scientific reports

Image quality assessment using deep learning in high b-value diffusion-weighted breast MRI

Lorenz A. Kapsner[1,2]✉, Eva L. Balbach[1], Lukas Folle[3], Frederik B. Laun[1], Armin M. Nagel[1], Andrzej Liebert[1], Julius Emons[4], Sabine Ohlmeyer[1], Michael Uder[1], Evelyn Wenkel[1] & Sebastian Bickelhaupt[1,5]

The objective of this IRB approved retrospective study was to apply deep learning to identify magnetic resonance imaging (MRI) artifacts on maximum intensity projections (MIP) of the breast, which were derived from diffusion weighted imaging (DWI) protocols. The dataset consisted of 1309 clinically indicated breast MRI examinations of 1158 individuals (median age [IQR]: 50 years [16.75 years]) acquired between March 2017 and June 2020, in which a DWI sequence with a high b-value equal to 1500 s/mm^2 was acquired. From these, 2D MIP images were computed and the left and right breast were cropped out as regions of interest (ROI). The presence of MRI image artifacts on the ROIs was rated by three independent observers. Artifact prevalence in the dataset was 37% (961 out of 2618 images). A DenseNet was trained with a fivefold cross-validation to identify artifacts on these images. In an independent holdout test dataset (n = 350 images) artifacts were detected by the neural network with an area under the precision-recall curve of 0.921 and a positive predictive value of 0.981. Our results show that a deep learning algorithm is capable to identify MRI artifacts in breast DWI-derived MIPs, which could help to improve quality assurance approaches for DWI sequences of breast examinations in the future.

Abbreviations

ADC	Apparent diffusion coefficient
AI	Artificial intelligence
AUPRC	Area under the precision-recall curve
AUROC	Area under the receiver operating characteristic curve
CAM	Class activation map
CNN	Convolutional neural network
CPU	Central processing unit
CV	Cross validation
DCE	Dynamic contrast enhanced
DWI	Diffusion weighted imaging
FGT	Fibroglandular breast tissue
GB	Gigabyte
GPU	Graphics processing unit
IQR	Interquartile range
IRB	Institutional review board
MIP	Maximum intensity projection
Mm	Millimeter
MRI	Magnetic resonance imaging
NPV	Negative predictive value

[1]Institute of Radiology, Universitätsklinikum Erlangen, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Maximiliansplatz 3, 91054 Erlangen, Germany. [2]Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Krankenhausstraße 12, 91054 Erlangen, Germany. [3]Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Martensstraße 3, 91058 Erlangen, Germany. [4]Department of Obstetrics and Gynaecology, Universitätsklinikum Erlangen, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Universitätsstraße 21-23, 91054 Erlangen, Germany. [5]German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. ✉email: lorenz.kapsner@uk-erlangen.de

PPV	Positive predictive value
PR	Precision-recall
RAM	Random-access memory
ROC	Receiver operating characteristic
ROI	Region of interest
s	Seconds
T	Tesla
UHE	University Hospital Erlangen

Mammography screening programs have successfully been implemented to reduce breast cancer-related mortality in females[1]. In breast imaging, magnetic resonance imaging (MRI) has mostly been used for screening in women with a hereditary breast cancer risk[2,3]. MRI examinations of the female breast are routinely performed using a multiparametric approach. Herein, MRI protocols consist of anatomical, non-contrast enhanced sequences and dynamic contrast enhanced (DCE) sequences after the administration of gadolinium containing intravenous contrast agents[4]. More recently, complementary MRI techniques such as diffusion weighted imaging (DWI) have evolved demonstrating a clinical potential for breast cancer screening[5]. DWI sequences reflect the random Brownian motion of water molecules within the tissue. The diffusion process herein has been suggested to correlate to distinct microstructural features of the tissue, e.g., the cellularity or microstructural complexity. With this correlation, DWI is of special interest in oncologic imaging allowing to detect and characterize alterations of diffusion processes within breast tissue[6,7]. Several studies have demonstrated the increased diagnostic accuracy of DWI in complementing the multiparametric MRI for the diagnosis of breast cancer[8].

Recently, abbreviated breast MRI protocols have been evaluated to improve the applicability of breast MRI in high-throughput settings, such as screening examinations[9,10]. Herein, mostly contrast enhanced protocols are considered. However, potential side effects of intravenous application of gadolinium containing contrast agents have been discussed in the last years[11–14], leading to the suspension of some linear contrast agents in Europe[15]. With this, increasing interest has emerged in non-contrast enhanced imaging techniques, such as DWI. Initial studies suggested that abbreviated non-contrast enhanced DWI MRI protocols might provide diagnostic value, however, mostly not reaching the outstanding sensitivity of DCE MRI due to the technical challenges of DWI[16]. While DWI can be performed on most state-of-the-art MRI scanners, achieving a high diagnostic quality and respective quality consistency over time remains a technical challenge in clinical routine. DWI sequences are prone to image artifacts, which may be introduced, for example, by patient motion, insufficient fat saturation, image distortion, and blurring[17]. This currently impedes the diagnostic assessment and limits the potential of DWI in clinical routine.

The application of DWI in breast imaging is gaining interest and first approaches are already investigating the stand-alone-value of the technique. In this context, both quantitative and artificial intelligence (AI) augmented evaluation techniques are becoming more important, for which advanced quality assurance and artifact assessment technologies would be beneficial.

Similar to the application in abbreviated breast DCE-MRI protocols, maximum intensity projections (MIP) can also be computed from DWI sequences in order to reduce the radiologist's initial reading time[16,18,19]. Since MIPs might accumulate (hyperintense appearing) artifacts from the single slices and thus impede the diagnostic assessment if used as an initial visualization approach for lesion detection, we here investigate the capability of a convolutional neural network (CNN) to detect artifacts occurring on high b-value DWI-derived MIPs in a large dataset as a preparatory groundwork for possible future application in abbreviated breast DWI-MRI protocols.

Results

Study cohort and demographics.

A total of 1309 clinically indicated breast MRI examinations fulfilled the inclusion criteria, corresponding to a total of n = 1158 patients (median age at first acquisition: 50 years [IQR: 16.75 years]) that were included in the study. Demographic data and sample characteristics are shown in Table 1. 1020 individuals of the study sample received one, 125 individuals received two, and 13 individuals received a total of three MRI examinations within the study period. The training dataset included 1134 examinations of 984 patients (median age at first acquisition: 50 years [IQR: 16 years]), resulting in a total of 2268 training images. The independent holdout test dataset included 175 examinations of 174 patients (median age at first acquisition: 50 years [IQR: 16 years]), resulting in 350 test images. No significant difference in the distribution of the age could be observed between the training cohort and the test cohort, neither when including only the first examination of each patient (p value: 0.66), nor when also including repeated studies (p value: 0.91).

Interrater agreement.

Regarding individual images, the interrater agreement between the three independent observers was Kappa = 0.577 ($p < 0.001$), corresponding to a moderate agreement according to Landis and Koch[20]. The interrater agreement between the three observers stratified by laterality was Kappa = 0.573 ($p < 0.001$) for images of the left breast and Kappa = 0.579 ($p < 0.001$) for images of the right breast.

Artifacts on DWI sequences.

According to the visual artifact assessment by the three observers, artifacts were present in 37% (961 out of 2618 images) of all images in the dataset. When considering both regions of interest (ROIs) together for each MRI examination, artifacts were present bilaterally in 26% (340), whereas unilateral artifacts occurred in 21.5% (281) of the examinations and a total of 52.6% (688) examinations were free from artifacts.

Variable	Overall sample	Training dataset	Test dataset
N patients	1158	984	174
Age			
Median age (IQR) [years]	50 (17)	50 (17)	50 (16)
Median age (IQR) at first acquisition [years]	50 (16.75)	50 (16)	50 (16)
N examinations	1309	1134	175
N repeated examinations per patient			
One examination	1020	847	173
Two examinations	125	124	1
Three examinations	13	13	0
N images	2618	2268	350
Left breast	1309	1134	175
Right breast	1309	1134	175
N artifacts (%)	961 (37%)	777 (34%)	184 (53%)
Left breast	466 (36%)	379 (33%)	87 (50%)
Right breast	495 (38%)	398 (35%)	97 (55%)

Table 1. Demographic data, sample characteristics, and target class distribution across the training dataset and the independent holdout test dataset. *IQR* interquartile range.

In the training dataset, artifacts were present in 34% (777 out of 2268 images), whereas in the test dataset, artifact prevalence was 53% (184 out of 350 images) of all images, corresponding to a statistically significant difference (*p* value: < 0.001).

Artifact detection using deep learning. Figure 1 shows the receiver operating characteristic (ROC) curve (left column) and the precision-recall (PR) curve (mid column) of the resulting DenseNet models from the fivefold cross-validation (CV) (row 1). Row 2 of Fig. 1 shows the corresponding performance curves of the ensemble of the 5 CV models computed using the predictions for the holdout test dataset. The training and validation loss curves for the models averaged over the 5 CV folds are shown in the right column of Fig. 1. The training performance measures for each model from the 5 CV-folds as well as the corresponding best epochs are given in supplemental Table S2. All 5 CV models were applied to predict the outcome in the independent holdout test dataset. On average, the DenseNet achieved an area under the PR curve of 0.915 (± 0.004) with a positive predictive value (PPV) of 0.953 (± 0.013) and a specificity of 0.970 (± 0.008) for the detection of significant artifacts on breast DWI MIPs in the independent holdout test dataset (Table 2). The DenseNet ensemble—created by calculating the arithmetic mean of the predicted probabilities of the 5 models for each image in the holdout test dataset and considering images with an averaged probability of > 0.5 to contain artifacts—showed an area under the PR curve of 0.921, with a PPV of 0.981 and a Specificity of 0.988, respectively (Table 2, column 8).

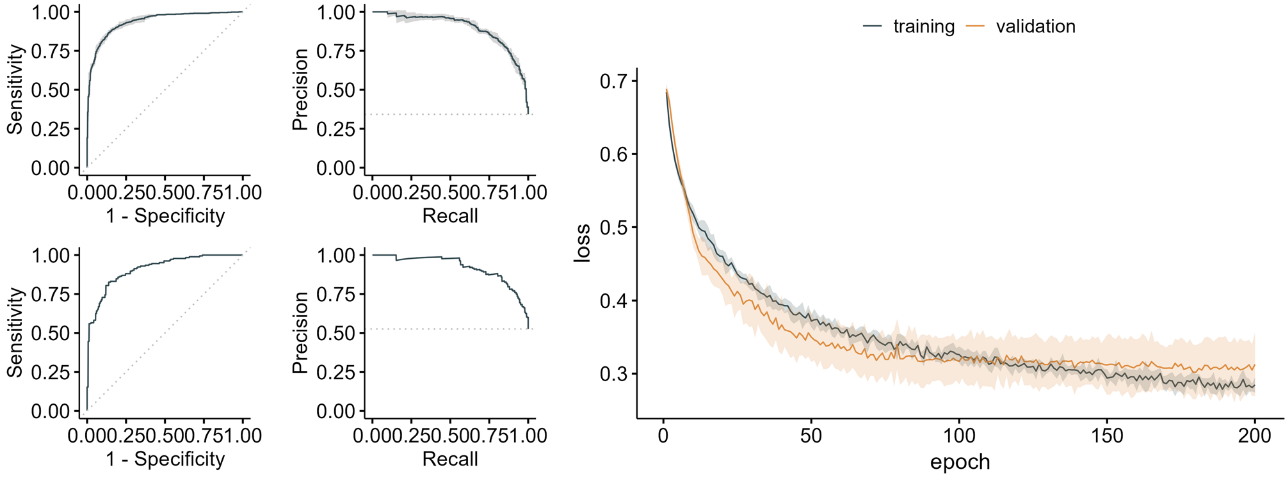

Figure 1. Deep learning results. The figure shows the receiver operating characteristic (ROC) curve (left column) and the precision-recall (PR) curve (mid column) and the loss curves (right column) for the DenseNet architecture. Row 1: ROC and PR curve averaged over the 5 cross-validation folds. Row 2: ROC and PR curve for the ensemble's prediction on the independent holdout test dataset. The training loss (dark blue) and the validation loss (yellow) curves are averaged over 5 CV folds.

Variable	M1	M2	M3	M4	M5	Mean (SD)	DenseNet ensemble
Accuracy	0.746	0.757	0.737	0.760	0.746	0.749 (±0.009)	0.763
AUROC	0.904	0.910	0.906	0.895	0.900	0.903 (±0.006)	0.910
AUPRC	0.917	0.921	0.914	0.912	0.912	0.915 (±0.004)	0.921
Sensitivity	0.538	0.571	0.533	0.560	0.549	0.550 (±0.016)	0.560
Specificity	0.976	0.964	0.964	0.982	0.964	0.970 (±0.008)	0.988
PPV	0.961	0.946	0.942	0.972	0.944	0.953 (±0.013)	0.981
NPV	0.656	0.669	0.650	0.668	0.658	0.660 (±0.008)	0.669

Table 2. Holdout test dataset performance. The table shows the performance measures of the 5 DenseNet cross-validation (CV) models (columns 2–6) along with their averaged performance (column 7) when applied to predict the outcome in the independent holdout test dataset (n = 350 images). Column 8 shows the performance of the DenseNet ensemble. The ensemble was created by calculating the arithmetic mean of the predicted probabilities of the 5 models for each image in the holdout test dataset and considering images with an averaged probability of > 0.5 to contain artifacts. Mean: (unweighted) average over 5 CV folds. SD: (unweighted) standard deviation over 5 CV folds. *AUROC* area under the ROC curve, *AUPRC* area under the precision-recall curve, *PPV* positive predictive value, *NPV* negative predictive value.

Examples of class activation maps (CAMs) for true positive, true negative, false positive and false negative predicted images are shown in Figs. 2, 3, 4 and 5. The CAMs were computed using the model with the highest area under the PR curve during training (i.e. CV fold 5, AUPRC = 0.911; see supplemental Table S2). The GradCAM++-results suggest that the network is capable to detect artifacts well (Fig. 2). From the generated CAM images can also be derived that in the absence of high signal intensities in the breast tissue, the whole organ seems to contribute to the class assignment for correctly classified artifact-free images, whereas in the presence of high signal intensities, the most important class-discriminative regions seem to correlate with areas that include blood vessels and fibroglandular breast tissue (FGT) (Fig. 3). The latter observation is quite in line with our previous results for the artifact detection in MRI-derived DCE-MIPs where a sharp demarcation of contrast agent-containing blood vessels from the surrounding breast tissue was considered to guide the neural

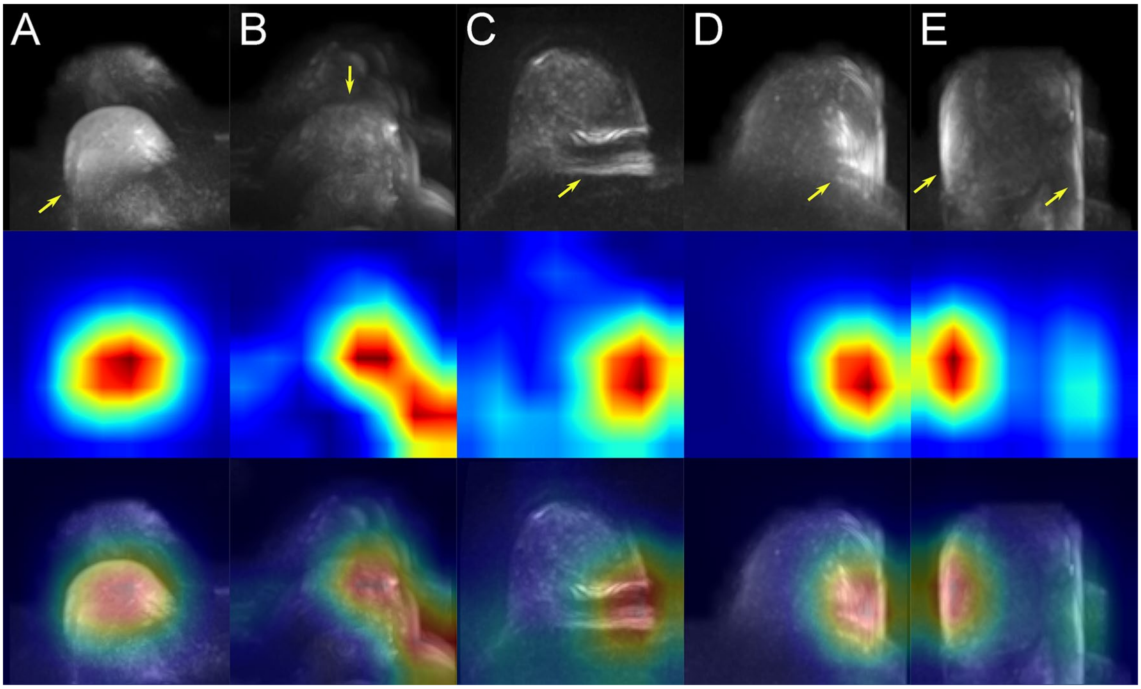

Figure 2. Class activation maps (examples): true positives. Original images are shown in row 1 (**A**–**E**). The Grad-CAM++ visualization for the predicted class (i.e. prediction/ground truth = 'artifact') are shown in row 2 and images of row 3 show the combined images. The heatmaps' color gradient shows from blue to red the relevance of each pixel for the inference of the respective class. Artifacts in DWI often originate from multiple technical and/or patient-related sources that may be interdepend and thus it is not always possible to attribute one specific artifact source. The arrows mark regions of artifacts within the images with possible contributing factors of insufficient fat suppression (e.g. visible in **A**), ghosting artifacts related to silicone implants (e.g. **B**), artifacts related to combinatory effects of distortion and insufficient fat suppression (e.g. visible in **C**) and related to remaining surface coil flares (e.g. visible in **D**, **E**).

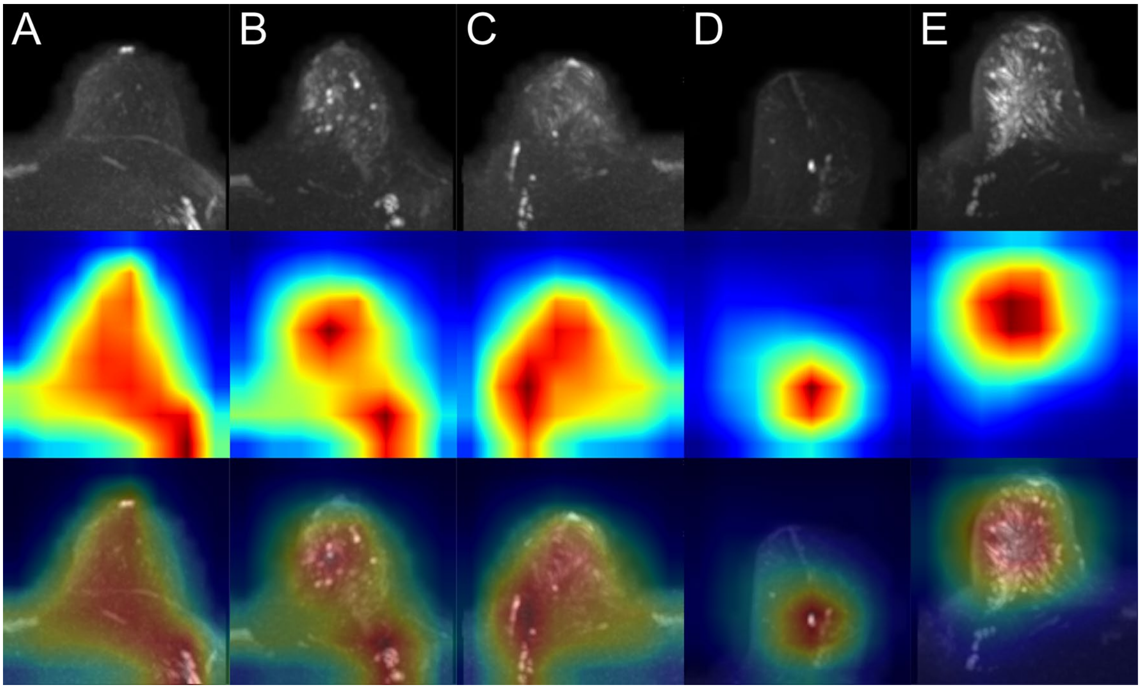

Figure 3. Class activation maps (examples): true negatives. Original images are shown in row 1 (**A**–**E**). The Grad-CAM++ visualization for the predicted class (i.e. prediction/ground truth = 'artifact-free') are shown in row 2 and images of row 3 show the combined images. The heatmaps' color gradient shows from blue to red the relevance of each pixel for the inference of the respective class.

network (NN) towards the negative class (both, true and false negative)[21]. For DWI MIPs, in addition to the lack of contrast agent administration, the demarcation of blood vessels and FGT is not as clear and sharp as in DCE MIPs. Nevertheless, the CAM results for the DWI MIPs let us assume that the mentioned attributes could be features used by the NN to distinguish between artifact-free and artifact-containing images. This is further underlined by the CAM results of the false negative classifications (Fig. 5), where image regions with high intensity values, such as areas containing blood vessels or FGT, seem to have guided the NN towards its (false) decision (i.e. falsely classifying them as artifact-free; rows 2–3 in Fig. 5), overseeing artifacts present in other image regions (yellow arrows, and rows 4–5 in Fig. 5). In contrast, the most important class-discriminative regions for false positive classifications seem to correlate with slightly blurry appearing image regions (rows 2–3 in Fig. 4). When providing the corresponding ground truth (i.e. artifact-free) to the computation of the CAM images, the class-discriminative regions are again overlapping with regions that contain blood vessels and FGT (rows 4–5 in Fig. 4).

Figure 6 shows DWI MIP ROIs of 15 clinical cases with BI-RADS 6 lesions from our dataset with various gradations of artifacts. Images A–E represent 5 cases without MRI artifacts. Images F–J are examples, where present artifacts have a moderate influence on the diagnostic evaluation, whereas images K–O contain examples with artifacts that would significantly impede a diagnostic evaluation. A corresponding graphic with BI-RADS 5 lesion is given in supplemental Fig. S1.

Discussion

Here we demonstrated the capability of an NN to detect MRI artifacts on qualitative DWI-derived MIPs. The DenseNet was trained on more than 2200 images of 1134 individual MRI examinations. The ensemble of the 5 CV models showed an area under the PR curve of 0.921 on the independent holdout test dataset with a PPV of 0.981 and a Specificity of 0.988, respectively. These results indicate that the ensemble classifier was able to detect artifact-containing images in the test dataset quite well.

MRI examinations of the female breast increasingly include DWI in the sequence protocol. DWI allows to detect suspicious focal and non-focal alterations of tissue diffusivity and to provide quantitative measures of derived parameters such as the apparent diffusion coefficient (ADC). One advantage of DWI is that it does not require the application of gadolinium containing intravenous contrast agents. The relation of gadolinium containing contrast agents to findings of deposition in the human body as well as in the environment has been under investigation over the past years[11–14, 22–25]. This has intensified the research on non-contrast enhanced MRI techniques of which DWI is of special interest. With the potential to perform a DWI MRI in only a couple of minutes and to avoid both ionizing irradiation and the application of contrast agents of any kind, DWI has also been investigated in breast MRI both as an expansion of DCE MRI and as a possible stand-alone application in the context of breast cancer screening[5–8].

Whilst the potential of DWI has been demonstrated in several studies, the sequence remains technically challenging and prone to artifacts[17]. This can impede the diagnostic evaluation of DWI images, which is further

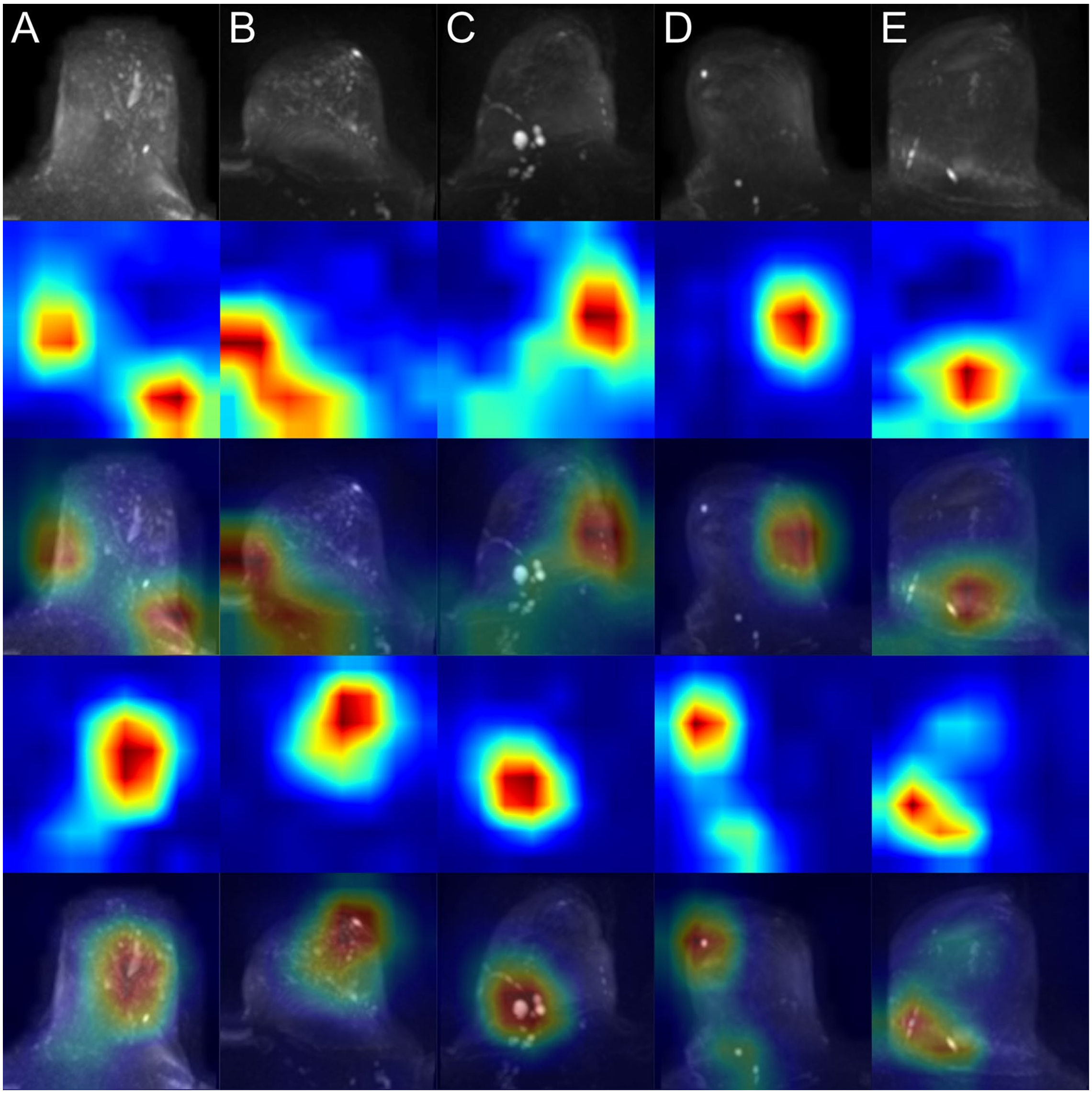

Figure 4. Class activation maps (examples): false positives. Original images are shown in row 1 (**A**–**E**). Rows 2–3 show the Grad-CAM++ visualization and the combined image for the predicted class (i.e. prediction = 'artifact'). Rows 4–5 show the Grad-CAM++ visualization and the combined image for the actual class (i.e. ground truth = 'artifact-free'). The heatmaps' color gradient shows from blue to red the relevance of each pixel for the inference of the respective class. A detailed interpretation of the original images and Grad-CAM++ visualizations of the (falsely) predicted class (rows 2–3) showed indeed corresponding areas of slight blurry and/or hyperintense appearing image regions, which, however, were rated as not significant by ≥ 2 out of three independent raters.

aggravated when image postprocessing techniques are applied, such as the generation of DWI derived MIPs for initial diagnostic assessment. The use of MIPs in breast MRI has largely evolved based on a publication by Kuhl et al. demonstrating the capability of DCE MRI derived MIPs to provide a high diagnostic accuracy while simultaneously reducing the reading times for radiologists[9]. Similar to DCE MRI, DWI also offers the possibility of generating MIPs from high b-value acquisitions that provide sufficient suppression of the FGT, leaving mainly the areas of potential interest visible in the image. Feasibility studies already demonstrated a high diagnostic accuracy when using the combination of high b-value DWI and MIPs in the initial reading of breast MRI[16, 18]. In general, reading schemes that involve MIPs depend on a high image quality to a large extent, because artifacts could potentially cover suspicious lesions on the 2D image and MIPs are particularly prone to artifacts as hyperintense artifacts may accumulate from the single slices into the MIP image. In the context of DWI, the generation of MIPs can be technically challenging as DWI itself is prone to image artifacts, even if the basic sequence has intensively been adjusted in order to avoid them.

Especially in the context of abbreviated MRI protocols, high image quality is particularly important since complementary image sequences might not be available to compensate for potential artifacts. In our dataset, artifacts were present in almost half of all examinations. Most common sources of artifacts include patient

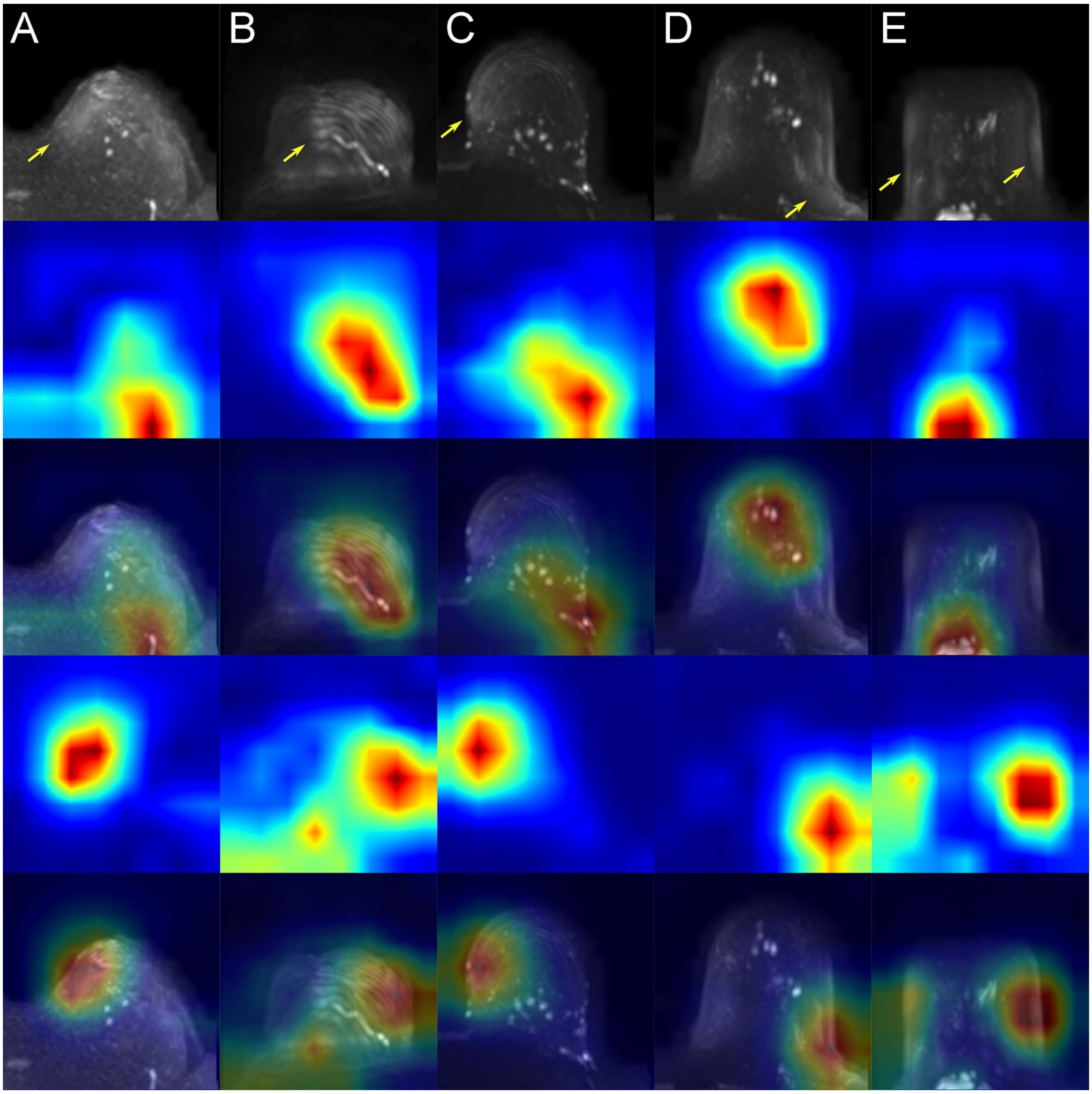

Figure 5. Class activation maps (examples): false negatives. Original images are shown in row 1 (**A**–**E**). Rows 2–3 show the Grad-CAM++ visualization and the combined image for the predicted class (i.e. prediction = 'artifact-free'). Rows 4–5 show the Grad-CAM++ visualization and the combined image for the actual class (i.e. ground truth = 'artifact'). The heatmaps' color gradient shows from blue to red the relevance of each pixel for the inference of the respective class. Artifacts in DWI often originate from multiple technical and/or patient-related sources that may be interdepend and thus it is not always possible to attribute one specific artifact source. The arrows mark regions of artifacts within the images with possible contributing factors of insufficient fat suppression (e.g. visible in **A**), artifacts emerging in the MIP corresponding to the repetition of thickened cutis projected into MIP (no technical artifact) (as visible in **B** and **C**, with the latter including artifacts of insufficient fat suppression), and artifacts associated to remaining surface coil flares (e.g. visible in **D**, **E**).

movement and insufficient fat suppression. Both types of artifacts can be difficult to avoid in advance, so an immediate assessment of image quality—potentially running directly on the MRI machine—could be of great importance, for example, to trigger actions that allow further handling of images with artifacts or even to repeat acquisitions. For example, the artifact detection algorithm could be implemented in a setting in which reading protocols include the initial assessment of a MIP image, such as abbreviated MRI protocols applied in a breast cancer screening context. During the image post-processing, the NN could label or select MIPs with sufficient image quality to be presented to the radiologist, as for artifact-containing MIPs it would mostly be unnecessary to open the MIP and instead single slice sequences could be read directly. Furthermore, hypothetically, the artifact-detection could also be applied during the ongoing examination and in the case of a poor image quality, for example, the acquisition of the DWI sequence could be repeated. One could also imagine to acquire DCE sequences only in the case of a detected poor image quality on the DWI MIP, however, all of these scenarios were not evaluated prospectively in our study. In addition to its application in clinical workflows, such an algorithm

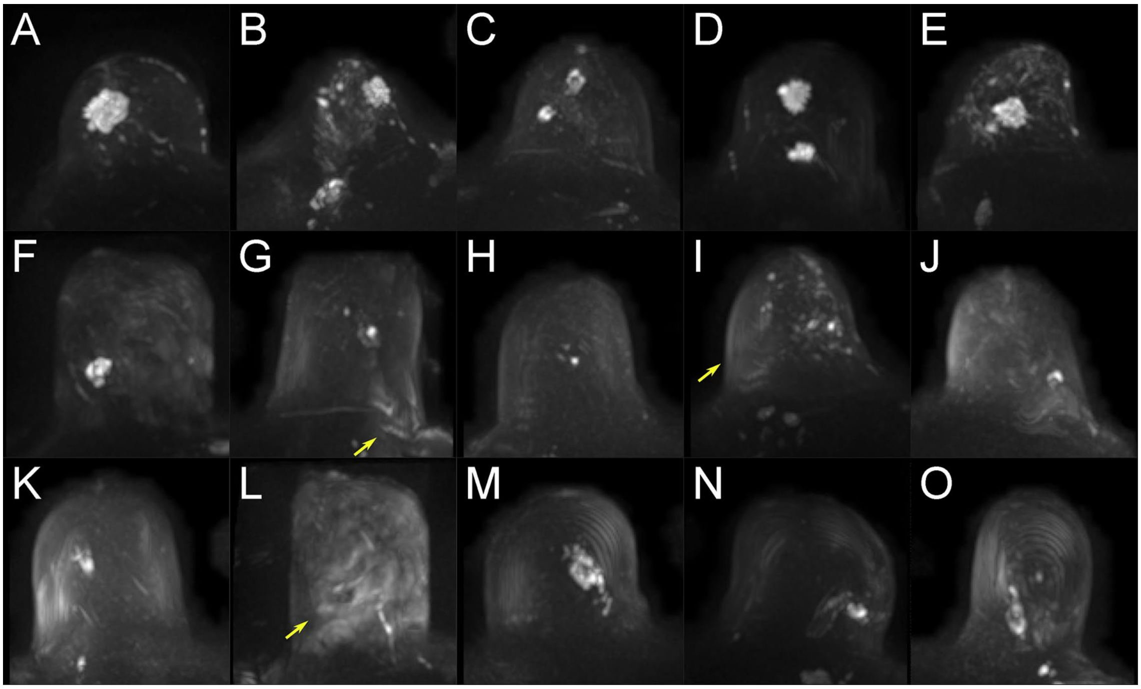

Figure 6. BI-RADS 6 lesions in clinical cases (examples). Each tile of the figure presents the left or right breast of one clinical case with a diagnosed BI-RADS 6 lesion. Row 1 (**A**–**E**) shows images without the presence of artifacts. Row 2 (**F**–**J**) shows images that contain artifacts with no or moderate influence on the diagnostic assessment. Row 3 (**K**–**O**) shows images with artifacts that significantly impede the diagnostic evaluation. Artifacts in DWI often originate from multiple technical and/or patient-related sources that may be interdepend and thus it is not always possible to attribute one specific artifact source. The arrows mark regions of artifacts within the images with possible contributing factors related to distortion (e.g. visible in **G**), insufficient fat suppression (e.g. visible in **I**, **L**), related to remaining surface coil flare (e.g. also visible in **I**), and pulsation related signal drops in DWI (e.g. also visible in **L**).

could also be used to detect artifacts in the curation and preparation of datasets for deep learning tasks, such as the automated detection of breast lesions.

In this feasibility study, we focused on applying the artifact detection directly on 2D MIP images, as they are a diagnostic tool increasingly investigated for its diagnostic value in abbreviated study protocols, and as outlined above, MIPs harbor a particular risk for artifact representation. The DenseNet NN architecture used in this study has also been applied previously by our group to detect MRI artifacts on DCE MIPs[21]. While the NN achieved a high area under the PR curve averaged over the 5 CV folds (supplemental Table S2) for detecting artifacts in DWI MIPs, not all artifacts were correctly classified and false positives occurred, especially in case of image regions that appeared blurred, as shown in the example CAM images in Fig. 4. Strictly speaking, the network here indeed detected slight artifacts in images with negative class labels, which, however, were rated as not significant by ≥ 2 out of the three independent raters.

The issue of MRI associated artifacts is well known and different solutions exist to address, for example, patient motion in MRI, which are summarized in a review article by Zaitsev et al.[26]. Common mitigation strategies for motion artifacts include, for example, *motion prevention* (e.g. training, breathhold, sedation in case of patients not able to comply with instructions such as children, etc.). *artifact reduction* (e.g. faster imaging, phase reordering, etc.), and *motion correction* (e.g. navigators, pro-/retrospective correction)[26]. Commonly employed post-processing techniques for motion correction include image registration, which aims at ensuring spatial alignment of separate images[27] and several algorithms have been adapted to breast MRI (e.g.[28–31]). While all of these approaches aim to improve the image quality of the acquired scan data, the NN presented here can be considered as a complementary method that could capture the remaining artifacts in MIPs after the aforementioned methods have been applied.

Our study has several limitations. First, the major limitation is the binary labeling used this study as artifacts can occur with a wide spectrum in terms of their severity and subsequent clinical relevance, which is most likely not represented satisfactorily by two classes. A multi-reader assessment with three independent raters and a best-of-n approach to define the target label was performed in order to establish a reliable ground truth for the presence of artifacts. However, the interrater agreement with a Kappa-statistic of 0.58 in our study—corresponding to a moderate strength of agreement according to Landis and Koch[20]—indicates certain challenges associated with artifact classifications in novel imaging techniques. Therefore, the development of acknowledged and objective assessment criteria to rate the severity of MRI artifacts would be of high interest, which would also allow for a repetition of our study with a more finely granulated labeling regarding artifact severity. Second, variations in the DWI sequence settings between the different MRI examinations resulted in heterogeneous image quality. However, this was not further explored in our study, and therefore the results are not suitable to make statements neither regarding the relationship between sequence settings and artifact susceptibility nor about the influence of the sequence settings on the artifact detection performance of the NN. Third, we only used the MIPs as input for the NN, so that the influence of other variables on the automated artifact detection, e.g. scanner-related features or demographic characteristics remains unclear. Future studies could extract more features from the data, such as patient age, body-mass-index, breast size, breast density, and patient age, and include these into the model building process as well. Furthermore, our study was performed using 2D DWI MIPs, which are currently of subordinated clinical relevance as compared to the 3D DWI sequences. As our study was intended as a preparatory groundwork to investigate the capability of a CNN to identify artifacts on DWI images, we considered the artifact detection on MIPs as a good starting point. Furthermore, considering the proneness to artifacts of DWI combined with the potential aggravation of artifacts when computing MIPs, we identified this imaging post-processing technique as one that would probably benefit strongly from an artifact detection algorithm, especially when using MIPs for the initial reading, e.g., as in abbreviated MRI protocols. However, future work should investigate and evaluate artifact detection on 3D DWI sequences as well and compare the results with the detection on MIPs and perhaps identify other potential areas of application. Future work in this field is important especially since MIPs accumulate hyperintense artifacts, whereas hypointense artifacts may go unnoticed since probably not being detectable due to the basic technical principle of MIPs. Another limitation is that there was no stratification on a patient level when creating the CV folds, potentially leading to images of the same patient being present in both, the CV folds' training and validation dataset. In theory, this could lead to an overly optimistic validation error if one would assume that the network would learn features from a MIP of a patient from the training dataset, which would help to better infer the artifact-class of either the contralateral breast or a MIP from another examination of the same patient that could potentially be included in the corresponding validation dataset. However, DWI artifacts originate mostly from technical issues (such as poor shimming, insufficient fat saturation, magnetic susceptibility differences or eddy currents[32]) or from patient movement, which are rather examination-specific features than dependent on patient characteristics and thus, such a stratification might not be important for the task at hand. To address this potential issue, an independent holdout test dataset was formed to get an unbiased final model evaluation. This test dataset contained only new patients that were not yet available in the training dataset. For the final model evaluation, the trained models were applied to predict the artifact class in this independent holdout test dataset. The consistency of the model performance results between the CV-training and the evaluation on the independent holdout test dataset indicates that the NN was indeed capable of learning features that are related to MRI artifacts on DWI-MIPs of the breast. Another restriction of this study is imposed by the use of high b-value DWI series. It remains unclear as by this study to what degree the results might be achievable as well at lower b-values. Furthermore, we did not have the possibility to apply our approach across different MRI vendors. Thus, future studies are needed to assess the aspect of generalizability of this algorithm to images of different b-values, image quality, and scanner systems. Last but not least, another limitation could be that the dataset represents a retrospective university hospital patient cohort and thus we cannot state to what degree similar patient- and artifact characteristics might be found, e.g., in a screening population and whether using such an artifact detection algorithm in clinical routine would actually lead to improvements in the reading process.

In conclusion, we here demonstrated an NN that detects artifacts in breast DWI-derived MIPs. The network was able to identify artifact-containing images in the independent holdout test dataset quite well and might serve as a starting point to develop more sophisticated quality assurance methods for breast MRI DWI sequences in the future.

Materials and methods

Study sample and ethics approval.

This retrospective study included breast MRI examinations from March 2017 to June 2020. The study was approved by the ethics committee of the Friedrich-Alexander-University (FAU) Erlangen-Nürnberg, waiving the need for informed consent. The authors declare that this research was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects. Inclusion criteria were the acquisition of a clinically indicated breast MRI at the Institute of Radiology of the University Hospital Erlangen (UHE) with at least one DWI sequence that was acquired with a high b-value equal to 1500 s/mm^2. For eligible examinations, additional series derived from the originally acquired DWI sequence (e.g. motion-correction or otherwise post-processed series) were excluded from the dataset. This cohort is partially overlapping with a previously reported study sample, in which we investigated the automated detection of artifacts using deep learning on DCE sequences[21]. However, the previous work included data from the years 2015 to 2019 and the DWI sequences have not been assessed there.

MRI protocol.

All MRI examinations were performed with a clinical indication at the hospital's routine MRI scanners (1.5 and 3 Tesla MRI; Model names: Magnetom Aera, Vida, and Skyra from Siemens Healthineers, Erlangen, Germany). The routine MRI protocols consisted of morphologic, T2-weighted, dynamic contrast-enhanced T1-weighted, and DWI sequences. The standard positioning of patients during the examination was in prone position with arms laterally to the body. DWI acquisitions were performed using different sequence types and ranges of b-values, commonly including b = 0, b = 750, and b = 1500 s/mm^2. Supplemental Table S1 gives a detailed overview of the different settings on which the DWI sequences of this study were based.

Data processing.

Imaging data were queried from the local routine picture archiving and communication system and transferred to evaluation workstations within the UHE Institute of Radiology. The DWI MRI sequences were processed in a similar manner as previously described for the DCE sequences[21]. An in-house developed Python script was used to represent the voxels with the highest intensity values along the z-axis (i.e. head-feet direction) on a transversal 2D image in order to compute a MIP from each individual qualitative DWI sequence. Quadratic tiles with a dimension of $1/2\ image_{width} \times 1/2\ image_{width}$ containing the left and right breast as ROIs were cropped out from the upper left and right parts of each MIP and saved as JPEG files for the visual artifact assessment by the three observers.

Visual artifact assessment. All processed images were labeled in binary manner by three independent observers (S.B., E.L.B., L.A.K.) with regard to the presence of significant artifacts (1 = artifacts present; 0 = no artifacts present). Artifacts on the DWI MIPs were visually evaluated, with sources and possible characteristics of artifacts derived from recent literature as found, e.g., in Partridge et al.[33]. The rating was performed regardless of whether or not the specific artifacts covered a significant breast lesion. All artifacts that could mask a lesion were therefore considered significant, regardless of the actual clinical relevance in the individual examination. A final label was computed for each image using the *best-of-n* approach, i.e., if $\geq$ 2 raters classified an image as to be of the positive class, the final label of the image was "artifacts present". This final label was used for the subsequent experiments and data analyses.

Image preprocessing and image augmentation. Image preprocessing was performed in Python (version 3.8.5) using *SimpleITK* version 2.0.2[34] following the procedure as previously described[21]. From each DWI volume, a MIP was computed as described above. To preprocess the cropped ROIs for training the deep learning networks, the images were further normalized (mean = 0, standard deviation = 1), resized to 256 × 256 pixels, and saved as *NumPy* arrays[35]. Image augmentation included random rotation (probability: 0.5, maximum angle: 180 degrees), random flip across x-axis and y-axis (probability: 0.5) and random zoom (probability: 0.5, minimum zoom: 0.5, maximum zoom: 1.5).

Deep learning. A DenseNet121[36] was trained to classify the presence or absence of artifacts on qualitative DWI MIPs of the left and right breast (i.e. a binary classification), utilizing the network architecture already implemented in the *monai* library version 0.4.0[37], which builds upon the *PyTorch* deep learning framework version 1.7.1[38]. The code was further organized with *PyTorchLightning* version 1.2.4[39], a wrapper for *PyTorch* that is tailored to application in research. The training of the deep learning network was carried out on a Tesla V100 graphical processing unit (GPU) with 32 GB memory and an Intel® Xeon® CPU E5-2698 v4 @2.20 GHz (20 cores) with 256 GB RAM. The methodology to carry out the experiments was aligned to our previous work[21]. A training dataset was formed from the examinations acquired up to and including the year 2019. An independent holdout test dataset set was formed from the examinations that were acquired in 2020 using only patients, which were not already included in the training dataset. The model parameters were optimized with a grid search on the training dataset, partitioning it by 80% to 20% for model training and evaluation (data not shown). Due to the observed class imbalance in the training dataset with the majority of images belonging to the negative class (i.e. no artifacts present), we primarily focused on the PR curve for evaluating the model performance, since it is said to more reliable in datasets with an imbalanced target class than the ROC curve[40]. The as such optimized model parameters were validated with a fivefold CV on the training dataset. The CV folds were generated in a stratified manner to ensure a similar artifact prevalence across the folds. For each fold, the training data was further randomly split into 80% that were actually used for training and 20% that were used for validation, i.e., monitoring loss and performance metrics during network training. Binary cross entropy with logits from *PyTorch* was employed as loss function. Class probabilities were calculated using the softmax function. We employed the 'Adam'[41] optimizer with a weight decay of $1e^{-5}$ and an initial learning rate of $\eta = 3e^{-5}$. The DenseNet121 network was further parameterized with a dropout probability of 10%. All models were trained for 200 epochs with a batch size of 128, resulting in 12 steps per epoch. For each CV model, the weights from the epoch with the lowest validation loss observed within 200 epochs were chosen to predict the respective CV fold's validation dataset and the observations in the independent holdout test dataset. In accordance with our previous work[21], we also created an ensemble from the 5 CV models to predict the artifact presence in the independent holdout test dataset. The ensemble was created by calculating the arithmetic mean of the predicted probabilities of the 5 CV models for each image in the test dataset and considering images with an averaged probability of $>$ 0.5 to contain artifacts.

Statistical analysis. The statistical analyses were performed with the R software version 4.2.1[42]. Summary statistics were computed in base R[42]. Fleiss' Kappa[43] was computed to test for interrater agreement of the artifact assessment between the three raters using the R package *irr* version 0.84.1[44]. Model metrics were calculated using the *mlr3measures* package version 0.5.0[45]. Graphics were created with the R packages *ggplot2* version 3.3.6[46], *ggpubr* version 0.4.0[47] and *precrec* version 0.12.9[48]. Wilcoxon's rank sum test (two-sided)[49, 50] was used for comparing the distribution of continuous variables between two groups. Differences between two categorical variables were assessed with the Chi-squared test[51]. Significance level was set to α=0.05. No correction for multiplicity was performed. Class activation maps (CAMs) depict so-called class-discriminative regions, which can be displayed as heatmaps that color-code image regions that are deemed important by the CNN classifier to identify the inferred class[52]. This method builds upon the finding that components of CNNs inherently have object detection capabilities[53] that allow to efficiently localize discriminative regions in the image, e.g. the pixels that discriminate between the categories a classifier was trained with[52]. The CAMs were generated for all images from the independent holdout test dataset, using the model that achieved the highest area under the PR curve during the fivefold CV. The CAM images were computed with the GradCAM++ algorithm[54] provided with the *monai* library[37] version 0.4.0, and the resulting images were assessed visually by an experienced board-certified radiologist (S.B.).

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to internal data transfer policies but are available from the corresponding author on reasonable request.

Code availability

The code used to perform the experiments is available at GitHub: https://github.com/kapsner/dwi_mip_artifact_classifier.

Received: 10 November 2022; Accepted: 20 June 2023
Published online: 29 June 2023

References

1. Hübner, J., Katalinic, A., Waldmann, A. & Kraywinkel, K. Long-term incidence and mortality trends for breast cancer in Germany. *Geburtshilfe Frauenheilkd.* **80**, 611–618. https://doi.org/10.1055/a-1160-5569 (2020).
2. Mann, R. M., Kuhl, C. K. & Moy, L. Contrast-enhanced MRI for breast cancer screening: Breast MRI for screening. *J. Magn. Reson. Imaging.* **50**, 377–390. https://doi.org/10.1002/jmri.26654 (2019).
3. Krassuski, L. M. *et al.* Decision aids for preventive treatment alternatives for BRCA1/2 mutation carriers: A systematic review. *Geburtshilfe Frauenheilkd.* **81**, 679–698. https://doi.org/10.1055/a-1326-1792 (2021).
4. Orel, S. G. & Schnall, M. D. MR imaging of the breast for the detection, diagnosis, and staging of breast cancer. *Radiology* **220**, 13–30. https://doi.org/10.1148/radiology.220.1.r01jl3113 (2001).
5. Amornsiripanitch, N. *et al.* Diffusion-weighted MRI for unenhanced breast cancer screening. *Radiology* **293**, 504–520. https://doi.org/10.1148/radiol.2019182789 (2019).
6. Sinha, S., Lucas-Quesada, F. A., Sinha, U., DeBruhl, N. & Bassett, L. W. In vivo diffusion-weighted MRI of the breast: Potential for lesion characterization. *J Magn. Reson. Imaging.* **15**, 693–704. https://doi.org/10.1002/jmri.10116 (2002).
7. Woodhams, R. *et al.* Diffusion-weighted imaging of the breast: Principles and clinical applications. *Radiographics* **31**, 1059–1084. https://doi.org/10.1148/rg.314105160 (2011).
8. Zhang, L. *et al.* Accuracy of combined dynamic contrast-enhanced magnetic resonance imaging and diffusion-weighted imaging for breast cancer detection: A meta-analysis. *Acta Radiol.* **57**, 651–660. https://doi.org/10.1177/0284185115597265 (2016).
9. Kuhl, C. K. *et al.* Abbreviated breast magnetic resonance imaging (MRI): First postcontrast subtracted images and maximum-intensity projection—A novel approach to breast cancer screening with MRI. *J. Clin. Oncol.* **32**, 2304–2310. https://doi.org/10.1200/JCO.2013.52.5386 (2014).
10. Deike-Hofmann, K. *et al.* Abbreviated MRI protocols in breast cancer diagnostics: Abbreviated breast MRI. *J. Magn. Reson. Imaging* **49**, 647–658. https://doi.org/10.1002/jmri.26525 (2019).
11. Errante, Y. *et al.* Progressive increase of T1 signal intensity of the dentate nucleus on unenhanced magnetic resonance images is associated with cumulative doses of intravenously administered gadodiamide in patients with normal renal function, suggesting dechelation. *Invest. Radiol.* **49**, 685–690. https://doi.org/10.1097/RLI.0000000000000072 (2014).
12. Kanda, T., Ishii, K., Kawaguchi, H., Kitajima, K. & Takenaka, D. High signal intensity in the dentate nucleus and globus pallidus on unenhanced T1-weighted MR images: Relationship with increasing cumulative dose of a gadolinium-based contrast material. *Radiology* **270**, 834–841. https://doi.org/10.1148/radiol.13131669 (2014).
13. McDonald, R. J. *et al.* Intracranial gadolinium deposition after contrast-enhanced MR imaging. *Radiology* **275**, 772–782. https://doi.org/10.1148/radiol.15150025 (2015).
14. Radbruch, A. *et al.* Gadolinium retention in the dentate nucleus and globus pallidus is dependent on the class of contrast agent. *Radiology* **275**, 783–791. https://doi.org/10.1148/radiol.2015150337 (2015).
15. European Medicines Agency, EMA's final opinion confirms restrictions on use of linear gadolinium agents in body scans. (2017) (accessed 27 April 2021). https://www.ema.europa.eu/en/documents/referral/gadolinium-article-31-referral-emas-final-opinion-confirms-restrictions-use-linear-gadolinium-agents_en.pdf.
16. Bickelhaupt, S. *et al.* Fast and noninvasive characterization of suspicious lesions detected at breast cancer X-ray screening: Capability of diffusion-weighted MR imaging with MIPs. *Radiology* **278**, 689–697. https://doi.org/10.1148/radiol.2015150425 (2016).
17. Le Bihan, D., Poupon, C., Amadon, A. & Lethimonnier, F. Artifacts and pitfalls in diffusion MRI. *J. Magn. Reson. Imaging.* **24**, 478–488. https://doi.org/10.1002/jmri.20683 (2006).
18. Bickelhaupt, S. *et al.* Maximum intensity breast diffusion MRI for BI-RADS 4 lesions detected on X-ray mammography. *Clin. Radiol.* **72**, 900-e1. https://doi.org/10.1016/j.crad.2017.05.017 (2017).
19. Kang, J. W. *et al.* Unenhanced magnetic resonance screening using fused diffusion-weighted imaging and maximum-intensity projection in patients with a personal history of breast cancer: Role of fused DWI for postoperative screening. *Breast Cancer Res. Treat.* **165**, 119–128. https://doi.org/10.1007/s10549-017-4322-5 (2017).
20. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159. https://doi.org/10.2307/2529310 (1977).
21. Kapsner, L. A. *et al.* Automated artifact detection in abbreviated dynamic contrast-enhanced (DCE) MRI-derived maximum intensity projections (MIPs) of the breast. *Eur. Radiol.* https://doi.org/10.1007/s00330-022-08626-5 (2022).
22. Brünjes, R. & Hofmann, T. Anthropogenic gadolinium in freshwater and drinking water systems. *Water Res.* **182**, 115966. https://doi.org/10.1016/j.watres.2020.115966 (2020).
23. Le Goff, S. *et al.* Compound-specific recording of gadolinium pollution in coastal waters by great scallops. *Sci. Rep.* **9**, 8015. https://doi.org/10.1038/s41598-019-44539-y (2019).
24. Lindner, U. *et al.* Analysis of Gadolinium-based contrast agents in tap water with a new hydrophilic interaction chromatography (ZIC-cHILIC) hyphenated with inductively coupled plasma mass spectrometry. *Anal. Bioanal. Chem.* **407**, 2415–2422. https://doi.org/10.1007/s00216-014-8368-5 (2015).
25. Rogowska, J., Olkowska, E., Ratajczyk, W. & Wolska, L. Gadolinium as a new emerging contaminant of aquatic environments. *Environ. Toxic Chem.* **37**, 1523–1534. https://doi.org/10.1002/etc.4116 (2018).
26. Zaitsev, M., Maclaren, J. & Herbst, M. Motion artifacts in MRI: A complex problem with many partial solutions: Motion artifacts and correction. *J. Magn. Reson. Imaging* **42**, 887–901. https://doi.org/10.1002/jmri.24850 (2015).
27. Maintz, J. A. & Viergever, M. A. An overview of medical image registration methods. In: *Symposium of the Belgian Hospital Physicists Association (SBPH/BVZF)* 1–22. https://dspace.library.uu.nl/handle/1874/18921 (Accessed 3 May 2023) (1998).
28. Arlinghaus, L. R. *et al.* Motion correction in diffusion-weighted MRI of the breast at 3T. *J. Magn. Reson. Imaging* **33**, 1063–1070. https://doi.org/10.1002/jmri.22562 (2011).

29. Boehler, T., Wirtz, S., Peitgen, H.-O. A combined algorithm for breast MRI motion correction, in: (eds Giger, M. L. & Karssemeijer, N.) 65141R (2007). https://doi.org/10.1117/12.708541.
30. Mattusch, C., Bick, U. & Michallek, F. Development and validation of a four-dimensional registration technique for DCE breast MRI. *Insights Imaging* **14**, 17. https://doi.org/10.1186/s13244-022-01362-w (2023).
31. Zuo, C. S., Jiang, A., Buff, B. L., Mahon, T. G. & Wong, T. Z. Automatic motion correction for breast MR imaging. *Radiology* **198**, 903–906. https://doi.org/10.1148/radiology.198.3.8628891 (1996).
32. Partridge, S. C. & McDonald, E. S. Diffusion weighted magnetic resonance imaging of the breast. *Magn. Reson. Imaging Clin. N. Am.* **21**, 601–624. https://doi.org/10.1016/j.mric.2013.04.007 (2013).
33. Partridge, S. C., Nissan, N., Rahbar, H., Kitsch, A. E. & Sigmund, E. E. Diffusion-weighted breast MRI: Clinical applications and emerging techniques: Diffusion-weighted breast MRI. *J. Magn. Reson. Imaging* **45**, 337–355. https://doi.org/10.1002/jmri.25479 (2017).
34. Lowekamp, B. C., Chen, D. T., Ibanez, L. & Blezek, D. The design of SimpleITK. *Front. Neuroinform.* https://doi.org/10.3389/fninf.2013.00045 (2013).
35. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362. https://doi.org/10.1038/s41586-020-2649-2 (2020).
36. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261–2269. https://doi.org/10.1109/CVPR.2017.243 (IEEE, 2017).
37. T.M. Consortium. Project MONAI. Zenodo https://doi.org/10.5281/zenodo.4323059 (2020).
38. Paszke, A., Gross, A., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. PyTorch: An imperative style, high-performance deep learning library. in (eds Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E. & Garnett, R.), *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 8026–8037 https://dl.acm.org/doi/pdf/10.5555/3454287.3455008. (Accessed 3 March 2021) (Curran Associates, Inc., 2019).
39. Falcon, W. *et al.* PyTorchLightning. Zenodo https://doi.org/10.5281/zenodo.3828935 (2021).
40. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC Plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432. https://doi.org/10.1371/journal.pone.0118432 (2015).
41. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization (2017) (accessed 13 January 2021). http://arxiv.org/abs/1412.6980.
42. R Core Team. R: A language and environment for statistical computing (2022) (accessed 12 October 2022). https://www.R-project.org/.
43. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378. https://doi.org/10.1037/h0031619 (1971).
44. Gamer, M., Lemon, J. & Singh, I. F. P. Irr: Various coefficients of interrater reliability and agreement (2019) (accessed 24 August 2021). https://CRAN.R-project.org/package=irr.
45. Lang, M. Mlr3measures: Performance Measures for 'Mlr3' (2022) (accessed 12 October 2022). https://CRAN.R-project.org/package=mlr3measures.
46. Wickham, H. Ggplot2: Elegant Graphics for Data Analysis (Springer, 2016). https://ggplot2.tidyverse.org.
47. Kassambara, A. Ggpubr: 'ggplot2' based publication ready plots. https://CRAN.R-project.org/package=ggpubr (Accessed 1 February 2022) (2020).
48. Saito, T. & Rehmsmeier, M. Precrec: Fast and accurate precision-recall and ROC curve calculations in R. *Bioinformatics* **33**(1), 145–147. https://doi.org/10.1093/bioinformatics/btw570 (2017).
49. Hollander, M. & Wolfe D. A. *Nonparametric Statistical Methods* 27–33 and 68–75 (John Wiley & Sons, 1973).
50. Bauer, D. F. Constructing confidence sets using rank statistics. *J. Am. Stat. Assoc.* **67**, 687–690. https://doi.org/10.1080/01621459.1972.10481279 (1972).
51. Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **50**, 157–175. https://doi.org/10.1080/14786440009463897 (1900).
52. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2921–2929. https://doi.org/10.1109/CVPR.2016.319 (IEEE, 2016).
53. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Object detectors emerge in deep scene CNNs (2015) (accessed 3 May 2023). http://arxiv.org/abs/1412.6856.
54. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* 839–847. https://doi.org/10.1109/WACV.2018.00097 (IEEE, 2018).

Acknowledgements

We thank the d.hip Digital Health Innovation Platform for their support and Martin Gründl (Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen) for his support in providing the data. This work was supported in part by the German Federal Ministry of Education and Research (BMBF) in association with the project SMART SELECT MRI under the funding number: FKZ 161B0976.

Author contributions

L.A.K.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing—original draft. E.L.B.: Data curation, Writing—review and editing. L.F.: Methodology, Writing—review and editing. F.B.L.: Supervision, Validation, Visualization, Writing—review and editing. A.M.N.: Supervision, Writing—review and editing. A.L.: Writing—review and editing. J.E.: Writing—review and editing. S.O.: Data curation, Writing—review and editing. M.U.: Funding acquisition, Resources, Writing—review and editing. E.W.: Data curation, Supervision, Writing—review and editing. S.B.: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing—review and editing.

Funding

Open Access funding enabled and organized by Projekt DEAL. We acknowledge financial support by Deutsche Forschungsgemeinschaft and Friedrich-Alexander-Universität Erlangen-Nürnberg within the funding programme "Open Access Publication Funding".

Competing interests

Sebastian Bickelhaupt reports financial support was provided by German Federal Ministry of Education and Research (BMBF) - FKZ 161B0976. Sebastian Bickelhaupt reports a relationship with Siemens Healthcare GmbH that includes: speaking and lecture fees. Eva Balbach reports a relationship with Siemens Healthcare GmbH that includes: speaking and lecture fees. Michael Uder reports a relationship with Siemens Healthcare GmbH that includes: member of the speakers bureau. The University Hospital Erlangen receives funding from Siemens Healthcare GmbH that is partially contributing to the job positions of Sebastian Bickelhaupt and Lorenz Kapsner. Julius Emons reports relationships with Pfizer Pharma GmbH, Novartis AG, and Eisai GmbH that include: speaking and lecture fees, participation on an advisory board, travel support (from Novartis AG). Lukas Folle, Frederik Laun, Armin Nagel, Andrzej Liebert, Sabine Ohlmeyer, and Evelyn Wenkel declare no potential competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-37342-3.

Correspondence and requests for materials should be addressed to L.A.K.

Reprints and permissions information is available at www.nature.com/reprints.