

Figure 19 shows the CNNs’ learning curves during the training and validating processes. The ROC curves and AUC values derived from fine-tuning each network on the mammographic dataset are presented in **Figure 20**. The AUC values above 0.80 in the training dataset achieved by MobileNet, VGG, and DenseNet architectures indicated their good discriminatory ability between BAC and non-BAC images. The performances could be further confirmed by assessing the independent test subset, where VGG16, MobileNet, and DenseNet201 achieved the most three accurate detections with AUC values of 0.79, 0.78, and 0.77, respectively. On the other hand, ResNet152V2 (0.67) and Xception (0.63) exhibited a comparatively lower performance, while ResNet101V2 demonstrated the worst result yielding an AUC of 0.51, close to a random chance classifier. Considering the convergence failure of ResNet101V2 also on the training and validation subsets, the network was eliminated from further analysis.

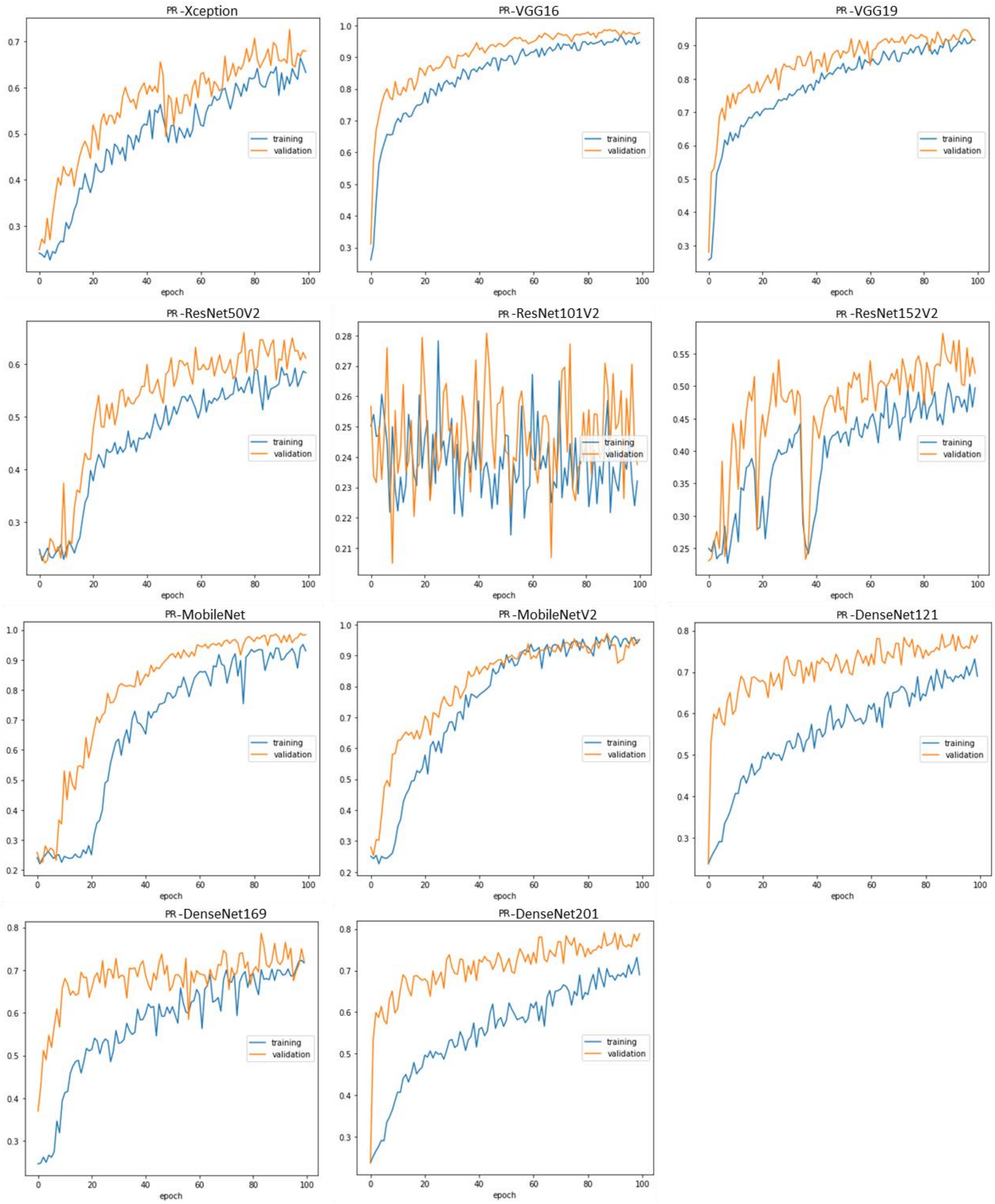


Figure 19 Learning curves (AUC-PR for training and validation) of the selected CNN architectures over 100 epochs. Due to the highly imbalanced dataset, the area under the precision-recall curve was monitored and the parameters pertaining to the maximum quantity generated the models configurations

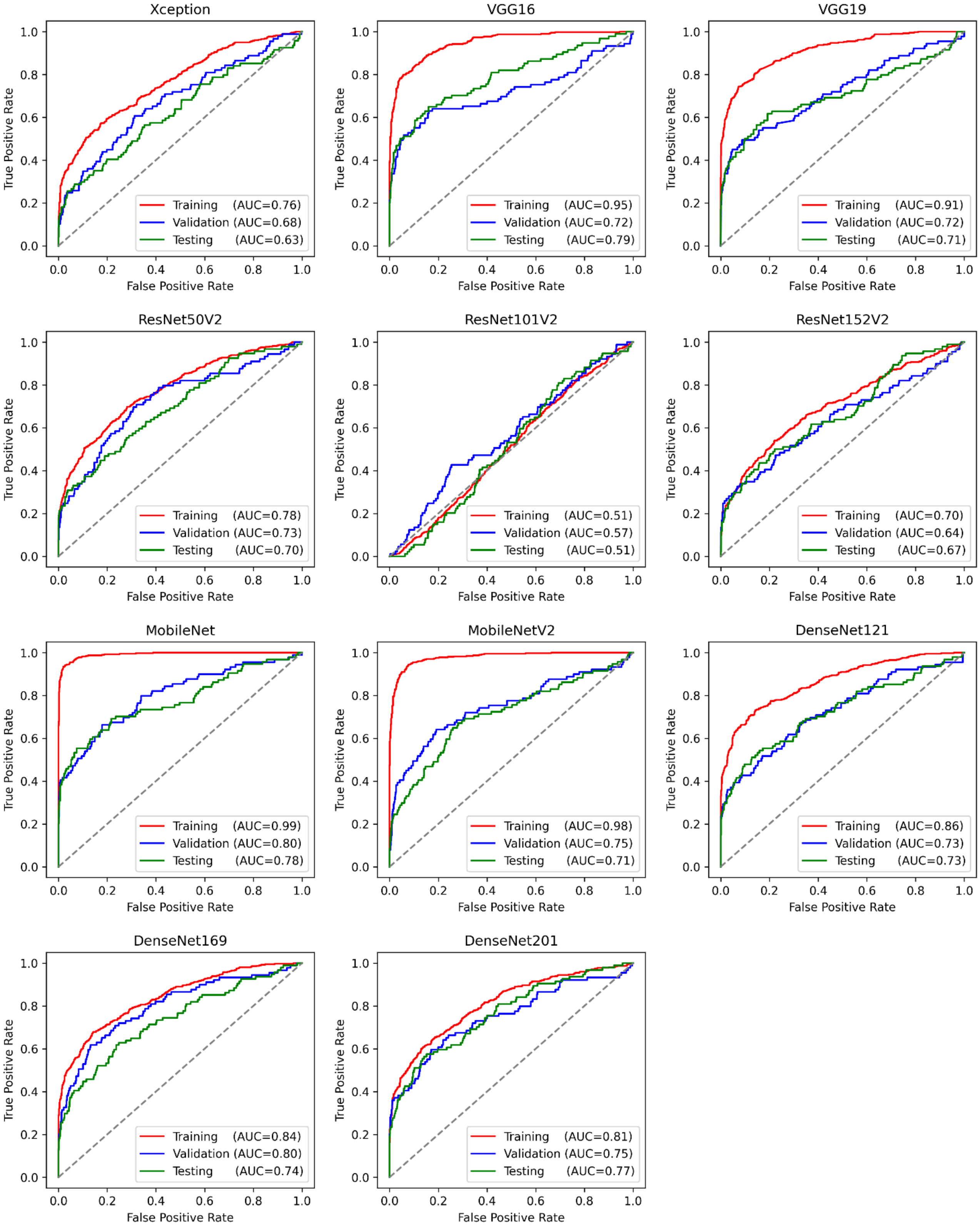


Figure 20 ROC curves and AUC values for each of the networks

Table 5 reports the quantitative prediction results of the networks at their optimal operating point. Among the models tested, VGG16 (0.53), MobileNet (0.51), and VGG19 (0.46) achieved the highest F1 scores, while ResNet50V2 (0.33), Xception (0.31), and ResNet152V2 (0.29) placed at the bottom. In terms of true positive detections, VGG16 ranked first correctly identifying 47/94 BAC images in the testing subset, higher than VGG19 and MobileNet each with 38/94 and 34/94 correct BAC detections. The architecture characteristics and the computational loads are summarized in **Table 6**. In general, fine-tuning each epoch of the pre-trained models on our mammographic dataset took between 241 seconds for lightweight MobileNet to 271 seconds for ResNet152V2 with the highest total number of parameters (around 59.5 million).

Table 5 Classification performances of the fine-tuned models

	Training					**Validation**					**Testing**				
	TN	**TP**	**FN**	**FP**	**F1**	**TN**	**TP**	**FN**	**FP**	**F1**	**TN**	**TP**	**FN**	**FP**	**F1**
Xception	1192	146	252	50	0.77	772	22	67	27	0.32	767	27	67	55	0.31
VGG16	1219	260	138	23	0.76	762	44	45	37	0.52	785	47	47	37	0.53
VGG19	1216	237	161	26	0.51	761	40	49	38	0.48	787	38	56	35	0.46
ResNet50V2	1209	125	273	33	0.61	784	22	67	15	0.35	801	23	71	21	0.33
ResNet152V2	1225	83	315	17	0.33	791	22	67	8	0.37	809	18	76	13	0.29
MobileNet	1242	247	151	0	0.62	793	36	53	6	0.55	817	34	60	5	0.51
MobileNetV2	1232	280	118	10	0.45	778	34	55	21	0.47	781	27	67	41	0.33
DenseNet121	1215	187	211	27	0.61	777	32	57	22	0.45	800	30	64	22	0.41

DenseNet169	1196	199	199	46	0.49	763	37	52	36	0.46	784	34	60	38	0.41
DenseNet201	1227	141	257	15	0.81	790	32	57	9	0.49	807	26	68	15	0.39

TN true negative, *TP* true positive, *FN* false negative, *FP* false positive, *F1* F_1 score

Table 6 Comparison of the deployed networks characteristics

Network	**Depth**	**Number of parameters (10^6)**		**Model size (MB)**	**Training time (s)/ epoch (s)**	**Testing time (ms)/ image**
		Total	**Trainable**			
Xception	36	22.04	4.34	117	251.6	39.4
VGG16	16	15.11	2.75	78.7	255.2	31.2
VGG19	19	20.42	2.75	99	262.6	38.4
ResNet50V2	50	24.74	2.23	111	242.5	28.0
ResNet152V2	152	59.51	2.23	245	271.2	61.8
MobileNet	28	3.88	1.71	28.1	241.1	15.9
MobileNetV2	53	3.04	1.20	21.2	245.9	18.9
DenseNet121	121	7.69	0.69	35.8	246.1	30.7
DenseNet169	169	13.62	1.02	61.4	249.3	39.3
DenseNet201	201	19.43	1.16	84.8	261.4	47.8

Several examples of the Grad-CAM++ heatmaps generated from image-level ground truth are presented in **Figure 21**, for an intuitive comparison of the best performances within various burden of BAC. The localization maps mainly emphasized the regions of BAC, while de-emphasizing the overall breast with varying extent of precision. Among them, the heatmaps created by the VGG architecture explicitly outperformed those by the others in the majority of examples and provided discriminative image regions of interest that could accurately localize the area related to BAC with finer-grained details. Additional examples of wrong predictions are presented in **Figure 22**. A visual assessment of the false negative detections revealed that variables such as dense tissue or faint BAC affected the models' accuracy in predicting the presence of BAC, but no consistent patterns were observed across different CNNs in the false positives.

The superiority of the VGG16 architecture in estimating BAC region was further supported by the Spearman's rank correlation analysis (Spearman $\rho = 0.68$, $p < 0.001$), performed in a subgroup of 56 exams comprising 94 BAC out of 112 total views (**Figure 23**). Meanwhile, the MobileNet ability to accurately visualize BAC areas within the images appeared inadequate and showed a poor correlation with the manually measured length, despite the good quantitative classification results.

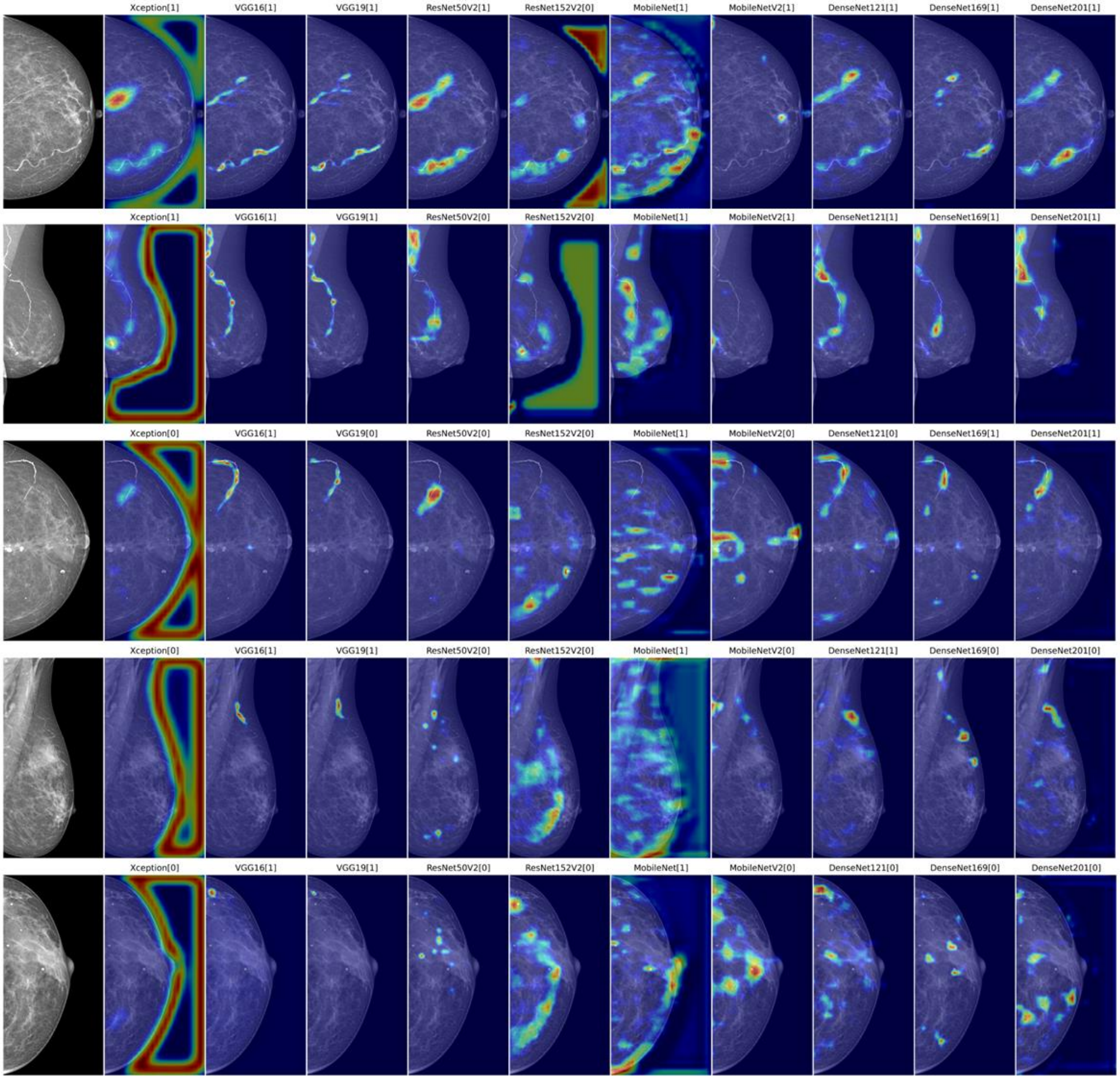


Figure 21 From left to right: original images (cropped to minimize the background), and examples of Grad-CAM++ heatmaps with the binary predicted labels (BAC:1 and non-BAC:0) generated from Xception, VGG16, VGG19, ResNet50V2, ResNet152V2, MobileNet, MobileNetV2, DenseNet121, DenseNet169, and DenseNet201. ResNet101V2 was excluded from the analysis due to its limited ability to effectively learn BAC features

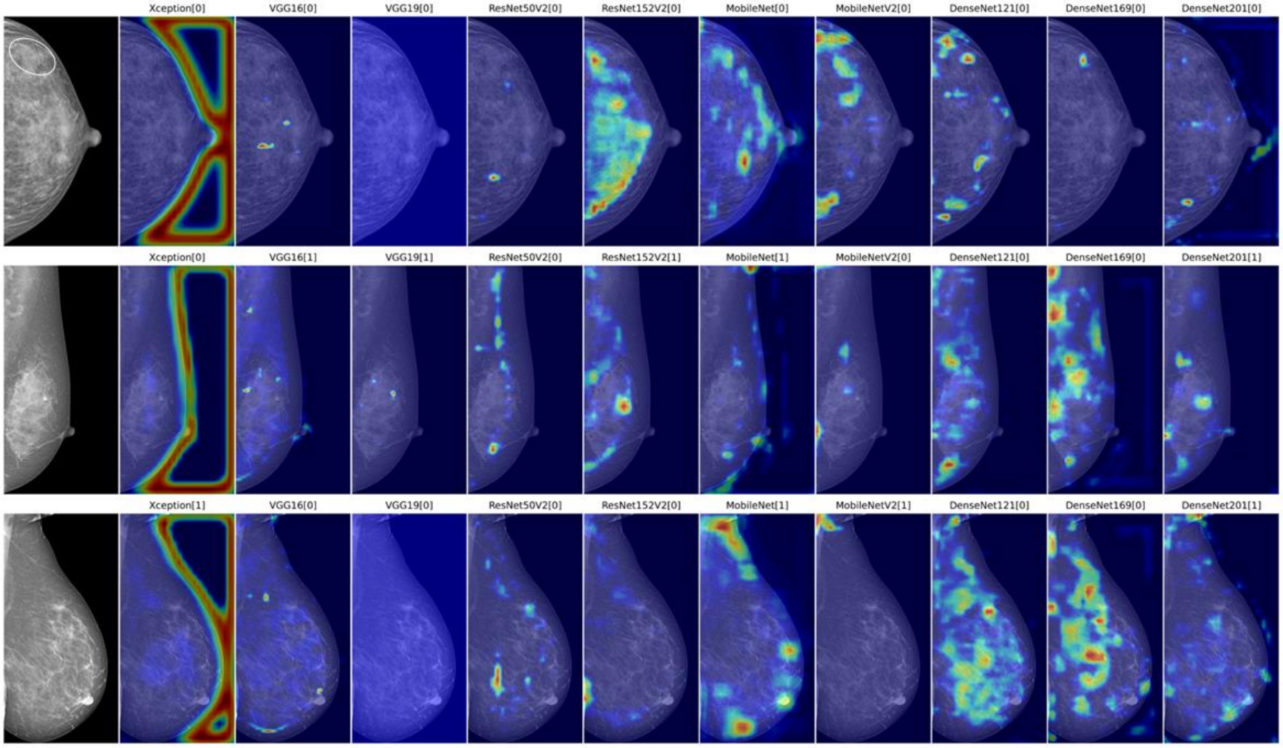


Figure 22 Examples of misclassifications. From top to bottom: a positive case with minor BAC concealed under dense breast tissue (circle) misclassified as negative, and two negative cases with benign calcifications and skinfolds mistaken as BAC by some CNNs

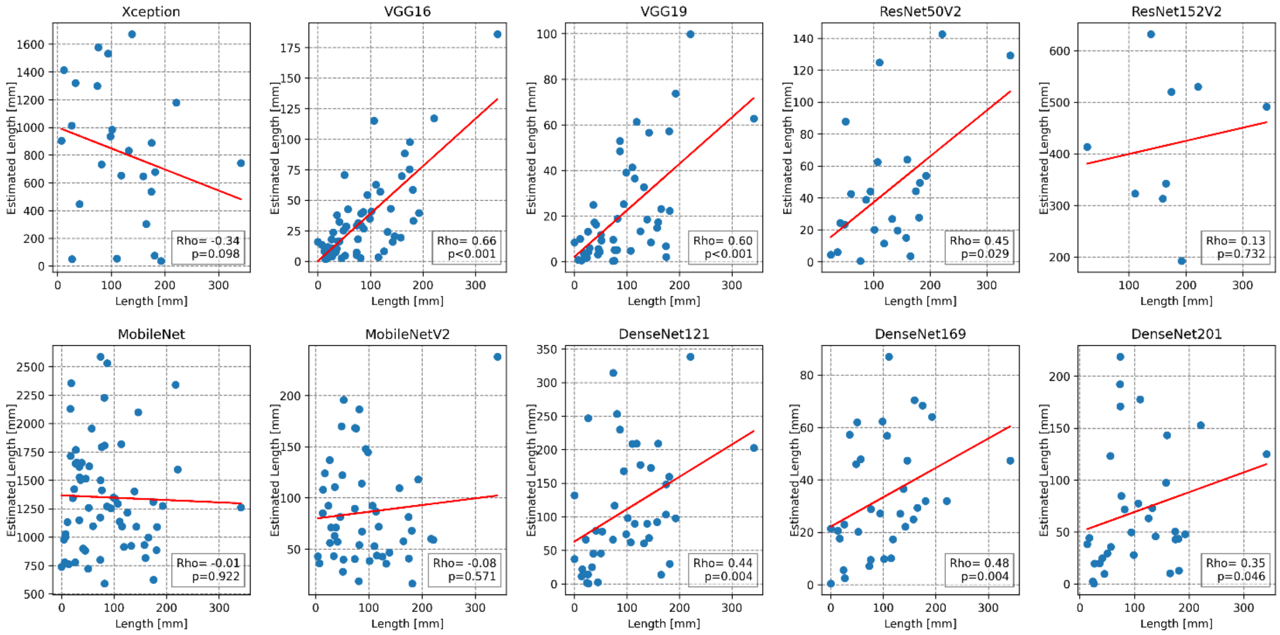


Figure 23 Scatterplots comparing the estimated BAC length (y-axis) and the manual length measurements (x-axis) in a subgroup of 56 women with 112 MLO views (red line, linear regression). Key statistics, including Spearman's rank correlation coefficient (rho) and p-value (p), are provided in the lower right corner of each plot

Discussion and Conclusions

In this work, we implemented different pretrained convolutional neural networks of varying depths and explored their performances for the automatic detection of BAC, a mammographic finding not related to breast cancer, which has been recently identified as a women-specific biomarker of cardiovascular risk. The performance ranking of the CNNs on the mammography dataset revealed that increasing depth and complexity may not necessarily improve the classification outcomes, as the best results were obtained by using relatively shallow models like VGG and MobileNet architectures in terms of higher AUC-ROC values. The highest F1 score and best visual explanation has been obtained by VGG16. When a biomarker like BAC is under consideration, these results play in favour of lightweight models be implemented quickly and efficiently even with limited hardware resources.

The use of AI networks, particularly DL-based approaches, has been explored in several studies as a solution to overcome the intrinsic limitations of manual BAC assessments [64], [66], [69]. Nonetheless, they predominantly relied on pixel-level segmentation, demanding meticulous manual annotation and often subject to observer variability. Therefore, the current study addressed the BAC classification problem based on a recently developed transfer learning-based weakly supervised framework that allows for estimation of calcified regions using only image-level annotations, thus further reducing operator dependency and radiologists' workload [23]. The shift toward transfer learning as a potential solution to the data scarcity problem, leveraged previously acquired knowledge of a well-established CNN network from large annotated open-source datasets and efficiently fine-tuned the relevant learned features for the specific BAC classification task in hand, rather than training from scratch [23], [61], [62].

According to our results in the testing subset, VGG16, MobileNet, and DenseNet201 exhibited the most accurate BAC detections with AUC-ROC values close to 0.80. In this setting, depth and complexity of the neural networks do not necessarily guarantee superior performance in classifying mammography images. Both VGG16 and MobileNet are relatively shallow networks. VGG16 is characterized by a straightforward sequential architecture with small 3×3 convolutional filters, allowing a more focused learning of relevant features, effective in various computer vision tasks [57]. MobileNet uses depth-wise separable convolutions that reduces the overall number of parameters, making it a lightweight and efficient model for mobile and embedded vision applications [102]. The other tested architectures, such as Xception and ResNetV2 [100], [101], are also recognized for their efficacy in attaining state-of-the-art results, though their performances may be influenced by the specific characteristics of the dataset and task at hand. The superiority of smaller networks to their deeper counterparts, when it comes to medical dataset often with limited number of samples, have also been observed in some other studies exploring DL techniques for a wide variety of diagnostic medical imaging applications such as chest x-ray classification or breast cancer diagnosis [108], [109].

The qualitative assessment of performances through generalized Grad-CAM complemented the quantitative analysis based on the AUC-ROC and F1 score metrics. Notably, the inherent simplicity and uniformity of the VGG16 architecture facilitated a more precise representation of the distinctive patterns associated with BAC on mammograms. These heatmaps hold potential for application in weakly-supervised segmentation, as we previously elaborated in [23], wherein BAC localization is achieved by a CNN trained only on image-level labels, without requiring pixel-by-pixel ground truth annotations. Consequently, an estimation of the BAC burden, as a by-product of the automatic detection framework, could be obtained by using simple thresholding and segmenting out the most intense pixels of the Grad-CAM++ heatmaps which encapsulated calcified areas of the original image. Furthermore, this visual approach introduces the prospect of integrating human expertise into the decision-making loop, as clinicians could contribute their insights to further refine the segmentation or improve the CNN model based on the visual cues provided by the heatmaps.

The comparability of our method and the other cited research may be limited as detailed BAC segmentations were mostly used to evaluate the outcomes [64], [66], [69]. The original study that proposed the novel weakly supervised BAC detection framework, achieved a promising performance by fine tuning VGG16 with an AUC-ROC of 0.94 in the testing subset and a strong correlation with manual BAC measurements (Spearman $\rho = 0.88$, $p < 0.001$) [23], surpassing all models in our analysis. Indeed, in the current experiment, a uniform transfer learning strategy followed by a harmonized hyperparameter set were adopted across all networks, which were probably not selected as precisely as in [23], since our priority was comparing architectures rather than optimizing each model. Furthermore, all models were evaluated on an independent testing subset reflecting real-world BAC prevalence of around 12%, as in the original research [23]. This realistic imbalanced subset ensures the CNNs' stability and robustness for future studies with BAC as the minority class, in contrast to the previous research that predominantly included BAC exams, risking model overfitting.

The present study has some limitations. First, the dataset included in this retrospective analysis was obtained from a single imaging center using two mammographic devices by the same manufacturer, which may introduce potential biases and constrain the generalizability of the findings. Second, while using a uniform training strategy across all neural network architectures enabled a fair comparison, it may limit the full potential of each model. Further research is warranted to explore customized configurations tailored to the unique characteristics of each architecture to exploit their maximum capabilities and optimize their performances. Lastly, the chosen metrics for performance evaluation provide robust insights, yet the clinical relevance of these metrics to real-world patient outcomes remains an area for future investigation.

In conclusion, this study demonstrated the efficacy of employing deep transfer learning-based approaches for BAC on mammograms, where networks such as VGG16 and MobileNet outperformed their deeper more complex counterparts. The competitive performance and notable computational efficiency of simpler networks highlighted the viability of adopting such models in clinical settings with substantial savings in both time and resources. Our extensive experiment and evaluations, both quantitative and qualitative, could provide valuable insights for researchers in selecting exemplary network architectures for automatic BAC detection and developing efficient tools for early CVD risk stratification in asymptomatic women. Further research is required to address the limitations and validate the models using a larger diverse study population, ultimately paving the way for integrating the models into clinical practices without any time loss for radiologists and fostering awareness of women's cardiovascular health in the context of widespread mammographic screening programs. Conversely, the use of mammographic images for cardiovascular risk stratification could be an added new motivation for participation to screening mammography programs, thus reinforcing its value also for secondary prevention of breast cancer in the female population [31]. As the field continues to evolve, a balance between diagnostic accuracy, computational efficiency, and real-world applicability will be crucial.

References

[1] C. W. Tsao *et al.*, “Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association,” *Circulation*, vol. 147, no. 8, Feb. 2023, doi: 10.1161/CIR.0000000000001123.

[2] A. Timmis *et al.*, “European Society of Cardiology: cardiovascular disease statistics 2021,” *Eur Heart J*, vol. 43, no. 8, pp. 716–799, Feb. 2022, doi: 10.1093/eurheartj/ehab892.

[3] M. Woodward, “Cardiovascular Disease and the Female Disadvantage,” *Int J Environ Res Public Health*, vol. 16, no. 7, p. 1165, Apr. 2019, doi: 10.3390/ijerph16071165.

[4] A. M. Möller-Leimkühler, “Gender differences in cardiovascular disease and comorbid depression.,” *Dialogues Clin Neurosci*, vol. 9, no. 1, pp. 71–83, Mar. 2007, doi: 10.31887/DCNS.2007.9.1/ammoeller.

[5] A. Iorga, C. M. Cunningham, S. Moazeni, G. Ruffenach, S. Umar, and M. Eghbali, “The protective role of estrogen and estrogen receptors in cardiovascular disease and the controversial use of estrogen therapy,” *Biol Sex Differ*, vol. 8, no. 1, p. 33, Dec. 2017, doi: 10.1186/s13293-017-0152-8.

[6] A. H. E. M. Maas *et al.*, “Cardiovascular health after menopause transition, pregnancy disorders, and other gynaecologic conditions: a consensus document from European cardiologists, gynaecologists, and endocrinologists,” *Eur Heart J*, vol. 42, no. 10, pp. 967–984, Mar. 2021, doi: 10.1093/eurheartj/ehaa1044.

[7] A. Timmis *et al.*, “European Society of Cardiology: Cardiovascular Disease Statistics 2017,” *Eur Heart J*, vol. 39, no. 7, pp. 508–579, Feb. 2018, doi: 10.1093/eurheartj/ehx628.

[8] B. Vogel *et al.*, “The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030,” *The Lancet*, vol. 397, no. 10292, pp. 2385–2438, Jun. 2021, doi: 10.1016/S0140-6736(21)00684-X.

[9] N. K. Wenger, “Transforming Cardiovascular Disease Prevention in Women: Time for the Pygmalion Construct to End,” *Cardiology*, vol. 130, no. 1, pp. 62–68, 2015, doi: 10.1159/000370018.

[10] R. M. Trimboli *et al.*, “Do we still need breast cancer screening in the era of targeted therapies and precision medicine?,” *Insights Imaging*, vol. 11, no. 1, p. 105, Dec. 2020, doi: 10.1186/s13244-020-00905-3.

[11] M. M. Gianino *et al.*, “Organized screening programmes for breast and cervical cancer in 17 EU countries: trajectories of attendance rates,” *BMC Public Health*, vol. 18, no. 1, p. 1236, Dec. 2018, doi: 10.1186/s12889-018-6155-5.

[12] 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, pp. 1094–1101, Feb. 2019, doi: 10.1161/CIRCULATIONAHA.118.038092.

[13] R. M. Trimboli, M. Codari, M. Guazzi, and F. Sardanelli, “Screening mammography beyond breast cancer: breast arterial calcifications as a sex-specific biomarker of cardiovascular risk,” *Eur J Radiol*, vol. 119, p. 108636, Oct. 2019, doi: 10.1016/j.ejrad.2019.08.005.

[14] L. Zazzeroni, G. Faggioli, and G. Pasquinelli, “Mechanisms of Arterial Calcification: The Role of Matrix Vesicles,” *European Journal of Vascular and Endovascular Surgery*, vol. 55, no. 3, pp. 425–432, Mar. 2018, doi: 10.1016/j.ejvs.2017.12.009.

[15] R. G. Micheletti, G. A. Fishbein, J. S. Currier, and M. C. Fishbein, “Mönckeberg Sclerosis Revisited: A Clarification of the Histologic Definition of Mönckeberg Sclerosis,” *Arch Pathol Lab Med*, vol. 132, no. 1, pp. 43–47, Jan. 2008, doi: 10.5858/2008-132-43-MSRACO.

[16] P. A. van Noord, D. Beijerinck, J. M. Kemmeren, and Y. van der Graaf, “Mammograms may convey more than breast cancer risk: breast arterial calcification and arterio-sclerotic related diseases in women of the DOM cohort.,” *Eur J Cancer Prev*, vol. 5, no. 6, pp. 483–7, Dec. 1996.

[17] A. H. E. M. Maas *et al.*, “Breast arterial calcifications are correlated with subsequent development of coronary artery calcifications, but their aetiology is predominantly different,” *Eur J Radiol*, vol. 63, no. 3, pp. 396–400, Sep. 2007, doi: 10.1016/j.ejrad.2007.02.009.

[18] B. B. NIELSEN and N. V. HOLM, “Calcification in breast arteries,” *Acta Pathologica Microbiologica Scandinavica Series A :Pathology*, vol. 93A, no. 1–6, pp. 13–16, Mar. 1985, doi: 10.1111/j.1699-0463.1985.tb03914.x.

[19] A. H. E. M. Maas and P. A. de Jong, “Mammograms to catch many birds with one stone,” *Eur Heart J*, vol. 42, no. 34, pp. 3371–3373, Sep. 2021, doi: 10.1093/eurheartj/ehab522.

[20] C. Iribarren *et al.*, “MultIethNic Study of BrEast ARterial Calcium Gradation and CardioVAscular Disease: cohort recruitment and baseline characteristics,” *Ann Epidemiol*, vol. 28, no. 1, pp. 41-47.e12, Jan. 2018, doi: 10.1016/j.annepidem.2017.11.007.

[21] C. Iribarren *et al.*, “Association of Breast Arterial Calcification Presence and Gradation with the Ankle-Brachial Index among Postmenopausal Women.,” *Eur J Cardiovasc Med*, vol. 5, no. 5, pp. 544–551, Nov. 2018.

[22] E. J. E. Hendriks *et al.*, “Breast Arterial Calcifications and Their Association With Incident Cardiovascular Disease and Diabetes,” *J Am Coll Cardiol*, vol. 65, no. 8, pp. 859–860, Mar. 2015, doi: 10.1016/j.jacc.2014.12.015.

[23] N. Mobini *et al.*, “Detection and quantification of breast arterial calcifications on mammograms: a deep learning approach,” *Eur Radiol*, vol. 33, no. 10, pp. 6746–6755, May 2023, doi: 10.1007/s00330-023-09668-z.

[24] E. J. E. Hendriks, P. A. de Jong, Y. van der Graaf, W. P. Th. 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, pp. 11–20, Mar. 2015, doi: 10.1016/j.atherosclerosis.2014.12.035.

[25] P. F. Schnatz, M. A. Rotter, S. Hadley, A. A. Currier, and D. M. O’Sullivan, “Hormonal therapy is associated with a lower prevalence of breast arterial calcification on mammography,” *Maturitas*, vol. 57, no. 2, pp. 154–160, Jun. 2007, doi: 10.1016/j.maturitas.2006.12.002.

[26] 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, p. e0236598, Jul. 2020, doi: 10.1371/journal.pone.0236598.

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

[28] Y. E. Yoon *et al.*, “Prediction of Subclinical Coronary Artery Disease With Breast Arterial Calcification and Low Bone Mass in Asymptomatic Women,” *JACC Cardiovasc Imaging*, vol. 12, no. 7, pp. 1202–1211, Jul. 2019, doi: 10.1016/j.jcmg.2018.07.004.

[29] C. Iribarren *et al.*, “Breast Arterial Calcification: a Novel Cardiovascular Risk Enhancer Among Postmenopausal Women,” *Circ Cardiovasc Imaging*, vol. 15, no. 3, Mar. 2022, doi: 10.1161/CIRCIMAGING.121.013526.

[30] P. F. Schnatz, K. A. Marakovits, and D. M. O’Sullivan, “The Association of Breast Arterial Calcification and Coronary Heart Disease,” *Obstetrics & Gynecology*, vol. 117, no. 2, pp. 233–241, Feb. 2011, doi: 10.1097/AOG.0b013e318206c8cb.

[31] V. Magni *et al.*, “Mammography biomarkers of cardiovascular and musculoskeletal health: A review,” *Maturitas*, vol. 167, pp. 75–81, Jan. 2023, doi: 10.1016/j.maturitas.2022.10.001.

[32] R. V. Parikh *et al.*, “Kidney function, proteinuria and breast arterial calcification in women without clinical cardiovascular disease: The MINERVA study,” *PLoS One*, vol. 14, no. 1, p. e0210973, Jan. 2019, doi: 10.1371/journal.pone.0210973.

[33] L. R. Margolies *et al.*, “Breast Arterial Calcification in the Mammogram Report: The Patient Perspective,” *American Journal of Roentgenology*, vol. 212, no. 1, pp. 209–214, Jan. 2019, doi: 10.2214/AJR.18.20171.

[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,” *Eur Radiol*, vol. 31, no. 2, pp. 958–966, Feb. 2021, doi: 10.1007/s00330-020-07136-6.

[35] R. M. Trimboli *et al.*, “Semiquantitative score of breast arterial calcifications on mammography (BAC-SS): intra- and inter-reader reproducibility,” *Quant Imaging Med Surg*, vol. 11, no. 5, pp. 2019–2027, May 2021, doi: 10.21037/qims-20-560.

[36] D. Ružičić *et al.*, “Novel Assessment Tool For Coronary Artery Disease Severity During Screening Mammography,” *Health Care Women Int*, vol. 39, no. 10, pp. 1075–1089, Oct. 2018, doi: 10.1080/07399332.2018.1463226.

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

[38] S. Molloi, T. Xu, J. Ducote, and C. Iribarren, “Quantification of breast arterial calcification using full field digital mammography,” *Med Phys*, vol. 35, no. 4, pp. 1428–1439, Apr. 2008, doi: 10.1118/1.2868756.

[39] M. Minsky, “Steps toward Artificial Intelligence,” *Proceedings of the IRE*, vol. 49, no. 1, pp. 8–30, Jan. 1961, doi: 10.1109/JRPROC.1961.287775.

[40] J. N. Nilsson, *Principles of artificial intelligence*. Springer Science & Business Media, 1980.

[41] S. J. Russell, P. Norvig, and E. Davis, *Artificial intelligence: a modern approach*. 2010.

[42] A. M. TURING, “Computing Machinery and Intelligence,” *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950, doi: 10.1093/mind/LIX.236.433.

[43] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” *IBM J Res Dev*, vol. 3, no. 3, pp. 210–229, Jul. 1959, doi: 10.1147/rd.33.0210.

[44] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[45] 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.

[46] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. Routledge, 1984. doi: 10.1201/9781315139470.

[47] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, Jul. 1998, doi: 10.1109/5254.708428.

[48] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, Aug. 1987, doi: 10.1016/0169-7439(87)80084-9.

[49] M. Haenlein and A. Kaplan, “A brief history of artificial intelligence: On the past, present, and future of artificial intelligence,” *Calif Manage Rev*, vol. 61, no. 4, pp. 5–14, Aug. 2019, doi: 10.1177/0008125619864925.

[50] F. Sardanelli, I. Castiglioni, A. Colarieti, S. Schiaffino, and G. Di Leo, “Artificial intelligence (AI) in biomedical research: discussion on authors’ declaration of AI in their

articles title," *Eur Radiol Exp*, vol. 7, no. 1, p. 2, Jan. 2023, doi: 10.1186/s41747-022-00316-7.

[51] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]," *IEEE Comput Intell Mag*, vol. 5, no. 4, pp. 13–18, Nov. 2010, doi: 10.1109/MCI.2010.938364.

[52] L. Deng, "Deep Learning: Methods and Applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014, doi: 10.1561/2000000039.

[53] I. Castiglioni *et al.*, "AI applications to medical images: From machine learning to deep learning," *Physica Medica*, vol. 83, pp. 9–24, Mar. 2021, doi: 10.1016/j.ejmp.2021.02.006.

[54] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.08458

[55] H.-C. Shin *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Trans Med Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems*, 2012.

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

[58] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[59] H. Greenspan, B. van Ginneken, and R. M. Summers, “Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique,” *IEEE Trans Med Imaging*, vol. 35, no. 5, pp. 1153–1159, May 2016, doi: 10.1109/TMI.2016.2553401.

[60] F. Pesapane, M. Codari, and F. Sardanelli, “Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine,” *Eur Radiol Exp*, vol. 2, no. 1, p. 35, Dec. 2018, doi: 10.1186/s41747-018-0061-6.

[61] 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.

[62] N. Tajbakhsh *et al.*, “Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?,” Jun. 2017, doi: 10.1109/TMI.2016.2535302.

[63] J.-G. Lee *et al.*, “Deep Learning in Medical Imaging: General Overview,” *Korean J Radiol*, vol. 18, no. 4, p. 570, 2017, doi: 10.3348/kjr.2017.18.4.570.

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

[65] K. Wang, N. Khan, and R. Highnam, “Automated Segmentation of Breast Arterial Calcifications from Digital Mammography,” in *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, IEEE, Dec. 2019, pp. 1–6. doi: 10.1109/IVCNZ48456.2019.8960956.

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

[67] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Aug. 2016, [Online]. Available: http://arxiv.org/abs/1608.06993

[68] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci Data*, vol. 4, no. 1, p. 170177, Dec. 2017, doi: 10.1038/sdata.2017.177.

[69] X. Guo *et al.*, "SCU-Net: A deep learning method for segmentation and quantification of breast arterial calcifications on mammograms," *Med Phys*, vol. 48, no. 10, pp. 5851–5861, Oct. 2021, doi: 10.1002/mp.15017.

[70] M. Alamir, M. Alghamdi, F. Collado-Mesa, and M. Abdel-Mottaleb, "Difference-of-Gaussian generative adversarial network for segmenting breast arterial calcifications in mammograms," *Expert Syst Appl*, vol. 217, p. 119506, May 2023, doi: 10.1016/j.eswa.2023.119506.

[71] K. Wang, M. Hill, S. Knowles-Barley, A. Tikhonov, L. Litchfield, and J. C. Bare, "Improving Segmentation of Breast Arterial Calcifications from Digital Mammography: Good Annotation is All You Need," 2023, pp. 134–150. doi: 10.1007/978-3-031-27066-6_10.

[72] S. S. Virani *et al.*, "Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association," *Circulation*, vol. 141, no. 9, Mar. 2020, doi: 10.1161/CIR.0000000000000757.

[73] A. H. E. M. Maas, “Maintaining cardiovascular health: An approach specific to women,” *Maturitas*, vol. 124, pp. 68–71, Jun. 2019, doi: 10.1016/j.maturitas.2019.03.021.

[74] U. N. Khot, “Prevalence of Conventional Risk Factors in Patients With Coronary Heart Disease,” *JAMA*, vol. 290, no. 7, p. 898, Aug. 2003, doi: 10.1001/jama.290.7.898.

[75] M. Zhao *et al.*, “Sex differences in cardiovascular medication prescription in primary care: A systematic review and meta-analysis,” *Journal of the American Heart Association*, vol. 9, no. 11. American Heart Association Inc., 2020. doi: 10.1161/JAHA.119.014742.

[76] J.-W. Suh and B. La Yun, “Breast Arterial Calcification: A Potential Surrogate Marker for Cardiovascular Disease,” *J Cardiovasc Imaging*, vol. 26, no. 3, p. 125, 2018, doi: 10.4250/jcvi.2018.26.e20.

[77] A. C. Moshyedi, A. H. Puthawala, R. J. Kurland, and D. H. O’Leary, “Breast arterial calcification: association with coronary artery disease. Work in progress.,” *Radiology*, vol. 194, no. 1, pp. 181–183, Jan. 1995, doi: 10.1148/radiology.194.1.7997548.

[78] L. Minssen *et al.*, “Breast arterial calcifications on mammography: a new marker of cardiovascular risk in asymptomatic middle age women?,” *Eur Radiol*, vol. 32, no. 7, pp. 4889–4897, Jul. 2022, doi: 10.1007/s00330-022-08571-3.

[79] M. A. Rotter, P. F. Schnatz, A. A. Currier, and D. M. O’Sullivan, “Breast arterial calcifications (BACs) found on screening mammography and their association with cardiovascular disease,” *Menopause*, pp. 276–281, Mar. 2008, doi: 10.1097/gme.0b013e3181405d0a.

[80] G. Litjens *et al.*, “State-of-the-Art Deep Learning in Cardiovascular Image Analysis,” *JACC Cardiovasc Imaging*, vol. 12, no. 8, pp. 1549–1565, Aug. 2019, doi: 10.1016/j.jcmg.2019.06.009.

[81] K. Fujiwara *et al.*, “Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis,” *Front Public Health*, vol. 8, May 2020, doi: 10.3389/fpubh.2020.00178.

[82] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-98074-4.

[83] Deepa S and Bharathi VS, “Efficient ROI Segmentation of Digital Mammogram Images using Otsu’s N thresholding method,” *International Journal of Engineering Research & Technology*, vol. 2, no. 1, pp. 1–6, Jan. 2013.

[84] N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” *IEEE Trans Syst Man Cybern*, vol. 9, no. 1, pp. 62–66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.

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

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

[87] 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,” in *2017 IEEE*

International Conference on Computer Vision (ICCV), IEEE, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[88] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Mar. 2018, pp. 839–847. doi: 10.1109/WACV.2018.00097.

[89] G. Di Leo and F. Sardanelli, "Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach," *Eur Radiol Exp*, vol. 4, no. 1, Dec. 2020, doi: 10.1186/s41747-020-0145-y.

[90] K. L. Wuensch and J. D. Evans, "Straightforward Statistics for the Behavioral Sciences.," *J Am Stat Assoc*, vol. 91, no. 436, p. 1750, Dec. 1996, doi: 10.2307/2291607.

[91] N. K. Wenger *et al.*, "Call to Action for Cardiovascular Disease in Women: Epidemiology, Awareness, Access, and Delivery of Equitable Health Care: A Presidential Advisory From the American Heart Association," *Circulation*, vol. 145, no. 23, Jun. 2022, doi: 10.1161/CIR.0000000000001071.

[92] R. M. Trimboli *et al.*, "Breast arterial calcifications on mammography: intra- and inter-observer reproducibility of a semi-automatic quantification tool," *Radiol Med*, vol. 123, no. 3, pp. 168–173, Mar. 2018, doi: 10.1007/s11547-017-0827-6.

[93] R. Khan and G. L. Masala, "Detecting Breast Arterial Calcifications in Mammograms with Transfer Learning," *Electronics (Basel)*, vol. 12, no. 1, p. 231, Jan. 2023, doi: 10.3390/electronics12010231.

[94] Q. Dong, S. Gong, and X. Zhu, “Imbalanced Deep Learning by Minority Class Incremental Rectification,” *IEEE Trans Pattern Anal Mach Intell*, vol. 41, no. 6, pp. 1367–1381, Jun. 2019, doi: 10.1109/TPAMI.2018.2832629.

[95] T. Chadashvili, D. Litmanovich, F. Hall, and P. J. Slanetz, “Do breast arterial calcifications on mammography predict elevated risk of coronary artery disease?,” *Eur J Radiol*, vol. 85, no. 6, pp. 1121–1124, Jun. 2016, doi: 10.1016/j.ejrad.2016.03.006.

[96] N. G. Galiano, N. Eiro, A. Martín, O. Fernández-Guinea, C. del B. Martínez, and F. J. Vizoso, “Relationship between Arterial Calcifications on Mammograms and Cardiovascular Events: A Twenty-Three Year Follow-Up Retrospective Cohort Study,” *Biomedicines*, vol. 10, no. 12, p. 3227, Dec. 2022, doi: 10.3390/biomedicines10123227.

[97] C. Iribarren, A. S. Go, I. Tolstykh, S. Sidney, S. C. Johnston, and D. B. Spring, “Breast Vascular Calcification and Risk of Coronary Heart Disease, Stroke, and Heart Failure,” *J Womens Health*, vol. 13, no. 4, pp. 381–389, May 2004, doi: 10.1089/154099904323087060.

[98] D. Mantas and C. Markopoulos, “Screening mammography: Usefulness beyond early detection of breast cancer,” *Atherosclerosis*, vol. 248, p. 1, May 2016, doi: 10.1016/j.atherosclerosis.2016.02.019.

[99] J.-Z. Cheng, C.-M. Chen, E. B. Cole, E. D. Pisano, and D. Shen, “Automated Delineation of Calcified Vessels in Mammography by Tracking With Uncertainty and Graphical Linking Techniques,” *IEEE Trans Med Imaging*, vol. 31, no. 11, pp. 2143–2155, Nov. 2012, doi: 10.1109/TMI.2012.2215880.

[100] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” Oct. 2016, [Online]. Available: http://arxiv.org/abs/1610.02357

[101] K. He, X. Zhang, S. Ren, and J. Sun, “Identity Mappings in Deep Residual Networks,” Mar. 2016, [Online]. Available: http://arxiv.org/abs/1603.05027

[102] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 2017, [Online]. Available: http://arxiv.org/abs/1704.04861

[103] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.04381

[104] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[105] L. Perez and J. Wang, “The Effectiveness of Data Augmentation in Image Classification using Deep Learning,” Dec. 2017.

[106] A. Mikolajczyk and M. Grochowski, “Data augmentation for improving deep learning in image classification problem,” in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, IEEE, May 2018, pp. 117–122. doi: 10.1109/IIPHDW.2018.8388338.

[107] G. Baselli, M. Codari, and F. Sardanelli, “Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way?,” *Eur Radiol Exp*, vol. 4, no. 1, Dec. 2020, doi: 10.1186/s41747-020-00159-0.

[108] K. K. Bressem, L. C. Adams, C. Erxleben, B. Hamm, S. M. Niehues, and J. L. Vahldiek, "Comparing different deep learning architectures for classification of chest radiographs," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-70479-z.

[109] L. Tsochatzidis, L. Costaridou, and I. Pratikakis, "Deep learning for breast cancer diagnosis from mammograms — A comparative study," *J Imaging*, vol. 5, no. 3, Mar. 2019, doi: 10.3390/jimaging5030037.

List of Tables

List of Figures