Dear Dr. Krupinski,

With interest I read the article titled “Prior to Initiation of Chemotherapy, Can We Predict Breast Tumor Response? Deep Learning Convolutional Neural Networks Approach Using a Breast MRI Tumor Dataset”, which appeared in the October 2018 issue of the Journal of Digital Imaging. I am writing to express my concerns regarding the validity of the results presented in this work. The manuscript reports on a deep learning (convolutional neural network, or “CNN”) model that was trained to predict response to breast cancer treatment with neoadjuvant chemotherapy (NAC) from pre-treatment breast MRI data. The proposed algorithm attains impressive performance on the unseen test data, with a mean Area Under the Receiver-Operating Characteristic curve (AUROC) of 0.98 and mean accuracy of 88% for predicting whether a patient would exhibit complete, partial, or no response to treatment (averaged in a five-fold cross-validation procedure).

A model with such strong prediction capabilities would have significant implications for future clinical decision-making procedures. However, similar studies have failed to reach performance levels anywhere close to the ones reported in this work. A patient’s response to neoadjuvant chemotherapy is known to depend on a large array of factors, making it unlikely that a single pretreatment MRI scan contains sufficient information to predict the outcome with such near-perfect levels of accuracy.

My concern is that these results can be largely explained by inadvertent “data leakage” between the training and test sets, and that the model may not generalize well to unseen data in practice. The cited work sets an extremely high benchmark for follow-up studies and may lead to unreasonable expectations for the use of similar deep learning models in clinical practice. I have reached out to the corresponding author to ask for clarification, but unfortunately have not yet received a reply so far.

I conjecture that data leakage occurred by accident in the procedure that was used to split the data into training, validation, and test sets. The full dataset was constructed by collecting all 2D MRI slices with at least 75 voxels of tumor tissue. In the section “CNN Training”, the authors describe that “the data was divided into 80% validation and 20% test. The validation test set was then divided into fivefold, and fivefold cross-validation was performed”. A crucial observation is that no mention is made of *stratifying the splits on a patient-level, as opposed to a slice-level*. If a straightforward, non-stratified random slice-level split was made, distinct MRI slices of the same patient would likely have occurred in both the training and test sets. Since nearby slices contain similar anatomy, the CNN would be able to discover a shortcut when predicting treatment response: it could simply learn to recognize each patient’s anatomy and “memorize” the corresponding treatment outcome from the training data, and subsequently present the known outcome when a similar-looking 2D scan *from the same patient* is encountered in the test set. We illustrate the visual similarity of adjacent MRI slices of the same patient in Figure 1.

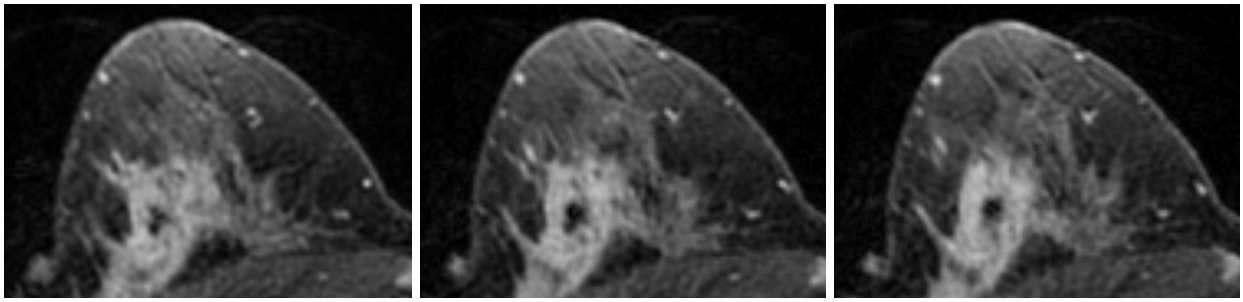

Figure 1. Three adjacent slices of a breast MRI scan (from the public Duke Breast Cancer MRI dataset) of a single patient. Although the images exhibit minor differences, the overall structure is highly similar.

In order to verify that the suspected data leakage can indeed explain the high accuracy and AUC scores, I ran a similar experiment on a subset of the public Duke Breast Cancer MRI dataset[1], consisting of 282 breast cancer patients who underwent neoadjuvant chemotherapy. Of these patients, 22% (n=61) had pathological complete response to treatment, and 78% (n=221) had residual disease. I trained a modern convolutional neural network (of ResNet18-type) to predict this (binary) treatment outcome based on 2D breast MRI slices: once on a “proper” patient-level split, and once on a “leaky” slice-level split. On average, each patient had 33 tumor-containing slices, and as a consequence nearly every patient had some of these slices appear in both the train and test sets for the “leaky” slice-level split. Clearly, no such mixing occurred for the patient-level split. We present the ROC curves of the evaluation of the trained model on the test data for both types of train/test splits in Figure 2. For the patient-level split, the model reached a test accuracy of 69% for the binary classification task, with an AUROC of 0.42. This is in stark contrast with the slice-level split, where the test accuracy was 97% with an AUROC of 0.99.

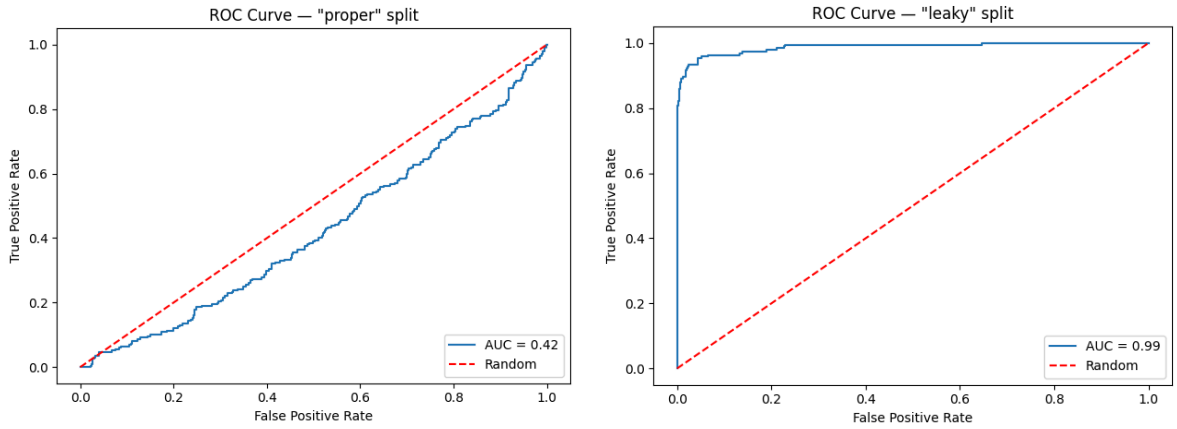

Figure 2. ROC curves for the binary treatment response prediction of 2D MRI slices using a ResNet18 convolutional neural network. The leftmost figure presents the results for a proper patient-stratified train/test split, and the rightmost figure for the non-stratified slice-level split.

Combined with the prior knowledge that predicting treatment response is a complex problem, this leads me to conjecture that the reported results can be explained by such data leakage between train and test set. Although I do not believe that this alleged leakage was intentional, I think it is important to raise the aforementioned concerns. Deep learning is an extremely powerful framework that holds much promise for the future of clinical decision-making, but we should remain critical in assessing its real-world performance. Furthermore, I believe this case study may serve as a valuable reminder to the research community that data leakage is a subtle issue that should always be carefully examined in any machine learning scenario.

Any clarifications from the original authors would be very welcome.

Sincerely,
Joren Brunekreef, PhD
Postdoctoral fellow, Netherlands Cancer Institute

Co-signed by:

Ritse Mann, MD, PhD
Group Leader Breast Imaging Group, Radboud University Medical Center & Netherlands Cancer Institute
Breast and interventional radiologist, Radboud University Medical Center

Eric Marcus, PhD
Senior postdoctoral fellow, AI for Oncology group, Netherlands Cancer Institute

Prof. Katja Pinker-Domenig, MD, PhD
Professor of Radiology, Weill Medical College of Cornell University
Director of Research, Memorial Sloan Kettering Cancer Center
Director of Breast MRI, Memorial Sloan Kettering Cancer Center

Prof. Jan-Jakob Sonke, PhD
Professor of Adaptive Radiotherapy, Netherlands Cancer Institute
Theme leader Image Guided Therapy, Netherlands Cancer Institute

Jonas Teuwen, PhD
Group leader AI for Oncology, Netherlands Cancer Institute

Bibliography

1. Saha, A. *et al.* A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *Br. J. Cancer* **119**, 508–516 (2018).