Deep Learning Framework for Managing Inter-Reader Variability in Background Parenchymal Enhancement Classification for Contrast-Enhanced Mammography

Elodie Ripaud[1,2,3], Clément Jailin[2], Pablo Milioni de Carvalho[1], Laurence Vancamberg[1], and Isabelle Bloch[3]

[1]GE HealthCare, Buc, France
[2]Université Paris-Saclay, CentraleSupélec, ENS Paris-Saclay, CNRS, LMPS, Gif-sur-Yvette, France
[3]Sorbonne Université, CNRS, LIP6, Paris, France
elodie.ripaud@gehealthcare.com

Abstract. Background parenchymal enhancement (BPE) classification for contrast-enhanced mammography (CEM) is highly affected by inter-reader variability. Traditional approaches aggregate expert annotations into a single consensus label to minimize individual subjectivity. By contrast, we propose a two-stage deep learning framework that explicitly models inter-reader variability through self-trained, reader-specific embeddings. In the first stage, the model learns discriminative image features while associating each reader with a dedicated embedding that captures their annotation signature, enabling personalized BPE classification. In the second stage, these embeddings can be calibrated using a small set of CEM cases selected through active learning and annotated by either a new reader or a consensus standard. This calibration process allows the model to adapt to new annotation styles with minimal supervision and without extensive retraining. This work leverages a multi-site CEM dataset of 7,734 images, non-exhaustively annotated by several readers. Calibrating reader-specific embeddings using a set of 40 cases offers an average accuracy of 73.5%, outperforming the proposed baseline method based on reader consensus. This approach enhances robustness and generalization in clinical environments characterized by heterogeneous labeling patterns.

Keywords: Background parenchymal enhancement · Contrast-enhanced mammography · Deep learning.

1 Introduction

Contrast-enhanced mammography (CEM) is a recent breast imaging technique based on a dual-energy X-ray mammography acquisition performed after injection of an intravenous iodinated contrast agent [20]. CEM combines a low energy (LE) and a high energy (HE) image to create a recombined image (REC) showing

contrast uptake and highlighting tumor angiogenesis. Compared to mammography alone, CEM has been shown to improve sensitivity and specificity for the detection of breast cancer [4]. Different levels of background parenchymal enhancement (BPE), indicating enhancement of normal fibroglandular tissue, are observed in CEM, as well as in contrast-enhanced magnetic resonance imaging (CE-MRI). The literature suggests that high BPE is associated with an increased risk of breast cancer [25] and may affect image interpretation, masking, or mimicking cancers [16]. BPE is visually assessed and reported by radiologists using the *Breast Imaging Reporting and Data System* (BI-RADS) [5] four-category scale: minimal, mild, moderate, and marked, as illustrated in Figure 1.

This classification task remains challenging due to notable variability among radiologists. Studies indicate heterogeneous results regarding inter-reader agreement in classifying BPE in CEM and CE-MRI, ranging from fair to substantial [2, 8, 12, 15, 17], based on Kappa interpretation [28]. Similarly, in digital mammography, breast density is classified into four categories. Despite efforts to standardize processes, human subjectivity continues to be observed [22]. Variability can be explained by several aspects:

1. Individual tendencies. Some readers may naturally tend to annotate more extremely or conservatively.
2. Training, education, and institutional guidelines [1].
3. Interpretation of image characteristics, particularly the overall enhancement patterns (*e.g.,* distinguishing between normal tissue and non-mass enhancement).
4. Prior case exposure. A reader used to high BPE cases (*e.g.,* depending on the patient population [21]) may reserve "marked" labels for extremes, while another may use them more often. This exposure bias causes consistent shifts along the BPE scale, despite agreement on case ordering.
5. Random uncertainties, mainly related to the consistency of individual reader assessments. This refers to intra-reader variability.

The literature suggests that fully automated approaches can lead to more standardized and reproducible BPE classification. Deep convolutional neural networks (CNN) and radiomics have been proposed, reaching accuracies of 67% to 75% [3, 6, 18, 23]. To train these models, studies typically create a single reference (consensus) by combining multiple expert annotations, sometimes inconsistently due to incomplete labeling. The most common method is majority voting. In addition, Park et al. developed a confidence-guided learning method for breast density classification, where radiologists' annotations are weighted by their confidence levels [19]. Li et al. introduced a method that leverages a small set of trusted data to jointly learn a data classifier and a label aggregator [13]. Probabilistic approaches have also been explored to estimate the true labels from multiple noisy annotations [11, 14] and model annotator expertise based on the data observed [29]. Another study investigated correcting reader bias in breast density assessment [26]. Scores on a 0–100 visual analog scale were standardized to align readers to a common distribution. Since not all readers assessed the same images, the method mitigated bias by comparing scores on shared cases.

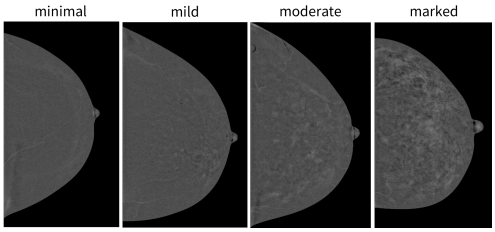


Fig. 1: BPE classification in REC images.

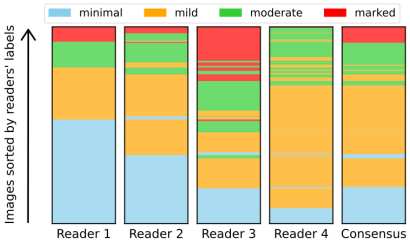


Fig. 2: BPE assessments of the test set sorted by labels from R_1 to R_4.

However, this approach may lack robustness when dealing with small sample sizes and using the four-category BI-RADS BPE scale, which does not capture variability between readers in such a nuanced way.

Moreover, the annotation process is resource-intensive and time-consuming, making it difficult to collect large-scale expert-labeled datasets. In practice, BPE and breast density classifiers are often trained using labels extracted from radiology reports written by different radiologists. It can lead to inconsistencies and affect the performance and reliability of the model predictions [27].

Rather than enforcing agreement through a single reference label, this work proposes to explicitly model inter-reader variability. We introduce a two-stage deep learning framework designed to handle heterogeneous and non-exhaustive multi-reader annotations. The approach leverages self-trained, reader-specific embeddings to capture individual annotation styles. In the first stage, the model learns discriminative image features while associating each reader with a dedicated embedding, enabling personalized BPE classifications. In the second stage, this embedding is calibrated using a small number of annotated examples, either from a new reader or a consensus. It enables standardized or site-specific BPE classifications with minimal annotation effort and no need for extensive retraining.

2 Database

This work leverages a CEM dataset of 1813 cases (7734 images) from various clinical sites. Each case consists of at least two bilateral views (craniocaudal and mediolateral oblique), acquired using different imaging systems: Senographe DS, Senographe Essential, and Senographe Pristina (GE HealthCare, Chicago, IL, USA). Several readers reviewed the dataset by assigning a BPE level per image: minimal, mild, moderate, or marked.

The dataset was divided into three subsets: training, validation, and test. Eight CEM experts annotated the training/validation data. Each reviewed a different mix of cases. Ultimately, each image was annotated by at least two readers, with the number of case readings per reader ranging from 108 to 876. The training-validation split was stratified considering both consensus and individual

BPE distributions. This resulted in a training set of 5832 LE/REC image pairs (1371 cases) and a validation set of 1011 pairs (239 cases). The test set was collected from a single clinical site and contains 891 LE/REC image pairs (203 cases). It was entirely annotated by four CEM experts, including two from the eight training readers, specifically Reader 1 (R_1) and Reader 2 (R_2), with 4 and 15 years of experience in CEM.

Preliminary analyses were conducted to investigate the inter-reader variability on the test dataset. Cohen's Kappa [28] shows fair inter-reader agreement ($\kappa = 0.26 \pm 0.15$), consistent with the literature findings. Kendall's Tau [7] was also employed to measure the ordinal association between readers' labels, resulting in $\tau = 0.69 \pm 0.05$. This indicates a significant alignment in readers' image rankings, as illustrated in Figure 2. The latter shows the BPE distribution for each reader and the consensus, defined as the majority vote. In case of a tie, the highest level is chosen. The images were hierarchically and stably sorted based on the labels assigned by each reader, in order from Reader 1 to 4. The different interpretations of the readers regarding the BPE scale are readily apparent. Reader 3 distributes his assessments uniformly across the scale, whereas Reader 2 assigns lower BPE levels, reserving higher categories for more extreme BPE cases. Reader 4, on the other hand, mainly uses the middle categories. Although BPE distributions vary across readers, they strongly agree on the relative ordering of images from low to high BPE. This consistency will be leveraged in our method by capturing the reader's signature.

3 Methods

An original two-stage deep learning framework is proposed. The first stage involves training a CNN backbone jointly with reader-specific embeddings, to enable personalized BPE assessments (Section 3.1). The second stage calibrates these embeddings using a small set of annotations to adapt the model output to new readers (Section 3.2).

3.1 BPE reader-specific classifier

Figure 3 (A) shows the BPE classification model, including the image feature extractor, reader-specific embeddings, and the classifier. We used ResNet-18 [9], initialized with ImageNet weights, to extract 512 features from LE/REC image pairs of 570×479 pixels. This architecture and image resolution have already demonstrated their effectiveness in BPE classification, offering a good balance between computational efficiency and classification accuracy [23]. In addition, an embedding layer was used to learn reader-specific representations. This layer is implemented as a trainable weight matrix, where each row corresponds to a reader ID and stores a continuous, dense vector (embedding) of size D. The embeddings encode in a low-dimensional space the annotation biases, outlined in the Introduction. For each reader R_i, the image features are concatenated with the corresponding embedding and fed into a classifier. It comprises two

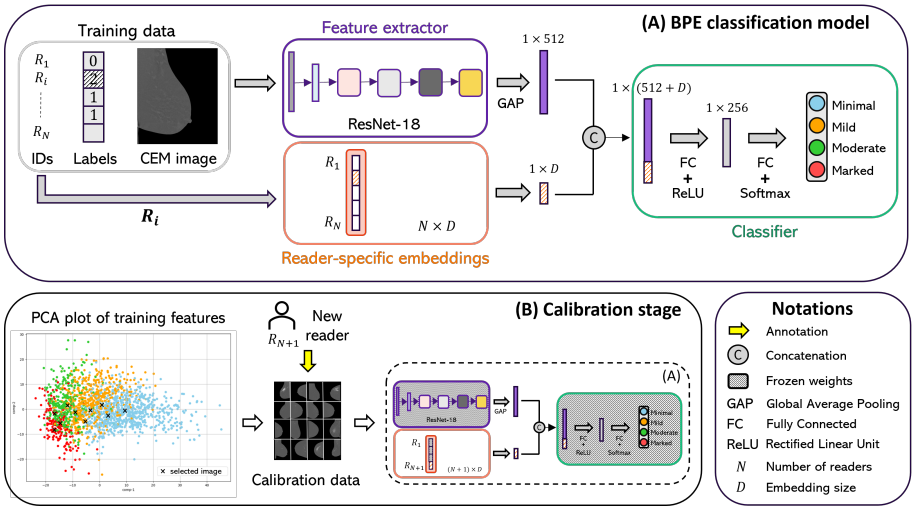


Fig. 3: Two-stage deep learning framework for BPE reader-specific classification.

Fully Connected layers, with ReLU and Softmax activation functions to output the four BPE class probabilities. Ultimately, a BPE score is calculated from the model outputs within the range $[0, 1]$ [24].

The pipeline is executed for N readers. Since no reader annotated the full dataset, we used a batch sampling strategy to ensure balanced reader representation in each batch of 32 images. The embeddings are jointly optimized with the network parameters via backpropagation using the SGD optimizer, by minimizing the average categorical cross-entropy loss across all readers [10]. The embedding weights were initialized from a normal distribution $\mathcal{N}(0, 1)$. Weight decay and classic geometrical data augmentation techniques (flips and rotations) were applied to prevent overfitting and enhance generalization.

We evaluated the effect of embedding size (D) using dimensions 1, 4, 8, 16, 32, and 64. With $N = 8$, the model was trained ten times to estimate the training uncertainty and then assessed on the test set using 4-class balanced accuracy. The mean value and the 95% confidence interval (CI) are reported. The results are computed for R_1 and R_2, as they are the only readers who annotated both training and test data.

3.2 Calibration

After training the model based on N readers using the embedding strategy for managing inter-reader variability, we want to adapt it to a new reader, hereafter called the reference reader. In a realistic scenario, this calibration stage relies on a limited set of annotations, as illustrated in Figure 3 (B).

The calibration data to be annotated by the reference reader was selected from the training and validation datasets using an active learning strategy

that combines representative-based and uncertainty-based sampling [30]. First, a principal component analysis was performed on the training/validation image features extracted by our backbone. K-means clustering was employed to ensure diversity and representativeness in sampling. For each image, the Euclidean distance to the centroid is calculated. To probe uncertainties and better capture the reference reader's bias, the variance of predictions from N reader-specific embeddings was also calculated. The aim is to identify samples that exhibit higher variability, especially those near BPE class boundaries. Representative-based and uncertainty-based sampling are combined to select relevant images by calculating a global score:

$$s_{\text{tot}} = \alpha \cdot \left(\sum_{i=1}^{d} (x_i - c_i)^2 \right)^{-1/2} + (1 - \alpha) \cdot \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2, \tag{1}$$

where α is the adjustment coefficient chosen experimentally, d is the dimension of the image feature vector x, c is the centroid of the cluster to which the image belongs, y denotes the BPE score produced by the model during the first training stage for a given reader, and $\bar{y}$ is the mean score across readers.

The images with the highest scores are ultimately included in the calibration dataset, along with views associated with the same exam. The reference reader then annotated the selected calibration data. The dataset was divided into training and validation sets in a 70/30 ratio, stratified based on the BPE level. To capture the signature of the reference reader, the feature extractor and classifier were frozen while the embedding layer was updated and retrained.

Calibration was performed using R_1 and R_2 as new reference readers. Our model was first trained with $N = 7$, *i.e.* excluding the reference reader. The impact of the calibration dataset size on model performance was evaluated on the test set using 4-class balanced accuracy. Our model was compared to the baseline method, which employs the architecture shown in Figure 3 (A) without the reader-specific embeddings part and was trained on the reader consensus (majority vote). For comparison with the calibration results, the final linear classification layers of the baseline model were specifically trained for the reference readers.

4 Results

Figure 4 shows the classifier performance on the test set depending on the embedding size for R_1 and R_2. The accuracy increases significantly from an embedding size of 1 to 32 for both readers, with less variability in performance. Stability and performance are likely due to the richer representation provided by larger embeddings. This helps the model to converge more consistently and capture more detailed information on the reader's behavior. For $D = 32$, the model achieves a 4-class balanced accuracy of 71.8% (95% CI: 70.4–73.3) for R_1 and 72.6% (95% CI: 71.5–73.6) for R_2. Our results are comparable to those reported

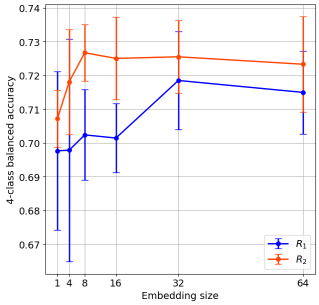


Fig. 4: Model accuracy for different embedding sizes.

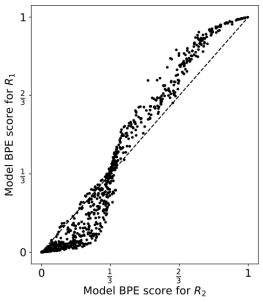


Fig. 5: Comparison of model R_1/R_2 scores.

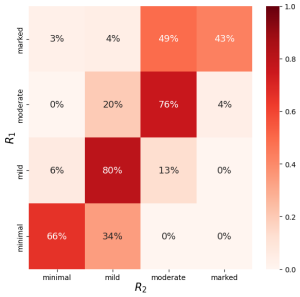


Fig. 6: Confusion matrix with R_1/R_2 labels.

by state-of-the-art BPE classifiers in CEM and CE-MRI [3, 6, 18, 23]. The implemented training strategy leverages images annotated by other readers while effectively encoding the signature of the readers. Unlike R_1, a smaller embedding size is sufficient to achieve maximum performance for R_2, stabilizing at $D = 8$. R_1 exhibits a BPE distribution that significantly differs from other readers by assigning more extreme BPE levels (minimal or marked), as shown in Figure 2. We hypothesize that the model must handle a greater diversity of annotation patterns from R_1, which requires adjusting the embedding size.

Figure 5 illustrates the model BPE scores for R_1 and R_2, using $D = 32$. It suggests a significant difference in interpretation between the two readers, which cannot be explained by a simple bijective application affecting the scores. It appears that the larger embeddings capture not only class shifts along the BPE scale but also other readers' labeling variations likely related to image features. The comparison of readers' model scores is consistent with their labels, as shown by the confusion matrix in Figure 6. Similar shifts are observed, indicating that the model faithfully captures variations in the readers' annotations in a more nuanced way on the BPE continuous scale. This highlights the model ability to adapt to individual differences in BPE assessments, preserving the internal consistency of each reader's annotations rather than enforcing a potentially misleading consensus.

Figure 7 shows the performance of the calibrated embeddings for different calibration dataset sizes, compared to the baseline model. An embedding size of 32 was chosen, as it demonstrated the best performance for both reference readers. No significant change in balanced accuracy is observed with the increasing number of calibration data. R_1 shows lower performance compared to R_2, consistent with the previous training results. For a dataset of 40 cases, our model achieves a 4-class balanced accuracy of 71.4% for R_1 and 75.5% for R_2, while the baseline model achieves 65.0% and 58.8%, respectively. Furthermore, the baseline classifier needs a larger dataset to enhance its performance. As the dataset size increased from 20 to 400 cases, the baseline balanced accuracy for R_1 improved significantly from 64.0% to 70.7%, and for R_2 from 43.3% to 70.9%.

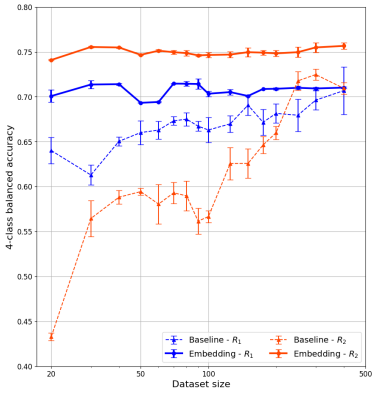
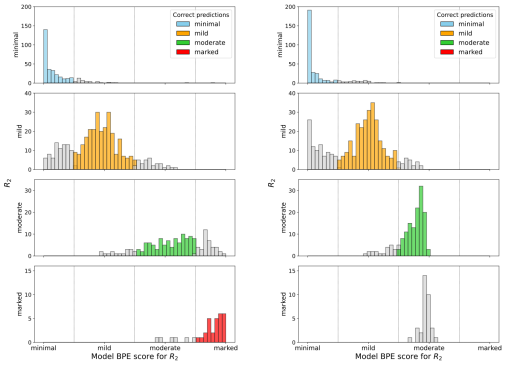


Fig. 7: Model balanced accuracy for different calibration dataset sizes.



(a) embedding-based (b) baseline

Fig. 8: BPE score distribution per target class for R_2-based calibration using 40 cases.

It reaches a performance comparable to that of the calibrated embeddings. During the calibration phase, the number of trainable parameters was lower for the embedding layer with 32 parameters compared to the baseline classifier with 1.32×10^5 parameters, preventing overfitting with a small set of annotations. The method based on calibrated embeddings, therefore, offers a more cost-effective solution. Its consistent performance across different calibration dataset sizes suggests robustness and the ability to generalize well, even with limited data. This underscores their potential in clinical applications where data scarcity is a major challenge.

Figure 8 shows the BPE score distribution per target class for R_2, from the (a) embedding-based and (b) baseline models, both trained on a 40-case dataset. The comparison of these models also reveals that the embeddings provide more reliable predictions, as evidenced by the distributions centered on the target class. By contrast, the baseline model incorrectly classified mild cases as minimal with high confidence and failed to predict any marked cases.

Embeddings appear less sensitive to the BPE class distribution in the calibration dataset. They enable an efficient reinterpretation of the BPE scale based on image features. By selecting a small, informative dataset, calibration helps adapt the model to a new reader. This applies whether the reader differs significantly from the consensus (like R_1) or shares similarities with it (like R_2). It supports both generalization and site-specific adaptation by aligning with a given BPE distribution. However, several improvements are needed. The application of this method to other datasets or clinical contexts remains to be explored. Additionally, the complexity of the embeddings can make the model's decisions more difficult to interpret. Exploring different active learning strategies for selecting calibration data could provide deeper insights into the effectiveness of our method. Finally, future work could aim to extend the evaluation to a larger pool

of readers. This would further strengthen the evidence for the model's adaptability to diverse annotation styles.

5 Conclusion

This paper presents a two-stage deep learning framework to address inter-reader variability in BPE classification for CEM. A CNN backbone is jointly trained with reader-specific embeddings to capture annotation signatures. These embeddings are then calibrated for new readers using a small, actively selected set of cases. Experiments demonstrate that this model outperforms the baseline approach, achieving an average balanced accuracy of 73.5% without extensive retraining, thereby enhancing robustness and generalization in clinical settings.

Acknowledgments. E. R. acknowledges the support of a CIFRE Ph.D. grant from ANRT # 2024/0145, and GE HealthCare.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Compliance with Ethical Standards. This research study was conducted retrospectively using anonymized human subject data made available by research partners.

References

1. Alomaim, W., O'Leary, D., Ryan, J., Rainford, L., Evanoff, M., Foley, S.: Variability of breast density classification between US and UK radiologists. Journal of Medical Imaging and Radiation Sciences **50**(1), 53–61 (2019)
2. Berg, W.A., Bandos, A.I., Zuley, M.L., Waheed, U.X.: Training radiologists to interpret contrast-enhanced mammography: toward a standardized lexicon. Journal of Breast Imaging **3**(2), 176–189 (2021)
3. Borkowski, K., Rossi, C., Ciritsis, A., Marcon, M., Hejduk, P., Stieb, S., Boss, A., Berger, N.: Fully automatic classification of breast MRI background parenchymal enhancement using a transfer learning approach. Medicine **99**(29) (2020)
4. Bozzini, A., Nicosia, L., Pruneri, G., Maisonneuve, P., Meneghetti, L., Renne, G., Vingiani, A., Cassano, E., Mastropasqua, M.G.: 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**, 723–731 (2020)
5. D'Orsi, C.J., Sickles, E.A., Mendelson, E.B., Morris, E.A., 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)
6. Eskreis-Winkler, S., Sutton, E.J., D'Alessio, D., Gallagher, K., Saphier, N., Stember, J., Martinez, D.F., Morris, E.A., Pinker, K.: Breast MRI background parenchymal enhancement categorization using deep learning: outperforming the radiologist. Journal of Magnetic Resonance Imaging **56**(4), 1068–1076 (2022)

7. Essam, F., El, H., Ali, S.R.H.: A comparison of the pearson, Spearman rank and Kendall tau correlation coefficients using quantitative variables. Asian Journal of Probability and Statistics pp. 36–48 (2022)
8. Grimm, L.J., Anderson, A.L., Baker, J.A., Johnson, K.S., Walsh, R., Yoon, S.C., Ghate, S.V.: Interobserver variability between breast imagers using the fifth edition of the BI-RADS MRI lexicon. American Journal of Roentgenology **204**(5), 1120–1124 (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Ho, Y., Wookey, S.: The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE Access **8**, 4806–4813 (2019)
11. Hong, C., Ghiassi, A., Zhou, Y., Birke, R., Chen, L.Y.: Online label aggregation: A variational Bayesian approach. In: Proceedings of the Web Conference. pp. 1904–1915 (2021)
12. Karimi, Z., Phillips, J., Slanetz, P., Lotfi, P., Dialani, V., Karimova, J., Mehta, T.: Factors associated with background parenchymal enhancement on contrast-enhanced mammography. American Journal of Roentgenology **216**(2), 340–348 (2021)
13. Li, S., Liu, T., Tan, J., Zeng, D., Ge, S.: Trustable co-label learning from multiple noisy annotators. IEEE Transactions on Multimedia **25**, 1045–1057 (2021)
14. López-Pérez, M., Morales-Álvarez, P., Cooper, L.A., Molina, R., Katsaggelos, A.K.: Deep Gaussian processes for classification with multiple noisy annotators. application to breast cancer tissue classification. IEEE Access **11**, 6922–6934 (2023)
15. Magni, V., Cozzi, A., Muscogiuri, G., Benedek, A., Rossini, G., Fanizza, M., Di Giulio, G., Sardanelli, F.: Background parenchymal enhancement on contrast-enhanced mammography: Associations with breast density and patient's characteristics. La Radiologia Medica **129**(9), 1303–1312 (2024)
16. del Mar Travieso-Aja, M., Naranjo-Santana, P., Fernández-Ruiz, C., Severino-Rondón, W., Maldonado-Saluzzi, D., Rodríguez, M.R., Vega-Benítez, V., Luzardo, O.: Factors affecting the precision of lesion sizing with contrast-enhanced spectral mammography. Clinical Radiology **73**(3), 296–303 (2018)
17. Melsaether, A., McDermott, M., Gupta, D., Pysarenko, K., Shaylor, S.D., Moy, L.: Inter-and intrareader agreement for categorization of background parenchymal enhancement at baseline and after training. American Journal of Roentgenology **203**(1), 209–215 (2014)
18. Nam, Y., Park, G.E., Kang, J., Kim, S.H.: Fully automatic assessment of background parenchymal enhancement on breast MRI using machine-learning models. Journal of Magnetic Resonance Imaging **53**(3), 818–826 (2021)
19. Park, S.Y., Sargent, D., Richmond, D.: Confidence-guided learning for breast density classification. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1623–1627 (2021)
20. Patel, B.K., Lobbes, M., Lewin, J.: Contrast enhanced spectral mammography: a review. In: Seminars in Ultrasound, CT and MRI. vol. 39, pp. 70–79. Elsevier (2018)
21. Portnow, L.H., Choridah, L., Kardinah, K., Handarini, T., Pijnappel, R., Bluekens, A.M., Duijm, L.E., Schoub, P.K., Smilg, P.S., Malek, L., et al.: International inter-observer variability of breast density assessment. Journal of the American College of Radiology **20**(7), 671–684 (2023)

22. Portnow, L.H., Georgian-Smith, D., Haider, I., Barrios, M., Bay, C.P., Nelson, K.P., Raza, S.: Persistent inter-observer variability of breast density assessment using bi-rads® 5th edition guidelines. Clinical Imaging **83**, 21–27 (2022)
23. Ripaud, E., Jailin, C., Quintana, G., de Carvalho, P.M., de la Rosa, R.S., Vancamberg, L.: Deep-learning model for background parenchymal enhancement classification in contrast-enhanced mammography. Physics in Medicine & Biology **69**(11), 115013 (2024)
24. Ripaud, E., Jailin, C., Quintana, G.I., de Carvalho, P.M., Mohamed, S., Moustafa, A.F., Kamal, R.M., Vancamberg, L.: Deep-learning-based background parenchymal enhancement quantification in contrast enhanced mammography: an application to neoadjuvant chemotherapy. In: 17th International Workshop on Breast Imaging (IWBI 2024). vol. 13174, pp. 471–477. SPIE (2024)
25. Sorin, V., Yagil, Y., Shalmon, A., Gotlieb, M., Faermann, R., Halshtok-Neiman, O., Sklair-Levy, M.: Background parenchymal enhancement at contrast-enhanced spectral mammography (CESM) as a breast cancer risk factor. Academic Radiology **27**(9), 1234–1240 (2020)
26. Sperrin, M., Bardwell, L., Sergeant, J.C., Astley, S., Buchan, I.: Correcting for rater bias in scores on a continuous scale, with application to breast density. Statistics in Medicine **32**(26), 4666–4678 (2013)
27. Squires, S., Harkness, E.F., Evans, D.G., Astley, S.M.: The effect of variable labels on deep learning models trained to predict breast density. Biomedical Physics & Engineering Express **9**(3), 035030 (2023)
28. Viera, A.J., Garrett, J.M., et al.: Understanding interobserver agreement: the kappa statistic. Family Medicine **37**(5), 360–363 (2005)
29. Yan, Y., Rosales, R., Fung, G., Subramanian, R., Dy, J.: Learning from multiple annotators with varying expertise. Machine learning **95**, 291–327 (2014)
30. Zhan, X., Wang, Q., Huang, K.h., Xiong, H., Dou, D., Chan, A.B.: A comparative survey of deep active learning. arXiv preprint arXiv:2203.13450 (2022)