Forecasting Artifacts in High b-Value DWI MIP Images of the Breast Using Deep Learning and T2-Weighted or Low b-Value Acquisitions

Andrzej Liebert[1,2], Dominique Hadler[1], Dominika Skwierawska[1], Lorenz A. Kapsner[1,3], Lukas Folle[4], Badhan Das [4], Hannes Schreiter[1], Sabine Ohlmeyer[1], Prof. Andreas Maier [4], Prof. Frederik B. Laun[1], Prof. Michael Uder[1], Prof. Evelyn Wenkel[5], Sebastian Bickelhaupt[1]

[1]Institute of Radiology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
[2] Institute of Computer Science, Polish Academy of Science, Warsaw, Poland
[3] Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany
[4] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany.
[5] Medizinische Fakultät, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany. Radiologie München, München, Germany

Andrzej.liebert@ipipan.waw.pl

Abstract. High b-value diffusion-weighted imaging (DWI) of the breast provides high sensitivity for cancer detection, but maximum-intensity-projection (MIP) images are frequently affected by artifacts. This study assessed whether high b-value DWI MIP artifacts can be forecast from preceding low b-value DWI or T2-weighted fat-saturated (T2w-FS) acquisitions. In total, 1,089 examinations from two 3.0 T systems and 195 examinations from the multi-vendor ACRIN-6698 dataset were retrospectively analyzed. Artifact severity was rated on a five-point scale by three readers; moderate-to-severe artifacts (score ≥3) occurred in 62% of examinations. A 3D DenseNet201 model was trained on internal data and evaluated in independent internal and external test datasets. Operating thresholds were selected in internal validation data to target high specificity. In the internal test dataset, forecasting from low b-value DWI and T2w-FS achieved AUROCs of 0.75 (CI: 0.68–0.82) and 0.74 (CI: 0.67–0.81), respectively, without a significant difference between methods (p=0.55). At the fixed thresholds, sensitivity/specificity were 0.38/0.93 for low b-value DWI and 0.49/0.87 for T2w-FS. In the external dataset, AUROCs were 0.63 (CI: 0.54–0.70) and 0.71 (CI: 0.64–0.78), respectively, with better discrimination for T2w-FS (p=0.013). Sensitivity/specificity were 0.21/0.89 for low b-value DWI and 0.61/0.68 for T2w-FS. Preceding breast MRI acquisitions contain information associated with subsequent high b-value DWI MIP artifacts. However, variable external performance indicates limited transferability and the need for multi-vendor training, recalibration, and prospective evaluation before clinical use.

Keywords: Magnetic resonance imaging, artifacts, deep learning, artifact prediction, quality assurance, diffusion weighted imaging.

1 Introduction

Diffusion weighted imaging (DWI) MRI is a technique allowing to map the random motion of water molecules in the tissue, thus detecting alterations of tissue diffusivity properties indicating potentially significant lesions. In breast imaging DWI has both been investigated in regards to its complementary value and the applicability in abbreviated protocols in breast MRI for certain indications [1,2].

Whilst DWI is of increasing interest in breast MRI, it is a technique known to be prone to the formation of artifacts. The reasons for this are manifold, with amongst them being fat saturation and shimming issues [3]. Artifacts in images can significantly impede the diagnostic assessment [3,4]. This becomes even more critical in case of abbreviated protocols as proposed for breast MRI with only limited complementary sequences covering the same region in different contrasts. Additionally, the maximum intensity projections (MIPs) used in some abbreviated approaches [5-7] might magnify the impediment caused by artifacts in individual slices since. This is caused by the high signal intensities of most artifacts, which are then transferred into the MIP images and which can potentially cover significant lesions. If artifacts are detected, the MRI acquisitions can be repeated aiming to adjust the protocol in order to reduce the artifacts.

However, before the artifacts even arise, they might be forecasted from preceding acquisitions, as it was previously shown for dynamic contrast enhanced acquisitions [8]. Such approach might allow to readjust the DWI acquisition parameters and scanner setting before the acquisition. removing the need for a repeated acquisition.

In this work, we investigate whether clinically relevant artifacts in breast MRI maximum-intensity-projection (MIP) images derived from high b-value DWI (b >750 s/mm²) can be forecast before the high b-value acquisition is assessed. Our contributions are threefold: first, we address artifact forecasting from preceding acquisitions rather than post-acquisition artifact detection; second, we compare two routinely available input sources, namely low b-value DWI (b = 50–100 s/mm²) acquired as part of the same multi-b-value protocol and T2-weighted fat-saturated (T2w-FS) imaging; and third, we evaluate the approach in independent internal and external multi-vendor breast MRI datasets. We hypothesized that anatomical and acquisition-related characteristics associated with subsequent high b-value DWI artifacts may already be reflected in the preceding low b-value DWI and T2w-FS acquisitions.

2 Material and Methods

2.1 Datasets

This IRB approved study used an in-house dataset consisting of breast MRI examinations performed between 2017 and 2020 at the University Hospital Erlangen using one of two 3T Siemens scanners: Magnetom Skyra Fit and Magnetom VIDA. MRI

examinations included in this study consisted of routine breast MRI examinations with clinical indications to perform the breast MRI including e.g. high-risk screening, assessment of neoadjuvant chemotherapy response, multifocal breast cancer, suspicion of relapsing breast cancer, and others. The breast MRI examination was performed as a routine multi-parametric breast MRI protocol including T2-weighted fat-saturated acquisition, and a multi-b-value DWI acquisition with the highest b-value being 1500 s/mm2. The full internal dataset of n=1089 cases was then split into an 80% training-validation (n=873) set and a 20% holdout test set (n=216). During the split it was ensured that no subjects occurring in the test-set were present also in the training-validation set. All internal MRI examinations from this study were previously included in other studies [8-11].

In order to test the generalizability of the algorithms on breast MRI acquisitions using other MRI manufacturers (GE Healthcare and Philips) the publicly available ACRIN-6698 [12] dataset was used. From this dataset, breast MRI examinations performed before the administration of neoadjuvant chemotherapy and which included both DWI and T2w-FS acquisitions were selected. This resulted in the creation of an external dataset which consisted of n=195 cases, with n=164 cases being acquired on a GE system (GE Signa n=150, Discovery 10, Optima n=4) and n=31 cases on a Philips system (Philips Ingenia n=3, Phlips Achieva n=28).

2.2 Visual image assessment for artifact classification

All data were transferred from the clinical picture archiving and communication system to research workstation. Herein, MIPs were derived from the highest b-value image acquisition of the DWI sequence (b=1500 s/mm2 for the internal data set and b=800 s/mm2 for the external data set) with an in-house Python (v3.10.) script using Simple ITK (v. 2.0.2) framework. The highest b-value acquisitions were used, as previous studies suggest these for generation of MIPs in order to assess the presence of significant lesions [6,7]. The MIPs were created by projecting the maximum voxel in-tensity along the head-feet axis of the volume, into a single transversal 2D image. After creation all MIPs were split into left and right side along the central head-feet axis to allow for an independent assessment of the presence of artifacts within each individual breast. The MIPs of each single breast were then assessed using 5-point Likert-like scale based on the presence of typical artifacts in breast DWI as previously described in literature e.g., by Partridge et al. [3,13]. The following classes were applied: 1 – no artifact at all, 2 – minor artifacts, but with-out the potential to impede the assessment, 3 – moderate artifacts, potentially impairing the visual assessment, 4 – pronounced artifacts, 5 – highly significant artifacts. Examples of MIP's affected by the different artifact strength classes are provided in Figure 1.

Classification was performed by two board-certified radiologists and a research assistant (Reader 1: S.B. > 10 years' experience, Reader 2: D.H. > 15 years of experience, Reader 3: D.S. > 3 years of breast MRI experience). The labels from the two breasts were then combined for each reader by choosing the maximum among the two breasts, and then the median score among the three readers was used as ground truth for the purpose of neural network training, validation and test. Pairwise inter-rater agreement

(IRA) was evaluated using Cohen's kappa (ƙ) with quadratic weighting and calculating 95% confidence intervals using the bootstrapping method with 10000 repetitions.

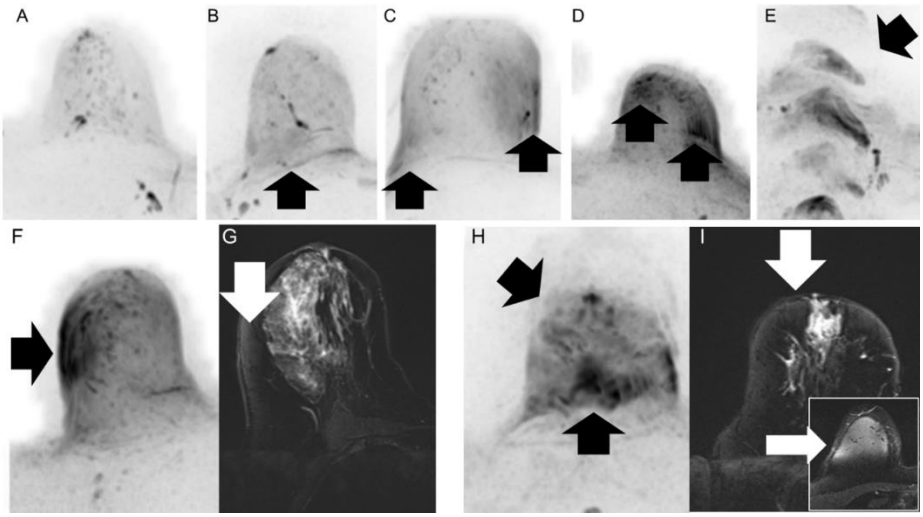

Fig. 1. A-E: Examples of artifacts in the different classes: A = no artifacts, B=minimal artifacts, C=moderate artifacts, D= pronounced artifacts, E=highly significant artifacts (black arrows mark the artifacts). F/G: Pronounced artifact in the high b-value DWI MIP (F, 1500s/mm2) (black arrow) with corresponding subtle signal intensity increase in the T2-weighted fat saturated image (G, white arrow). H/I: Highly significant artifacts in the high b-value DWI MIP (1500s/mm^2) (black arrows) with corresponding (rather subtle) blurry appearance in the T2-weighted fat saturated image (I) and a locally inhomogeneity in the fat saturation in a skin fold underneath the breast (small image and small arrow).

2.3 Neural Network architecture, training and evaluation

3D-DensetNet201[14] network was implemented using PyTorch and MONAI framework with a softmax activation at the end of the network. As input data either the volumetric T2w-FS acquisitions or a low b-value acquisition (b-value=50 s/mm2 for internal data set and b-value=100 s/mm2 for external dataset) was used. As targets during the training the five multi-class examination level labels were used. The network was trained for 200 epochs in 5-fold stratified cross-validation procedure [15,16], using a cross-entropy loss function. Training was performed on a single workstation using a RTX6000 nVidia GPU (24GB RAM). Training hyperparameters were optimized using the first training and validation split.

An ensemble classifier was created by averaging the probability values for each class provided by the best performing models from each of the 5 cross-validation folds.

For the evaluation of the performance, the artifact scores 3, 4 and 5 were considered as indicating the potential for a significant artifact to occur, thus the probabilities from these scores were added together and the probabilities of scores 1 and 2 were also added up. Receiver operating characteristic (ROC) curves were used together with their area under the curve (AuROC), and sensitivity and specificity were used for evaluation of

performance based on these dichotomized results. Based on the RoC a threshold for the probability of the forecasting ensemble was identified on the validation set, for which the sensitivity achieved a value of at least 0.9. Statistical differences in the performance for different input data was evaluated using DeLong tests on the AuROC's separately for both test sets with a p-value of 0.05 considered to indicate significant difference.

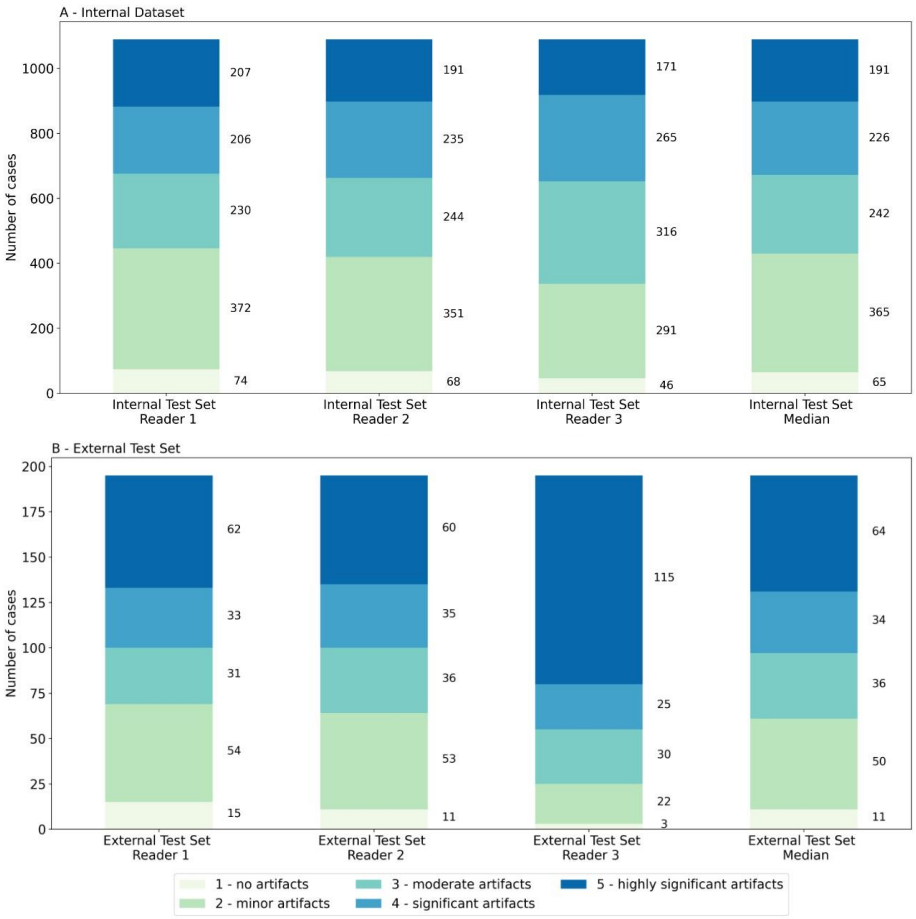

Fig. 2. Distribution of artifact strength scores provided by the three readers and median score of these reader for both datasets.

3 Results

3.1 DWI MIP artifact labelling

Figure 2 shows the distribution of scores provided by the readers on the examination level for both the full internal and external dataset. Pairwise IRA was almost perfect between R1 and R2 for both internal (ƙ=0.96 CI: 0.96-0.97) and external dataset (ƙ=0.94 CI: 0.92-0.96). IRA between R3 and either R1 and R2 was substantial for the

internal (vs R1 ƙ=0.77 CI: 0.74-0.79, vs R2 ƙ=0.78 CI: 0.75-0.81) and only moderate for the external dataset (vs R1 ƙ=0.55 CI: 0.46-0.63, vs R2 ƙ=0.57 CI: 0.48-0.67)
The least common artifact strength given was a score of 1 with 5.9% (n=76). Among all examinations, the most common artifact strength was 2 – i.e., minor artifacts without the potential to impede assessment with 32.3% (n=415) of all examinations. Moderate artifacts (label 3) were identified in 21.7% (n=278) of all cases. In 20.2% (n=260) and 19.9% (n=255) of cases, pronounced (artifact strength 4) and highly significant (artifact strength 5) artifacts were identified respectively. This resulted in significant artifacts (artifact scores 3, 4 and 5) being attributed to 61.8% (n=793) of all DWI MIPs.

3.2 Forecasting performance on internal and external dataset.

Figure 3 shows the ROC for forecasting using low-b-value, or T2w acquisitions for both internal and external data set. Based on the validation set ROC a fixed operating point was identified for both methods (DWI: 0.759, T2w: 0.696). On the internal test set, forecasting using the low-b-value DWI achieved an AuROC of 0.75 (CI: 0.68-0.81) while forecasting on the T2w-FS acquisition achieved an Au-ROC of 0.74 (CI:0.67-0.82) without significant difference in between the methods (p=0.55). At the fixed operating point, this translated into sensitivities of 0.38 (CI: 0.27-0.48), 0.49 (CI: 0.37-0.60) respectively for the low b-values and T2w-FS acquisitions at specificities of 0.93 (CI: 0.88-0.97) and 0.87 (CI: 0.81-0.92) accordingly.
In the external dataset forecasting using low b-value DWI achieved AuROC of 0.62 (CI: 0.54-0.70) and using T2w-FS acquisitions achieved an Au-ROC of 0.71 (CI: 0.64-0.78) with a significant difference between the methods (p =0.013). At the fixed operating point this translated into forecasting sensitivity of 0.21 (CI: 0.13-0.30) at a specificity level of 0.89 (CI:0.83-0.95) when using low b-values. When using T2w-FS at the chosen fixed operating point a sensitivity of 0.61 (CI: 0.52-0.70) and specificity of 0.68 (CI: 0.59-0.77) was achieved. Average computational time for a forecasting a single examination was approx. 4.1 sec.

4 Discussion

Image artifacts are relevant sources of issues in medical imaging – as well in breast MRI examinations. Current strategies mostly aim at detecting images directly after the image acquisition, allowing to repeat the acquisition in case of detected artifacts. Considering the high standardization of image acquisition workflows, it would be desirable to be able to forecast the likelihood of artifacts prior starting acquisition of a specific sequence – during an ongoing examination. This would allow to intervene by adjusting sequence settings before artifacts are actually occurring.

In this study we were able to demonstrate the feasibility to forecast the artifact formation in high b-value DWI acquisition derived MIPs in breast MRI using a neural network wand either the preceding low b-value DWI or a T2w-FS acquisition from the same MRI examination. The achieved performance of such approach showed that at a

high specificity >0.9 such approach could forecast in an internal dataset about 40% of the artifacts using either input acquisition.

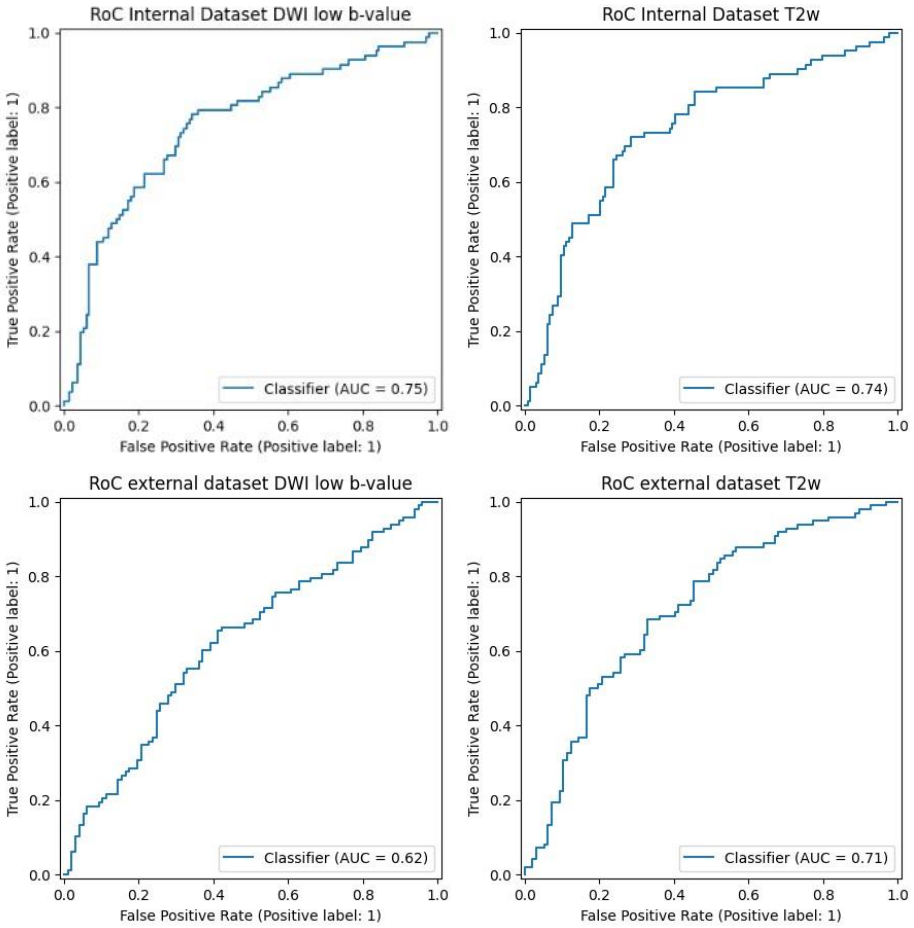

Fig. 3. Receiver operator curves for the Internal and external data set for the two types of acquisitions used for the forecasting of artifact occurrence in high b-value DWI MIPs.

However, external validation revealed limited and input-dependent transferability. For forecasting based on low b-value DWI, sensitivity decreased from 0.38 in the internal dataset to 0.21 in the external dataset, while specificity remained comparatively high (0.89). In contrast, the T2w-FS-based approach achieved a higher sensitivity in the external dataset (0.61), but this was accompanied by a marked decrease in specificity to 0.68. These findings indicate that, although preceding acquisitions contain information associated with subsequent high b-value DWI MIP artifacts, the calibration and clinical operating characteristics of the models did not generalize consistently to the external dataset. Differences in scanner vendors, acquisition protocols, image characteristics, and artifact patterns may have contributed to this reduced transferability.

Our study indicates that the underlying technical challenges inducing artifacts in the DWI MIPs, such as DWI include fat suppression and magnetic susceptibility artifacts, might already be reflected even in the preceding T2w-FS image volumes. Whilst features affecting the likelihood of artifacts occurrence in DWI MIPs seem to be detectable in volumetric data, this should not be confused with indicating an insufficient image quality in all individual slices, which was not specifically assessed in our study. Still technical features and patient related aspects with regards to e.g. the effectiveness of fat saturation techniques seem to be extractable and usable to predict artifacts in the DWI MIPs, even using the T2w-FS acquisition, which might be reasonable considering the T2w-FS and DWI sequence techniques. An investigation of the correlation between the artifact strength on slice level and on the MIP, images should be performed in the future as this task is beyond the scope of this manuscript.

Forecasting of breast MRI artifact occurrence in DCE subtraction MIP images was recently proposed achieving a sensitivity of 30% at a 90% specificity level [8]. The performance of our solution was slightly better in our internal dataset and slightly worse in the external dataset. However, the previous forecasting approach was not tested in an external dataset.

It should be noted that forecasting of artifacts implies as well that some predictions inevitable are inaccurate. This can be the case when a patient was moving during the T2w-FS acquisition but fully lying still during the DWI acquisition and vice versa as such events cannot be correctly identified by a predictive AI algorithm. Again, it thus needs to be emphasized that the approach does not aim at detecting an already existing artifact after an acquisition of a contrast has been performed but to forecast a likelihood for an artifact emerging during the next part of the scan, based on the already acquired part of the scan process.

The approach as developed should be implemented as part of the scan process itself. It's fast computation time approx. 4 seconds allows it to be included into a tightly scheduled workflow of the clinical routine. This would allow to forecast during examination and give direct feedback on the workstation of the MRI scanner to the technicians. Such as scanner-near implementation would allow to adapt the sequence in real-time providing timely feedback and allowing for a sufficient timespan to implement sequence adjustments before starting the next acquisition during the examination.

There are several limitations to this study. First the results as presented are of limited cross-vendor and field-strength generalizability even though the approach was tested in an external test set. The forecasting network was trained on a large dataset of $n \geq 800$ breast MRI examination which was acquired however on only two 3T systems of a single vendor. The systems also did not vary in field strength, which might affect the observed contrast and as such influence the performance of the networks. The significantly lower forecasting performance indicates that for future adaptation of the method to data originating from other vendors and sites training using larger and more diverse datasets will be necessary. Further, our study did not include individual classifications of artifacts type and source, artifact volume and strength and potential strategies to mitigate these sources of artifacts by the technicians. Future studies might include as well analyzing different sequence settings to investigate in-sequence transferability (e.g. between b-values) and in-between sequences (e.g. standard EPI vs. resolve etc.). Such

information will be of importance in order to be able to derive adjustment strategies that can then be used to avoid an artifact predicted in an effective manner. Such a derived adjustment strategy based on the prediction of an artifact was beyond the scope of this initial stud and accordingly not investigated, thus we cannot draw any conclusion to which degree the respective artifacts could have actually been avoided by adjusting sequence parameters or whether the artifacts might have partially been "unavoidable". However, the aim of this study was to investigate the feasibility of such a technique with regards to our hypothesis, thus the chosen methods seem appropriate for the preliminary evaluation and can support further research in this area. Future studies thus might as well investigate at what score or likelihood an intervention is actually desirable and clinically necessary. In our study we tried to address this question by binarizing the scoring, however this scientific approach might be further improved in the future, including as well more detailed information on the type of the artifact.

In conclusion we could demonstrate the feasibility of predicting potentially significant artifacts in high b-value DWI MIPs based on an AI-analysis of the preceding T2w-FS acquisitions or low b-value DWI images from the same patient examination. Further studies seem thus justified to investigate the potential of the method in more details.

Acknowledgments. This work was co-financed by the German Federal Ministry of Education and Research (BMBF) in association with the project SMART SELECT MRI (funding number: FKZ 161B0976), by the Bavarian State Ministry of Science and the Arts in the framework of the bidt Graduate Center for Postdocs and by the Polish National Agency for Academic Exchange within Polish Returns Programme.

References

1. Amornsiripanitch N, Bickelhaupt S, Shin HJ, Dang M, Rahbar H, Pinker K, Partridge SC. Diffusion-weighted MRI for Unenhanced Breast Cancer Screening. Radiology 2019;293(3):504-520. doi: 10.1148/radiol.2019182789
2. Partridge SC, Rahbar H, Murthy R, Chai X, Kurland BF, DeMartini WB, Lehman CD. Improved diagnostic accuracy of breast MRI through combined apparent diffusion coefficients and dynamic contrast-enhanced kinetics. Magn Reson Med 2011;65(6):1759-1767. doi: 10.1002/mrm.22762
3. Partridge SC, McDonald ES. Diffusion weighted magnetic resonance imaging of the breast: protocol optimization, interpretation, and clinical applications. Magn Reson Imaging Clin N Am 2013;21(3):601-624. doi: 10.1016/j.mric.2013.04.007
4. Partridge SC, Nissan N, Rahbar H, Kitsch AE, Sigmund EE. Diffusion-weighted breast MRI: Clinical applications and emerging techniques. J Magn Reson Imaging 2017;45(2):337-355. doi: 10.1002/jmri.25479
5. Kuhl CK, Schrading S, Strobel K, Schild HH, Hilgers R-D, Bieling HB. Abbreviated Breast Magnetic Resonance Imaging (MRI): First Postcontrast Subtracted Images and Maximum-Intensity Projection—A Novel Approach to Breast Cancer Screening With MRI. Journal of Clinical Oncology 2014;32(22):2304-2310. doi: 10.1200/jco.2013.52.5386
6. Bickelhaupt S, Laun FB, Tesdorff J, Lederer W, Daniel H, Stieber A, Delorme S, Schlemmer HP. Fast and Noninvasive Characterization of Suspicious Lesions Detected at Breast Cancer X-Ray Screening: Capability of Diffusion-weighted MR Imaging with MIPs. Radiology 2016;278(3):689-697. doi: 10.1148/radiol.2015150425
7. Ohlmeyer S, Laun FB, Bickelhaupt S, Palm T, Janka R, Weiland E, Uder M, Wenkel E. Ultra-High b-Value Diffusion-Weighted Imaging-Based Abbreviated Protocols for Breast Cancer Detection. Invest Radiol 2021;56(10):629-636. doi: 10.1097/rli.0000000000000784
8. Liebert A, Das BK, Kapsner LA, Eberle J, Skwierawska D, Folle L, Schreiter H, Laun FB, Ohlmeyer S, Uder M, Wenkel E, Bickelhaupt S. Smart forecasting of artifacts in contrast-enhanced breast MRI before contrast agent administration. Eur Radiol 2024;34(7):4752–4763. doi: 10.1007/s00330-023-10469-7
9. Kapsner LA, Balbach EL, Folle L, Laun FB, Nagel AM, Liebert A, Emons J, Ohlmeyer S, Uder M, Wenkel E, Bickelhaupt S. Image quality assessment using deep learning in high b-value diffusion-weighted breast MRI. Scientific reports 2023;13(1):10549. doi: 10.1038/s41598-023-37342-3
10. Kapsner LA, Ohlmeyer S, Folle L, Laun FB, Nagel AM, Liebert A, Schreiter H, Beckmann MW, Uder M, Wenkel E, Bickelhaupt S. Automated artifact detection in abbreviated dynamic contrast-enhanced (DCE) MRI-derived maximum intensity projections (MIPs) of the breast. Eur Radiol 2022;32(9):5997–6007. doi: 10.1007/s00330-022-08626-5
11. Liebert A, Schreiter H, Kapsner LA, Eberle J, Ehring CM, Hadler D, Brock L, Erber R, Emons J, Laun FB, Uder M, Wenkel E, Ohlmeyer S, Bickelhaupt S. Impact of non-contrast-enhanced imaging input sequences on the generation of virtual contrast-

enhanced breast MRI scans using neural network. Eur Radiol 2024:2024.2005.2003.24306067. doi: 10.1007/s00330-024-11142-3
12. Newitt DC, Partridge S, Zhang Z, Gibbs J, Chenevert T, Rosen M, Bolan P, Marques H, Romanoff J, Cimino L. ACRIN 6698/I-SPY2 breast DWI. The Cancer Imaging Archive 2021.
13. Partridge SC, Nissan N, Rahbar H, Kitsch AE, Sigmund EE. Diffusion-weighted breast MRI: Clinical applications and emerging techniques. Journal of magnetic resonance imaging : JMRI 2017;45(2):337-355. doi: 10.1002/jmri.25479
14. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition2017; p. 4700-4708.
15. Jung Y, Hu J. AK-fold averaging cross-validation procedure. Journal of nonparametric statistics 2015;27(2):167-179.
16. Bey R, Goussault R, Grolleau F, Benchoufi M, Porcher R. Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. Journal of the American Medical Informatics Association 2020;27(8):1244-1251.