Deep-learning model for background parenchymal enhancement classification in contrast-enhanced mammography

E Ripaud, C Jailin, G I Quintana, P Milioni de Carvalho, R Sanchez de la Rosa, L Vancamberg

Deep-Learning Model for Background Parenchymal Enhancement Classification in Contrast-Enhanced Mammography

E. Ripaud[a], C. Jailin[a], G. I. Quintana[a], P. Milioni de Carvalho[a], R. Sanchez de la Rosa[a], and L. Vancamberg[a] *

[a]GE HealthCare, Buc, France

Abstract

Background: Breast Background Parenchymal Enhancement (BPE) is correlated with the risk of breast cancer. BPE level is currently assessed by radiologists in Contrast-Enhanced Mammography (CEM) using 4 classes: minimal, mild, moderate and marked, as described in *Breast Imaging Reporting and Data System* (BI-RADS). However, BPE classification remains subject to intra- and inter-reader variability. Fully automated methods to assess BPE level have already been developed in breast Contrast-Enhanced MRI (CE-MRI) and have been shown to provide accurate and repeatable BPE level classification. However, to our knowledge, no BPE level classification tool is available in the literature for CEM.

Materials & Methods: A BPE level classification tool based on Deep Learning (DL) has been trained and optimized on 7012 CEM image pairs (low-energy and recombined images) and evaluated on a dataset of 1013 image pairs. The impact of image resolution, backbone architecture and loss function were analyzed, as well as the influence of lesion presence and type on BPE assessment. The evaluation of the model performance was conducted using different metrics including 4-class balanced accuracy and mean absolute error. The results of the optimized model for a binary classification: minimal/mild versus moderate/marked, were also investigated.

Results: The optimized model achieved a 4-class balanced accuracy of 71.5% (95% CI: 71.2–71.9) with 98.8% of classification errors between adjacent classes. For binary classification, the accuracy reached 93.0%. A slight decrease in model accuracy is observed in the presence of lesions, but it is not statistically significant, suggesting that our model is robust to the presence of lesions in the image for a classification task. Visual assessment also confirms that the model is more affected by non-mass enhancements than by mass-like enhancements.

Conclusion: The proposed BPE classification tool for CEM achieves similar results than what is published in the literature for CE-MRI.

Keywords: Contrast-Enhanced Mammography, Background Parenchymal Enhancement, Deep Learning, Breast Imaging

1 Introduction

Contrast-Enhanced Mammography (CEM) is a recent imaging technique exploiting the process of tumor angiogenesis using intravenous injection of an iodine contrast agent. It has been demonstrated to improve breast cancer detection and characterization [1]. In CEM, two 2D projection

*corresponding author: laurence.vancamberg@gehealthcare.com

images of the breast are acquired at different energies of the X-ray beam; these low-energy (LE) and high-energy (HE) images are used to create a so-called recombined image (REC) showing contrast uptake [2]. Although this imaging modality has shown greater sensitivity and specificity for breast cancer detection compared to standard mammography, especially in women with dense breasts [3, 4], its diagnostic performance remains comparable to contrast-enhanced magnetic resonance imaging (CE-MRI) [2, 5]. Nonetheless, CEM is cost effective and some studies have shown a strong preference of patients for CEM over MRI [6, 7].

Background Parenchymal Enhancement (BPE) refers to the enhancement of the normal fibro glandular tissue (FGT) of the breast after contrast administration. BPE has been first described in breast CE-MRI and included in the *Breast Imaging Reporting and Data System* [8] (BI-RADS) 5^{th} edition in 2013. Its counterpart in CEM was introduced with the BI-RADS supplement dedicated to CEM in 2022 [9]. The literature suggests that BPE is associated with increased risk of breast cancer [10, 11]. It is a key feature visually assessed and reported by radiologists at study-level, using four classes: minimal, mild, moderate, and marked. Some studies have shown a correlation between BPE and clinical factors such as age, breast density, menstruation status and menstrual cycle timing [12, 13]. In addition, high BPE (moderate/marked) can mask lesions and thus affect image reading and tumor extent assessment [14, 15, 16]. For all those reasons, assessing BPE in CEM exam is essential, but as no automatic classifier exists, this task is currently extremely reader-dependent.

The classification of BPE using four levels is highly affected by intra- and inter-reader variability. Based on kappa statistic (κ), radiologists have demonstrated moderate agreement in classifying BPE ($\kappa = 0.43$, 95% CI: 0.05–0.69) after training on CEM [17]. In breast CE-MRI, inter-reader agreement was fair [18, 19], but improved with training until a moderate agreement [19]. These results highlight the importance of guidelines and standardized BPE levels in a reference atlas. To overcome the problem of human variability, fully-automatic methods have been proposed in breast CE-MRI on 3D volumes, notably using a deep convolutional neural network (CNN) architecture [20, 21] such as VGG [22, 23] or a radiomics feature extraction approach with breast and FGT segmentation [24]. These models have shown accuracy values ranging from 67 to 75% for 4-class classification and 79 to 91.5% for binary classification. A similar classification tool does not exist in CEM. Even if CE-MRI differs from CEM in particular from its 3D aspect, automated MRI approaches can be used as a source of inspiration, as both imaging modalities are used to detect contrast uptakes coming from angiogenesis. In addition, a study has shown a substantial agreement ($\kappa = 0.66$, 95% CI: 0.61-0.70) for BPE classification between CEM and breast MRI [25], highlighting similarities between these two applications [26].

Another four-level classification exists in mammography: the classification of breast density [27]. In the case of breast density, the literature has shown that deep learning (DL) tools can help reduce reader variability [28]. DL models such as ResNet-50 and EfficientNet-B0 have already shown to provide accurate and standardized assessment using a consensus between 2 to 3 radiologists as the reference standard [29, 30, 31, 32, 33], and can therefore be used as a source of inspiration. Compared with BPE, the image pattern recognition task remains different (*i.e.,* different physics, physiological phenomena, input images and image rendering).

As for BPE in CE-MRI and breast density in mammography, a BPE level classification tool in CEM would improve consistency between radiologists. In this work, a DL-based tool assessing BPE of a CEM image pair (LE/REC images) is developed. The purpose of this tool is to replicate a consensus-based radiologist assessment and automate the classification of BPE, thereby improving clinical effectiveness. The model was trained on a large CEM database labeled by multiple

independent readers. To the best of our knowledge, this study is the first to conduct the evaluation of an Artificial Intelligence (AI) model to assess the BPE level in CEM.

2 Materials

2.1 Database

The database used for this study contains CEM images from patients acquired between June 2019 and December 2022 from five clinical sites worldwide, in Europe, North America, Africa and Asia. In total, 2023 patients were enrolled, contributing to a dataset comprising 9073 image pairs (low-energy and recombined images).

CEM exams were performed with three different acquisition systems from GE HealthCare (GEHC): Senographe Pristina™, Senographe Essential™ and Senographe DS™ (GE HealthCare, Chicago, IL, USA).
Each patient case consisted of CEM images, including low-energy (LE) images, high-energy (HE) images and recombined (REC) images. REC images were obtained using SenoBright™ HD with NIRA, the latest GEHC recombination algorithm effective in reducing CEM-related artifacts [34]. For all patients, CEM images were acquired for the left and right breasts mainly with craniocaudal (CC) and mediolateral oblique (MLO) projections. Potential retakes and additional views such as mediolateral (ML), lateromedial (LM), lateral (LAT) and exaggerated CC (XCC) were also included in the database, but represent only 3% of the images. One to four exams were collected per patient.

2.2 Annotation and data split

The complete database was split into 3 datasets: a training, a validation and a test set of 1581, 222 and 220 patients, respectively, corresponding to 7012, 1048 and 1013 LE/REC image pairs.

To ensure labelling accuracy and reduce individual biases, multiple CEM experts previously trained on a set of images performed the annotation using the same software. Each reader assigned a BPE level to each CEM image pair: minimal, mild, moderate or marked.

Two methods are defined below to determine the ground truth GT associated with each CEM image pair. Both methods use the ground-truth probability P_i^{GT} associated with class $i \in \{\text{minimal}, \text{mild}, \text{moderate}, \text{marked}\}$, defined as follows:

$$P_i^{GT} = \frac{1}{N_{\text{reader}}} \sum_{n=1}^{N_{\text{reader}}} \begin{cases} 1 & \text{if reader } n \text{ assigns the class } i \\ 0 & \text{else} \end{cases} \tag{1}$$

where N_{reader} is the number of readers.

1. **Distribution Method**: This approach uses the variety of labels given by different readers for each image. It offers a more detailed insight than just a single discrete label by using a probability distribution of BPE levels as ground truth.

$$GT_{\text{distribution}} = [P_{\text{minimal}}^{GT}, P_{\text{mild}}^{GT}, P_{\text{moderate}}^{GT}, P_{\text{marked}}^{GT}] \tag{2}$$

2. **Categorical Method**: Here, the most common label among the readers is chosen as the consensus. In case of a tie, the image gets the highest BPE level.

$$GT_{\text{categorical}} = \arg\max_i P_i^{GT} \tag{3}$$

As a result, every image was associated with both a discrete BPE level label (categorical method) and a distribution of BPE labels (distribution method).

For the training and validation sets, the BPE annotation was performed by three independent CEM-expert scientists. The split between training and validation was stratified patient-wise based on the BPE level, regardless of clinical site or acquisition system. The BPE distribution is reported in Table 1. It can be noted that a perfectly distributed dataset was not possible as the database does not include enough high BPE cases. As some images (CC/MLO, left/right breast) inside a CEM exam from the same patient might have different assigned BPE, the category reported per patient in the table corresponds to the majority class on all views.

The test set was collected in a later stage from a single clinical site and annotated by a radiologist, specializing in breast imaging and CEM for over 5 years, and by three CEM-expert scientists with 4, 6 and 10 years of experience. The Cohen's kappa statistic was used to measure the inter-rater reliability. The overall inter-rater agreement was fair ($\kappa = 0.25$, 95% CI: 0.08-0.42). Ultimately, the test set includes 57 minimal cases, 106 mild cases, 42 moderate cases and 15 marked cases as reported in Table 1.

In addition, to analyze the influence of lesion presence and type, a radiologist conducted lesion location annotation and lesion classification of the test dataset, distinguishing between mass and non-mass enhancement. The annotation process consists of marking rectangular boxes on the lesion areas of each view, and provides the approximate lesion size. The test dataset was then divided into three categories: images without lesions, images containing mass-like lesions, and images containing non-mass-like lesions. Since the BPE texture is visually closer to a non-mass enhancement, CEM image pairs containing both mass and non-mass were classified as non-mass. The retakes and additional views were excluded from this analysis. Table 2 presents the stratification of the test dataset, including number of images and annotation length per lesion type. When an image contains multiple lesions, the reported annotation length corresponds to the sum of the lesion sizes.

Dataset	**Number of patients**	**BPE per patient**			
		minimal	**mild**	**moderate**	**marked**
Training	1581 (7012 images)	472 (30%)	750 (47%)	270 (17%)	89 (6%)
Validation	222 (1048 images)	73 (33%)	68 (31%)	43 (19%)	38 (17%)
Test	220 (1013 images)	57 (26%)	106 (48%)	42 (19%)	15 (7%)

Table 1: BPE distribution and number of patients per dataset.

Lesion type	**Number of images**	**Annotation length [mm]** mean ± std (min-max)
None	422	NA
Mass	277	61 ± 33 (16-171)
Non-mass	134	109 ± 44 (30-299)

Table 2: Stratification of the test dataset by lesion type. The number of images and the average annotation length are reported.

3 Methods

3.1 Model architecture and training

An architecture based on a deep convolutional neural network was implemented. The AI-pipeline consists in three different steps: a pre-processing stage, an image feature extraction stage, and a classification stage. Figure 1 shows the overall process of the BPE level classification.

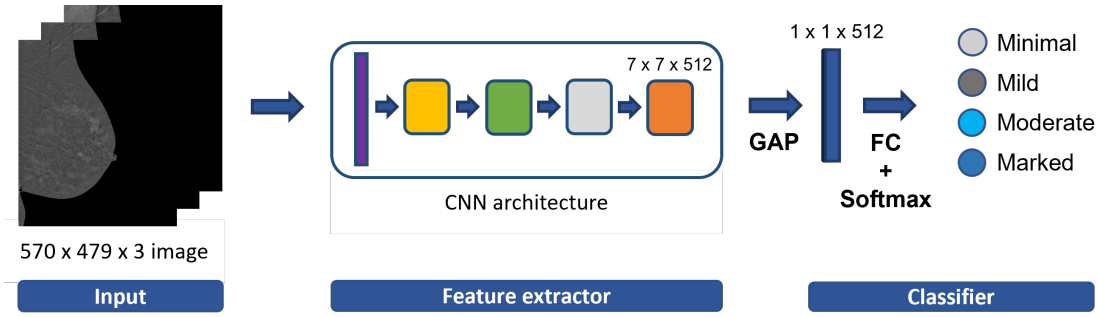

Figure 1: DL-based BPE classifier. Example using a ResNet-18 architecture and an image size of 570×479 pixels.

Pre-processing Pre-processing was applied to the images, including thresholding on the recombined image, then normalization on all channels. To meet the input requirement of pre-trained DL models, the REC image was stacked twice to obtain, with the LE image, 3 channels. Since the database contains images of different sizes, zero padding was used to set a single image size of 2850×2394 pixels and preserve a constant image resolution in the dataset (100 μm). All images were bi-linearly resized. The impact of resizing was analyzed by training the AI-pipeline with four different sizes of image: (285×239), (407×342), (570×479), and (950×798) pixels corresponding respectively to a resolution of 1000, 700, 500, and 300 μm. The image resolutions are to be compared with the length of the BPE patterns. For this purpose, an order of magnitude of the BPE granular texture length can be extracted and compared with the initial image size. The method proposed here is a texture length extraction obtained by thresholding the auto-correlation function at a chosen threshold (here 0.5):

$$L = 2 \cdot \max \{x \mid \forall x, |(f * f)(x)| > 0.5\} \tag{4}$$

Feature extraction Several deep-learning models were investigated for feature extraction including ResNet-18 [35], MobileNetV3-Small [36], DenseNet-121 [37] and VGG-16 [38]. Models were initially pre-trained on ImageNet. VGG-16 is a standard CNN model already used in different studies including the fully automatic classification of breast CE-MRI background parenchymal enhancement [22], and has 138×10^6 trainable parameters. The three other tested models, ResNet-18, DenseNet-121 and MobileNetV3-Small are lighter with respectively 11.7×10^6, 7.9×10^6 and 2.5×10^6 parameters and thus less prone to overfitting [39].

To prevent over-fitting, data augmentation strategies were performed, including random horizontal and vertical flips, rotations with random angle from 0 to 180 degrees and random zooms. As in our training dataset the BPE classes are unbalanced, class weighting method was also used.

Classification The final classification step consists in a Global Average Pooling layer (GAP) and a top Fully Connected (FC) layer with a Softmax activation function, as shown in Figure 1. The last layer outputs 4 probability scores P_i, one for each following class: minimal, mild, moderate and marked. To provide a final unique classification, the class with the highest score was selected. The model was trained to classify each CEM image pair independently.

Loss function As explained in paragraph 2.2, each CEM image pair is associated with a discrete label and a BPE class distribution. To leverage both the discrete nature required for a clinical BPE application and the continuous information within the distribution, we propose a custom weighted loss function $\mathcal{L}_{\text{tot}}$, defined as:

$$\mathcal{L}_{\text{tot}} = \alpha_{\text{CCE}}\mathcal{L}_{\text{CCE}} + \alpha_{\text{c\_MSE}}\mathcal{L}_{\text{c\_MSE}} + \alpha_{\text{KL}}\mathcal{L}_{\text{KL}} \tag{5}$$

- $\mathcal{L}_{\text{CCE}}$ is the categorical cross-entropy loss, the standard loss used for multi-class classification [40].
- $\mathcal{L}_{\text{c\_MSE}}$, a custom mean squared error (MSE) loss, is a regression loss. It is defined as the mean of weighted squared errors. This loss leverages both the discrete label and the BPE class distribution. The error between ground truth and output scores is weighted by class. Each weight corresponds to the number of classes of deviation from the discrete ground truth label. Classification errors are all the more penalized as the number of gap classes is high. This loss is used to leverage the hierarchical order of BPE levels. The $\mathcal{L}_{\text{c\_MSE}}$ loss is defined as:

$$\mathcal{L}_{\text{c\_MSE}} = \sum_{i=1}^{n_C} w_i \cdot (P_i^{GT} - P_i)^2, \; w_i = |GT - i| \tag{6}$$

  where n_C denotes the number of classes (here $n_C = 4$), w_i is the weight associated with class i and GT is the ground truth class. The ground-truth probability of class i is referred to as P_i^{GT} and the probability of class i predicted by the model as P_i.
- $\mathcal{L}_{\text{KL}}$ is the Kullback-Leibler (KL) divergence loss, it measures how two probability distributions are different from each other [41]. The model output is compared to the probability distribution obtained from multiple annotations per image.

The loss weights α_{CCE}, $\alpha_{\text{c\_MSE}}$, and α_{KL} are set such as $\alpha_{\text{CCE}} + \alpha_{\text{c\_MSE}} + \alpha_{\text{KL}} = 1$. Different combinations of loss weights were tested to optimize the model.

For cases failing between two adjacent classes (*e.g.,* mild/moderate) where ground truth is distributed across these two classes, KL divergence loss and custom MSE loss will encourage the model to represent this distribution. On the opposite, cross-entropy loss will penalize the model if it does not output the correct major class.

3.2 Evaluation protocol

To perform the evaluation, we defined a reference model (called hereafter reference model). The latter model corresponds to: a ResNet-18 architecture, an image resizing of 570×479 pixels and the loss function defined in Equation 5 with $\alpha_{\text{CCE}} = 1$, $\alpha_{\text{c\_MSE}} = 0$, and $\alpha_{\text{KL}} = 0$. All performances of other model's variants (different image resolutions, architectures, loss weights) will be compared to this reference model performance. The CNN architecture and loss function chosen for the reference model correspond to classics in the literature, while the choice of input image size

is based on preliminary tests.

For each experiment, the model was trained 5 times to estimate the training uncertainty. For all metrics, related 95% confidence interval is presented. In addition, p-values were computed with the Welch's t-test to assess whether there is a statistically significant difference in performance between each model variant and the reference model. Two notations can be reported: n.a. (not applicable) for the reference model and n.s. (not significant) in the absence of a statistically significant difference (p-value > 0.05).

Evaluation metrics Because the BPE model outputs both a probability distribution (4 scores) and a discrete final BPE label, multiple metrics were used to evaluate the results: the 4-class balanced accuracy and the mean absolute error (MAE).

To obtain the MAE, the classification problem was reconfigured into a representation on a [0,1] scale, considering the class hierarchy [minimal, mild, moderate, marked]. For a particular BPE class distribution, the center of gravity $\mathcal{G}$, a value between [0,1], has been chosen to represent the central tendency of the distribution. It is defined as the following weighted sum:

$$\mathcal{G} = \sum_{i=1}^{n_C} \alpha_i \cdot P_i \tag{7}$$

where n_C denotes the number of classes (here $n_C = 4$), P_i is the probability associated with class i and α_i is the weight associated with class i, such as $\alpha_0 = 0$, $\alpha_1 = 1/3$, $\alpha_2 = 2/3$, $\alpha_2 = 1$. This implies that a 100% probability of belonging to minimal, mild, moderate or marked class corresponds respectively to a center of gravity $\mathcal{G} = 0$, $\mathcal{G} = 1/3$, $\mathcal{G} = 2/3$, $\mathcal{G} = 1$. The mean absolute error between predicted and ground truth distributions was then computed by calculating the difference between the centers of gravity of the two distributions.

For the reference model, additional metrics are reported: area under the curve (AUC) of one-vs-rest receiver operating characteristic curves, confusion matrix, accuracy at one class apart, and accuracy for binary classification, *i.e.,* low BPE (minimal/mild) versus high BPE (moderate/marked). These are standard metrics for a classification problem [42].

Lesion impact on BPE classification evaluation Contrast-enhancing lesions could be interpreted by the DL model as BPE, causing a decrease in the performance of the BPE level classification. The impact of the presence and type of lesions on the reference model classification performance was therefore analyzed. The model performance was evaluated on three subsets: one containing no lesion, another containing mass-like lesions and the third containing non-mass-like lesions, using both 4-class balanced accuracy and MAE. As shown in Table 2, images containing non-mass lesions are associated with a higher lesion size. Indeed, their reported annotation average size is 109 mm compared with 61 mm for masses. To exclude potential bias in our results related to lesion size, a second analysis was conducted, considering only images with lesion sizes greater than 5 cm.
Besides, to obtain visual explanations of the CNN classification decision, the class activation maps were extracted from the last convolutional layer using the Gradient-weighted Class Activation Mapping (Grad-CAM) [43] method. Grad-CAM quantifies the gradients of the target class output with respect to the feature maps of the last convolutional layer in the backbone. These gradients represent how sensitive the output is to changes in the feature maps. By weighting the feature maps with these gradients, Grad-CAM generates a heat-map that indicates the importance of each spatial location in the feature maps for the final prediction. Additionally, a heat-map can be

obtained for each of the four outputs, to understand which parts of the image contributed to the output scores of each class.

4 Results

Performance of the reference model This section presents the classification results of the reference model described in 3.2.
The 4-class balanced accuracy is 71.5% (95% CI: 71.2–71.9) and the mean absolute error is 0.094 (95% CI: 0.090–0.099). Figure 2 presents the corresponding confusion matrix. Most classification errors occur between adjacent classes, except for 4 images on average (out of 1013). The accuracy at one class apart is 99.7%. The best classification performance was achieved for the extreme classes, *i.e.,* minimal and marked, with average success rates of 95.6% and 80.2%, respectively. Only 14 images labeled as minimal have been incorrectly classified as mild. The one-vs-rest AUC of minimal, mild, moderate and marked classes, is respectively 0.95 (95% CI: 0.948–0.951), 0.88 (95% CI: 0.868–0.891), 0.89 (95% CI: 0.883–0.896) and 0.97 (95% CI: 0.963–0.972). This highlights that images labeled as minimal or marked are more easily classified by the model. In addition, the model has more difficulty distinguishing between certain categories, in particular between minimal/mild and between moderate/marked classes, as shown in Figure 2. Indeed, for binary classification, the average accuracy of the model is of 93.0%.

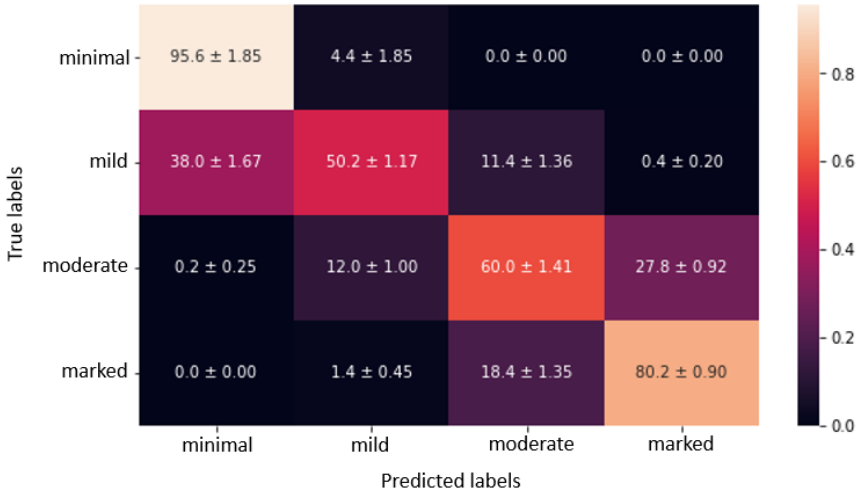

Figure 2: Confusion matrix of the reference model containing mean and standard deviation values in percentage.

Optimal input image resolution The classification results for different input image resolutions are presented in this section. The 4-class balanced accuracy and its corresponding confidence interval (CI) at 95% are presented in Figure 3 and reported in Table 3 together with the mean absolute error. We observe that increasing image size from 285×239 to 407×342 pixels improves the average 4-class balanced accuracy from 70.4% to 72.0% and reduces the associated variance. At size 570×479 pixels, the performances are not statistically different, but the variance still decreases. For higher image resolution, the performances both in terms of accuracy and MAE are significantly reduced. Therefore the image size of 570×479 pixels is best suited to combine performance and low variability of results.

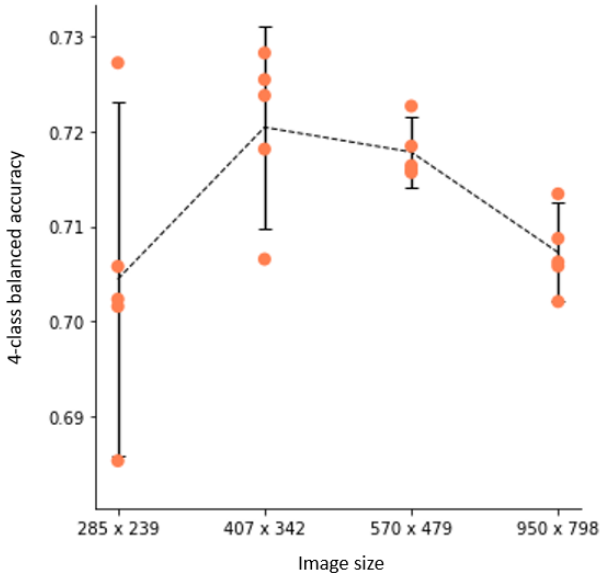

Figure 3: 4-class balanced accuracy for different input image resolutions. The orange bullets represent the five trainings. The black dotted line links the mean accuracies. The black whiskers indicate the 95% CI.

	4-class balanced accuracy	MAE
285×239	70.4% [68.5–72.3] (n.s.)	0.092 [0.087–0.097] (n.s.)
407×342	72.0% [70.9–73.1] (n.s.)	0.087 [0.083–0.091] (n.s.)
570×479	71.8% [71.4–72.2] (n.a.)	0.090 [0.088–0.092] (n.a.)
950×798	70.7% [70.2–71.2] (0.002)	0.095 [0.091–0.099] (0.028)

Table 3: Evaluation metrics for different input image sizes. Mean value, 95% CI and p-value are reported. n.a.: not applicable, n.s.: not significant.

Backbone evaluation The evaluation of different DL backbone architectures is reported in Table 4. In terms of accuracy, no statistically significant difference was found between ResNet-18 and MobileNetV3-Small. ResNet-18 shows a 4-class balanced accuracy of 71.5% (95% CI: 70.5–72.6) and MobileNetV3-Small of 72.0% (95% CI: 70.6–73.4). In addition, slightly lower results are obtained for VGG-16 and DenseNet-121 models with an average accuracy, respectively, of 70.3% (95% CI: 70.1–70.5) and 69.8% (95% CI: 68.4–71.2). ResNet-18 was finally chosen to maximize accuracy while having the minimum variability.

Evaluation of different loss weight combinations Different loss weights were experimented from the loss function defined in Equation 5. The corresponding evaluation metrics, including 4-class balanced accuracy and MAE, are reported in Table 5. The three first lines correspond to the use of a single type of loss (either categorical cross-entropy, KL divergence or custom MSE), and the other lines combine different losses.
Other combinations of loss weights were tested, but only relevant results are reported. We excluded the case where $\alpha_{\mathrm{CCE}} = 0$ in the loss combination experiments as the BPE assessment remains a classification issue for which the categorical cross-entropy loss is best suited. Regarding the three first lines, the model performance is significantly weaker for the custom MSE loss. Indeed, the 4-class balanced accuracy is of 67.5% (95% CI: 64.8–70.2) against 71.5% (95% CI: 71.2–71.9) for the

	4-class balanced accuracy
ResNet-18	71.5% [70.5–72.6] (n.a.)
VGG-16	70.3% [70.1–70.5] (0.033)
MobileNetV3	72.0% [70.6–73.4] (n.s.)
DenseNet-121	69.8% [68.4–71.2] (0.030)

Table 4: 4-class balanced accuracy for different feature extractor architectures. Mean accuracy, 95% CI and p-value are reported. n.a.: not applicable, n.s.: not significant.

α_{CCE}	$\alpha_{\text{c\_MSE}}$	α_{KL}	4-class balanced accuracy	MAE
1	0	0	71.5% [71.2–71.9] (n.a.)	0.094 [0.090–0.099] (n.a.)
0	1	0	67.5% [64.8–70.2] (0.014)	0.086 [0.084–0.088] (0.004)
0	0	1	71.1% [70.4–71.8] (n.s.)	0.086 [0.083–0.090] (0.003)
0.5	0.5	0	71.2% [70.2–72.2] (n.s.)	0.091 [0.088–0.094] (n.s.)
0.25	0	0.75	70.9% [69.9–71.9] (n.s.)	0.089 [0.084–0.094] (n.s.)

Table 5: Evaluation metrics for different loss weight combinations. Mean value, 95% CI and p-value are reported. n.a.: not applicable, n.s.: not significant.

categorical cross-entropy loss and 71.1% (95% CI: 70.4–71.8) for the KL divergence loss. In terms of mean absolute error, the cross-entropy loss is less effective with a value of 0.094 against 0.086 for the two other losses. By combining cross-entropy loss and custom MSE loss, the results are better but not significantly different from those of cross-entropy loss alone. For the KL divergence loss, the increased weight of the cross-entropy loss does not appear to improve the performance of the model, on the contrary the MAE tends to increase. Ultimately, loss combinations do not lead to significantly better results than the categorical cross-entropy loss alone as used by the reference model.

Influence of the presence of lesions on classification results Table 6 indicates the 4-class balanced accuracy and MAE results for the test dataset stratified by lesion presence and type. The model performs better in the absence of lesions with a 4-class balanced accuracy of 72.7% (95% CI: 71.0–74.4) against 72.3% (95% CI: 69.7–74.9) in the presence of masses and 71.0% (95% CI: 67.1–75.0) in the presence of non-masses. Nevertheless, the absence of any significant statistical difference between these results suggests that the model is robust to the presence of lesions for a classification task. As for the mean absolute error, it is significantly lower in the absence of lesions, with a value of 0.085. The difference in MAE between mass and non-mass lesions is also statistically significant, suggesting that the DL model performs better in predicting a BPE score distribution when the image contains mass enhancement rather than non-mass enhancement.

		4-class balanced accuracy	MAE
Absence of lesions		72.7% [71.0–74.4] (n.a.)	0.085 [0.079–0.091] (n.a.)
Presence of lesions	Mass	72.3% [69.7–74.9] (n.s.)	0.096 [0.094–0.099] (0.001)
	Non-mass	71.0% [67.1–75.0] (n.s.)	0.108 [0.106–0.111] (8e-06)

Table 6: Evaluation metrics for a stratified test dataset: absence of lesions vs presence of lesions (mass/non-mass). Mean value, 95% CI and p-value are reported. n.a.: not applicable, n.s.: not significant.

Table 2 highlights that non-masses have a higher average size than masses. Given this difference between lesion types and in order to determine whether it affects previous results, lesions smaller

	4-class balanced accuracy	MAE
Mass	72.4% [70.0–74.8] (n.a.)	0.101 [0.099–0.103] (n.a.)
Non-mass	70.0% [65.6–74.5] (n.s.)	0.110 [0.107–0.113] (7e-05)

Table 7: Mass vs Non-mass for lesion size > 5 cm. Mean value, 95% CI and p-value are reported. n.a.: not applicable, n.s.: not significant.

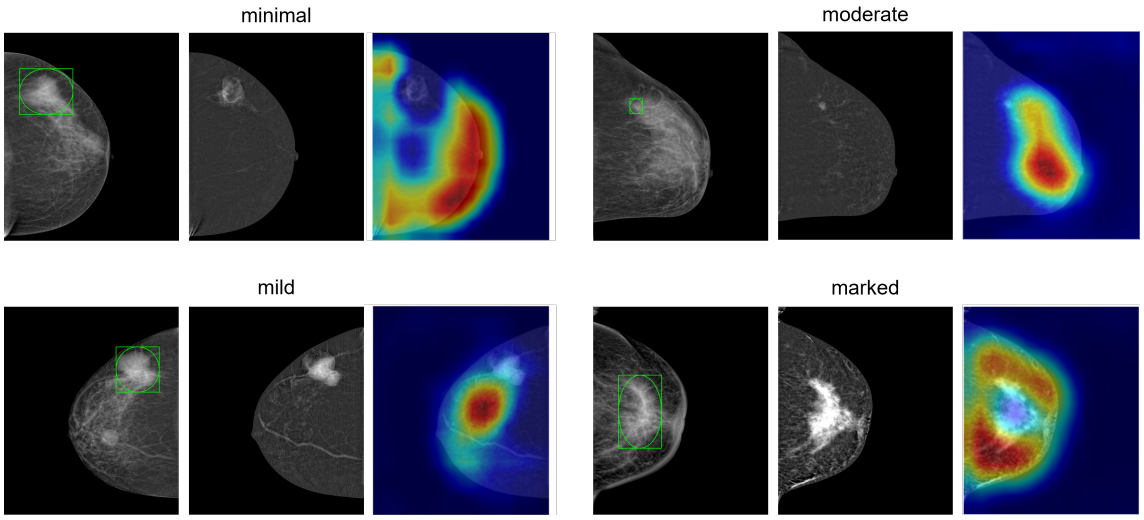

Figure 4: Pairs of low-energy and recombined images and Grad-CAM results for different BPE levels on cases showing lesion contrast uptake. Each green box corresponds to the ground truth cancer annotation that appears to be taken into account in the Grad-CAM heatmap.

than 5 cm were excluded from the test dataset for a second analysis. Two sets were then obtained: one set with masses (151 images) and another with non-masses (127 images), all larger than 5 cm. The results presented in Table 7 lead to the same conclusion: there is no statistical difference in terms of classification performance (4-class balanced accuracy) between mass and non-mass, while there is a statistical difference when considering the MAE metric. Consequently, this confirms that the results of Table 6 are not biased by the size of the lesions, and the observed differences are indeed due to the type of finding.

CEM image pairs are shown below to visually support the quantitative analysis of the lesion type most impacting the BPE classification. Figure 4 shows LE images, REC images and Gradient-weighted class activation maps (Grad-CAMs) for four cases comprising a mass-like lesion contrast uptake. On each LE image, the lesion location annotation was conducted. The heat-maps reflect higher neuron activation for the predicted class, knowing that each example was correctly classified, and gives an indication on where the DL focuses its attention. The attention seems to be well located on the clinically relevant region, *i.e.,* not on the lesion area. The minimal case presents a heat-map focused on the flat area. The mild, moderate and marked cases are well characterized by features extracted in the area where BPE is indeed seen. Each heat-map avoids the lesion. The Grad-CAMs seem to illustrate that the model does not take into account small and large masses in the BPE assessment.
Regarding misclassification, Figure 5 shows the Grad-CAMs for two other cases with lesions. The first case was labeled as mild and classified as moderate. For the predicted class, the heat-map is located both on the lesion and another region of the breast. This lesion is not as distinct as in

the previous cases. However, the score outputted by the model for the mild class is not null. The second case presents a non-mass-like enhancement lesion, the model seems to identify patterns specific to the lesion as BPE. While the ground truth is minimal, this image was classified as marked by the DL tool.

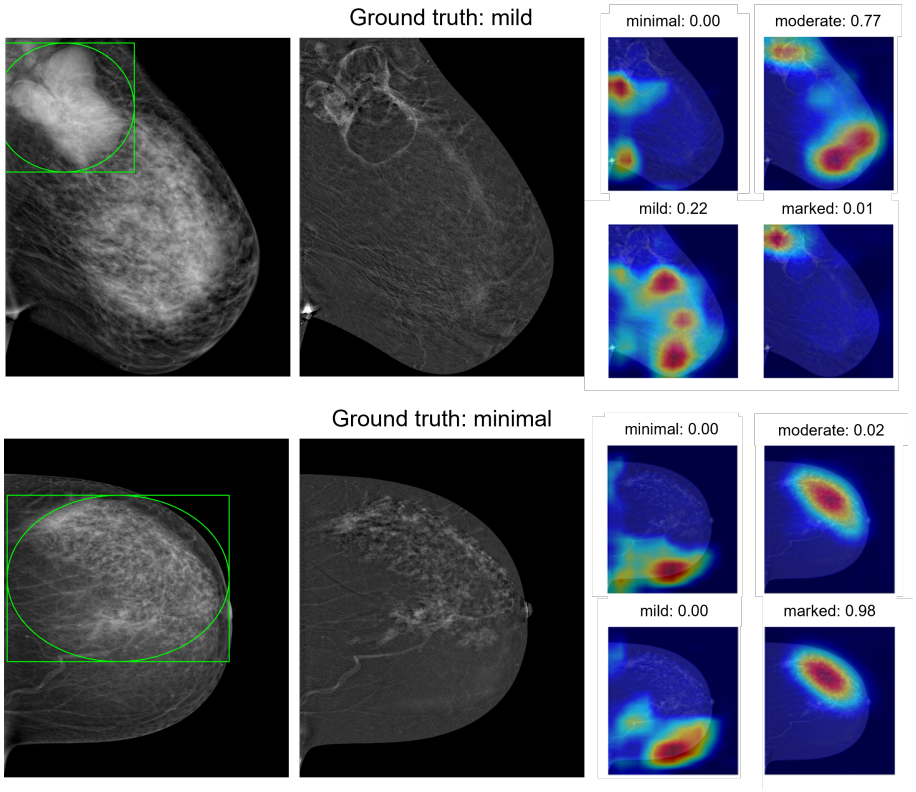

Figure 5: Pairs of low-energy and recombined images and Grad-CAMs for all 4 outputs on two cases incorrectly classified. The probability distribution scores are indicated, as well as the ground truth label. Each green box corresponds to the ground truth cancer annotation.

5 Discussion

The reference model reached an accuracy of 71.5% for 4-class classification and 93.0% for binary classification. There is no basis for comparison in CEM literature, but in breast CE-MRI, BPE classification models have been implemented with accuracy values ranging from 67 to 75% for 4-class classification and 79 to 91.5% for binary classification [22, 23, 24]. Our results fall within these ranges, and are even better for binary classification. In addition, the accuracy at one class apart was of 99.7%. This value shows that the majority of classification errors are between adjacent classes. Similar conclusions have been discussed in another BPE breast CE-MRI study [22]. The inter-reader agreement on the test set was described as fair ($\kappa = 0.25$, 95% CI: 0.08-0.42). Such a gap between readers confirms the need for standardization and may explain why exploring and adapting different parameters is not enough to significantly increase the success rate of the model.

Several variants of this model were tested. For the image size analysis, it can be seen that an image too small does not allow the model to optimally identify the different levels of BPE given the loss of information. But, an image too large does not seem to be necessary for this classification task.

The image resolutions are to be compared with the length of the BPE patterns. This length (computed by auto-correlation length presented in 3.1) is 3 to 4 mm. At the lowest image resolution, the BPE pattern is hence represented by 3-4 pixels. Consequently, keeping a resolution adapted to the CNN model scales is important for correctly encoding the BPE patterns. For large images, there may be noise affecting the pattern recognition by the CNN. Also, the network architecture may not be suitable for extracting the BPE pattern size in high-resolution images.

ResNet-18 shows results equivalent to MobileNetV3-Small and better than VGG-16 and DenseNet-121 on the monitored metrics. The number of trainable parameters of the MobileNetV3-Small model (2.5×10^6) is significantly lower, making them less prone to overfitting, as opposed to VGG-16 which has 138×10^6 parameters. In addition, ResNet and MobileNet are less computationally demanding and would be more adapted to standard clinical hardware. It is worth noting that the choice of the optimal image resolution and optimal loss combination was performed using ResNet-18 and fixed for the architecture comparison. Therefore, they may be not as optimal for other architectures as for ResNet-18. A complete study will require to optimize those parameters for all architectures.

On the optimal loss analysis, in terms of 4-class balanced accuracy, the categorical cross-entropy gives the best performance. On the other hand, KL divergence loss achieves better results than cross-entropy loss regarding the MAE, even though they demonstrate similar accuracy. This is expected given that the KL divergence compares predicted and ground truth probability distributions and that the MAE corresponds to the difference between the centers of gravity of these same two distributions. In addition, to leverage the hierarchical order of BPE levels, a custom MSE loss was tested but did not show any improvements. The combination of different losses did not improve either the performance of the model. Depending on the target BPE application, some losses are more appropriate, *e.g.,* categorical cross-entropy is more suitable for displaying a single BPE category, while to show the 4 output scores, KL divergence is better.

Regarding the influence of the presence and type of lesions, the model appears robust to the presence of lesions for a classification task. When the DL model predicts a BPE score distribution, a significant decrease in performance is observed in the presence of lesions and between mass and non-mass lesion types. As the decrease in model accuracy in the presence of non-masses is not due to the size of the lesions, it is probably be caused by the resemblance between the BPE texture and non-mass enhancement. A multi-view BPE classifier could help to distinguish BPE from lesions and would improve the performance of a such tool.

This study has several limitations and improvement opportunities. First, the proposed model is an image-level classification model. For future work, the objective is to combine all exam views to obtain a relevant BPE level prediction and to entirely and precisely replicate the clinical task at study level.
One of the limitations of this study is the imbalance of our dataset in terms of BPE. Indeed, the test set comprises 15 marked cases, compared to 106 mild cases. The primary reason for this imbalance is that marked cases represent less than 5% of clinical cases [44]. A balanced database would be optimal to obtain a better performing model.
Furthermore, a study has shown that the label variability has a considerable impact on the evaluation of the quality of models [45], so improving the quality of the labels is essential to achieve better performance of our classification tool. The poor agreement between readers on training and test sets may affect model performance. Ultimately, multiple expert radiologists performing BPE assessment of the training, validation and test datasets under clinical conditions through a more controlled reading procedure would improve the model's performance.

6 Conclusion

This study presents a BPE level classification tool for CEM based on deep learning. Several variants of a reference model were tested by modifying image resolution, backbone architecture and loss function to obtain an optimized classifier. Ultimately, the reference model achieved an accuracy of 71.5% for the 4-class classification problem and 93.0% for the binary classification. Most classification errors occur between adjacent classes. In addition, the classification model has demonstrated robustness in the presence of lesions in the image. To our knowledge, this study is the first to conduct an evaluation of such a BPE level classifier in CEM.

Acknowledgement

The authors would like to warmly acknowledge Sara Mohamed, Ann-Katherine Carton and Viviane Devauges from *GE HealthCare* for their help in the data annotation.

Compliance with Ethical Standards

This research study was conducted retrospectively using anonymized human subject data made available by research partners. Applicable law and standards of ethic have been respected.

References

[1] M. S. Jochelson, M. B. Lobbes, Contrast-enhanced mammography: state of the art, Radiology 299 (1) (2021) 36–48.

[2] B. K. Patel, M. Lobbes, J. Lewin, Contrast enhanced spectral mammography: a review, in: Seminars in Ultrasound, CT and MRI, Vol. 39, Elsevier, 2018, pp. 70–79.

[3] E. M. Fallenberg, F. F. Schmitzberger, H. Amer, B. Ingold-Heppner, C. Balleyguier, F. Diekmann, F. Engelken, R. M. Mann, D. M. Renz, U. Bick, et al., Contrast-enhanced spectral mammography vs. mammography and MRI–clinical performance in a multi-reader evaluation, European radiology 27 (2017) 2752–2764.

[4] A. Bozzini, L. Nicosia, G. Pruneri, P. Maisonneuve, L. Meneghetti, G. Renne, A. Vingiani, E. Cassano, M. G. Mastropasqua, Clinical performance of contrast-enhanced spectral mammography in pre-surgical evaluation of breast malignant lesions in dense breasts: a single center study, Breast Cancer Research and Treatment 184 (2020) 723–731.

[5] J. M. Lewin, B. K. Patel, A. Tanna, Contrast-enhanced mammography: a scientific review, Journal of Breast Imaging 2 (1) (2020) 7–15.

[6] M. M. Hobbs, D. B. Taylor, S. Buzynski, R. E. Peake, Contrast-enhanced spectral mammography (CESM) and contrast enhanced MRI (CEMRI): Patient preferences and tolerance, Journal of medical imaging and radiation oncology 59 (3) (2015) 300–305.

[7] W. A. Berg, A. I. Bandos, M. G. Sava, Analytic Hierarchy Process Analysis of Patient Preferences for Contrast-Enhanced Mammography (CEM) versus MRI as Supplemental Screening Options for Breast Cancer, Journal of the American College of Radiology (2023).

[8] A. C. of Radiology, C. J. D'Orsi, E. A. Sickles, E. B. Mendelson, E. A. Morris, et al., ACR BI-RADS Atlas: breast imaging reporting and data system; mammography, ultrasound, magnetic resonance imaging, follow-up and outcome monitoring, data dictionary, ACR, American College of Radiology, 2013.

[9] C. Lee, J. Phillips, J. Sung, J. Lewin, M. Newell, CONTRAST ENHANCED MAMMOGRAPHY (CEM)(A Supplement to ACR BI-RADS® Mammography 2013), American College of Radiology: Reston, VA, USA (2022).

[10] V. Sorin, Y. Yagil, A. Shalmon, M. Gotlieb, R. Faermann, O. Halshtok-Neiman, M. Sklair-Levy, Background parenchymal enhancement at contrast-enhanced spectral mammography (CESM) as a breast cancer risk factor, Academic radiology 27 (9) (2020) 1234–1240.

[11] S. Savaridas, D. Taylor, D. Gunawardana, M. Phillips, Could parenchymal enhancement on contrast-enhanced spectral mammography (CESM) represent a new breast cancer risk factor? Correlation with known radiology risk factors, Clinical radiology 72 (12) (2017) 1085–e1.

[12] Z. Karimi, J. Phillips, P. Slanetz, P. Lotfi, V. Dialani, J. Karimova, T. Mehta, Factors associated with background parenchymal enhancement on contrast-enhanced mammography, American Journal of Roentgenology 216 (2) (2021) 340–348.

[13] S. Zhao, X. Zhang, H. Zhong, Y. Qin, Y. Li, B. Song, J. Huang, J. Yu, Background parenchymal enhancement on contrast-enhanced spectral mammography: influence of age, breast density, menstruation status, and menstrual cycle timing, Scientific Reports 10 (1) (2020) 8608.

[14] M. del Mar Travieso-Aja, P. Naranjo-Santana, C. Fernández-Ruiz, W. Severino-Rondón, D. Maldonado-Saluzzi, M. R. Rodríguez, V. Vega-Benítez, O. Luzardo, Factors affecting the precision of lesion sizing with contrast-enhanced spectral mammography, Clinical Radiology 73 (3) (2018) 296–303.

[15] B. A. Varghese, M. Perkins, S. Cen, X. Lei, J. Fields, J. Jamie, B. Desai, M. Thomas, D. H. Hwang, S. Lee, et al., Cem radiomics for distinguishing lesion from background parenchymal enhancement in patients with invasive breast cancer, in: 18th International Symposium on Medical Information Processing and Analysis, Vol. 12567, SPIE, 2023, pp. 134–146.

[16] C. Jailin, S. Mohamed, P. Iordache, P. Milioni De Carvalho, S. Yehia, Ahmed, E. Abdullah, Abdel Sattar, A. Farouk Ibrahim, Moustafa, M. Mohammed, Gomaa, R. Mohammed, Kamal, L. Vancamberg, AI-based cancer detection model for Contrast-Enhanced Mammography, Bioengineering (2023).

[17] W. A. Berg, A. I. Bandos, M. L. Zuley, U. X. Waheed, Training radiologists to interpret contrast-enhanced mammography: toward a standardized lexicon, Journal of Breast Imaging 3 (2) (2021) 176–189.

[18] L. J. Grimm, A. L. Anderson, J. A. Baker, K. S. Johnson, R. Walsh, S. C. Yoon, S. V. Ghate, Interobserver variability between breast imagers using the fifth edition of the BI-RADS MRI lexicon, American Journal of Roentgenology 204 (5) (2015) 1120–1124.

[19] A. Melsaether, M. McDermott, D. Gupta, K. Pysarenko, S. D. Shaylor, L. Moy, Inter-and intrareader agreement for categorization of background parenchymal enhancement at baseline and after training, American Journal of Roentgenology 203 (1) (2014) 209–215.

[20] X. Ma, J. Wang, X. Zheng, Z. Liu, W. Long, Y. Zhang, J. Wei, Y. Lu, Automated fibroglandular tissue segmentation in breast mri using generative adversarial networks, Physics in Medicine & Biology 65 (10) (2020) 105006.

[21] R. Ha, P. Chang, E. Mema, S. Mutasa, J. Karcich, R. T. Wynn, M. Z. Liu, S. Jambawalikar, Fully automated convolutional neural network method for quantification of breast MRI fibroglandular tissue and background parenchymal enhancement, Journal of digital imaging 32 (2019) 141–147.

[22] K. Borkowski, C. Rossi, A. Ciritsis, M. Marcon, P. Hejduk, S. Stieb, A. Boss, N. Berger, Fully automatic classification of breast mri background parenchymal enhancement using a transfer learning approach, Medicine 99 (29) (2020).

[23] S. Eskreis-Winkler, E. J. Sutton, D. D'Alessio, K. Gallagher, N. Saphier, J. Stember, D. F. Martinez, E. A. Morris, K. Pinker, Breast MRI background parenchymal enhancement categorization using deep learning: outperforming the radiologist, Journal of Magnetic Resonance Imaging 56 (4) (2022) 1068–1076.

[24] Y. Nam, G. E. Park, J. Kang, S. H. Kim, Fully automatic assessment of background parenchymal enhancement on breast mri using machine-learning models, Journal of Magnetic Resonance Imaging 53 (3) (2021) 818–826.

[25] J. Sogani, E. A. Morris, J. B. Kaplan, D. D'Alessio, D. Goldman, C. S. Moskowitz, M. S. Jochelson, Comparison of background parenchymal enhancement at contrast-enhanced spectral mammography and breast MR imaging, Radiology 282 (1) (2017) 63–73.

[26] E. Luczynska, M. Pawlak, T. Piegza, T. J. Popiela, S. Heinze, S. Dyczek, W. Rudnicki, Analysis of background parenchymal enhancement (BPE) on contrast enhanced spectral mammography compared with magnetic resonance imaging, Ginekologia polska 92 (2) (2021) 92–97.

[27] S. D. Pawar, P. T. Joshi, V. A. Savalkar, K. K. Sharma, S. G. Sapate, Past, Present and Future of Automated Mammographic Density Measurement for Breast Cancer Risk Prediction, in: Journal of Physics: Conference Series, Vol. 2327, IOP Publishing, 2022, p. 012076.

[28] A. T. Watanabe, T. Retson, J. Wang, R. Mantey, C. Chim, H. Karimabadi, Mammographic breast density model using semi-supervised learning reduces inter-/intra-reader variability, Diagnostics 13 (16) (2023) 2694.

[29] A. Ciritsis, C. Rossi, I. Vittoria De Martini, M. Eberhard, M. Marcon, A. S. Becker, N. Berger, A. Boss, Determination of mammographic breast density using a deep convolutional neural network, The British journal of radiology 92 (1093) (2019) 20180691.

[30] B. Rigaud, O. O. Weaver, J. B. Dennison, M. Awais, B. M. Anderson, T.-Y. D. Chiang, W. T. Yang, J. W. Leung, S. M. Hanash, K. K. Brock, Deep Learning Models for Automated Assessment of Breast Density Using Multiple Mammographic Image Types, Cancers 14 (20) (2022) 5003.

[31] S. Li, J. Wei, H.-P. Chan, M. A. Helvie, M. A. Roubidoux, Y. Lu, C. Zhou, L. M. Hadjiiski, R. K. Samala, Computer-aided assessment of breast density: comparison of supervised deep learning and feature-based statistical learning, Physics in Medicine & Biology 63 (2) (2018) 025005.

[32] R. Sexauer, P. Hejduk, K. Borkowski, C. Ruppert, T. Weikert, S. Dellas, N. Schmidt, Diagnostic accuracy of automated ACR BI-RADS breast density classification using deep convolutional neural networks, European Radiology (2023) 1–8.

[33] V. Magni, M. Interlenghi, A. Cozzi, M. Alì, C. Salvatore, A. A. Azzena, D. Capra, S. Carriero, G. Della Pepa, D. Fazzini, et al., Development and validation of an AI-driven mammographic breast density classification tool based on radiologist consensus, Radiology: Artificial Intelligence 4 (2) (2022) e210199.

[34] G. Gennaro, E. Baldan, E. Bezzon, F. Caumo, Artifact reduction in contrast-enhanced mammography, Insights into Imaging 13 (1) (2022) 1–11.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[36] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314–1324.

[37] 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, 2017, pp. 4700–4708.

[38] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[39] J. Lever, M. Krzywinski, N. Altman, Points of significance: model selection and overfitting, Nature methods 13 (9) (2016) 703–705.

[40] Y. Ho, S. Wookey, The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling, IEEE access 8 (2019) 4806–4813.

[41] F. Pérez-Cruz, Kullback-Leibler divergence estimation of continuous distributions, in: 2008 IEEE international symposium on information theory, IEEE, 2008, pp. 1666–1670.

[42] L. Maier-Hein, B. Menze, et al., Metrics reloaded: Pitfalls and recommendations for image analysis validation, arXiv. org (2206.01653) (2022).

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[44] M. L. Yang, C. Bhimani, R. Roth, P. Germaine, Contrast enhanced mammography: focus on frequently encountered benign and malignant diagnoses, Cancer Imaging 23 (1) (2023) 10.

[45] S. Squires, E. F. Harkness, D. G. Evans, S. M. Astley, The effect of variable labels on deep learning models trained to predict breast density, Biomedical Physics & Engineering Express 9 (3) (2023) 035030.