

Fully automatic classification of breast MRI background parenchymal enhancement using a transfer learning approach

Karol Borkowski, PhD^{*} , Cristina Rossi, PhD , Alexander Ciritsis, PhD , Magda Marcon, PhD, Patryk Hejduk, MSc, Sonja Stieb, MD, Andreas Boss, PD, Nicole Berger, MD 

Abstract

Marked enhancement of the fibroglandular tissue on contrast-enhanced breast magnetic resonance imaging (MRI) may affect lesion detection and classification and is suggested to be associated with higher risk of developing breast cancer. The background parenchymal enhancement (BPE) is qualitatively classified according to the BI-RADS atlas into the categories “minimal,” “mild,” “moderate,” and “marked.” The purpose of this study was to train a deep convolutional neural network (dCNN) for standardized and automatic classification of BPE categories.

This IRB-approved retrospective study included 11,769 single MR images from 149 patients. The MR images were derived from the subtraction between the first post-contrast volume and the native T1-weighted images. A hierarchic approach was implemented relying on 2 dCNN models for detection of MR-slices imaging breast tissue and for BPE classification, respectively. Data annotation was performed by 2 board-certified radiologists. The consensus of the 2 radiologists was chosen as reference for BPE classification. The clinical performances of the single readers and of the dCNN were statistically compared using the quadratic Cohen’s kappa.

Slices depicting the breast were classified with training, validation, and real-world (test) accuracies of 98%, 96%, and 97%, respectively. Over the 4 classes, the BPE classification was reached with mean accuracies of 74% for training, 75% for the validation, and 75% for the real word dataset. As compared to the reference, the inter-reader reliabilities for the radiologists were 0.780 (reader 1) and 0.679 (reader 2). On the other hand, the reliability for the dCNN model was 0.815.

Automatic classification of BPE can be performed with high accuracy and support the standardization of tissue classification in MRI.

Abbreviations: BPE = background parenchymal enhancement, CAM = class activation map, dCNN = deep convolutional neural network, FGT = fibroglandular tissue, Grad-CAM = Gradient-weighted Class Activation Mapping, MRI = magnetic resonance imaging.

Keywords: accuracy, AI, BPE, breast parenchymal enhancement, dCNN, deep learning, MRI

Editor: YX Sun.

KB and CR equally contributed to the work.

This work was supported by the Clinical Research Priority Program “AI in oncological imaging.”

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, University of Zurich, Switzerland.

* Correspondence: Karol Borkowski, Institute of Diagnostic and Interventional Radiology, University Hospital Zürich, Rämistr. 100, 8091 Zürich, Switzerland (e-mail: karol.borkowski@usz.ch).

Copyright © 2020 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial License 4.0 (CCBY-NC), where it is permissible to download, share, remix, transform, and buildup the work provided it is properly cited. The work cannot be used commercially without permission from the journal.

How to cite this article: 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* 2020;99:29(e21243).

Received: 6 February 2020 / Received in final form: 4 June 2020 / Accepted: 11 June 2020

<http://dx.doi.org/10.1097/MD.0000000000021243>

1. Introduction

Magnetic resonance imaging (MRI) is an established technique for breast imaging and it is used for evaluation of the breast tissue in high-risk patients, pre-operative staging, monitoring of chemotherapy effect, evaluation of women with breast implants, or occult primary breast cancer.^[1]

After administration of the contrast agent, both, lesions and normal fibroglandular tissue (FGT), may enhance.^[2] In some subjects, the enhancement of the FGT, that is, the background parenchymal enhancement (BPE), may present asymmetric and non-diffusive distribution, as well as a suspicious dynamic response. In those cases, the BPE can affect the diagnostic accuracy of the lesion Breast Imaging-Reporting and Data System of the American College of Radiology (ACR BI-RADS) classification.^[3–5]

Not only technical factors (e.g., concentration of the contrast agent, T1-weighted contrast of the sequence)^[6] but also the vascular mammary anatomy and the hormonal status are known to affect the BPE levels.^[7,8] In young patients and patients undergoing hormonal therapy, BPE is more markedly expressed than in other patients.^[9] In order to account for the monthly hormonal changes of the breast, breast MRI is preferably performed during the 7th to 14th day of the menstrual

cycle.^[10,11] Moreover, to achieve a better standardization of the BPE classification, radiologists are requested to rate the BPE according to the BI-RADS classification^[12] as minimal, mild, moderate, or marked. However, the visual rating of the BPE is prone to be reader-dependent; in a study from 2015, Grimm et al reported a fair mean inter-observer reliability in the BPE classification (Cohen's kappa, $k=0.28$).^[13]

Besides the relevance in the definition of the diagnostic accuracy of breast MRI, few studies have claimed the association between BPE and breast cancer risk.^[14–16]

To overcome the problem of the human variability in the classification of the BPE, automatic or semi-automatic methods have been proposed. The reported technical solutions propose a volumetric or quantitative computation of the FGT enhancement.^[17–19] Although those methods aim at an objective evaluation of the BPE, the association between quantitative parenchymal enhancement (QPE) and the BPE is only fair, due to the lacking possibility of accounting for the intensity of the enhancement or for the presence of spotted BPE patterns.^[17]

Similarly, to the case of the BPE, also the human visual classification of the mammographic breast density is reader-dependent. In the case of the mammographic breast density, deep learning has shown to provide clinically valuable classification by relying on image pattern recognition.^[20–22]

In this study, we propose the use of a deep convolutional neural network (dCNN) for the classification of the BPE in MRI. Performances of the algorithm were clinically validated by comparing the classification of the algorithm on a real-world data set with the consensus of 2 board-certified radiologists.

The breast MRI measurement is usually performed with some margin, so before the BPE classification, the slices containing breasts must be selected. To this end, we propose an auxiliary model that recognizes slices depicting the breast.

2. Materials and methods

2.1. Study design and population

This retrospective study was approved by the local Ethics Committee. All patients undergoing a breast MRI at our institution from September 2013 to October 2015 were considered for analysis. A total number of 149 patients was included. The mean patient age \pm standard deviation was 49 ± 6 years. Each patient was examined once.

2.2. Image acquisition

All breast MRI examinations were performed with the patient in prone position using a 3-T unit (MAGNETOM Skyra, Siemens Medical Solution, Erlangen, Germany) and a dedicated 4-channel breast coil. For each patient, the imaging protocol included an axial T2-weighted short-tau inversion recovery sequence (TR 5600 ms or 8970 ms, TE 70 ms, inversion time 150 ms, flip angle 150°, voxel 1.3 mm \times 0.6 mm \times 0.3 mm or 0.7 mm \times 0.7 mm \times 2 mm) and an axial diffusion-weighted sequence (TR/TE 4300/89 ms, voxel 2.7 mm \times 2.7 mm \times 4 mm, b -values 0, 500, 1000 s/mm²) before contrast agent injection. Thereafter, a dynamic protocol consisting in the acquisition of a T1-weighted gradient-echo three-dimensional fast low-angle shot sequences (TR/TE 11/4.89 ms, voxel 0.8 mm \times 0.8 mm \times 1.3 mm) before and after contrast agent administration (0, 1, 2, 3, 4, and 5 minutes)

was acquired. The dose of the contrast agent was adapted to the weight of the patient (0.1 mmol/kg).

2.3. Dataset preparation

The retrieved dataset consisting in 149 studies and 11,769 MR images was used for training 2 models: the breast detection model for the recognition of slices depicting the breast, and the BPE model for the BPE classification. The dataset contained 1169 slices depicting breast with implants and 699 without depicted breast. For the breast detection model, the whole retrieved dataset was split into 3 categories: “breast,” “no-breast,” and “implants.” Each category was randomly split into training, validation, and test set at a ratio of 70%, 20%, and 10%, respectively.

For the BPE model, 9902 single-slice MR images from 124 patients without breast implants were selected and annotated in terms of BPE according to the BI-RADS atlas (3613 as minimal, 4282 as mild, 1556 as moderate, and 451 as marked). To balance the number of images belonging to each BPE category, the data were augmented by a random shifting and zooming in the range of $\pm 5\%$ and by horizontal flipping. The catalog structure containing the data sets is presented in Figure 1. The images were preprocessed before feeding them into the neural networks by cutting-off the one-third of their bottom part, which does not contain breast, and by normalization the values to 0 to 1 range.

For the of BPE model, subsets of 87, 25, and 12 patients were randomly selected for the training-, validation-, and test-partitions, respectively. For the test partition, a subset of 100 images (25 for each BPE category) was chosen for the evaluation. The set partitions had been selected before the model was trained.

The images used for the training of the BPE model were annotated in consensus by 2 radiologists with more than 5 years experience in breast imaging. In this study, this assessment is regarded as the gold standard of the BPE class assessment and will be referred to as “reference.” BPE scores were assigned slice-wise based on the image volume resulting from the subtraction of the native fat-suppressed T1-weighted images from the first post-contrast volume. In the case of BPE asymmetry between the left and right side, the higher level of BPE was assigned.

2.4. Model architecture and training

Both models were implemented by means of a deep convolutional neural network. The network consisted of 2 densely connected layers on top of the convolutional part of the VGG16^[23] network trained on the ImageNet dataset, which has been already successfully applied for the assessment of medical breast images.^[24]

Both models were trained on the NVIDIA GeForce GTX1080 graphical processing unit for 100 and 120 epochs, respectively, using the Adam optimizer.^[25] To avoid overfitting, the training process was stopped as soon as the loss function calculated for the validation set had raised or the difference between the accuracy for the training and validation set had exceeded 3 percent points. Moreover, the model was saved after an epoch, after which the validation accuracy was the highest.

For the breast detection model the categorical cross-entropy loss function was applied. For the BPE model, a custom loss function was applied in order to take advantage of the graduating categories ($A < B < C < D$). To this end, the cross-entropy value

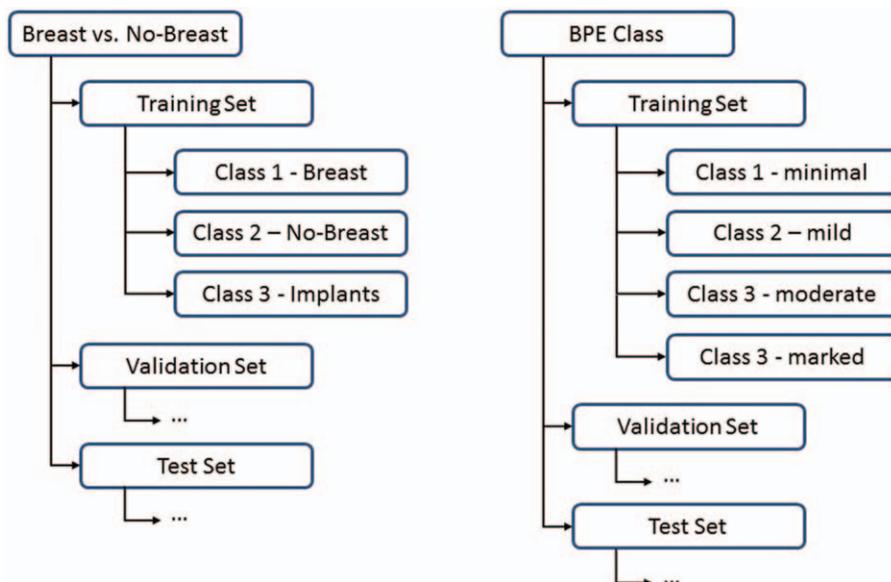


Figure 1. The catalog structure corresponding to the breast detection model (left) and the BPE model (right).

for each sample was multiplied by the value of the mean square error.

Nevertheless, the cross-entropy do not take advantage of the gradation of the BPE categories ($A < B < C < D$). To take this fact into account this fact, a custom loss function was applied. Namely, the value of the cross-entropy loss for each sample was multiplied by the value of the mean square error.

2.5. Statistical and clinical validation

For each model, the performances over the validation dataset were quantified in terms of the metrics of the confusion matrix. The performances of each model over the real-world dataset were compared with the reference. Performances were expressed in terms of accuracy, precision, recall, and F1-score. For each model, the confusion matrix was computed.

The output of the algorithm on the real-world data was also used for the clinical validation. In this case, each experienced radiologist was requested to perform the classification on the real-world dataset. The radiologists were blind to the results of the previously performed consensus classification and of the algorithm classification.

Based on the 3 classifications over the same dataset and on the consensus decision taken as a gold standard, the inter-rater reliability was assessed by means of the quadratic Cohen’s Kappa coefficient (κ).^[26]

3. Results

3.1. Statistical validation

The statistical validation of the breast detection and of the BPE models is reported in Tables 1 and 2, respectively. The corresponding learning curves are presented in Figure 2. In the case of the breast model, the training was stopped after 100 epochs, when the accuracy for the validation set reached the plateau of 97.5%. For the BPE model, the training was stopped

after 150 epochs. The highest accuracy for the validation set was achieved after the 67-th epoch, so the state of the model at that stage was used.

After the training, both models were validated on the real-world datasets. The accuracy, precision, recall, and F1-score for each class of the breast and the BPE models are reported in Tables 1 and 2, respectively. In the case of the breast model, the overall accuracy was equal to 96%. Only 1 image that presented breast has been erroneously classified to the “no-breast” category and 2 images without breasts have been classified as depicting the breast. The best performance has been achieved in the case of the breast with implants. All the images from the test set that presented implants have been correctly recognized and none has been erroneously assigned to this group. The corresponding confusion matrix is presented in Figure 3 in the form of a heat-map.

In the case of the BPE model, the overall accuracy was equal to 75%. Almost all misclassifications occurred only between adjacent classes, for example, mild and moderate. The confusion matrix corresponding to this model is presented in Figure 4.

In most cases, the confidence of the assignment to a particular class was greater than 99% in the case of the breast model and significantly greater than 90% in the case of the BPE model.

For the BPE model, the T1-weighted native image of 1 representative subject was superimposed onto the class activation map (CAM) implemented using the Gradient-weighted Class Activation Mapping (Grad-CAM) approach^[27] (Fig. 5). The

Table 1 Accuracy, precision, recall, and the F1-score of the breast detection model evaluated on the real-world data.

	Accuracy	Precision	Recall	F1-score
Breast	0.96	0.92	0.96	0.94
No-breast	0.92	0.96	0.92	0.94
Implants	1.0	1.0	1.00	1.00

Table 2

Accuracy, precision, recall, and the F1-score of the BPE (background parenchymal enhancement) class model evaluated on the real-world data.

	Accuracy	Precision	Recall	F1-score
Minimal	0.84	0.84	0.84	0.84
Mild	0.80	0.59	0.80	0.68
Moderate	0.48	0.71	0.48	0.57
Marked	0.88	0.92	0.88	0.90

CAM map indicates the regions on which the prediction has been based on.

3.2. Clinical validation

Reader 1 assessed 69 real-world images in accordance with the reference, while reader 2 correctly assessed 52 images. Therefore,

the accuracy of the human experts for the dataset of 100 images was 69% and 52%, respectively.

Figure 6 presents the inter-rater reliability expressed by the Cohen's Kappa for the predictions of the BPE model, assessments done independently by 2 expert human readers (reader 1 and reader 2) and the reference in every possible combination. The lowest value of the kappa (0.679 ± 0.18) was obtained for the reliability of reader 1 and the reference, while the highest one (0.815 ± 0.13) was obtained for the model and the reference. The statistics of the assessments done by each human expert and the model are reported in Table 3. The “ \pm ” sign indicates the confidence interval.

Figure 7 shows the same results presented in terms of each BPE class separately, where the four-class classification problem was translated to four one vs all classifications. The presented kappa coefficients were obtained with regard to the reference. The reliability of both readers was different for different BPE classes and ranged between 0.47 and 0.71 for the first and 0.24 and 0.49

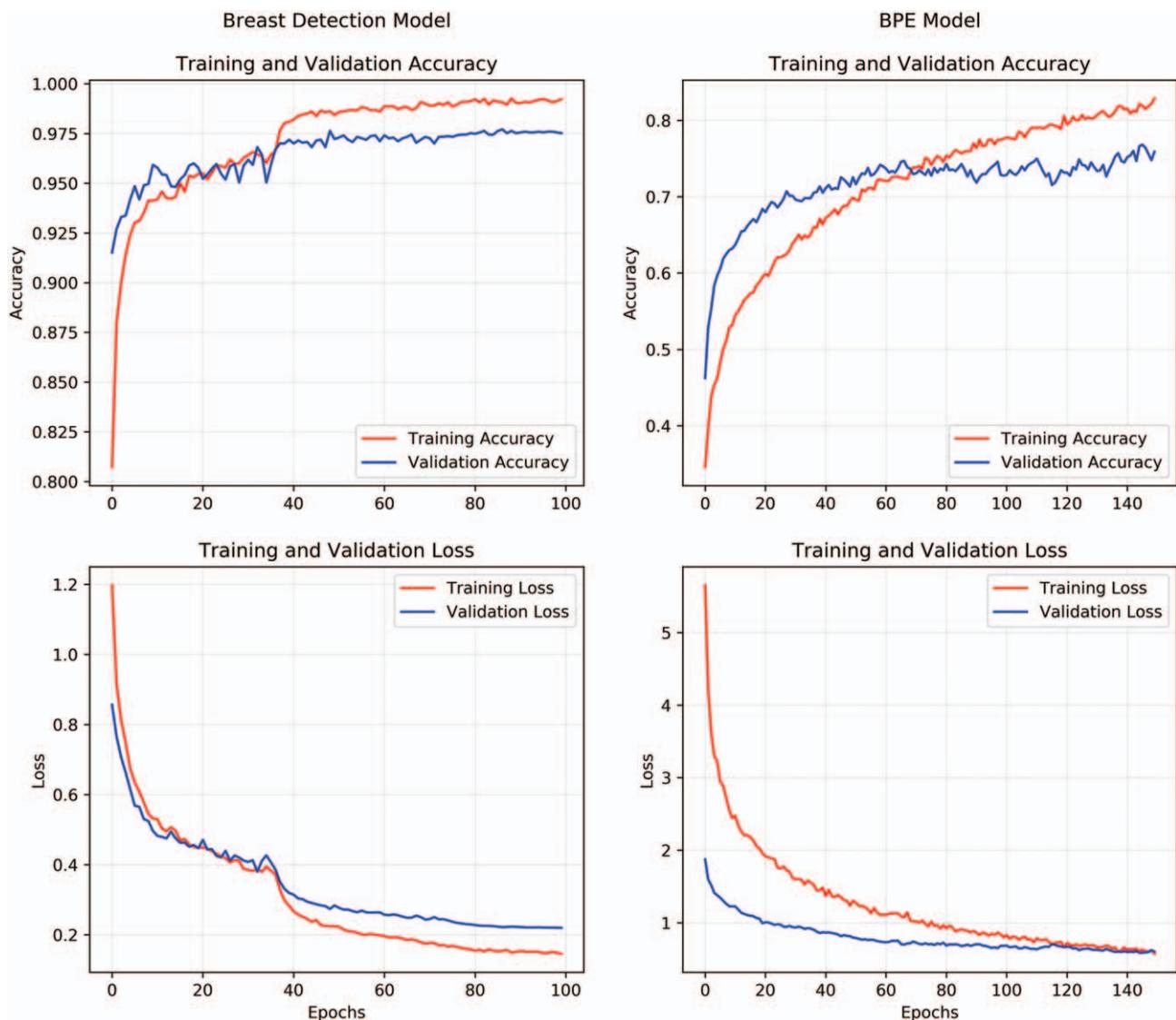


Figure 2. The loss function (bottom) and accuracy (top) plots for the training (red) and validation (blue) set depicting the learning process of the breast detection model (left) and BPE model (right).



Figure 3. A confusion matrix for the validation of the breast detection model using the real-world dataset.

for the second reader. The reliability of the second reader was lower than of the first one in the case of all BPE classes except the minimal enhancement. The reliability of the model was higher than of the both readers in the case of all classes except the moderate enhancement.

4. Discussion

In this study, we implemented a fully automatic approach for the classification of BPE categories according to the Breast Imaging Reporting and Data System atlas and validated its clinical use by comparing the performances of the algorithm with those of consenting expert human readers.

The approach relies on the use of a transfer learning method. To obtain a slice-wise classification of the BPE, a hierarchical approach was implemented, which consisted in the use of 2 computational models: the first intended to detect slices imaging the breast and the second performing the actual BPE classification. The rationale behind the study is that the use of a dCNN algorithm trained on thousands of data labeled by consenting

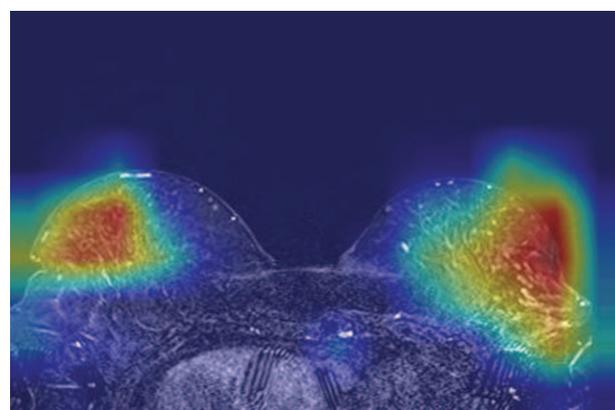


Figure 5. An exemplary input image with superimposed the class activation map. The dark red regions correspond to the area that highly contributed to the final model prediction.

radiologists expert in breast imaging may allow a standardized BPE classification. Although the problem of the automatic assessment of BPE has been already addressed by Ha et al,^[28] so far quantitative assessment of BPE based on tissue segmentation has been proposed. To the best of our knowledge, the deep learning approach able to mimic the human evaluation of the whole image pattern for qualitative BPE classification has never been published before.

The breast detection and the BPE models were trained for 150 and 120 epochs, what allowed achieving the accuracy for the validation set equal to 97% and 90%, respectively. The accuracies obtained for the real-world data were similar, what indicates that the models are not over-fitted. The learning curves for the breast detection model (Fig. 2) achieve the first plateau after about 35 epochs. Nevertheless, the reduction of the learning rate during the training enabled to increase the accuracy by additional 2%, yielding a characteristic inflection of the accuracy and loss curves. In the case of the BPE class model, the accuracy plateau was not reached, so the dropout of the learning rate was

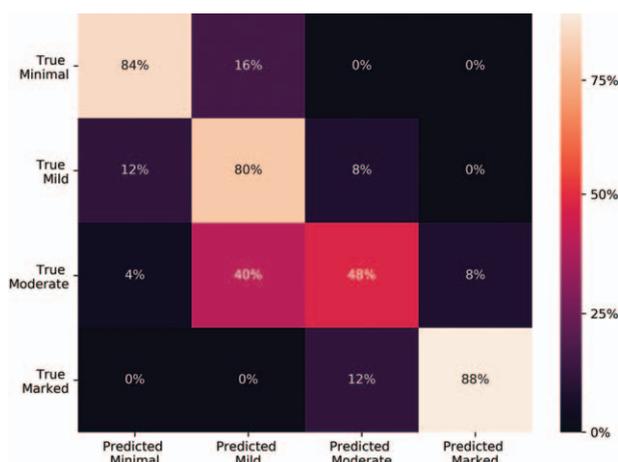


Figure 4. A confusion matrix for the validation of the BPE class model using the real-world dataset.

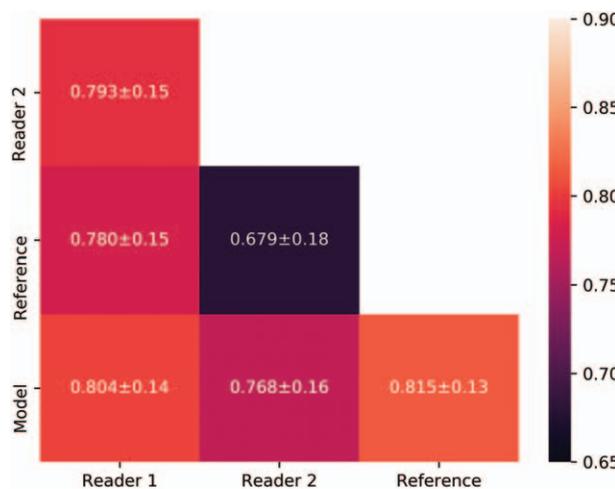


Figure 6. The values of the Cohen's kappa calculated for the predictions of the model, answer of both human readers and the consensus decision in each possible combination.

BPE class	Minimal	Mild	Moderate	Marked
Reader 1	29%	34%	29%	17%
Reader 2	26%	33%	17%	24%
Model	26%	21%	26%	27%

not applied. The model was saved after the 67-th epoch, when it achieved the highest validation accuracy. As a comparison, the accuracy of human readers was 69% and 52%. Such discrepancies between experienced radiologists confirm the need of the standardization.

As shown in Figure 6, almost all misclassifications done by the BPE model occurred only for adjacent categories. Since the BPE classification guidelines are subject to a human interpretation, there is no ground truth behind a given image and this kind of disagreement is common between different radiologists and even between 2 assessments of the same specialist.^[29] Therefore, a more relevant way to assess the model is to compare the inter-rater reliability expressed by the Cohen's kappa coefficient. The kappa coefficient obtained for the agreement between both experts and each expert with the model 0.793 ± 0.15 , 0.804 ± 0.14 , and 0.768 ± 0.16 , respectively. These values are consistent with values assessed in other studies, which ranged between 0.73 and 0.93,^[29–31] and confirm that expert readers achieve an almost perfect agreement with the consenting classification. As compared to the reference, the inter-rater reliability of the model is higher (0.815) than that of the experienced radiologists (0.78 and 0.679). These findings suggest that deep convolutional neural

networks are a reliable and standardized tool in the assessment of the background parenchymal enhancement in MRI. The class activation map presented in Figure 5 shows that the BPE model classifies the images based on the image region that contain the most important information and ignores the background, what confirms the above conclusion.

The possible source of the bias in the assessment of the BPE class is that the different BPE classes are not equally common. The lower enhancement classes occur more often than the higher ones, as can be seen from the review of 650 breast MRI examinations described by Abramovici and Mainiero^[32] and in other studies.^[30,32] This fact is reflected in the statistics of the radiologists' answers, as reported in Table 3. The readers classified the images to the lower classes more frequently, even though the BPE classes in the test dataset were equally represented. Since the neural network model has been trained on a balanced dataset, it is free from this kind of bias.

The main limitation of our study is a limited number of studies, what has been mitigated by application of the transfer learning and data augmentation; possible bias of the human experts in the BPE class assignment, what to some extent, has been mitigated by taking the consensus decision as the reference; and finally, the fact that all studies were performed in 1 institution using the same MRI scanner. Validation of the model using images from other institutions is proposed for the future study. Another limitation is the relatively small size of the real-world dataset. This representative set was a trade-off between robust statistics and the limited reading-time of human experts. However, in the case of a balanced class distribution, the potential bias is expected to be less severe.^[33]

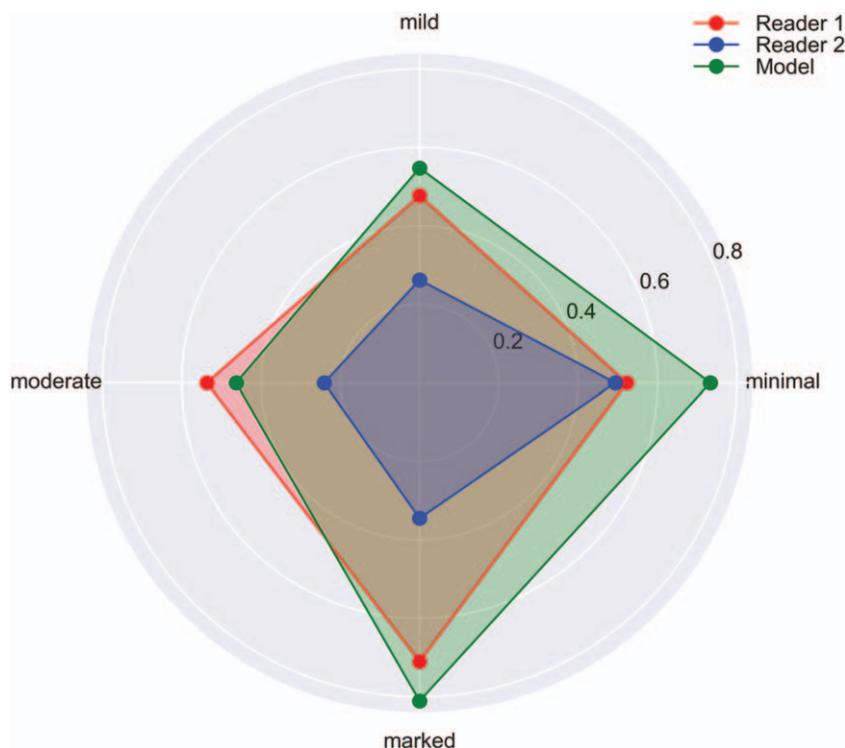


Figure 7. The values of the Cohen's kappa for both radiologists (blue and red) and the model (green) calculated for each BPE class separately. The results were with regard to the reference.

5. Conclusion

In conclusion, the MRI breast images can be effectively classified according to their background parenchyma enhancement, by means of a deep convolutional neural network. The neural network is at least as accurate as an experienced radiologist. Moreover, their predictions are standardized and not influenced by the effect of the intra-reader discrepancy. The convolutional part of the VGG16 network can serve as an effective feature extractor for breast MRI, even though it was not trained on medical images.

Author contributions

Conceptualization: Karol Borkowski, Cristina Rossi, Nicole Berger.

Investigation: Karol Borkowski, Cristina Rossi, Alexander Ciritsis, Magda Marcon, Patryk Hejduk, MSc, Sonja Stieb, Andreas Boss.

Methodology: Karol Borkowski, Alexander Ciritsis, Cristina Rossi.

Project administration: Andreas Boss, Nicole Berger.

Validation: Karol Borkowski, Cristina Rossi, Magda Marcon, Nicole Berger.

Writing – original draft: Karol Borkowski.

Writing – review and editing: Karol Borkowski, Cristina Rossi, Nicole Berger.

References

- Mann RM, Balleyguier C, Baltzer PA, et al. Breast MRI: EUSOBI recommendations for women's information. *Eur Radiol* 2015;25:3669–78.
- Morris EA. Diagnostic breast MR imaging: current status and future directions. *Radiol Clin N Am* 2007;45:863–80. vii.
- Teifke A, Hlawatsch A, Beier T, et al. Undetected malignancies of the breast: dynamic contrast-enhanced MR imaging at 1.0T. *Radiology* 2002;224:881–8.
- Shimauchi A, Jansen SA, Abe H, et al. Breast cancers not detected at MRI: review of false-negative lesions. *Am J Roentgenol* 2010;194:1674–9.
- Uematsu T, Kasami M, Watanabe J. Does the degree of background enhancement in breast MRI affect the detection and staging of breast cancer? *Eur Radiol* 2011;21:2261–7.
- Kuhl C. The current status of breast MR imaging. Part I. Choice of technique, image interpretation, diagnostic accuracy, and transfer to clinical practice. *Radiology* 2007;244:356–78.
- Müller-Schimpfle M, Ohmenhäuser K, Stoll P, et al. Menstrual cycle and age: influence on parenchymal contrast medium enhancement in MR imaging of the breast. *Radiology* 1997;203:145–9.
- Delille J-P, Slanetz PJ, Yeh ED, et al. Physiologic changes in breast magnetic resonance imaging during the menstrual cycle: perfusion imaging, signal enhancement, and influence of the T1 relaxation time of breast tissue. *Breast J* 2005;11:236–41.
- Dontchos BN, Rahbar H, Partridge SC, et al. Influence of menstrual cycle timing on screening breast MRI background parenchymal enhancement and diagnostic performance in premenopausal women. *J Breast Imaging* 2019;1:205–11.
- Kang SS, Ko EY, Han B-K, et al. Background parenchymal enhancement on breast MRI: influence of menstrual cycle and breast composition. *J Magn Reson Imaging* 2014;39:526–34.
- Kuhl CK, Bieling HB, Gieseke J, et al. Healthy premenopausal breast parenchyma in dynamic contrast-enhanced MR imaging of the breast: normal contrast medium enhancement and cyclical-phase dependency. *Radiology* 1997;203:137–44.
- Sickles EA, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS[®] Mammography. In: ACR BI-RADS[®] Atlas, Breast Imaging Reporting and Data System. Reston, VA, American College of Radiology; 2013.
- Grimm LJ, Anderson AL, Baker JA, et al. Interobserver variability between breast imagers using the fifth edition of the BI-RADS MRI Lexicon. *Am J Roentgenol* 2015;204:1120–4.
- King V, Brooks JD, Bernstein JL, et al. Background parenchymal enhancement at breast MR imaging and breast cancer risk. *Radiology* 2011;260:50–60.
- Hattangadi J, Park C, Rembert J, et al. Breast stromal enhancement on MRI is associated with response to neoadjuvant chemotherapy. *Am J Roentgenol* 2008;190:1630–6.
- Dontchos BN, Rahbar H, Partridge SC, et al. Are qualitative assessments of background parenchymal enhancement, amount of fibroglandular tissue on MR images, and mammographic density associated with breast cancer risk? *Radiology* 2015;276:371–80.
- Pujara AC, Mikheev A, Rusinek H, et al. Comparison between qualitative and quantitative assessment of background parenchymal enhancement on breast MRI. *J Magn Reson Imaging* 2018;47:1685–91.
- Jiang L, Hu X, Xiao Q, et al. Fully automated segmentation of whole breast using dynamic programming in dynamic contrast enhanced MR images. *Med Phys* 2017;44:2400–14.
- Xu X, Fu L, Chen Y, et al. Breast region segmentation using convolutional neural network in dynamic contrast enhanced MRI. Conference proceedings: . . . Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference 2018;750–3.
- Yi PH, Lin A, Wei J, et al. Deep-learning-based semantic labeling for 2D mammography and comparison of complexity for machine learning tasks. *J Digit Imaging* 2019;32:565–70.
- Ciritsis A, Rossi C, Vittoria De Martini I, et al. Determination of mammographic breast density using a deep convolutional neural network. *Br J Radiol* 2018;92:20180691.
- Li S, Wei J, Chan H-P, et al. Computer-aided assessment of breast density: comparison of supervised deep learning and feature-based statistical learning. *Phys Med Biol* 2018;63:025005.
- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *CoRR* 2014;abs/1409.1556.
- Fick DA. V. On liquid diffusion. *Lond Edinb Phil Mag J Sci* 1855;10:30–9.
- Diederik P, Kingma JLB. ADAM: a method for stochastic optimization. *Int Conf Learn Represent* 2015.
- Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–20.
- Silvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV) 2017 IEEE International Conference on Computer Vision (ICCV), 2017/10/2017 2017;618–26.
- Ha R, Chang P, Mema E, et al. Fully automated convolutional neural network method for quantification of breast mri fibroglandular tissue and background parenchymal enhancement. *J Digit Imaging* 2019; 32:141–7.
- Dilorenzo G, Telegrafo M, La Forgia D, et al. Breast MRI background parenchymal enhancement as an imaging bridge to molecular cancer subtype. *Eur J Radiol* 2019;113:148–52.
- Brooks JD, Sung JS, Pike MC, et al. MRI background parenchymal enhancement, breast density and serum hormones in postmenopausal women. *Int J Cancer* 2018;143:823–30.
- Preibsch H, Wanner L, Bahrs SD, et al. Background parenchymal enhancement in breast MRI before and after neoadjuvant chemotherapy: correlation with tumor response. *Eur Radiol* 2016;26:1590–6.
- Abramovici G, Maimieri MB. Screening breast MR imaging: comparison of interpretation of baseline and annual follow-up studies. *Radiology* 2011;259:85–91.
- Provost F, Fawcett T. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining Newport Beach, CA: AAAI Press; 1997. 43–48.