Artificial intelligence in the Automated Triaging of Breast Screening MRI Examinations

By Dr. KGA Gilhuijs

INTRODUCTION

Breast density is a risk factor for developing breast cancer [1-4]. In fact, women in the highest breast-density category (i.e., BI-RADS class D) are at a risk of developing cancer that is three to six times greater than that of women with almost entirely fatty breasts [5,6]. Not only are women with dense breasts at increased risk, but the density of their breast tissue also complicates the detection of cancers by conventional screening mammography. In contrast to screening mammography, breast MRI allows functional information to be obtained. A recent randomized controlled trial (DENSE: ClinicalTrials.gov: NCT01315015) investigated whether MRI has complementary value in a mammography-screening population of women with extremely dense breasts. The results of the multi-institutional DENSE trial demonstrated clearly that supplemental screening with MRI does indeed help detect cancers at an earlier stage and significantly reduces the rate of interval cancers [7]. However, if, in order to take advantage of this finding, supplemental MRI were to be implemented on a broad scale, the workload of breast MRI radiologists would be considerably increased — nearly 82.000 women would be eligible for supplemental MRI in the Netherlands alone [8]. Of these women, the vast majority (~90%) of their breast MRI examinations would show only normal anatomical variation.

An unmet need thus exists for a technology that could automatically identify and dismiss those MRI scans with normal breast anatomy, thus enabling radiologists to prioritize other cases and reduce their workload. This is a challenging task because "normal" spans a wide range of phenotypical patterns of enhancement on breast MRI — particularly in this population of women with dense breasts — which may partly overlap with patterns associated with early signs of breast cancer.

In our study published recently in Radiology [9], we investigated the feasibility of automated triaging using deep learning to exclude/dismiss the largest number possible of breast MRI examinations without lesions while still identifying all cases with cancer. In this short review we summarize the key aspects of this study and reflect on its potential value.

The Author
Dr. K.G.A. Gilhuijs,
University Medical Center Utrecht,
Image Sciences Institute,
Utrecht, The Netherlands.
Email: k.g.a.gilhuijs@umcutrecht.nl

STUDY PARTICIPANTS, MRI, AND RADIOLOGISTS

Our study constituted a secondary analysis of the MRI data from the first round of the DENSE trial which included 4783 MRI examinations from eight hospitals in the Netherlands between December 2011 and January 2016. Participants were recruited from the Dutch national screening program, which offers biennial mammography to women between 50 and 75 years of age. Of the 4783 examinations, 4581 (95.8%) with complete digital data were available for Artificial Intelligence (AI) work-up.

In the DENSE trial, a fixed MRI protocol was used among the eight participating hospitals. Although this protocol involved multiparametric MRI [10], our study only used a subset of the data, namely what is typically used in abbreviated MRI: the pre-contrast T1-weighted MRI and the first post-contrast T1-weighted MRI, both acquired at high spatial resolution. Although the remaining MRI sequences would add more information to the AI, this choice of abbreviated MRI was made deliberately in order to achieve a solution to the triaging problem that is likely to be more widely applicable. MRI units from two different manufacturers were used (Philips and Siemens), both with 3-T field strength. The use of contrast agent was standardized on Gadobutrol.

The MRI examinations were scored by trained breast MRI radiologists using the BI-RADS MRI lexicon [11]. Only lesions scored BI-RADS 3 were read twice by two different radiologists.

ARTIFICIAL INTELLIGENCE

A method was developed based on deep learning technology to automatically establish whether MRI breast images contain lesions. Deep learning is a machine-learning method — a form of artificial intelligence — that mimics the behavior of the human brain, albeit far from matching its ability. It simulates a neural network with multiple layers, allowing it to "learn" image labels (e.g. "lesions present" vs "lesions not present") from a large set of training images. Deep-learning networks in this context are basically non-linear statistical classification methods.

In the development process, several design considerations had to be taken into account. First, although 4581 MRI examinations certainly constitute a large sample size, only 77 examinations of these were associated with the presence of malignant disease. Training a deep-learning network to discriminate between MRI examinations of breasts with verified malignant disease and those without would therefore be extremely challenging to accomplish from a purely statistical point of view. Instead, we focused on deep learning to discriminate between MRI examinations of breasts with lesions (785 examinations, BI-RADS 2 through 5) and those without lesions (3796 examinations, BI-RADS 1), and optimized the operating threshold to include all malignant lesions.

Second, due to the design of bilateral breast MRI coils, the left and right breast are depicted in the same field of view.

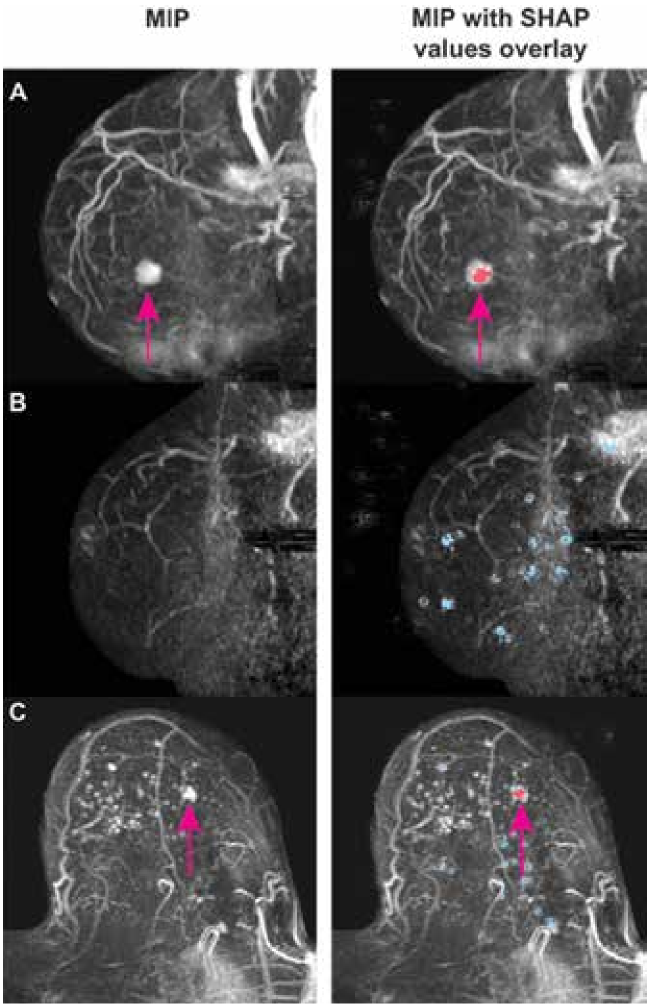

Figure 1. Examples of deep Shapley additive explanations (SHAP) overlay images. Maximum intensity projection (MIP) images are on left, and MIP images with the SHAP overlay are on right. Positive SHAP values (red) show areas that contribute to a high probability of lesion presence, negative SHAP values (blue) show locations with reduced probability. (A) Sagittal MIP images of contrast-enhanced breast MRI scan of an invasive ductal carcinoma in a 57-year-old woman with Breast Imaging Reporting and Data System (BI-RADS) category 4. The deep learning (DL) model yielded a probability of lesion presence of 90%. Positive SHAP values (red) are shown to coincide with the location of the lesion (arrows). (B) Sagittal MIP images of contrast-enhanced breast MRI scan of a breast without lesions in a 53-year-old woman with BI-RADS 1 score. The DLmodel yielded a probability of lesion presence of 11%. Negative SHAP values (blue) are diffusely distributed in the breast region. (C) Transverse MIP images of contrast-enhanced breast MRI scan of a ductal carcinoma in situ in a 65-year-old woman with BI-RADS 4 score. The DL model yielded a probability of lesion presence of 32%—the lowest probability value among all breasts with malignant disease in our study. Positive SHAP values (red) are shown to coincide with the location of the lesion (arrows). Image reproduced, courtesy RSNA, from ref [9].

To make optimal use of the available sample size, deep learning was restricted to regions tightly boxed around the left and right breast separately. This was done automatically. The deep-learning results per breast were then combined into one result per bilateral examination.

PREPROCESSING

Prior to deep learning, several automated preprocessing steps had to be performed on the three-dimensional (3D) MRI data. First, automated cropping to single out the left breast and the right breast. Second, automated deformable registration to correct for soft-tissue deformation that may have occurred between the precontrast and postcontrast MRI [12]. Third, automated calculation of maximum intensity projection images (MIPs) in three orthogonal directions (i.e., coronal, sagittal, and transversal views). Hence, subsequent deep learning was performed on MIP images in the three orthogonal views rather than on the original 3D MRI data.

DEEP LEARNING

A deep-learning network has a large number of configuration parameters such as number of layers, layer size, loss function, learning rate, etc., whose values need to be optimized. We also automated this process to avoid subjectivity in the configuration and found that the VGG deep-learning architecture was optimal for the triaging task [13]. The process of optimization has been reported in detail [9].

To train and validate the deep-learning network, an internal-external validation process was used [14]: the data from one hospital were left out and used as an independent test set, while the data from the other hospitals were combined to create the deep-learning model. The combined data were randomly divided (at the screening-participant level) into a training set (80% of the data) and a validation set (20% of the data). This process was repeated until each of the eight hospitals had been once the independent test set. Thus, the internal-external validation process created eight deep-learning networks.

The areas under the Receiver-Operating Characteristics curve (AUC) of the models applied to the test set were established and their mean AUC was used as estimate of overall prospective performance of the AI.

In addition to the AUC, each model had an operating threshold at which all malignant lesions were triaged to radiologist reading. This operating threshold was validated in the test sets and the fraction of examinations without lesions correctly dismissed was averaged over the eight hospitals.

Because deep-learning networks are often perceived as black boxes – so potentially limiting their acceptance in the clinic – "*explainable artificial intelligence*" was added in the form of deep Shapley additive explanations (SHAP) [15] [9]. SHAP visualizes the contribution of each image pixel to the prediction result using a color overlay [Figure 1]. It provides insight into what the deep-learning network has "seen" and hence whether the resulting prediction is plausible.

OVERVIEW OF KEY RESULTS

The deep learning yielded an average area under the ROC curve of 0.83 (95% CI 0.80, 0.85) [Figure 2]. This would exclude from radiologist reading 39.7% (95% CI: 30.0, 49.4) of the MRI examinations without lesions while dismissing none of the 77 cancers. At this operating threshold, the network would triage to radiologist reading 90.7% (95% CI 86.7- 94.7) of the MRI examinations with lesions (including the 77 cancers). Among the benign lesions triaged were fibroadenoma, indeterminate mass lesions and cysts.

Some aspects were found to affect the triaging process. BI-RADS 2 lesions were more often dismissed than BI-RADS 4 and 5 lesions.

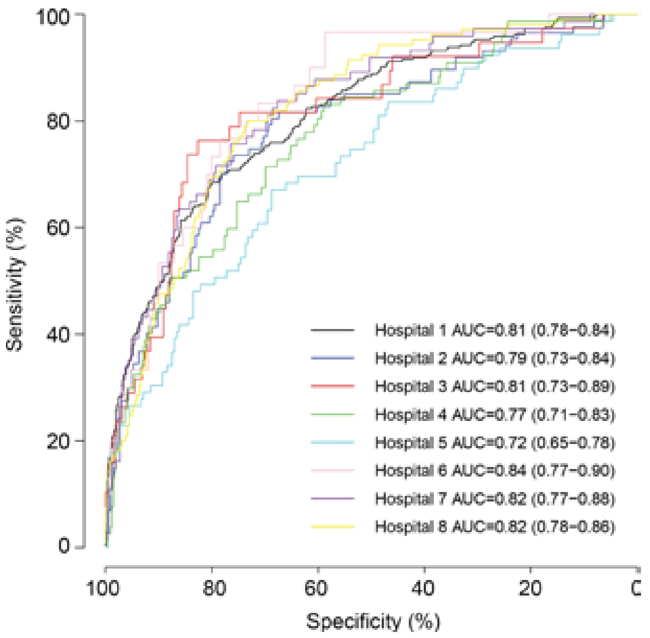

Figure 2: Receiver operating characteristics curves from the eight hospitals in the DENSE trial. Each curve is the result of testing on one of the participating hospitals using internal-external validation. Curves show the sensitivity and specificity of the method in the differentiation between MRI scans with lesions and those without lesions. Numbers in parentheses are the 95% confidence interval. AUC = area under the receiver operating characteristic curve. Image reproduced, courtesy RSNA, from ref [9].

In fact, none of the BI-RADS-5 lesions was dismissed. In addition, non-mass enhancing lesions were more often dismissed than mass-enhancing lesions. Breast parenchymal enhancement (BPE) was found to affect the triaging process only when lesions were absent (minimal BPE leading to more dismissals than severe BPE), but not when lesions were present.

FUTURE PERSPECTIVES

In our study, a method based on deep learning was developed to automatically inspect MRI screening examinations of women with extremely dense breasts in order to dismiss the largest number of normal scans without excluding any cases of malignant disease. With the approach, it was shown that approximately 40% of the breast screening examinations could thus be safely dismissed.

Although these results are promising, some aspects of the approach need to be considered. The AI was developed specifically for the Dutch screening population of women with extremely dense breasts. It is still unknown how well the AI would generalize to other populations or other screening indications such as women at high lifetime risk of developing breast cancer. It is however resonable to supppose that the AI could be further generalized by adding MRI examinations from other populations to the training process.

Although the study demonstrates that an autonomous AI-based system can be constructed to triage thousands of breast MRI examinations without dismissing any cancer, current guidelines and regulations make it likely that practical implementation of the approach in clinical routine will start as a method to assist the radiologist rather than as an independent, unsupervised tool. For instance, the worklist of radiologists could be automatically prioritized by assigning AI-triaged cases to certain radiologists or certain time slots of the day, while AI-dismissed cases could be offered for review to other radiologists and/or other time slots. Such an implementation would have the advantage of being able to collect large amounts of evidence across many more hospitals, populations and MRI units. Ultimately it is inevitable that — after extensive review of additional accumulated evidence — fully autonomous implementation will actually also require adjustments to existing laws and regulations. This is a process that is currently well underway for other critical non-imaging AI applications, such as those governing self-driving cars.

ACKNOWLEDGEMENTS

I thank the DENSE trial study group for making the trial data available for this study. The contributions of all the participants in our subsequent study are gratefully acknowledged. A full list of disclosures and acknowledgements is provided in our original publication [9].

REFERENCES

1. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992;184(3):613–617.
https://doi.org/10.1148/radiology.184.3.1509041

2. Boyd NF, Guo H, Martin LJ, *et al.* Mammographic density and the risk and detection of breast cancer. N Engl J Med 2007;356(3):227–236.
https://doi.org/10.1158/1055-9965.EPI-09-0881

3. Kerlikowske K. The mammogram that cried Wolfe. N Engl J Med 2007;356(3):297–300.
https://doi.org/10.1056/NEJMe068244

4. Wanders JO, Holland K, Veldhuis WB, *et al.* Volumetric breast density affects performance of digital screening mammography. Breast Cancer Res Treat 2017;162(1):95–103.
https://doi.org/10.1007/s10549-016-4090-7

5. Price ER, Hargreaves J, Lipson JA, *et al.* The California breast density information group: a collaborative response to the issues of breast density, breast cancer risk, and breast density notification legislation. Radiology 2013;269(3):887–892.
https://doi.org/10.1148/radiol.13131217

6. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev 2006;15(6):1159–1169. *https://doi.org/10.1158/1055-9965.EPI-06-0034.*

7. Bakker MF, de Lange SV, Pijnappel RM, et al. Supplemental MRI Screening for Women with Extremely Dense Breast Tissue. N Engl J Med 2019;381(22):2091–2102.
https://doi.org/10.1056/NEJMoa1903986

8. Monitor population screening for breast cancer. Integraal kankercentrum Nederland web site. Published online 2016. Accessed August 22, 2019.
https://iknl.nl/kankersoorten/borstkanker/onderzoek/monitor-bevolkingsonderzoek

9. Verburg E, van Gils CH, van der Velden BHM, Bakker MF, Pijnappel RM, Veldhuis WB, Gilhuijs KGA. Deep Learning for Automated Triaging of 4581 Breast MRI Examinations from the DENSE Trial. Radiology. 2022;302(1):29-36.
https://doi.org/10.1148/radiol.2021203960

10. Emaus MJ, Bakker MF, Peeters PH, *et al.* MR Imaging as an Additional Screening Modality for the Detection of Breast Cancer in Women Aged 50-75 Years with Extremely Dense Breasts: The DENSE Trial Study Design. Radiology 2015;277(2):527–537.
https://doi.org/10.1148/radiol.2015141827

11. Morris EA, Comstock, CE, Lee CH, *et al.* ACR BI-RADS Magnetic Resonance Imaging. In: 2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. Reston, Va: American College of Radiology, 2013.

12. Verburg E, van Gils CH, Bakker MF, Viergever MA, Pijnappel RM, Veldhuis WB, Gilhuijs KGA. Computer-Aided Diagnosis in Multiparametric Magnetic Resonance Imaging Screening of Women With Extremely Dense Breasts to Reduce False-Positive Diagnoses. Invest Radiol. 2020;55(7):438-444. *https://doi.org/10.1097/RLI.0000000000000656*

13. Simonyan K, Zisserman A. Very deep convolutional networks for largescale image recognition. arXiv preprint arXiv:1409.1556. *https://arxiv.org/abs/1409.1556.* Published September 4, 2014. Accessed October 29, 2018.

14. Steyerberg EW, Mushkudiani N, Perel P, *et al.* Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med 2008;5(8):e165; discussion e165.
https://doi.org/10.1371/journal.pmed.0050165

15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017) [book online]. Proceedings of the 2017 Conference on Neural Information Processing Systems. San Diego, Calif: Neural Information Processing Systems Foundation, 2017.
https://arxiv.org/abs/1705.07874.