3D CNN for breast density classification in microwave imaging

Biomedical Signal Processing and Control 120 (2026) 110185

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

3D CNN for breast density classification in microwave imaging

Alessandra Ronca [a,b,*], Mario Badia [c], Navid Ghavami [c], Moein Movafagh [d], Mehran Taghipour-Gorijkolaie [d], Gianluigi Tiberi [c,d], Luca Zilberti [b], Alessandro Arduino [b]

[a] Politecnico di Torino, Torino, Italy
[b] Istituto Nazionale di Ricerca Metrologica, Torino, Italy
[c] UBT-Umbria Bioengineering Technologies Srl, Perugia, Italy
[d] London South Bank University, London, UK

ARTICLE INFO

Keywords:
Breast density
Convolutional neural network
Deep learning
Electrical properties
Microwave imaging

ABSTRACT

Breast density assessment plays a crucial role in mammography interpretation and breast cancer risk evaluation. Traditionally, this process relies on subjective radiological assessment, which can lead to variability in results. Recently, deep learning algorithms have shown great promise in automating breast density classification based on mammography images. In this study, we propose a deep learning-based approach for breast density classification using microwave imaging technology. We developed and trained a custom convolutional neural network using raw measurement data from a microwave breast test device to classify patients' breasts as dense or non-dense, with radiologist assessments based on the BI-RADS classification system serving as reference. The dataset comprised 6709 measurements: earlier data (first 5920 measurements) were split into training (75 %) and validation (25 %) sets, while the most recent data (last 789 measurements) formed an independent test set. The measurements were collected using different devices from multiple sites. The model architecture was selected on the basis of a device-stratified Monte Carlo cross-validation, ensuring that data from different acquisition devices were proportionally represented in each fold. The selected model achieved a total accuracy of 80.1 % ± 0.7 % on the test set, with 79.4 % ± 1.8 % and 80.8 % ± 2.4 % accuracy for low- and high-density cases, respectively. Furthermore, bilateral consistency was identified as a key indicator of classification reliability, reaching 84 % of performance accuracy when verified. This work demonstrates that it is possible to aid clinical decision-making in screening programs with quantitative results based on non-ionizing radiation.

1. Introduction

In modern clinical practice, personalized medicine approaches are becoming increasingly important for patient care. During breast cancer screening, breast density (BD) represents a critical patient-specific characteristic that provides valuable information for risk assessment [1]. High BD has the highest attributable risk of developing cancer, even with respect to breast cancer gene mutations (BRCA) [2]. BD is a comparison of the relative amounts of fat versus fibroglandular tissue in the breast [1] and its assessment depends on the radiologist's visual quantification of the amount of fibroglandular tissue from mammography, magnetic resonance imaging (MRI), or ultrasound images. According to BI-RADS [1], BD is classified into four levels denoted by capital roman letters: A, the breasts are almost entirely fatty; B, there are scattered areas of fibroglandular density; C, the breasts are heterogeneously dense; D, the breasts are extremely dense.

BD affects the sensitivity of a mammography screening program [3]. From a statistical analysis on about 25 000 patients, the overall sensitivity of a screening program for breast cancer detection resulted in 79.9 %; however, by dividing the patients based on the BD assessed by radiologists, the sensitivity increased for A and B classes to, respectively, 100 % and 83.9 %, and decreased for C and D classes to, respectively, 72.9 % and 50.0 % [3]. Dense breasts pose a challenge, because the fibroglandular tissue can mask radiological signs of cancer. Optimizing screening protocols based on BD could improve diagnostic accuracy and early cancer detection.

1.1. Current approaches to breast density classification

Currently, there are no universally agreed methods for BD classification [4]. These methods can be subjective (BI-RADS [1], Hand Delineation), where the radiologist assigns a BD category based on their

* Corresponding author at: Politecnico di Torino, Torino, Italy.
E-mail address: alessandra.ronca@polito.it (A. Ronca).

https://doi.org/10.1016/j.bspc.2026.110185
Received 26 November 2025; Received in revised form 9 March 2026; Accepted 28 March 2026

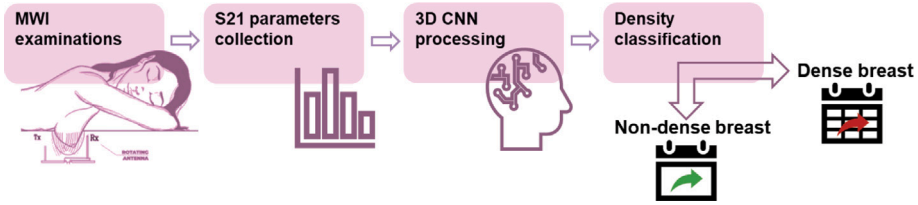

Fig. 1. Proposed MWI-based screening workflow. Women undergo MWI examination using the MammoWave device, from which S_{21} scattering parameters are extracted and processed through a 3D CNN for automated BD classification. Based on the binary classification outcome (non-dense vs. dense), women could be assigned to differentiated and personalized screening pathways.

own visual evaluation of the images, or semi-automatic (ImageJ [5], Cumulus [6]), or fully automated (VOLPARA [7], Quantra [8]).

Recently, the possibility of using deep learning (DL) technologies to classify the BD have become the subject of scientific investigation [9–11]. Properly engineered DL models can perform BD evaluations with accuracy comparable to experienced radiologists: a 94 % accuracy for dense or non-dense classification and a 90 % accuracy for the four classes assessment are shown in [12]. In [13], accuracies of 80 % in four-class BD classification and 93 % in binary classification are obtained on a multi-site dataset. A hybrid machine learning and DL approach for breast segmentation, BD estimation and density-based risk assessment is presented in [14], and showed a strong agreement with Cumulus. Finally, an outstanding approach with an overall accuracy of 98.75 % was presented in [9]. DL has been used for BD assessment also in breast MRI, obtaining an accuracy of 67 % and 89 % for four-class and binary classifications, respectively [15].

These DL models leverage extensive datasets, available for conventional breast imaging technologies with decades-worth of accumulated data that can be used. Similar DL models have never been proposed for BD classification in microwave imaging (MWI).

1.2. Breast density classification in microwave imaging

Over the past decade, significant research efforts in MWI have positioned it as a promising alternative to conventional breast examination technologies. MWI is non-invasive, cost-effective, and could complement other techniques, such as mammography or ultrasound, particularly for women with high BD, where mammography may be less effective [16].

In a typical MWI system, an electromagnetic (EM) signal is emitted from a source and directed towards the breast. Upon interaction with the breast, the scattered signal is collected by a receiving system. The fundamental principle enabling tumor detection is the electrical property (EP) contrast between malignant and healthy breast tissue, which introduces characteristics within the scattered signal that can be interpreted.

The EP contrast is large also between fat and fibroglandular tissue, suggesting a direct correlation between the average EP values of a breast and its BD. The average permittivity of two subjects' breasts, one in class B and the other in class D according to VOLPARA, have been estimated in [17] starting from MWI results and a significantly larger average permittivity was observed in the subject in class D. This fact is further demonstrated in this work through an analysis of anatomical breast models.

The shared dependency of both averaged EP values and BD on breast tissue composition enables the potential use of EPs as a novel method to characterize BD. In particular, since the measured EM signal depends on the target EPs, it is reasonable to assume that different features are present in the MWI signal for each BD class. Indeed, a MWI device used to assess the BD by analyzing its 3D output images has already been presented in the literature [18]. Precisely, a binary classification (classes A and B denoted as non-dense, and classes C and D as dense) with an accuracy of 76 % in a group of healthy volunteers has been obtained through a statistical analysis of the image histograms. The accuracy decreased to 70 % in presence of pathological cases.

Smith et al. [19] developed an alternative handheld device based on microwaves to assess BD. The device acquires twelve measurements per breast across different angular positions and employs an undisclosed algorithm to analyze the received microwave signals to determine the BD. The authors performed a clinical validation involving 557 breasts and reported a good agreement with radiologist classification (Cohen's kappa = 0.659). In particular, they observed an accuracy of 70.9 % for four-class classification and of 91.9 % for binary classification.

On the one hand, BD assessment based on MWI output [18] is limited by its spatial resolution, which is too low to enable accurate volumetric quantification of different tissue types within the breast, as achieved by other imaging technologies. On the other hand, it has been proven that BD classification based on EM measurements is feasible with high accuracy levels [19], although through a design exclusive for BD classification. The approach proposed in this paper integrates BD classification based on EM measurements with a MWI device, avoiding the elaboration of the imaging output and potentially serving as a support mechanism to incorporate prior information into the image reconstruction algorithm itself.

Precisely, a three-dimensional (3D) convolutional neural network (CNN) for BD classification will be presented in this paper. The 3D CNN was trained and validated on a suitable amount of measured S-parameters by a MWI device for breast cancer detection, aiming to classify the patient's BD by extracting information from raw data. To the best of authors' knowledge, this is the first time a DL model operating on the raw S_{21} parameters of an MWI system is proposed for BD classification. The considered MWI system is the MammoWave device (UBT Srl, Perugia, Italy) [20]. The clinical implementation of the proposed approach is shown in the flowchart in Fig. 1.

2. Methods

2.1. Breast EPs and density analysis

To investigate the relationship between breast tissue EPs and BD, 110 digital single-breast models were utilized. The digital models were generated using data acquired from a public repository containing models in MHA format derived from MRI images for medical imaging research [21]. The images were processed using 3D Slicer (https://www.slicer.org/) and converted to 3D voxel models, where six tissues were distinguished: skin, muscle, fibroglandular tissue, fat, and, if present, benign or malignant tumors. EP values of healthy tissues were assigned according to the database of the IT'IS Foundation [22]. For benign and malignant tumors, the EPs at different frequencies were linearly interpolated from 0.5 GHz to 8.0 GHz based on the values reported in [23].

Fig. 2 illustrates the process for a breast model, highlighting in particular the region of interest (ROI) chosen for the analysis (Fig. 2c), and the distribution of the assigned EPs (Fig. 2d). The chosen ROI comprises the part of the breast mainly exposed to MWI radiation, whereas it excludes the chest area.

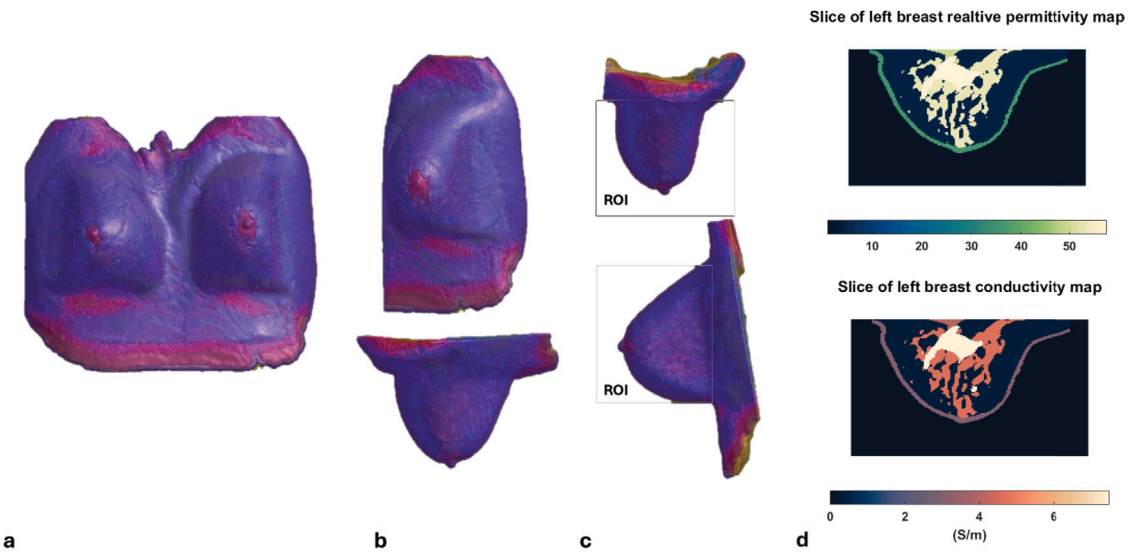

Fig. 2. Example of a virtual breast model: (a) the breast's anatomical model, (b) the sliced left breast, (c) the part of the model in the ROI chosen for the analysis, (d) the maps of relative permittivity and conductivity (S/m) of a cross-section of the breast model (Colormap: [24]).

For each model, both BD and spatial average EPs were calculated. The BD was obtained as the ratio between the volume occupied by fibroglandular tissue and the total volume of the model within the ROI, following a common volume-based approach [25]. Finally, the correlation between EP values and BDs was investigated.

2.2. Data collection through MammoWave device

MammoWave is an MWI system operating in the frequency range from 1 GHz to 9 GHz. It employs a multi-bistatic scanning approach with a stationary transmitting horn antenna, that irradiates the breast under examination, and a moving receiving Vivaldi antenna that collects the scattered signals. Both antennas operate in the air without matching medium and are vertically positioned at the same height.

During operation, the transmitter is positioned at 10 different locations, organized as five pairs centered at 0°, 72°, 144°, 216°, and 288° around the breast holder. Each pair of locations consists of two slightly tilted positions almost 4.5° from the central angle to enable signal subtraction or averaging. For each transmitting position, the receiving antenna rotates a full 360° around the breast, collecting signals at 80 positions with 4.5° increments. The antennas connect to a 2-port Copper Mountain vector network analyzer (VNA) that collects the S_{21} parameters.

A complete scan takes approximately 8 min. During the examination, the patient lies prone on the MammoWave bed with a breast positioned within a cup inside a cylindrical hub. In the conventional application of MWI, the received S_{21} parameters are processed by an algorithm based on Huygens' principle [20].

MammoWave has been involved in three clinical trials:

1. "A multicentric, single arm, prospective, stratified clinical investigation to evaluate the ability of MammoWave in breast lesions detection" (NCT04253366), activated in 3 clinical centers in Italy and Spain. The study is completed with 353 volunteers.
2. "A Clinical Investigation to Confirm the Ability of MammoWave in Breast Lesions Detection" (NCT05300464), activated in 3 clinical centers in Italy and Spain. In this ongoing study, more than 400 volunteers have been enrolled [26].
3. "A Clinical Investigation to Evaluate Microwave Imaging Via MammoWave in a Population-based Screening Program for Early Breast Cancer Detection" (NCT06291896). This study recruits 10 000 volunteers in 10 European centers and was approved by the relevant Ethics Committees of Italy, Spain, Portugal, Poland, and Switzerland [27].

In all the clinical trials, all volunteers, after signing the informed consent, underwent MammoWave examination, in addition to the conventional breast examination path (used as reference standard), during which data related to the mammographic BD was collected. For this investigation, we retrospectively used 6709 breast data (collected within the aforementioned ongoing clinical trials) from women with a mean age of 58.3 years ± 7.6 years (range: 33 years to 77 years). Occurrences of each BD class are almost uniformly distributed within the considered age range.

Although the technicians are trained to follow standardized acquisition procedures, breast positioning as well as possible different system drifts can slightly affect MammoWave acquisition.

2.3. CNN input datasets

Measured S_{21} parameters are collected in real-valued tensors with dimension 1601×10×80×2: the frequency sampling of 5 MHz gives 1601 data composed by 10 transmitter locations, each with 80 receiver locations, split into real and imaginary part. Before providing this data to the developed CNN, some pre-processing was performed.

Measurements coming from transmitters in paired locations were averaged. This helped reducing input data without significant loss of information, given the small differences in antenna positions and consequently in the acquired data.

Only 9 frequencies uniformly distributed in the frequency range were selected to limit the amount of input data while maintaining a good variability in the signals. The frequency selection strategy was validated by statistically comparing models with higher sampling densities and employing Grad-CAM [28] attention analysis.

The input data to the CNN ends up in a 3D structure composed of 9 frequencies, 5 transmitter locations, and 80 receiver locations ordered in the three dimensions. The real and imaginary components are concatenated as two channels, resulting in a 4D input tensor of dimensions $2 \times 9 \times 5 \times 80$.

The collected data were split into two datasets. One dataset contains the first 5920 acquired cases. They consist of 552 samples for class A, 2472 for class B, 2266 for class C, and 630 for class D. This distribution is in accordance with the BD distribution observed in the general population, where the frequency of breast densities is reported as 10 % A, 40 % B, 40 % C, and 10 % D [3]. This set was used for the CNN training and validation. The other dataset contains the last 789 acquired samples and was used to test the trained model on blind

Table 1
Distribution of the device data. The columns 'Non-dense' report the percentage of volunteers in each dataset with non-dense BD according to the radiologists' classification.

Device	Training (75%) and validation (25%)	Non-dense	Blind data	Non-dense
1	398	50.0 %	74	62.2 %
2	707	42.0 %	246	33.3 %
3	752	33.2 %	1	0.0 %
4	693	52 %	74	68.9 %
5	952	72.2 %	182	71.4 %
6	1428	50.9 %	64	43.8 %
7	965	51.1 %	0	–
8	25	56.0 %	148	45.3 %
9	0	–	274	48.9 %
All	5920	50.9 %	1063	50.6 %

data, simulating a clinical scenario. This chronological separation of the available data is significant because the second dataset might contain errors or system drifts not present in the training and validation dataset, providing a more realistic assessment of the model's applicability in clinics.

Conventional breast examination path allowed to classify the 6079 breasts as follows: 35 breasts with histology-confirmed tumor, and the remaining 6044 breasts with no lesion or benign lesions.

As illustrated in Section 2.2, the data were collected across different acquisition sites (identified by an anonymized label from 1 to 9), each equipped with the same model of the MammoWave device. First and second dataset comes from devices 1 to 8, and 274 additional data measured by device 9 were classified as dense or non-dense by the CNN to assess the ability of the method to classify unseen data from a newly introduced device.

The distribution of the data across devices and the percentages of non-dense BD classified by radiologists in each site are reported in Table 1. BD classification is the result of the assessment of two independent radiologists. When a consensus was not reached, a third independent radiologist was consulted to determine the final classification. For each patient, the final assessment is available, whereas the single outcome of each radiologist and the possible intervention of the third radiologist are not available information. Automated software tools were not used for BD assessment as they were not available in all the involved sites.

2.4. Device stratified Monte Carlo cross-validation

To ensure robust model evaluation and hyperparameter optimization, a device stratified Monte Carlo cross-validation approach was implemented. The first dataset, used for training and validation of the CNN, comprised measurements from eight distinct devices (cf. Table 1). To maintain the representativeness of each device across the training, stratified sampling that preserves the distribution of data across each device in both training and validation sets was employed.

The training procedure of each model was repeated 10 times, each time with an independent training-validation partition of the first dataset. To attribute the variability between trained model performances exclusively to the model architecture and parameters, the same 10 partitions of the dataset are used in all the cross-validation studies. The following metrics are computed by applying each trained model to the corresponding validation set:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

Table 2
Hyperparameters tested and final selection. Tabular CL values refer to the deepest convolutional block, the other blocks have CL=2. Bold values are the starting choice for each parameter.

Parameter	Tested values	Selected value
CB - CL	5 - 2 5 - 1 4 - 2 4 - 1 3 - 2 3 - 1 2 - 2 2 - 1	3 - 1
FCL	16, **32**, 64, 128, 256, 512	32
CH	(16, 32, 64) **(32, 64, 128)** (64, 128, 256) (128, 256, 512)	(32, 64, 128)
AF	ReLU, PReLU	PReLU
LR	0.01, 0.005, **0.001**, 0.0005, 0.0001	0.01
DR	0.0, 0.2, **0.3**, 0.5	0.2
wLF	BCE, **Cross-entropy**	Cross-entropy
Optimizer	Adam, **AdamW**	Adam

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN} \tag{5}$$

Here, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively, under the null hypothesis that the breast under examination is dense (i.e., non-dense class is negative and dense class is positive). The area under the receiver operating characteristic (ROC) curve (AUC-ROC) [29] is evaluated as an additional metric. For each metric, both mean and standard deviation were computed across the 10 trained models.

2.5. Architecture selection process

To identify the optimal network architecture and hyperparameter configuration, a step-wise analysis was conducted using the cross-validation. Precisely, the hyperparameters listed in Table 2 were optimized one by one after setting initially all the parameters to the bold values of the second column of the table, as described below.

The initial network configuration processed the input through 3 convolutional blocks (CB). The first and second blocks contain 2 convolutional layers (CL) with kernel sizes of $3 \times 3 \times 3$ and strides of $1 \times 1 \times 1$, while the third block contains 1 convolutional layer with kernel sizes of $3 \times 1 \times 3$, and strides same as before. All the blocks have a dropout rate (DR) of 0.3. The first convolutional block transforms the 2 input channels (CH) into 16 feature maps, the second block increases these to 32 feature maps, and the third block produces 64 feature maps. Each convolutional block includes a maximum pooling layer with a pool size of $2 \times 2 \times 2$, reducing the central dimension to 1 at the last block. Kernel size, stride, and pooling were tailored to match the input data dimensions. Activation function (AF) at each block is the parametric rectified linear unit (PReLU) [30]. The extracted features are then flattened and passed through 3 fully connected layers (FCL), with 1280, 32, and 2 neurons, respectively. The weighted cross-entropy loss function (wLF) is used to quantify the discrepancy between the model's predictions and the actual labels. The AdamW optimizer [31] (weight decay=1×10^{-2}, betas=(0.9,0.999)) was used to train the model with an initial learning rate (LR) of 1×10^{-3}. A dynamic scheduling strategy was employed to adapt the LR during training: the LR is reduced by a factor of 0.2 whenever the validation metric plateaus, i.e., when no

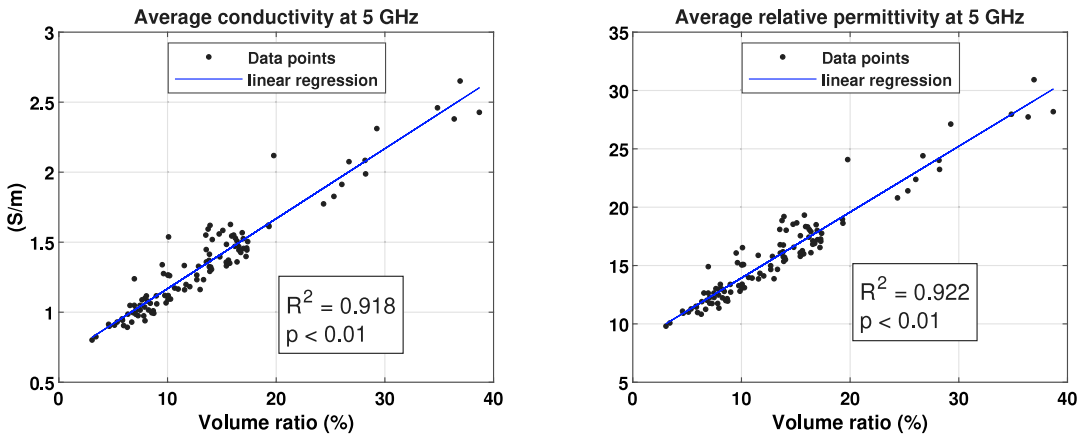

Fig. 3. Average conductivity (left) and relative permittivity (right) at 5 GHz against BD.

improvement is observed for 10 consecutive epochs. A minimum LR threshold of 1×10^{-6} was set to prevent excessive reduction.

The selection process relied on a qualitative observation of three metrics on validation dataset: accuracy, which indicates the ability of networks to classify unseen data, F1-score, which describes the balance in the model performance, and AUC-ROC.

2.6. Performance evaluation

Subsequently to the choice of the CNN architecture, the same stratified Monte Carlo cross-validation procedure was applied to the final selected model architecture on the second dataset to verify the model's robustness. All the metrics described in Section 2.4 were used to describe the binary classification results. In the end, three complementary analyses were conducted: first, the classification approach was extended to four-classes, with the same CNN architecture employed to assess whether the model could maintain valuable discriminative power; second, a per-device analysis was performed to evaluate the generalizability of the classification model across different acquisition sites, and to identify potential device-specific biases; third, a bilateral breast analysis was conducted to leverage the paired nature of breast measurements. All three analyses utilized the same performance metrics previously described to ensure consistency and comparability of results.

3. Results

3.1. EPs and BD correlation

Fig. 3 shows the average EP values at 5 GHz against the BD for 110 breast models. It illustrates that a highly significant statistical linear relationship exists between the two variables, that was consistently observed across the entire frequency range (1 GHz to 9 GHz with 1 GHz steps).

3.2. Data preprocessing analysis

Previous measurements show that the maximum absolute difference between the averaged signals from paired transmitters and the signals from the corresponding central transmitter was at most 3.56 % of the respective signal ranges. The mean absolute error (MAE) varied from 0.26 to 0.50 %, while the root mean square error (RMSE) ranged from 0.35 to 0.72 % across the five pairs.

The attention analysis based on Grad-CAM on the test dataset revealed that the model's focus was distributed relatively uniform across the 9 selected frequencies, with attention weights ranging from about 0.79 to 0.86.

With regards to the frequency resolution, with 9 frequency points (1 GHz intervals) the validation accuracy was 77.2 ± 0.8 %, with 17 frequency points (0.5 GHz intervals) the validation accuracy was 74.6 ± 1.5 %, with 33 frequency points (0.25 GHz intervals) the validation accuracy was 75.5 ± 1.9 %, and with 65 frequency points (0.125 GHz intervals) the validation accuracy was 74.6 ± 1.9 %. Training time almost doubled at each frequency resolution increment. The statistical analysis, applying Wilcoxon signed-rank test with Bonferroni correction, revealed significantly superior performance for the 9 frequency points selection compared to the higher frequency resolution configurations. Pairwise comparisons for most of the metrics between 17, 33, and 65 points did not show significant differences. All the values extracted from the statistical analysis have been reported in Supplementary Materials.

3.3. CNN architecture selection and performance

The final architecture of the CNN implemented in this work is presented in Fig. 4.

With respect to the initial choice described in Section 2.5, the Adam optimizer is used to train the model with an initial learning rate of 1×10^{-2}, and the dropout value is chosen equal to 0.2. Table 2 presents an overview of the hyperparameters listed in the testing order and, for each of them, the optimal choice. Supplementary Material Figs. S1–S8 report the metric values evaluated for each final parameter choice.

Fig. 5 shows training loss and validation accuracy trends of the model with the selected architecture and parameters. For the 10 selected splits of the dataset, the accuracy and loss trends exhibit a consistent pattern, reaching a plateau after around 50 epochs. Performance metrics for the classification on validation dataset are reported in the Supplementary Material Table S1. The performance metrics were extracted at epoch 53, to ensure that the reported results reflect a stable model performance, while avoiding potential fluctuations due to earlier training instability or overfitting at later stages.

The CNNs required about 84 s for training using an NVIDIA A100 40 GB GPU and about 5 s for inference using CPU (Intel Core i7-4790 CPU).

3.4. CNN evaluation in a clinical scenario

The second dataset contains 789 data samples measured with the MammoWave devices numbered from 1 to 8 in Table 1. Classification performance on this dataset presented a mean accuracy value of 74.4 % with a standard deviation equal to 1.1 % (all metrics in Supplementary Material Table S2). Performance metrics for the classification on the additional dataset from device 9 are reported in the Supplementary Material Table S3.

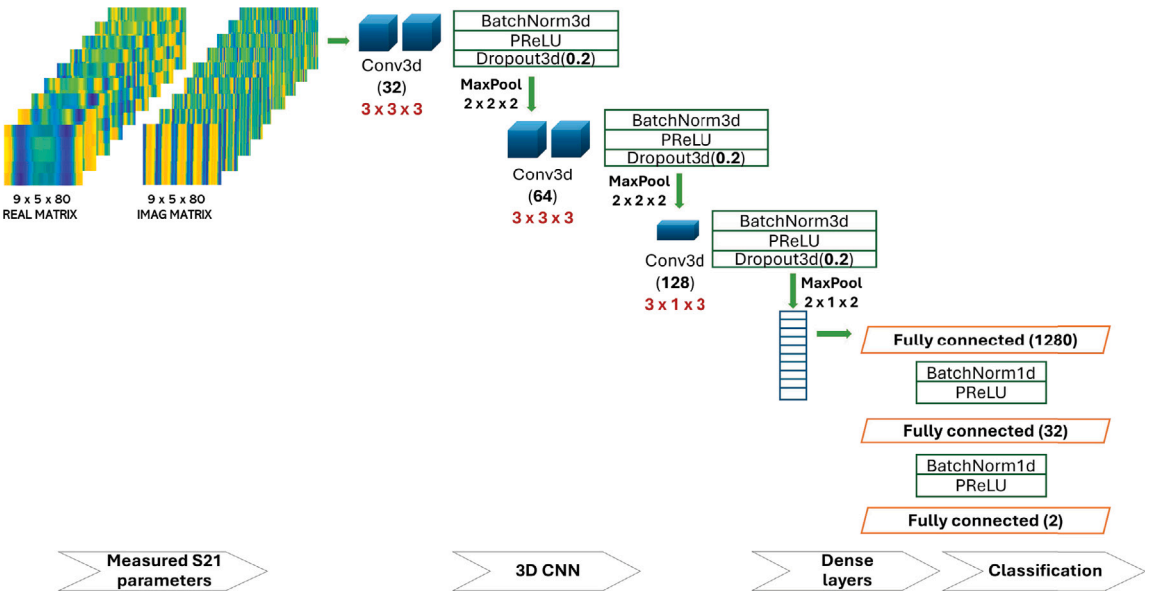

Fig. 4. Architecture of the CNN.

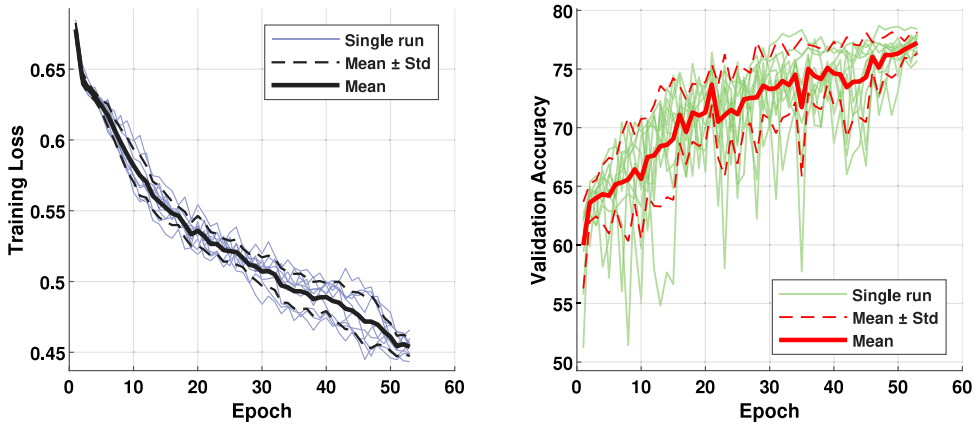

Fig. 5. Model training loss and validation accuracy across multiple training runs.

Fig. 6 reports the accuracy of the model in the classification of the breasts in the second dataset device by device.

Final model performance evaluation required some data exclusion: device 7 does not contribute to the analysis because there are no data to take under consideration in the dataset shifted in time; device 3 contained only a single sample in the test set, potentially distorting aggregate metrics; similarly, device 8 presented a severe class imbalance between training and test sets because of the lack of data in the first dataset. On the final filtered test dataset, mean accuracy value was 80.1 % with a standard deviation equal to 0.7 % (all metrics in Supplementary Material Table S4).

It is interesting to check the accuracy in the classification of each of the four BD classes (Supplementary Material Table S5). The classification of breasts in class A as non-dense has an accuracy of 91.4 % ± 0.8 %, and for breasts in class D as dense 96.5 % ± 1.5 %. The accuracy for classification of breasts in class B as non-dense and breasts in class C as dense is equal to 77.0 % ± 2.1 % and 75.2 % ± 3.1 %, respectively.

Fig. 7 shows values of the descriptive metrics of classification performance for the validation and the filtered test dataset.

Performance metrics per-device for the classification on the filtered test dataset are reported in the Supplementary Material Table S6.

The four-class breast density classification achieved an overall accuracy of about 50.2 ± 2.5 % (all metrics in Supplementary Material Table S7). Class-wise accuracy varied, with higher performance for A and D of around 82% and around 67% for B and C (Supplementary Material Table S8).

The second dataset includes coupled measurements from 313 volunteers, with some additional cases involving data from only one breast. Among all the breast pairs classified according to BD by the proposed CNN, 272 ± 4 exhibited the same BD value for both the left and right breasts, and 84.11 % ± 0.78 % of them are correctly classified. About this paired data classification, results in the sub-classes A or B for non-dense breast and C or D for dense breast were analyzed and reported in Table 3.

4. Discussion

The device-stratified Monte Carlo cross validation approach was deemed sufficient to capture the central tendency of the proposed model performance. The classifier achieves a mean accuracy of 80.1 % ± 0.7 % (Supplementary Material Table S4), indicating solid overall performance.

The obtained sensitivity of 80.8 % ± 2.4 % (Supplementary Material Table S4) is particularly relevant, because it reflects the model's ability to correctly identify dense breasts, that are the clinically more significant category due to its association with increased breast cancer risk and reduced mammographic sensitivity to tumor detection.

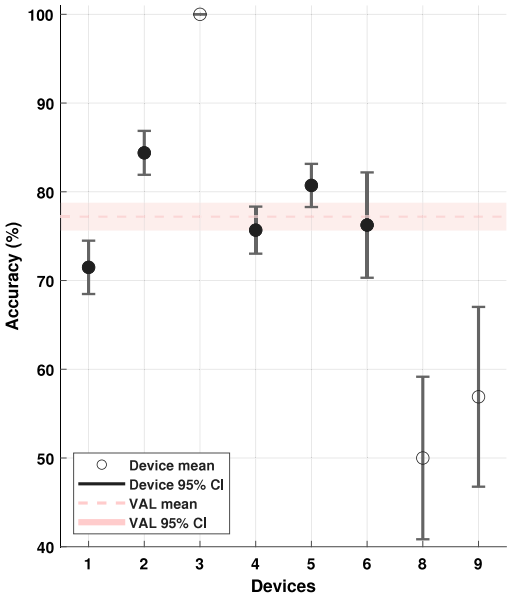

Fig. 6. Per-device accuracy across Monte Carlo cross-validation. Numbers on the horizontal axis, from 1 to 9 (with 7 excluded), refer to the different devices as previously labeled. Black filled point corresponds to devices in the filtered test dataset. Pink band represents the classification accuracy mean and 95% confidence interval on validation dataset (VAL).

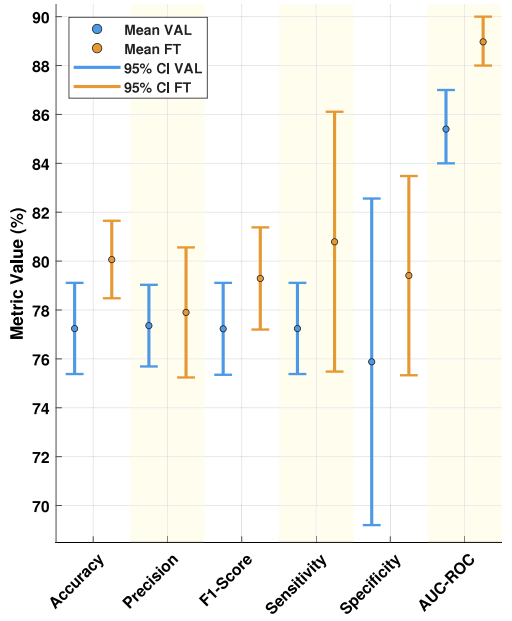

Fig. 7. Performance metrics across the validation (VAL) and filtered test (FT) datasets.

Furthermore, the specificity value of 79.4 % ± 1.8 % (Supplementary Material Table S4) shows that the model also performs well in correctly identifying non-dense cases, contributing to overall robustness. Although some false positives are unavoidable, the achieved specificity helps mitigate their impact, including psychological distress for patients and unnecessary resource consumption resulting from additional examinations.

Results in each of the four BD classes confirm that the classifier robustly identifies cases at the BD spectrum's edges (A and D classes), which is clinically crucial as these categories require different screening strategies and risk assessments [32]. For B and C classes, the accuracy

Table 3
Classification accuracy (mean ± std) across Monte Carlo cross-validation models and density classes on the filtered test dataset considering paired samples. "Matching classification" refers to cases where both samples in a pair receive the same and correct prediction. "Matching misclassification" indicates that both samples receive the same but incorrect prediction. "Mismatch in prediction" occurs when the classifier assigns different labels to the two paired samples.

Class	No. of couples	Matching classification (%)	Matching misclassification (%)	Mismatch in prediction (%)
A	28	87.86 ± 1.84	3.57 ± 0.00	8.57 ± 1.84
B	137	70.51 ± 2.63	16.50 ± 2.13	12.99 ± 1.71
C	112	66.61 ± 4.02	17.59 ± 2.98	15.80 ± 2.70
D	36	92.50 ± 3.22	0.00 ± 0.00	7.50 ± 3.22

is moderately lower but balanced. This is consistent with the increased difficulty in distinguishing intermediate BD, where tissue characteristics are less distinct and may overlap between classes. Moreover, inter-reader variability may also contribute to labeling inconsistencies in these borderline cases, potentially introducing noise into the training data that affects the model performances. About this aspect, the literature documents differences in inter-reader agreement depending on the classification scale used [8]: it varies with a Cohen's kappa from 0.38 to 0.65 on a four-category scale, while reaching a value of 0.85 on a two-category scale. The uncertainty related to the intermediary classes effectively confuses the CNN during training, leading to less precise feature extraction.

While standard imaging-based methods, such as FDA-approved X-ray tools and MRI-based AI systems, report binary classification accuracies above 90 % [25] and 89 % [15], respectively, the proposed approach based on raw MWI signals achieves lower accuracy. However, compared to X-ray mammography, MWI operates with non-ionizing radiation, ensuring enhanced patient safety [33], while compared to MRI-based systems, MWI offers a significantly more cost-effective and accessible solution for population-wide screening programs.

Moreover, the approach is fundamentally different, as it relies directly on the signal acquired by the MWI device, therefore S parameters and not images. Despite containing no visually interpretable information for clinicians, these raw signals enable direct automated classification of samples into respective BD classes.

This paper corroborates the findings presented in [17], which initially suggested a potential correlation between EPs and BD. This research builds upon these earlier observations by establishing a quantifiable linear relationship between these two variables. While our correlation analysis between EPs and BD is based on digital anatomical models, this choice is supported by extensive in-vivo evidence. Previous studies have shown that breast EPs are strongly associated with tissue composition, particularly adipose and fibroglandular content, which are the main determinants of BD [34,35].

With respect to the classification methodology employed in [18], that reported an accuracy of 76 %, the present study demonstrates superior performances, which are particularly significant given the heterogeneous nature of the dataset, which comprises multiple sites and devices, and a population composed of both healthy breasts and breasts with benign or malignant pathologies. In addition, the dataset employed in this study is substantially larger than the one considered in [18], contributing to a more reliable and generalizable assessment of the model's performances.

The work presented in [19] lays an important groundwork for extracting BD information from microwave EM measurements within the same MammoWave frequency range. However, the limited disclosure of validation strategy and algorithm details, and a totally different hardware architecture with respect to MammoWave, hinder a direct comparison with the proposed model.

The BD class is not limited to be an important clinical information of a patient, but becomes a valuable feature to improve the classification

of benign and malignant masses [36], and in setting input parameters for image reconstruction algorithms. For instance, when the reconstruction is based on Huygens' principle [37], knowledge of the BD helps in selecting optimal values for the constant EPs of the domain. Similarly, in reconstructions based on contrast source inversion [38], knowledge of the BD class would suggest a consistent initial guess for the iterative procedure that leads to quantitative maps of the EPs [38,39].

4.1. Four-class BD assessment

While the overall classification accuracy remains relatively low, the weighted AUC-ROC score (77.0 % ± 1.3 %) suggests that the predicted class probabilities are informative and discriminative to a certain extent. These results motivate further investigation into the predicted probability scores to identify potential systematic trends or confidence patterns associated with each BI-RADS class, potentially guiding improved decision-making strategies.

Although the distribution of data in the four classes reflects the population distribution, the observed suboptimal performance can be attributed both to the inherent class imbalance (A and D: 10 %; B and C: 40 %), mitigated using a weighted cross-entropy loss function, and the intrinsic challenge of distinguishing among four categories. Classes B and C are particularly prone to misclassification, consistent with the known difficulty in clinical BD categorization due to their overlapping features. Moreover, as explained before, the inter-reader variability, which particularly affects the four-classes BD assessment [8], acts as a source of label noise that will be reflected in the lower model's performance with respect to binary classification.

4.2. Per-device analysis

As shown in Fig. 6, the per-device accuracy highlights that the devices included in the final dataset yield metrics that are consistent with validation dataset metrics and representative of the model's behavior under the discussed variability.

The analysis revealed that classification performances are affected by both dataset composition and device-specific factors.

For instance, device 5, despite having a balanced training set, achieved high specificity but low sensitivity, likely due to a non-dense-skewed test set. Device 2 showed the opposite behavior (high sensitivity and low specificity), consistent with a dense-biased test dataset.

The notably low performances on device 9, not used in the model training, underscore the model's limited ability to generalize to data from unseen devices. The primary source of the low performance could be the noise present in the labeling process. To this end, an international dataset is collected from different sites that follow local training protocols for BD assessment. This leads to high intra-center agreement but also inter-center (and inter-country) variability [4]. Since the CNN is trained on labels that reflect these diverse local biases, a device from a new center might present a labeling logic that slightly shifts from the average training distribution. This fact was already suggested by performance metrics on device 8, and emphasizes the importance of incorporating a representative amount of data from any newly deployed device into the training pipeline through an incremental fine-tuning.

4.3. Data preprocessing analysis

The preprocessing pipeline led to a reduction of the frequency samples from 1601 to 9, uniformly sampled in the frequency range from 1 GHz to 9 GHz. No significant improvements in performances when increasing the sampling size can be attributed to the inherent redundancy in high-resolution spectral measurements. While the EPs of breast tissues exhibit systematic frequency-dependent trends over the microwave range, these variations are dominated by broad relaxation processes that produce smooth changes rather than narrow-band resonances or discontinuities [40]. Consequently, adjacent frequency points in the original dataset are highly correlated, providing limited additional discriminative information for the binary classification task.

The almost uniform distribution of model's attention across frequencies validates the initial hypothesis that broadband sampling across the entire measurement spectrum would capture the essential EM signatures for tissue differentiation.

The model was retrained using frequency points weighted according to these attention values. It resulted in a classification whose performances remained substantially similar to the uniform frequency sampling (accuracy equal to 75.5 ± 2.67 % on the stratified Monte Carlo cross validation), confirming that the initial sampling strategy was near-optimal for the classification task.

In conclusion, the reduction from 1601 to 9 frequency points represents a 178-fold decrease in frequency-domain dimensionality, translating to significant computational savings during both training and inference.

The proposed model architecture relies on *a posteriori* validation through attention analysis and does not incorporate the frequency importance directly into the model. To improve the architecture in the future, an attention mechanism specifically designed for 3D feature maps [41] could be analyzed and implemented.

In the context of preprocessing, normalization of the input data prior to training was avoided. Although the input values span both positive and negative ranges, normalizing them to a fixed distribution (e.g., using z-score) was found to reduce model performance in preliminary experiments (validation accuracy equal to 69.51 ± 1.84 %). Moreover, a substantial reduction in standard deviation across all metrics, most notably in specificity (from 9.42% to 2.96%), demonstrates that the absolute scale of the input data may encode task-relevant information. The comparison between the performance metrics with and without input data normalization is presented in Supplementary Material Table S9.

4.4. Additional architecture design considerations

Several well-established DL architectures have demonstrated success in biomedical and spectral-spatial analysis, including ResNet [12], DenseNet [42,43], vision transformers (ViTs) [44,45], and CNN-based models for hyperspectral imaging (HSI) [46–48].

Prior to adopting a DL approach, simpler machine learning methods for classification tasks [49,50] were explored and tested using the same device-stratified Monte Carlo cross-validation protocol directly to the flattened raw input data: logistic regression reached a mean accuracy of 72.79% ± 0.77% on the validation dataset; support vector machine (SVM) achieved a mean accuracy of 71.15% ± 0.96% on the validation dataset. With regards to the DL approach, a 2D CNN was evaluated as a preliminary baseline to assess the impact of spatial feature extraction. In this configuration, the input was reshaped into a 5×80 grid where the 9 frequency points in real and imaginary parts were expanded into 18 input channels. Using the same cross-validation protocol and with a hyperparameter setup consistent with the final 3D CNN model for a fair comparison, the 2D CNN achieved a mean accuracy of 73.92 ± 3.15% and an AUC-ROC of 82.77 ± 1.87%. The comparison between their classification is reported in the Table S10 in Supplementary Material. The 3D CNN achieved superior classification performance across all the metrics, with a standard deviation systematically lower with respect to the 2D CNN, indicating a more robust classification varying the folds.

In the literature, 3D CNNs have been successfully employed to process videos [51], demonstrating the effectiveness of treating temporal sequences through the third dimension. This conceptual framework motivated the proposed 3D CNN architecture for S-parameters, treating frequency samples analogously to temporal frames in video processing. Following the findings in [51], where homogeneous $3 \times 3 \times 3$ convolutional kernels across all layers proved optimal for spatiotemporal feature learning, we adopted the same kernel size throughout our network. The real and imaginary components of S_{21} parameters are

treated as dual channels, analogous to RGB channels in image processing. However, due to the significantly different input dimensions ($2 \times 9 \times 5 \times 80$ vs. $3 \times 16 \times 112 \times 112$ in [51]), a shallower network architecture with fewer filters was designed testing optimal hyperparameters for our classification purpose.

The implemented architecture is lightweight (< 2M parameters) compared to well-established DL architectures (ResNet-18/34, EfficientNet), whose strong performance in BD assessment is typically achieved through transfer learning from large-scale datasets followed by fine-tuning [12,13,15].

4.5. Paired data classification strategy

Within the test dataset, the majority of the data originates from both the left and right breasts of the same patient. Therefore, it is possible to incorporate a BD assessment by considering paired data belonging to the same individual. According to the radiologists' BD assessment, both breasts of the same patient are consistently assigned to the same class.

A study conducted on 400 young patients and 100 of their mothers investigated the symmetry in BD in right and left breasts [52]. For each patient, water content MRI images of the breast were obtained, from which the BD were deduced. In the young group of patients, with an averaged age of 20.8 years, mean percentage of water content was 0.84 % higher in the right compared with the left breast. In mothers group, with an average age of 49.6 years, there were no significant difference in the mean percentage of water in both breasts. No significant differences in BD are presented for left and right breast, and this opens the way to analyze our results considering every data as a couple of breasts of the same patient.

Thanks to the correlation between the tissue water content and its EPs in the microwave frequency range [34], the symmetry in the breast water content reported in [52] implies a symmetry in the breast EPs as well.

For about 87 % of the breast pairs, the proposed CNN classified both the left and right breasts in the same BD class. In the remaining cases, where one breast was classified as dense and the other as non-dense, assuming that the two breasts should belong to the same BD class, it can be stated with certainty that the model misclassified one of the two breasts.

Hence, inconsistencies between left and right breasts in the CNN classification may serve as an indicator of unreliable assessment, supporting the advisability of repeating the examination. This allows to limit the evaluation of the performance of the proposed CNN to the cases in which the classification of the pair of breasts is consistent. In this case, the model achieves an almost perfect classification performance for breasts in classes A and D. Conversely, the cases in which both pairs are misclassified by the proposed model predominantly involve borderline density cases (classes B or C).

We explored whether incorporating bilateral information could enhance model performances, modifying the input structure to simultaneously process data from both breasts of the same patient, maintaining the assumption of consistent BD classification across bilateral samples. However, the model trained on paired breast inputs achieved a validation accuracy of 73.08 % ± 2.1 %. It should be noted that the paired approach inherently reduces the available dataset size by approximately half, which may contribute to the observed performance drop. Nevertheless, the bilateral pairing should theoretically provide complementary information that the network could exploit for improved classification accuracy [53], and the exploration of strategies to help the network to manage this type of data becomes a potential area for further investigation.

Based on the presented findings, the single-breast classification approach was maintained as our primary methodology, while the bilateral consistency assumption enables the definition of a protocol where consistent bilateral classification confirms the result's reliability and improves the accuracy of the BD assessment, while inconsistent results suggest the need for a re-acquisition.

5. Conclusion

In this study, a model to classify the BD from MWI raw data was proposed. The collected results demonstrate the feasibility of direct EM-based BD classification.

Furthermore, the introduction of a bilateral analysis strategy provides a novel framework for clinical quality assurance. By leveraging the physiological symmetry of BD, the system can identify inconsistent measurements that may require re-acquisition, while providing a high-confidence classification when bilateral concordance is achieved.

Considering that MWI imposes no age limitations and allows repeated use without exposure risks [33], this approach offers significant advantages. Access to such clinical information could enable patients to establish personalized and efficient screening pathways from an early age, potentially improving early detection outcomes and overall quality of care.

CRediT authorship contribution statement

Alessandra Ronca: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mario Badia:** Writing – review & editing, Visualization, Validation, Resources, Methodology, Conceptualization. **Navid Ghavami:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Moein Movafagh:** Writing – review & editing, Visualization, Validation, Supervision, Software, Methodology, Investigation, Data curation, Conceptualization. **Mehran Taghipour-Gorijkolaie:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Gianluigi Tiberi:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization. **Luca Zilberti:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Conceptualization. **Alessandro Arduino:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gianluigi Tiberi reports a relationship with Umbria Bioengineering Technologies that includes: employment and equity or stocks. Mario Badia reports a relationship with Umbria Bioengineering Technologies that includes: employment. Navid Ghavami reports a relationship with Umbria Bioengineering Technologies that includes: employment. Mario Badia, Navid Ghavami, and Gianluigi Tiberi report a relationship with Umbria Bioengineering Technologies that includes employment, and equity or stocks. For all the other authors there is no financial/personal interest or belief that could affect their objectivity. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the funding received by the MammoScreen project, co-funded by the European Union's Horizon research and innovation framework programme, Grant agreement 101097079. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union, HaDEA or the UK R&I agency. Neither the European Union nor the granting authority can be held responsible for them.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bspc.2026.110185.

Data availability

The authors do not have permission to share data.

References

[1] D.A. Spak, J.S. Plaxco, L. Santiago, M.J. Dryden, B.E. Dogan, Bi-rads® fifth edition: a summary of changes, Diagn. Interv. Imaging 98 (3) (2017) 179–190, http://dx.doi.org/10.1016/j.diii.2017.01.001.

[2] C.M. Vachon, C.H. van Gils, T.A. Sellers, K. Ghosh, S. Pruthi, K.R. Brandt, V.S. Pankratz, Mammographic density, breast cancer risk and risk prediction, Breast Cancer Res.: BCR 9 (6) (2007) 217, http://dx.doi.org/10.1186/bcr1829.

[3] S. Weigel, W. Heindel, J. Heidrich, H.W. Hense, O. Heidinger, Digital mammography screening: sensitivity of the programme dependent on breast density, Eur. Radiol. 27 (7) (2017) 2744–2751, http://dx.doi.org/10.1007/s00330-016-4636-4.

[4] W. Alomaim, D. O'Leary, J. Ryan, L. Rainford, M. Evanoff, S. Foley, Subjective versus quantitative methods of assessing breast density, Diagn. (Basel, Switzerland) 10 (5) (2020) 331, http://dx.doi.org/10.3390/diagnostics10050331.

[5] ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, [Available online on March 19 2025], https://imagej.net/ij/.

[6] Li J., L. Szekely, L. Eriksson, et al., High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer, Breast Cancer Res 14 (R114) (2012) http://dx.doi.org/10.1186/bcr3238.

[7] N. Singh, P. Joshi, D.K. Singh, et al., Volumetric breast density evaluation using fully automated volpara software, its comparison with BIRADS density types and correlation with the risk of malignancy, Egypt J Radiol Nucl Med 53 (2022) 118, http://dx.doi.org/10.1186/s43055-022-00796-y.

[8] E.U. Ekpo, C. Mello-Thoms, M. Rickard, P.C. Brennan, M.F. McEntee, Breast density (bd) assessment with digital breast tomosynthesis (dbt): agreement between Quantra™ and 5th edition, BI-RADS®, The Breast (ISSN: 0960-9776) 30 (2016) 185–190, http://dx.doi.org/10.1016/j.breast.2016.10.003.

[9] N. Saffari, H.A. Rashwan, M. Abdel-Nasser, V. Kumar Singh, M. Arenas, E. Mangina, B. Herrera, D. Puig, Fully automated breast density segmentation and classification using deep learning, Diagn. (Basel) 10 (11) (2020) 988, http://dx.doi.org/10.3390/diagnostics10110988.

[10] K. Alhusari, S. Dhou, Machine learning-based approaches for breast density estimation from mammograms: A comprehensive review, J. Imaging 11 (2) (2025) 38, http://dx.doi.org/10.3390/jimaging11020038.

[11] N.C. da Rocha, A.M.P. Barbosa, Y.O. Schnr, et al., Enhancing breast density assessment in mammograms through artificial intelligence, J. Digit. Imaging. Inform. Med. (2025) http://dx.doi.org/10.1007/s10278-025-01657-6.

[12] C.D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, R. Barzilay, Mammographic breast density assessment using deep learning: Clinical implementation, Radiology 290 (1) (2019) 52–58, http://dx.doi.org/10.1148/radiol.2018180694.

[13] T.P. Matthews, S. Singh, B. Mombourquette, J. Su, M.P. Shah, S. Pedemonte, A. Long, D. Maffit, J. Gurney, R.M. Hoil, N. Ghare, D. Smith, S.M. Moore, S.C. Marks, R.L. Wahl, A multisite study of a breast density deep learning model for full-field digital mammography and synthetic mammography, Radiol Artif Intell 3 (1) (2020) e200015, http://dx.doi.org/10.1148/ryai.2020200015.

[14] O. Haji Maghsoudi, A. Gastounioti, C. Scott, L. Pantalone, F.F. Wu, E.A. Cohen, S. Winham, E.F. Conant, C. Vachon, D. Kontos, Deep-LIBRA: An artificial-intelligence method for robust quantification of breast density with independent validation in breast cancer risk assessment, Med Image Anal 73 (2021) 102138, http://dx.doi.org/10.1016/j.media.2021.102138.

[15] X. Jing, M. Wielema, A.G. Monroy-Gonzalez, T.R.G. Stams, S.V.K. Mahesh, M. Oudkerk, P.E. Sijens, M.D. Dorrius, P.M.A. van Ooijen, Automated breast density assessment in MRI using deep learning and radiomics: Strategies for reducing inter-observer variability, J. Magn. Reson. Imaging 60 (1) (2024) 80–91, http://dx.doi.org/10.1002/jmri.29058.

[16] A. de Jesus Aragão, D. Carvalho, B. Sanches, W.A.M. van Noije, A review on microwave imaging systems for breast cancer detection, IEEE Access 12 (2024) 190611–190628, http://dx.doi.org/10.1109/ACCESS.2024.3516762.

[17] P. Mojabi, J. Bourqui, Z. Wang, et al., Microwave imaging for breast density assessment: initial investigation, in: Proc. URSI Int. Symp. Electromagnetic Theory (URSI EMTS), Vancouver, Canada May 23–26, 2023, http://dx.doi.org/10.36227/techrxiv.172418011.13654621/v1.

[18] A. Iriarte, C. Bore G. de vargas, et al., Estimation of breast density using radio wave radar imaging techniques, in: Proc. European Congress of Radiology, ECR, Vienna, Austria, Feb 28-Mar 4, 2018, http://dx.doi.org/10.1594/ecr2018/C-1623.

[19] C.G. Graff, N. Sharma, M. Fletcher, F. Eskandari, L. Cornock, D. Gibbins, Breast density scoring via a novel microwave energy device: Agreement with mammographic assessment in the leeds mi scan® breast density trial, in: Scientific Exhibit, ECR 2025, 2025, http://dx.doi.org/10.26044/ecr2025/C-27261, Poster C-27261.

[20] A. Vispa, L. Sani, M. Paoli, A. Bigotti, G. Raspa, N. Ghavami, S. Caschera, M. Ghavami, M. Duranti, G. Tiberi, UWB device for breast microwave imaging: phantom and clinical validations, Measurement 146 (2019) 582–589, http://dx.doi.org/10.1016/j.measurement.2019.05.109.

[21] A. Pelicano, M.C.T. Gonçalves, D.M. Godinho, T. Castela, M.L. Orvalho, N.A.M. Araújo, E. Porter, R.C. Conceição, Development of 3D MRI-based anatomically realistic models of breast tissues and tumors for microwave imaging diagnosis, Sensors 21 (24) (2021) 8265, http://dx.doi.org/10.3390/s21248265.

[22] C. Baumgartner, P.A. Hasgall, F. Di Gennaro, E. Neufeld, B. Lloyd, M.C. Gosselin, et al., IT'IS database for thermal and electromagnetic parameters of biological tissues, version 4.2, 2024, http://dx.doi.org/10.13099/VIP21000-04-2, [Online]. Available: https://itis.swiss/virtual-population/tissue-properties/database/dielectric-properties, (Accessed 17 March 2025).

[23] Y. Cheng, M. Fu, Dielectric properties for non-invasive detection of normal, benign, and malignant breast tissues using microwave theories, Thorac Cancer 9 (4) (2018) 459–465, http://dx.doi.org/10.1111/1759-7714.12605.

[24] F. Crameri, Scientific colour maps (8.0.1), 2023, http://dx.doi.org/10.5281/zenodo.8409685, Zenodo.

[25] J.S. Chalfant, A.C. Hoyt, Breast density: Current knowledge, assessment methods, and clinical implications, J. Breast Imaging 4 (4) (2022) 357–370, http://dx.doi.org/10.1093/jbi/wbac028.

[26] L. Papini, et al., MammoWave clinical trials within RadioSpin project: results obtained using microwave images' features approaches with thresholds, in: 2024 IEEE International Symposium on Antennas and Propagation and INC/USNC-URSI Radio Science Meeting (AP-S/INC-USNC-URSI), Firenze, Italy, 2024, pp. 1519–1520, http://dx.doi.org/10.1109/AP-S/INC-USNC-URSI52054.2024.10686131.

[27] D. Álvarez Sánchez-Bayuela, et al., Microwave imaging for breast cancer screening: protocol for an open, multicentric, interventional, prospective, non-randomised clinical investigation to evaluate cancer detection capabilities of MammoWave system on an asymptomatic population across multiple European countries, BMJ Open 14 (11) (2024) e088431, http://dx.doi.org/10.1136/bmjopen-2024-088431.

[28] J. Gildenblat, contributors, Pytorch library for CAM methods, 2021, GitHub, [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam.

[29] M.R.J. Junge, J.R. Dettori, ROC solid: Receiver operator characteristic (ROC) curves as a foundation for better diagnostic tests, Glob. Spine J. 8 (4) (2018) 424–429, http://dx.doi.org/10.1177/2192568218778294.

[30] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: 2015 IEEE International Conference on Computer Vision, ICCV, Santiago, Chile,, 2015, pp. 1026–1034, http://dx.doi.org/10.1109/ICCV.2015.123.

[31] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2017, http://dx.doi.org/10.48550/arXiv.1711.05101, arXiv preprint arXiv:1711.05101.

[32] R.M. Mann, A. Athanasiou, P.A.T. Baltzer, J. Camps-Herrero, P. Clauser, E.M. Fallenberg, et al., Breast cancer screening in women with extremely dense breasts: recommendations of the European society of breast imaging (EUSOBI), Eur. Radiol. 32 (6) (2022) 4036–4045, http://dx.doi.org/10.1007/s00330-022-08617-6.

[33] A. Ronca, L. Zilberti, O. Bottauscio, G. Tiberi, A. Arduino, Safety assessment of microwave breast imaging: Heating analysis on digital breast phantoms, Appl. Sci. 15 (8) (2025) 4262, http://dx.doi.org/10.3390/app15084262.

[34] M. Lazebnik, D. Popovic, L. McCartney, C.B. Watkins, M.J. Lindstrom, J. Harter, S. Sewall, T. Ogilvie, A. Magliocco, T.M. Breslin, et al., A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries, Phys. Med. Biol. 52 (20) (2007) 6093–6115, http://dx.doi.org/10.1088/0031-9155/52/20/002.

[35] K. Sasaki, E. Porter, E.A. Rashed, L. Farrugia, G. Schmid, Measurement and image-based estimation of dielectric properties of biological tissues — Past, present, and future, Phys. Med. Biol. 67 (14) (2022) 14TR01, http://dx.doi.org/10.1088/1361-6560/ac7b64.

[36] M.-L. Huang, T.-Y. Lin, Considering breast density for the classification of benign and malignant mammograms, Biomed. Signal Process. Control. 67 (2021) 102564, http://dx.doi.org/10.1016/j.bspc.2021.102564.

[37] L. Sani, A. Vispa, R. Loretoni, M. Duranti, N. Ghavami, D. Alvarez Sánchez-Bayuela, S. Caschera, M. Paoli, A. Bigotti, M. Badia, M. Scorsipa, G. Raspa, M. Ghavami, G. Tiberi, Breast lesion detection through MammoWave device: Empirical detection capability assessment of microwave images' parameters, PLoS One 16 (4) (2021) e0250005, http://dx.doi.org/10.1371/journal.pone.0250005.

[38] A. Ronca, A. Arduino, L. Zilberti, O. Bottauscio, G. Tiberi, Assessment of the feasibility of breast lesion detection with contrast source inversion for microwave tomography: A virtual experiment, in: 2024 18th European Conference on Antennas and Propagation (EuCAP), Glasgow, United Kingdom, 2024, pp. 1–5, http://dx.doi.org/10.23919/EuCAP60739.2024.10501105.

[39] A. Ronca, A. Arduino, L. Zilberti, O. Bottauscio, G. Tiberi, 3D GPU-based implementation of the contrast source inversion for breast lesion detection, in: 2024 International Conference on Electromagnetics in Advanced Applications, ICEAA, Lisbon, Portugal, 2024, pp. 262–262, http://dx.doi.org/10.1109/ICEAA61917.2024.10701936.

[40] S. Gabriel, R.W. Lau, C. Gabriel, The dielectric properties of biological tissues: II. Measurements in the frequency range 10 Hz to 20 GHz, Phys. Med. Biol. 41 (11) (1996) 2251–2269, http://dx.doi.org/10.1088/0031-9155/41/11/002.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, http://dx.doi.org/10.48550/arXiv.1706.03762, arXiv preprint arXiv:1706.03762.

[42] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, HI, USA, 2017, pp. 2261–2269.

[43] J. Stawiaski, A pretrained DenseNet encoder for brain tumor segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018, in: Lecture Notes in Computer Science, Vol. 11384, Springer, Cham, 2019.

[44] A. Halder, S. Gharami, P. Sadhu, et al., Implementing vision transformer for classifying 2D biomedical images, Sci. Rep. 14 (12567) (2024) http://dx.doi.org/10.1038/s41598-024-63094-9.

[45] M. Naas, H. Mzoughi, I. Njeh, M. BenSlima, An explainable AI for breast cancer classification using vision transformer (ViT), Biomed. Signal Process. Control. 108 (2025) 108011, http://dx.doi.org/10.1016/j.bspc.2025.108011.

[46] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, J.A. Benediktsson, Deep learning for hyperspectral image classification: An overview, IEEE Trans. Geosci. Remote Sens. 57 (9) (2019) 6690–6709, http://dx.doi.org/10.1109/TGRS.2019.2907932.

[47] C.L. Lai, R. Karmakar, A. Mukundan, R.K. Natarajan, S.C. Lu, C.Y. Wang, H.C. Wang, Advancing hyperspectral imaging and machine learning tools toward clinical adoption in tissue diagnostics: A comprehensive review, APL Bioeng. 8 (4) (2024) 041504, http://dx.doi.org/10.1063/5.0240444.

[48] M. Peker, Classification of hyperspectral imagery using a fully complex-valued wavelet neural network with deep convolutional features, Expert Syst. Appl. 173 (2021) 114708, http://dx.doi.org/10.1016/j.eswa.2021.114708.

[49] A. Ameen, I.E. Fattoh, T. Abd El-Hafeez, et al., Advances in ECG and PCG-based cardiovascular disease classification: A review of deep learning and machine learning methods, J. Big Data 11 (2024) 159, http://dx.doi.org/10.1186/s40537-024-01011-7.

[50] M. Taghipour-Gorjikolaie, B. Khalesi, B. Khalid, N. Ghavami, M. Badia, M. Ghavami, G. Tiberi, Breast density classification using frequency-based features in microwave imaging, Sci. Rep. 15 (2025) 45445, http://dx.doi.org/10.1038/s41598-025-28629-8.

[51] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, Santiago, Chile, 2015, pp. 4489–4497, http://dx.doi.org/10.1109/ICCV.2015.510.

[52] S. Hennessey, E. Huszti, A. Gunasekura, A. Salleh, L. Martin, S. Minkin, S. Chavez, N.F. Boyd, Bilateral symmetry of breast tissue composition by magnetic resonance in young women and adults, Cancer Causes Control 25 (4) (2014) 491–497, http://dx.doi.org/10.1007/s10552-014-0351-0.

[53] W. Zhao, et al., BASCNet: Bilateral adaptive spatial and channel attention network for breast density classification in the mammogram, Biomed. Signal Process. Control. 70 (2021) 103073, http://dx.doi.org/10.1016/j.bspc.2021.103073.