Diagnostic and Interventional Imaging (2019) **100**, 219—225

ORIGINAL ARTICLE

Detection and characterization of MRI breast lesions using deep learning

P. Herent[a], **B. Schmauch**[a], **P. Jehanno**[a,*], **O. Dehaene**[b], **C. Saillard**[a], **C. Balleyguier**[c], **J. Arfi-Rouche**[c], **S. Jégou**[a]

[a] *Owkin Inc, Research and Development Laboratory, 75003 Paris, France*
[b] *École Centrale d'Electronique (ECE), 75015 Paris, France*
[c] *Radiology Department, Institut Gustave-Roussy, 94805 Villejuif, France*

KEYWORDS
Magnetic resonance imaging (MRI);
Breast lesion detection;
Convolution neural networks;
Transfer learning;
Attention model

Abstract
Purpose: The purpose of this study was to assess the potential of a deep learning model to discriminate between benign and malignant breast lesions using magnetic resonance imaging (MRI) and characterize different histological subtypes of breast lesions.
Materials and methods: We developed a deep learning model that simultaneously learns to detect lesions and characterize them. We created a lesion-characterization model based on a single two-dimensional T1-weighted fat suppressed MR image obtained after intravenous administration of a gadolinium chelate selected by radiologists. The data included 335 MR images from 335 patients, representing 17 different histological subtypes of breast lesions grouped into four categories (mammary gland, benign lesions, invasive ductal carcinoma and other malignant lesions). Algorithm performance was evaluated on an independent test set of 168 MR images using weighted sums of the area under the curve (AUC) scores.
Results: We obtained a cross-validation score of 0.817 weighted average receiver operating characteristic (ROC)-AUC on the training set computed as the mean of three-shuffle three-fold cross-validation. Our model reached a weighted mean AUC of 0.816 on the independent challenge test set.
Conclusion: This study shows good performance of a supervised-attention model with deep learning for breast MRI. This method should be validated on a larger and independent cohort.

As the number of radiological examinations steadily increases, so does the complexity of their interpretation and demands on providers [1]. Radiologists are exposed to decision fatigue, which can lead to a high frequency of medical

* Corresponding author.
E-mail address: paul.jehanno@owkin.com (P. Jehanno).

https://doi.org/10.1016/j.diii.2019.02.008

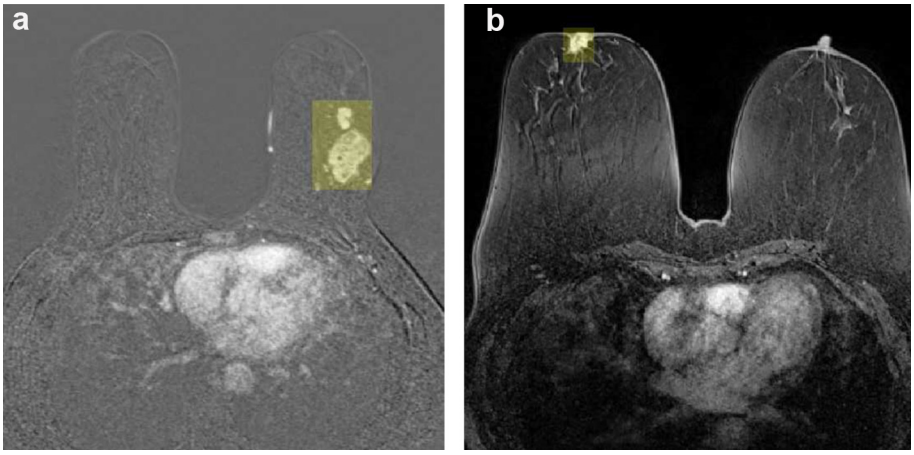

Figure 1. T1-weighted MR images in the axial plane obtained after intravenous administration of a gadolinium chelate. A. Binary mask is superimposed over the MR image that shows invasive ductal carcinoma. The lesion shows heterogeneous enhancement. B. MR image shows one invasive ductal carcinoma.

errors including missed, incorrect, or delayed diagnoses [2]. In addition, interpretations by radiologists are prone to high intra- and inter-individual variability [3,4].

Today, breast magnetic resonance imaging (MRI) is used for many indications in the management of breast cancer. The main indications of MRI include screening of patients with a high risk of developing breast cancer [5], determining the extent of disease, assessing positive margins, monitoring the response to neo-adjuvant chemotherapy, evaluation of metastatic axillary lymphadenopathy of unknown primary [6,7]. Breast MRI is a multiparametric examination. In this regard, a classical protocol includes several sequences including dynamic T1-weighted gradient echo images obtained before and after intravenous administration of a gadolinium chelate, T2-weighted images or short-tau inversion-recovery (STIR) images, and diffusion-weighted (DW) images [8].

Deep learning is a subtype of machine learning that uses layers of artificial neurons, called neural networks [9], and has demonstrated superior performance compared to standard computer vision algorithms [10]. Deep learning in radiology has the potential to substantially alter each step of the medical imaging pipeline such as image reconstruction [11], image segmentation, and final interpretation [12,13]. Most studies on breast imaging and deep learning have focused on mammography [14,15]. Less evidence is available concerning breast MRI. However, a comparison of human performance to that of radiomics algorithms and convolutional neural network (CNN) in breast lesion characterization with MRI showed that interpretation by a radiologist was better (AUC = 0.98) than CNN (AUC = 0.88) or radiomics (AUC = 0.81) [16].

Managing artificial intelligence projects for clinical practice requires a combination of expertise between data scientists and radiologists. Here, we present an innovative tool for imaging interpretation of breast MRI examinations that could increase safety and reliability in the near future. This study was the result of a collaboration made possible by a data challenge organized during the *Journées Francophones de Radiologie* in Paris in October 2018.

The purpose of this study was to assess the potential of a deep learning model to discriminate between benign and malignant breast lesions using MRI, and characterize different histological subtypes of breast lesions.

Materials and methods

Preprocessing

The dataset consisted of anonymized two-dimensional T1-weighted gadolinium chelate-enhanced MR images of the breast provided during the *Journées Francophones de Radiologie* 2018. Despite the standardization already performed by the challenge organizers, the data were highly heterogeneous in scale (Fig. 1). All images were resized to 240×345 to have the same image size for all images.

Automatic feature extraction

To extract features from images, we used a 50-layer residual neural network (ResNet-50) [17], pretrained on the ImageNet dataset, from which we removed the last two layers. This network was designed for color images, thus each grayscale image was copied three times to simulate the red, green, and blue channels. For an input image of $3 \times 240 \times 345$, the network produced a feature map with a dimension of $2048 \times 8 \times 11$. A first simple approach consisted of averaging this representation over the spatial dimensions, as shown in Eq. (1):

$$x_k = \frac{1}{8 \times 11} \sum_{i,j} x_{kij} \quad (1)$$

This technique yielded a feature vector size of 2048 for each image, which was fed to a single densely connected layer with five neurons for each classification task (malignancy, normal tissue, other benign lesions, invasive ductal carcinoma (IDC), and other malignant lesions). The main drawback of this approach was that it does not differentiate regions of little interest, such as the thorax or background.

Supervised attention mechanism

One challenge of this task was the heterogeneity in the appearance and size of breast lesions. We facilitated learning by decomposing classification into two steps: (i), detection of abnormalities present on MR images and (ii), classification of these lesions.

These two steps were simultaneously performed by two branches of the same model. For the first, we created and used additional labels for localization. These labels consisted of bounding boxes surrounding the lesions. These annotations did not require precise characterization. They were rapidly performed by a 5th year resident in radiology (P.H.), who had limited experience in breast MRI (Fig. 2).

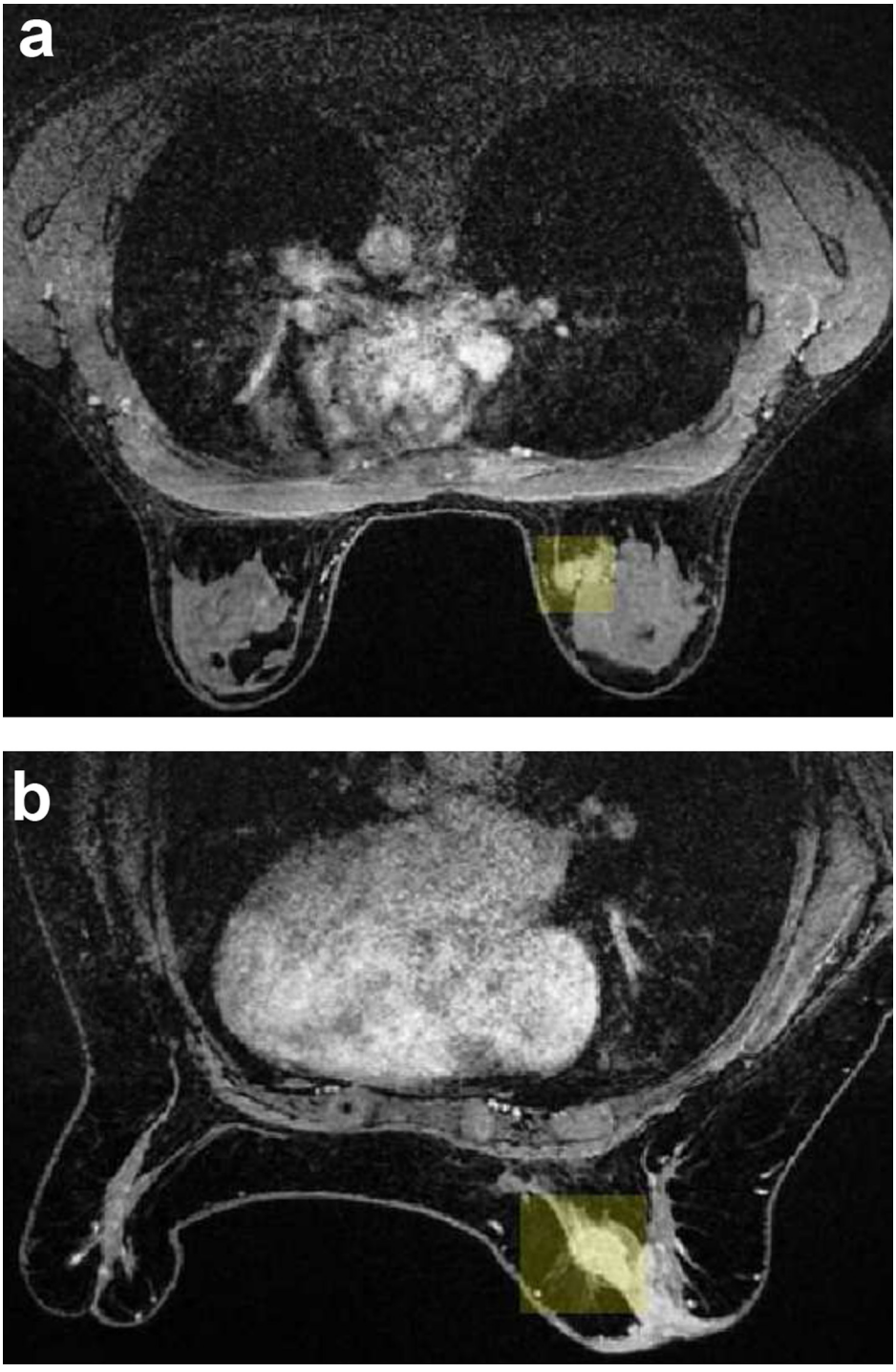

Figure 2. T1-weighted MR images in the axial plane obtained after intravenous administration of a gadolinium chelate. Annotation masks (in yellow) are superimposed over the MR images that show (a) a proliferating lesion and (b) an invasive ductal carcinoma. The annotations were made by a radiologist using a dedicated tool which enables to draw bounding box.

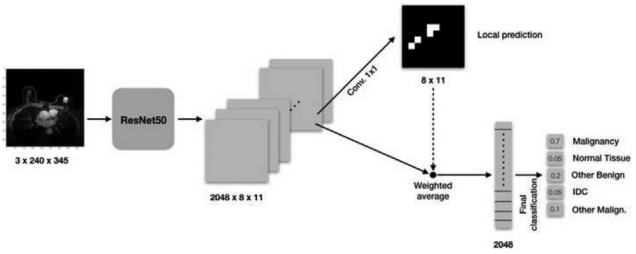

Figure 3. Architecture of our network. Each image of size 240×345 is fed into the ResNet50 Neural Network which produces 2048 images of size 8×11. They are then fed to the upper branch also called the ''attention block'' of the algorithm that learns to detect anomalies in the image. The latter are also fed to a second branch that averages features maps over the selected areas. Finally, the 2048 features are fitted to a logistic regression that outputs a score ranging from 0 to 1 for each category of focal lesion. This score can be interpreted as the probability of existence of such lesion in the image.

For each image, a binary mask of the same size, indicating the presence or absence of lesions was generated. The size of this mask was reduced to match the output dimensions of the ResNet (i.e., 8×11 pixels for an input image with spatial dimensions of 240×345).

The localization module was a single 1×1 convolution, applied to the output of the ResNet. This transformed the 2048×8 11 representation into a single image with the dimensions 8×11, to which we applied a sigmoid function to generate a prediction between 0 and 1. This module was trained to reproduce the binary mask generated from the annotations. This local prediction was then used to guide the main module responsible for determining the presence and characterization of lesions in the image. We used the local prediction to compute a weighted average of the final feature map, in which pij is the local prediction for pixel (i, j), as shown in Eq. (2):

$$x_k = \frac{\sum_{i,j} \rho_{ij} x_{kij}}{\sum_{i,j} \rho_{ij}} \tag{2}$$

When the module predicted a uniform probability of a lesion over the entire image, the formula was equivalent to the spatial average of the simple model. Conversely, when the module predicted the presence of a lesion in a single pixel with high confidence, only the feature vector extracted from this pixel was used for the final prediction.

The final prediction was performed by a densely connected layer with five neurons, one for each prediction: lesion malignancy classification, normal tissue, other benign

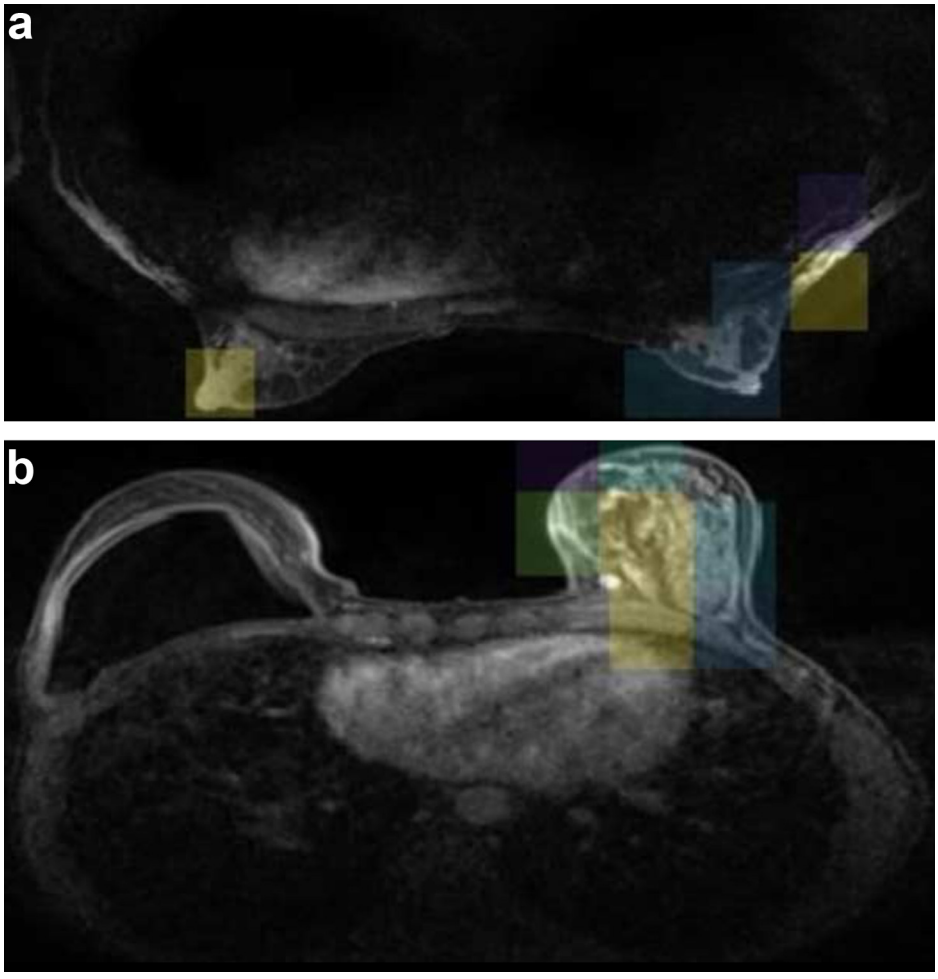

Figure 4. Two examples of attention maps generated by the model for (a) glandular tissue and (b) an invasive ductal carcinoma. This shows that the trained model can detect without human intervention the lesions or normal tissue on new image it was not trained on.

lesion, IDC, and other malignant lesions. The architecture of the model is shown in Fig. 3.

Furthermore, this attention mechanism allows interpretation of the model's predictions. To do so, we took the 8×11 attention map $\frac{\rho_{ij}}{\sum_{i,j}\rho_{ij}}$ and resized it to the original image dimensions (i.e., 240×345). This map could be superimposed over the image to see the areas considered by the model to make its decision, as shown in Fig. 4.

Implementation

Our model was trained simultaneously on the three tasks evaluated in this challenge (lesion detection, diagnosis of malignancy, and lesion classification). This multitask technique limited overfitting. However, the tasks were not learned at the same pace. Thus, we saved three copies of the weights, chosen depending on the performance of the model on a validation set. When the model reached its best AUC for lesion detection, we saved the first copy of its weights, which was used only for this task. We used stochastic gradient descent with Nesterov momentum to train the models. The results were highly variable due to the small amount of data, and thus we performed three-fold cross-validation, repeated over three different splits of the data. We repeated nine experiences during which we randomly selected 223 images (out of the 335 training set images) to train our neural network, and estimated its performance computing an AUC over the 112 images left. We then computed the mean scores over those nine different experiences to evaluate our model before executing it on the test set provided by the organizers of the challenge.

Results

The number of each lesion type provided by the challenge organizers is detailed in Table 1. The number of breast lesions provided in the training set for each lesion category is presented in Table 2 and was used to determine the final score, following Eq. (3):

$$\text{score} = 0.6 \times \text{AUC}_{\frac{\text{benign}}{\text{malignant}}} + 0.4 \times 1/4 \sum_{\text{lesion}} \sum_{\text{subgroups}} \text{AUC}_{\text{lesion subgroup}} \quad (3)$$

The average ROC-AUC scores achieved by our model on the repeated cross-validation (i.e., three times three-fold) on each lesion subgroup and the weighted sum according to the evaluation score of the challenge, determined using equation 1, are shown in Table 3 and corresponding ROC curves are provided in Fig. 5. The same model achieved a weighted AUC of 0.816 on the independent challenge test set.

Discussion

Our results ranked first in this challenge. This result is promising, given the relatively small amount of data with a training dataset consisting of only 335 images. The use of a supervised attention model was doubly beneficial. First,

Table 1 Number of breast lesions provided in the training dataset.

Lesion type	
Mammary gland	104 (31%)
Sclerosing adenosis	3 (0.9%)
Radial scar	2 (0.6%)
Fibroadenoma	24 (7%)
Galactophoritis	5 (1.5%)
Atypical hyperplasia	4 (1.2%)
Cyst	23 (6.9%)
PASH	1 (0.3%)
Papilloma	1 (0.3%)
Cytosteatonecrosis	13 (3.9%)
Intra-mammary lymph node	24 (7%)
Other proliferating lesion	8 (2.4%)
Invasive ductal carcinoma	82 (24.5%)
Invasive lobular carcinoma	16 (4.8%)
Triple negative cancer	18 (5.4%)
Intraductal carcinoma	5 (1.5%)
Mucinous carcinoma	2 (0.6%)
Total	335 (100%)

PASH indicates pseudoangiomatous stromal hyperplasia.

Table 2 Number of breast lesions provided in the training set for each lesion category.

Lesion type	Training set	Test set
Mammary gland	104 (31.1%)	[a]
Other benign lesions	108 (32.2%)	[a]
Invasive ductal carcinoma	82 (24.5%)	[a]
Other malignant lesion	41 (12.2%)	[a]
Total	335	168

[a] Note: the test set was used by the data challenge organizers to evaluate the study algorithms. Therefore, lesion types were not known for the test set.

Table 3 Detailed AUC scores according to breast lesion type.

Lesion group	AUC
Malignancy	0.869 (0.027)
Mammary gland	0.728 (0.046)
Other benign lesions	0.659 (0.048)
Invasive ductal carcinoma	0.805 (0.039)
Other malignant lesions	0.761 (0.065)
Overall score (weighted sum)	0.817 (0.036)

Data are presented as means. Numbers in parentheses are standard deviations. AUC values were computed as the mean of three shuffled three-fold cross-validation on the training set. Nine experiments in which two third of the images from the training set were randomly used to train our algorithm and the scores were computed from the results on the last third. Mean scores over those nine experiments are shown and standard deviations are provided between brackets.

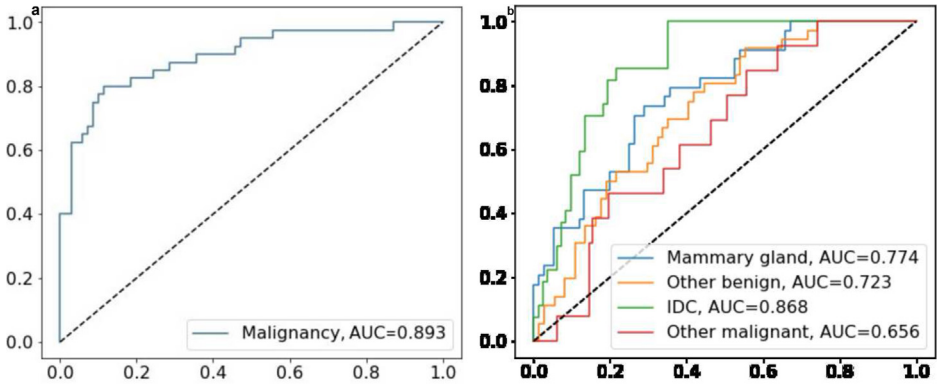

Figure 5. Diagrams show ROC curves for (a) benign vs. malignant classification and (b) lesion classification obtained with one model. The scores for lesion classification are heterogeneous, ranging from 0.656 for detection of the subgroup ''other malignant lesions'' to 0.868 for the detection of invasive ductal carcinoma. These differences could be explained by the difference of samples number for each lesion category: there were 82 examples of intraductal carcinoma in the training set, and only 41 in the category ''other malignant lesions''. The model performs better when we feed more examples to it.

expert-labelling on image strongly improved interpretability of the results. The generated heat maps made it possible to better understand how the model performed, including misclassification. Furthermore, the bounding boxes annotations of the radiologist significantly increased the performance of the model. Annotating datasets is a time-consuming task using classical research tools, and such tools are not amenable to a radiological workflow. One of the challenges in the deep-learning era in medical imaging is to build efficient tools to develop strong models (similar to human performance and clinically relevant) based on annotations from radiologists' workflow [18]. We have developed a tool that enabled rapid labeling which improved the performance without being time-consuming. With this tool, the whole dataset was labeled in less than an hour. Since this challenge, some papers applied CNN beyond characterization of lesions on breast MRI, such as predicting molecular subtype of breast cancer [19], or response to neoadjuvant therapy [20]. This confirms that beyond the current hype, machine learning has many applications to come in cancer management that could change the clinical decisions and become an important tool for the physician.

Breast lesions are clinically characterized with ACR Bi-RAds on multiparametric MRI, and a recent study using CNN showed good performance on multiple sequences [16] with an AUC of 0.89. Even if our study is not directly comparable (we used one slice of only one MRI sequence), our result encourages us to apply our method on three-dimensional sequences and compare to the CNN method proposed [16] in order to check if our attention method improves significantly the performances on a new cohort.

In conclusion, the validation of an algorithm on an independent dataset, extended to other sequences with additional 3D information rather than a single 2D image, is an essential step in judging the generalizability of the model. Further studies are required to show the interest of such a technique and demonstrate a clinically implementable workflow for lesion classification, especially with BI-RADS. The use of larger databases and multiparametric MRI are likely to further increase the accuracy of the model.

Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans as well as in accordance with the EU Directive 2010/63/EU for animal experiments.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article. The authors also confirm that the personal details of the patients and/or volunteers have been removed.

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Disclosure of interest

Paul Herent is a resident radiologist at APHP and is a part-time consultant at Owkin. Benoit Schmauch, Simon Jégou, Paul Jehanno and Charlie Saillard are employees at Owkin. Olivier Dehaene was a data scientist intern at Owkin at the

time the Data Challenge of JFR 2018 occurred and is now a full-time employee at Owkin.

References

[1] Reiner BI, Krupinski E. The insidious problem of fatigue in medical imaging practice. J Digit Imaging 2012;25:3—6.

[2] Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015;35:1668—76.

[3] Muenzel D, Engels HP, Bruegel M, Kehl V, Rummeny EJ, Metz S. Intra- and inter-observer variability in measurement of target lesions: implication on response evaluation according to RECIST 1.1. Radiol Oncol 2012;46:8—18.

[4] Suzuki C, Torkzad MR, Jacobsson H, Astrom G, Sundin A, Hatschek T, et al. Interobserver and intraobserver variability in the response evaluation of cancer therapy according to RECIST and WHO-criteria. Acta Oncol 2010;49:509—14.

[5] Monticciolo DL, Newell MS, Moy L, Niell B, Monsees B, Sickles EA. Breast cancer screening in women at higher-than-average risk: recommendations from the ACR. J Am Coll Radiol 2018;15:408—14.

[6] Morris EA, Comstock CE, Lee CH. ACR BI-RADS® magnetic resonance imaging. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA. Am Coll Radiol 2013.

[7] Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS® fifth edition: a summary of changes. Diagn Interv Imaging 2017;98:179—90.

[8] Thakran S, Gupta PK, Kabra V, Saha I, Jain P, Gupta RK, et al. Characterization of breast lesion using T1-perfusion magnetic resonance imaging: qualitative vs. quantitative analysis. Diagn Interv Imaging 2018;99:633—42.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436—44.

[10] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comp Sci 2015;115:211—52.

[11] Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. Nature 2018;555:487—92.

[12] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500—10.

[13] Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. Jpn J Radiol 2018;36:257—72.

[14] Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol 2017;52:434—40.

[15] Kooi T, Litjens G, van Ginneken B, Gubern-Merida A, Sanchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303—12.

[16] Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. Radiology 2019;290:290—7.

[17] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016 [abs/1512.03385].

[18] Abajian AC, Levy M, Rubin DL. Informatics in radiology: improving clinical work flow through an AIM database: a sample web-based lesion tracking application. Radiographics 2012;32:1543—52.

[19] Ha R, Mutasa S, Karcich J, Gupta N, Pascual Van Sant E, Nemer J, et al. Predicting breast cancer molecular subtype with MRI dataset utilizing convolutional neural network algorithm. J Digit Imaging 2019, http://dx.doi.org/10.1007/s10278-019-00179-2.

[20] Ha R, Chin C, Karcich J, Liu MZ, Chang P, Mutasa S, et al. Prior to initiation of chemotherapy, can we predict breast tumor response? Deep learning convolutional neural networks approach using a breast MRI tumor Dataset. J Digit Imaging 2018, http://dx.doi.org/10.1007/s10278-018-0144-1.