OPEN ACCESS
International Journal of Neural Systems, Vol. 34, No. 10 (2024) 2450052 (15 pages)
© The Author(s)
DOI: 10.1142/S0129065724500527

Motion Artifact Detection for T1-Weighted Brain MR Images Using Convolutional Neural Networks

Erik Roecher[*,**], Lucas Mösch[*], Jana Zweerings[*], Frank O. Thiele[†],
Svenja Caspers[‡,§], Arnim Johannes Gaebler[*,¶,‖], Patrick Eisner[*],
Pegah Sarkheil[*] and Klaus Mathiak[*,¶]

[*]*Department of Psychiatry, Psychotherapy and*
Psychosomatics, Faculty of Medicine
RWTH Aachen, Germany

[†]*Philips Healthcare, Aachen, Germany*

[‡]*Institute for Anatomy I, Medical Faculty & University*
Hospital Düsseldorf, Heinrich Heine University Düsseldorf
Düsseldorf, Germany

[§]*Institute of Neuroscience and Medicine (INM-1)*
Research Centre Jülich, Jülich, Germany

[¶]*JARA-BRAIN, Jülich Aachen Research Alliance (JARA)*
Translational Brain Medicine, Germany

[‖]*Institute of Neurophysiology, Faculty of Medicine*
RWTH Aachen, Germany
[**]*erik.roecher@rwth-aachen.de*

Received 4 March 2024
Accepted 17 June 2024
Published Online 12 July 2024

Quality assessment (QA) of magnetic resonance imaging (MRI) encompasses several factors such as noise, contrast, homogeneity, and imaging artifacts. Quality evaluation is often not standardized and relies on the expertise, and vigilance of the personnel, posing limitations especially with large datasets. Machine learning based on convolutional neural networks (CNNs) is a promising approach to address these challenges by performing automated inspection of MR images. In this study, a CNN for the detection of random head motion artifacts (RHM) in T1-weighted MRI as one aspect of image quality is proposed. A two-step approach aimed to first identify images exhibiting pronounced motion artifacts, and second to evaluate the feasibility of a more detailed three-class classification. The utilized dataset consisted of 420 T1-weighted whole-brain image volumes with isotropic resolution. Human experts assigned each volume to one of three classes of artifact prominence. Results demonstrate an accuracy of 95% for the identification of images with pronounced artifact load. The addition of an intermediate class retained an accuracy of 76%. The findings highlight the potential of CNN-based approaches to increase the efficiency of *post-hoc* QAs in large datasets by flagging images with potentially relevant artifact loads for closer inspection.

Keywords: Structural MRI; quality assessment; CNN; motion artifacts; DNN; clinical image acquisition.

[**] Corresponding author.

1. Introduction

Magnetic resonance imaging (MRI) is central to many clinical and research applications due to high spatial resolution of the assessed structures. These high-resolution images are susceptible to noise, contrast, inhomogeneity, and further imaging artifacts that may impact image quality. Thus, image quality is not readily quantifiable within a single metric and the definition of acceptable image quality heavily depends on the specific application (i.e. the brain regions of interest, sample size or analysis type). Nevertheless, quality assurance (QA) is a key aspect of reliable medical diagnosis and analysis of imaging data. Quality assessment (QA) should be reproducible, objective, and accessible.[1] In this study, a deep learning algorithm for the classification of visually detectable motion artifacts as one key aspect of image quality is established and optimized. The presented approach shows high accuracy for the identification of images with pronounced motion artifacts and thus, can be used to “flag” respective images for further visual inspection. In addition, the pipeline was integrated as a first use case for data quality assurance at a local database for MR imaging.

Random head motion artifacts (RHM) are among the most common visual artifacts encountered in clinical routine imaging and MR studies. RHM typically appear as circular-shaped patterns of wave-like changes in brightness in T1-weighted images. These artifacts are foremost caused by increased head motions of participants due to extensive breathing, body motion, or illness-related unrest. RHM poses several potential problems for the utilization of the acquired data. Most studies investigating MRI motion artifacts focus on functional MR imaging.[2,3] Siegel and colleagues demonstrated a statistically significant improvement of BOLD signal in task-based fMRI when correcting for motion, while Power and colleagues linked subject motion to changes in the time courses of resting state functional connectivity.[2,3] For structural imaging, Reuter and colleagues systematically investigated the impact of head motion on morphometric estimates of brain structures and showed that reduced image quality affects grey matter volume and thickness estimates.[4] The authors conclude that head motion introduces a bias in the data that can potentially lead to the overestimation of effects when groups with different movement susceptibility are compared. In a similar vein, Blumenthal and colleagues showed that greater motion artifacts are associated with smaller estimates of grey matter volume.[5] Duffy and colleagues indicate significant changes in cortical thickness estimates after motion correction.[6] Here, more widespread cortical thinning was found after applying their convolutional neural network (CNN) based correction. An extensive review of clustering approaches for MRI by Mirzaei and Adeli highlights the impact of image artifacts and noise on the performance of clustering algorithms.[7] While time series measurements such as resting-state provide movement parameters, which allow estimates for movement intensity,[8] there are no comparable indicators available in structural imaging. The application of fiducial markers such as reflective markers tracked by infrared cameras or radio frequency coil markers tracked during image acquisition has been proposed to monitor motion[9] and provide motion compensation.[10,11] However, fiducial markers require a prior setup, rendering a retrospective analysis, e.g. of large databases, impossible and are furthermore unable to fully eliminate motion artifacts. Yet, data analysis may be sensitive to these artifacts,[4] which are easily missed in short visual inspections. Thus, automatic, and hence, user-independent procedures for structural imaging are a promising approach to present a fast and accurate feedback of specific aspects of data quality that can prompt further inspection, and if indicated, immediate repetition of structural imaging.

In the last decade, artificial neural networks and deep learning have gained major attention in the computer vision community and beyond.[12–14] These methods surpassed traditional machine learning methods in a variety of tasks, especially in complex pattern recognition, and yield great potential in medical imaging.[17–20] Accordingly, CNNs provide a promising algorithmic foundation for the automated classification of RHM that manifests as a complex pattern in MR images. So far, most studies focus on either an overall image quality classification incorporating motion, noise and contrast into a single quality metric or the correction of artifacts. Bottani *et al.* applied a CNN approach on a set of 5500 manually labeled T1-weighted images to distinguish good, intermediate, and bad quality.[21] Image quality labels combine noise, motion, and contrast in a single label. The accuracy of the detection of low-quality images was 83%. The authors note that with their annotation process it may be difficult to distinguish where image degradation appeared. Especially motion and noise may be confused. Keshavan and colleagues used the Healthy Brain Network dataset with an existing quality label and adapted a VGG16 image classification CNN.[22] The authors predicted expert quality ratings using available crowd-sourced ratings on a scale from 0 (fail) to 1 (pass) with a high match indicated by an AUC of 0.99. Few studies applied deep learning to detect motion artifacts specifically. Küstner *et al.* successfully trained a CNN on patches of cranial T1-weighted images to detect volitionally induced motion artifacts on 16 healthy volunteers.[23] Participants were instructed to deliberately tilt their heads to induce motion artifacts (motion-affected dataset) or to remain still (motion-free dataset). Due to the limited sample size, they applied a leave-one-out approach with a detection rate of 100% and cross-validated with abdominal images, leading to an accuracy of 82%.[23] Extending their work, they applied a Generative Adversarial Network (GAN) for motion artifact correction.[24] However, the authors stress that anatomical features may be altered, or hidden, and clinical application needs to be evaluated. Further, retrospective head motion artifact correction using deep learning 3D CNN was proposed by Duffy *et al.* on structural images of 864 participants.[6] Al-Masni and colleagues introduced stacked U-Nets, trained on the clinical data of 83 participants with artificially created motion artifacts for artifact correction in structural images.[25] Although retrospective correction of motion artifacts may be a suitable solution for existing datasets, methods that support initial artifact detection can prompt repetition of scans if feasible and are thus promising in terms of reducing the need for *post-hoc* artificial data enhancement that may be associated with the risk of distortion of the originally acquired data.

In this study, an investigation was conducted whether CNNs are feasible candidates for RHM detection in structural MRI in a large population-based dataset, aimed to identify and flag potentially critical images. The applicability of deep learning for the classification of images showing no visible versus pronounced motion artifacts was explored and relevant artificial neural network topologies for this classification task were established. Further, we explored the possibility of extending our approach by integrating an intermediate class of moderate artifact prominence. The generalizability of our trained neural networks was assessed on an independent dataset and applicability was demonstrated on our in-house neurofeedback database.

2. Methods

2.1. *Overview*

The main objective of the presented approach was to identify images exhibiting pronounced motion artifacts. Automated detection of motion artifacts in MRI can be used to flag potentially critical images and prompt suitable actions by clinicians or study staff. Further, if applied during the scanning it would allow for the rapid and automatic identification of images that might require reacquisition. To achieve this goal, a set of CNNs was implemented, performing a two-class classification of T1-weighted MR images with no visible artifacts and images with pronounced motion artifacts (as detailed in Sec. 2.6.). These networks were trained to achieve high accuracy and low false-negative rates. Further extending this work, we investigated the feasibility of a more detailed classification by introducing an additional intermediate-artifact-load class. This three-class classification task is performed by an ensemble of CNNs, which is also detailed in Sec. 2.6.

2.2. *Dataset*

Training dataset

For training and validation of the networks, an existing dataset of T1-weighted images originating from the 1000BRAINS study[26] was used. The primary aim of this study was to investigate the structural and functional variabilities in the human brain during aging. A population-based study was chosen to ensure a realistic representation of the data quality.

Out of the available data (for details, see Ref. 26), 420 T1-weighted brain images (see below) with an isotropic resolution of one millimeter were selected for a balanced representation of artifact prominence (scanning parameters: repetition time TR = 2.25 s, echo time TE = 3.03 ms, TI = 900 ms, field of view FoV = $256 \times 256\,\mathrm{mm}^2$, flip angle = 9°). All images were acquired using the same 3T whole-body MR scanner (Tim-TRIO, Siemens Medical Systems, Erlangen, Germany).

An additional set of T1-weighted images of 19 participants was acquired in our in-house three Tesla MR Scanner (Tim-TRIO, Siemens Medical Systems, Erlangen, Germany) with identical imaging parameters as for the 1000BRAINS study. In a series of measurements, participants were instructed to either refrain from moving to ensure high image quality or to apply several methods to induce motion artifacts such as pronounced breathing, severe eye movements, and using one foot to "write" their name in the air to ensure prolonged physical activity. Accordingly, the acquired data contains images of the same participant with different levels of artifact prominence. This approach was chosen to reduce the risk of learning based on individual subject-specific anatomical characteristics rather than actual artifact detection.

Generalizability dataset

To assess the generalizability of the approach, we evaluated the performance of the network on an independent dataset that was acquired as part of a randomized clinical trial, hence, including both, data of patients ($N = 53$) and healthy controls ($N = 22$). This generalizability dataset consisted of images acquired with identical sequence parameters and labeled following the same expert rating approach as in the training dataset. Thereof, 75 images were selected and distributed equally across the three classes to prevent class imbalances.

Neurofeedback database

As a use case for the classification of motion artifacts, we applied the created networks on an independent research dataset containing both patients and healthy individuals ($N = 87$ patients; $N = 94$ healthy individuals). The data was not labelled by experts. The parameters for the high-resolution T1 images for the included datasets were: 1 mm isotropic voxel resolution, repetition time TR = 2.00 s, echo time TE = 3.03 ms, TI = 900 ms, field of view FoV = $256 \times 256\,\mathrm{mm}^2$, flip angle = 9° for 139 participants. For the remaining 42 participants the scanning parameters slightly differed: 1 mm isotropic voxel resolution, repetition time TR = 2.30 s echo time TE = 2.98 ms, field of view FOV = $256 \times 256\,\mathrm{mm}^2$, TI = 900 ms, flip angle = 9°. All images were acquired using the same 3T whole-body MR scanner (Tim-TRIO, Siemens Medical Systems, Erlangen, Germany).

2.3. *Image volume annotation*

Manual image QA of all images was performed by two trained experts under the supervision of the QA team of the Psychiatric Imaging Network Germany (PING).[27] Each image volume was assigned to one of three classes indicating RHM prominence. Class 0 was defined as images without visibly detectable motion artifacts throughout the brain; Class 1 reflected images with moderate motion artifacts in at least one image slice; Class 2 comprised images with pronounced motion artifacts. Importantly, image volumes retained only one single annotation for the entire volume, even when artifacts were only visible within isolated slices. In case of interrater conflict regarding image assignment (i.e. the same image volume was assigned to different classes by the raters), a supervisory decision was obtained from a QA expert of the PING consortium resulting in one annotation of each image volume. For an exemplary image in each of the three classes see Fig. 1. The additional image volumes acquired in our in-house MR scanner were rated following the same procedure. Inter-rater reliability between rater A and rater B prior to a supervisory decision by trained QA expert on the two-class annotation yielded a Cohen's kappa of $\kappa = 0.72$ and for the three-class annotation a Cohen's kappa of $\kappa = 0.52$.[28]

Annotated images were randomly selected from the whole pool of images to create a dataset with a matched number of images per class to prevent trained networks to favor a specific class. This final dataset was consistent throughout all further analysis, training, and validation. Each class in this dataset contained 140 volumes resulting in a dataset with a total number of 420 image volumes. Mean age of participants included was 61.04y (± 12.74y SD)

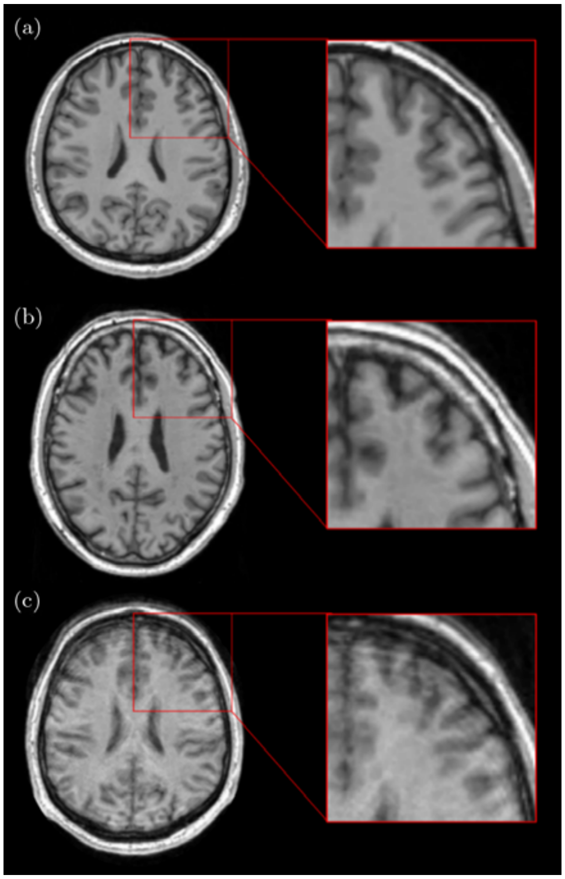


Fig. 1. Exemplary illustration of the three-class classification of head motion artifact prominence in transversal image slices. Motion artifacts can be most readily identified at contrast edges. (a) Top row: Class 0 denotes no visible motion artifacts. (b) Middle row: Class 1 images show moderate motion artifacts in at least one slice and (c) Bottom row: Class 2 contains images with pronounced motion artifacts visible in all slices. Right column shows zoomed in representations of each example.

with a slightly increasing mean age alongside increasing artifact prominence (57.25y ± 12.65y SD; 61.29y ± 13.05y SD; 64.97y ± 11.25y SD). In total, 231 male and 189 female participants were included.

2.4. *Pre-processing and data augmentation*

Head localization, rotation, and individual head sizes and shapes differ in standard MRI acquisition. Further, the general image intensity might differ between participants. These factors may introduce biases in the machine learning process. Therefore, consistent pre-processing steps and anatomical normalization were performed before images entered network training. These steps included bias field correction using SPM 12[29] and co-registration to a MNI template using an Euler transform (translation and rotation). Co-registration was applied to ensure that heads are centered and aligned in all T1w images. Images retained their 1 mm isotropic voxel resolution and received the MNI reference image resolution of 181 × 217 × 181 voxel. These processed 3D images represented each subject.

Image augmentation techniques were applied online during the training of individual networks to compensate for overfitting. These data augmentation techniques included rotation, scaling, and shearing. Each technique was applied by a random degree within a pre-set range with the value ranges detailed in the results section.

2.5. *Network training parameter*

The core training framework was written in the programming language Python (version 3.7.3[30]) using the machine learning software libraries Tensorflow (version 1.14.0[31]) and Keras (version 2.2.4). If not reported otherwise, default parameters of Tensorflow and Keras were applied.

Training and evaluation of all networks was performed using five-fold cross-validation. Therefore, the full dataset was split into five consistent parts with an equal number of images in each class. Each part served once as a validation set and was part of the training set in the remaining four cross-validation trainings. The weights and bias variables of each trained network were initialized using He normal distribution.[32] These initialized variables were stored and reused in each of the five trainings, which ensured an equal starting point for each training-split.

All convolutional layers employ a Rectified Linear Unit (ReLU) activation function.[33] The batch size was set to 5 with an initial learning rate of $1e-5$ and a reduction until $1e-6$ during training for all trainings, unless otherwise specified. Optimization was performed by an Adam optimizer, using a cross-entropy loss function. The training was stopped after twenty epochs without decreased loss. Reported results represent the calculated mean over all five cross-validation runs. Hyperparameter values reported here were a result of a grid search approach and were chosen as the best-performing parameters as measured by prediction accuracy using cross-validation.

2.6. *Network topology*

The investigation aimed to separate the T1-weighted images based on the prominence of visible motion artifacts. A two-step approach was implemented to achieve this goal. First, an initial network was trained on stacks of consecutive image slices. Each stack inherited the label of its parent image volume. Accordingly, the output of this network was a prediction of artifact load in a single stack (Fig. 2). Second, a subsequent network was trained to predict overall artifact load from all stacks belonging to one image volume. This workflow is illustrated in Fig. 3.

Stack network

From each 3D volume, stacks were extracted as consecutive transversal slices between slice 90 and slice 150. This range ensures optimal brain and hence, artifact coverage in all slice stacks. The number of slices in each stack acts as a hyperparameter during the optimization of the CNNs. It was a fixed parameter that determined the amount of image data fed into the CNNs simultaneously. Smaller stacks may lead to stacks containing no artifacts for images that contain artifacts in other slices while larger stacks may lead to decreased learning performance due to increased complexity of network input. Accordingly, this parameter balanced the probability of artifact detection within a single stack and learning performance. Stack size was set to 5 for two-class classifications and either 10 or 20 for three-class classifications. The stride size between stacks was one.

All slices of a stack were zero-padded to a size of 232×232 and individually min-max normalized $I_{\mathrm{norm}} = \frac{I - I_{\mathrm{min}}}{I_{\mathrm{max}} - I_{\mathrm{min}}}$, with I_{norm} being the normalized voxel intensity of a Voxel I and I_{min} and I_{max} being the minimum and maximum values in the stack. Augmentation techniques were applied equally to all slices in a stack (e.g. all slices in a stack are rotated by the same degree).

Our stack networks consisted of multiple blocks of two successive 2D convolutional (conv2D) layers with an identical number of channels followed by one max-pooling layer. Convolutional layers utilize image convolution with trainable filter kernels, which are trained to highlight relevant features and structures in their input. The kernel size of both conv2D layers was set to 3×3 and the kernel size of the max-pooling layer was set to 2×2. This results in a factor two resolution reduction after each block. The number of blocks and the number of channels in each block varied across different networks. The last block's output was flattened and fed into a set of fully connected (FC) layers, each with a dropout-rate of 0.4.[34] The final layer of a stack network was a softmax layer with either two or three neurons, depending on the number of classes, which outputs a prediction probability for each class. For an overview of all network parameters see Table 1.

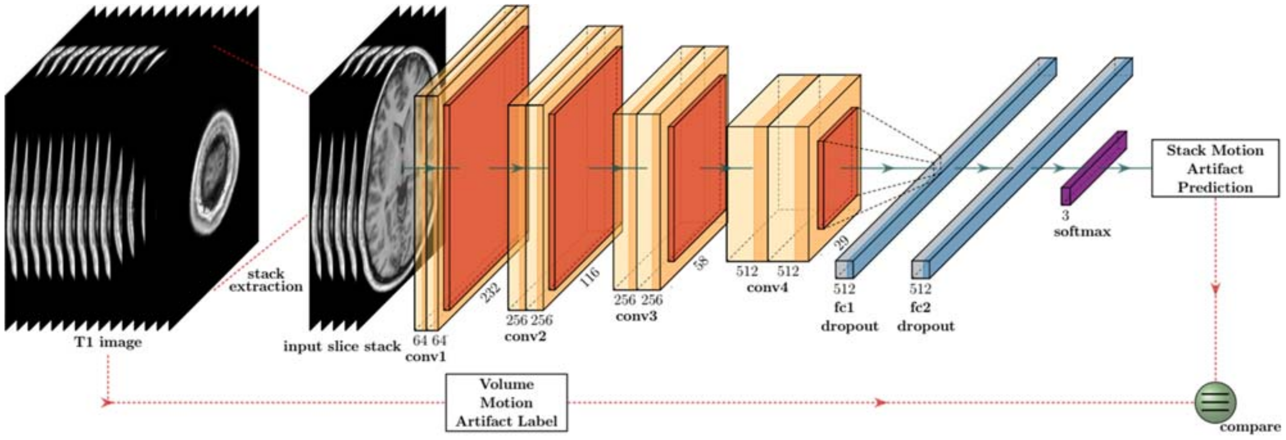


Fig. 2. The figure illustrates the slice-stack extraction from a T1-weighted volume and the CNNs network topology. Extracted slice stacks are fed into a sequential CNN consisting of four pairs of convolutional layers. Each pair used identical parameters and was followed by one max-pooling layer. The final convolutional layer was flattened and fed into two FC layer and a final softmax classification layer, resulting in a prediction of motion artifact prominence for each respective stack. Depicted values for layer sizes are exemplary. For specific values see Tables 1 and 2.

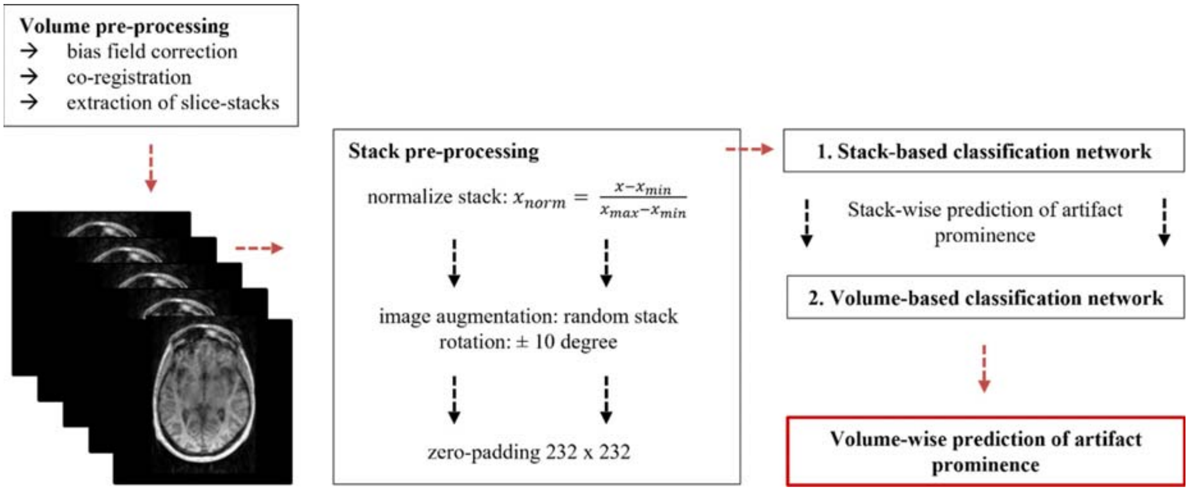

Volume pre-processing
→ bias field correction
→ co-registration
→ extraction of slice-stacks
Stack pre-processing
normalize stack: $x_{norm} = \frac{x-x_{min}}{x_{max}-x_{min}}$
image augmentation: random stack rotation: ± 10 degree
zero-padding 232 x 232
1. Stack-based classification network
Stack-wise prediction of artifact prominence
2. Volume-based classification network
Volume-wise prediction of artifact prominence

Fig. 3. Illustration of the pre-processing and data flow. Image volumes were bias field corrected, co-registered and slice stacks of varying size were extracted. The extracted stacks were normalized and underwent image augmentation including 10° rotation. Zero-padding to size of 232×232 was performed. Stacks entered a stack-based classification network, which outputs one prediction probability for each class. All stack predictions of a single image volume then entered a second network, which returns a final prediction probability for the image volume.

Class weights were introduced to improve the initially observed class imbalance in network predictions. Weighting factors of 2.0, 4.0, and 1.0 for images with no visible, moderate, and more pronounced artifact prominence were selected. These values were chosen to emphasize learning of a correct classification of our intermediate class by assigning a larger weighting factor to the class's loss term during training. Choosing a larger weight leads to a larger impact of that specific class on the overall loss during training. Class weights were used exclusively during the training of stack networks performing a three-class classification. Class weighting was not applied during the validation phase and had no impact on model predictions.

Volume network

The volume network was trained to pool all stack predictions into a final volume-wise prediction of artifact prominence. The input of this network consisted of a matrix sized 55×2 for the two-class classification and 55×3 for the three-class classification, respectively. These matrices contained the predicted classification probabilities for 55 stacks extracted from one image volume (stacks of 5 slices extracted in a slice range of 90–150 with a stride size of 1).

A volume network for the two-class classification consisted of multiple FC layers with a dropout rate of 0.35 and a final softmax classification layer, which outputs a probability for each class. Network parameters are listed in Table 1.

Additionally, an ensemble network for the three-class classification was implemented. This network received outputs from three different stack networks simultaneously, to facilitate recognition of underlying patterns in stack network predictions.

3. Results

The results are grouped according to the two- and three-class classification approaches. For each classification results are listed separately for stack and volume networks.

3.1. *Two-class classification*

The two-class classification aimed at distinguishing images with no visible artifacts from those with more pronounced artifacts.

Stack network

On the slice-stack level, the network had an average classification accuracy across the five splits in the cross-validation approach of 0.92 (± 0.03). Accordingly, the probability that the network assigns the correct class to a given stack was 92%. Since stacks were evaluated separately, their predicted class could

Table 1. Stack and volume network three-class classification.

Stack-network two-class classification					
Stack parameter	Slice interval	90–150	Rotation range	5	
	Stack size	5	Normalization	min–max	
	Stack stride	1	Optimizer	Adam	
	Slice axis	Transversal	Learning rate	$1e-5$	
	Padding shape	232 × 232	Loss	Categorial cross-entropy	
	Resample slice	No	Initializer	He-normal	
Network parameter	Conv layer	2 × 4	Stack-accuracy	mean (SD)	0.92 (0.03)
	Conv channel	64–256–128–256		split 1	0.96
	FC dropout	0.4		split 2	0.94
	FC layer	2		split 3	0.89
	FC neurons	512–512		split 4	0.94
				split 5	0.88
Volume-network two-class classification					
Volume network	Input: stack pred	110: 55 stacks á 2 classes	FC layer input	2 × 25	
	Dropout rate	0.25	Learning rate	$1e-5$	
	Optimizer	Adam	Loss	Categorical cross-entropy	
	Accuracy Statistics		Accuracy volume network		
Volume-accuracy	mean (SD)	0.93 (0.03)	mean (SD)	0.95 (0.02)	
	split 1	0.96	split 1	0.98	
	split 2	0.95	split 2	0.96	
	split 3	0.89	split 3	0.91	
	split 4	0.96	split 4	0.96	
	split 5	0.89	split 5	0.95	

The Table provides an overview over network parameter and network accuracy of the two-class classification for the stack network (top half), as well as network parameter and network accuracy of the volume network (bottom half). Accuracy is presented as mean value and standard deviation (SD) across all five splits with a comparison of computing the mean across all stack predictions (left side) with the accuracy of the volume network (right side).

vary across a volume even though the training label was the same across the entire volume.

Volume network

Of final relevance is the accuracy for each volume integrated across the respective slice stacks. In the volume network, the predictions from the stacks were aggregated to form a single quality prediction for the whole volume (see Table 1). The accuracy of the quality prediction for the volume was 0.95 (± 0.01). Importantly, simply computing the mean of the predictions from the stack network resulted in a lower accuracy (0.93). Hence, the network seems to recognize underlying patterns that extend predictions solely based on the stacks.

3.2. *Three-class classification*

The same method as for the two-class classification was chosen with the addition of an intermediate class comprising images with moderate artifact load.

Stack network

At the first level, the predictors from slice stacks were computed. Three different stack networks were modelled with different parameters for number of slices and resampling (see Table 2). The networks had very similar accuracy (Network 1: 0.69 ± 0.03; 2: 0.68 ± 0.03; and 3: 0.69 ± 0.02). Even when considering the theoretical chance level of 0.33, the predictions were clearly lower than in the two-class model.

Table 2. Stack networks three-class classification.

		Stack-networks three-class classification		
		Network 1	Network 2	Network 3
Stack parameter	Slice interval	90–150	90–150	90–150
	Stack size	20	0	20
	Stack stride	1	1	1
	Slice axis	transversal	transversal	transversal
	Padding shape	232 × 232	232 × 232	232 × 232
	Resample slice	116 × 116	No	No
Training parameter	Rotation	10	10	10
	Normalization	min–max	min–max	min–max
	Optimizer	Adam	Adam	Adam
	Learning rate	1e−5	1e−5	1e−5
	Loss	Categorial cross-entropy	Categorial cross-entropy	Categorial cross-entropy
	Initializer	He-normal	He-normal	He-normal
Network parameter	Conv layer	2 × 4	2 × 4	2 × 4
	Conv channel	64–256–256–512	64–256–256–512	64–256–256– 512
	FC dropout	0.4	0.4	0.4
	FC layer	2	2	2
	FC neurons	512–512	512–512	512–512
Stack-accuracy	mean (SD)	0.69 (0.03)	0.68 (0.03)	0.69 (0.02)
	split 1	0.70	0.70	0.69
	split 2	0.68	0.68	0.67
	split 3	0.69	0.65	0.68
	split 4	0.73	0.72	0.72
	split 5	0.63	0.63	0.66

The table provides an overview over network parameters and prediction accuracy of the three-class classification of three stack networks. Each column represents one individual stack network. The first section shows values chosen for the stack input of the networks. The second section shows parameters chosen for training, while the third section presents the topology of the individual networks including convolutional (conv) layer sizes and FC layer sizes. Finally, the last section displays the accuracy of each stack network per split and as mean value and SD across all five splits.

Volume network

In parallel to the stack network computations, one volume network was trained on the output of all three stack networks simultaneously. Hence, one single-volume network with the input of three different stack networks was created. Stack network predictions were preprocessed on individual network paths in the volume network, comprising of two FC layers with 64 neurons. Preprocessing paths were merged and further processed by two FC layers with 256 and 128 neurons, respectively, and a final softmax classification layer. During training, dropout was applied to all FC layers with a dropout rate of 0.35. This ensemble network reached a final network performance of 0.76 (± 0.04).

Notably, the accuracy of this volume network was much higher compared to simply computing the mean across all stack predictions of each stack network, and then forming a majority vote based on the mean across the network predictions to assign a class (0.67; ± 0.06; see Table 3). Hence, our volume network improved accuracy by 10%. In addition, precision and recall was calculated by class. Precision of a class X was given by $\text{precision}_{\text{Class}X} = \text{TP}_{\text{Class}X}/(\text{TP}_{\text{Class}X} + \text{FP}_{\text{Class}X})$ and recall was given by $\text{recall}_{\text{Class}X} = \text{TP}_{\text{Class}X}/(\text{TP}_{\text{Class}X} + \text{FN}_{\text{Class}X})$. TP and FP were all true positives and false positives within the prediction of class X and FN were all false negatives within class X. Precision by class for the ensemble network was 0.75, 0.68, and 0.83 (class 0: no visible motion artifacts, class 1: moderate motion artifacts, class 2: pronounced motion

Table 3. Volume network three-class classification.

Volume-network three-class classification			
Input: stack predictions	120: 40 stacks á 3 classes	150: 50 stacks á 3 classes	120: 40 stacks á 3 classes
FC layer input	2 × 64	2 × 64	2 × 64
Merged input FC		256–128	
Dropout rate	0.35	Optimizer	Adam
Learning rate	1e−5	Loss	Categorical cross-entropy
Accuracy Statistics		Accuracy volume network	
mean (SD)	0.67 (0.06)	mean (SD)	0.76 (0.04)
split 1	0.68	split 1	0.76
split 2	0.67	split 2	0.76
split 3	0.64	split 3	0.75
split 4	0.77	split 4	0.82
split 5	0.58	split 5	0.69

The table provides an overview over network parameters and prediction accuracy of the three-class classification of the volume network. The upper half of the Table shows input size, layer size and parameter used for training of the FC network. The bottom half of the table compares computing the mean across all stack predictions of each volume network, and then forming a majority vote based on the mean across the stack network predictions (left side) with the accuracy of an ensemble volume network (right side). Results are displayed per split and as mean value and SD across all five splits.

artifacts) and recall by class was 0.81, 0.59, 0.88 in the same order as precision.

3.3. *Generalizability*

For the two-class classification our findings indicate good generalizability on our independent dataset (for details see Sec. 2.2. generalizability dataset). On average across all five splits, 24 images with pronounced artifact load were correctly classified. In addition, 15 of 25 images without visible artifacts were identified. This two-class classification CNN was not trained on images with moderate artifact load and can therefore not predict the intermediate class. However, it is noteworthy that out of the 25 images with moderate artifact load, 17 were classified as images with artifact load and 8 as images without visible artifacts. This indicates that the network has a high sensitivity for artifact detection. Accordingly, image classification is likely conservative.

For the three-class classification, accuracy on volume level was 49% across splits with the best performance split of 60%. Accordingly, generalizability is above statistical and empirical chance level. Notably, 23.2 out of 25 images belonging to class 2 were identified correctly, indicating a high accuracy for heavy artifact detection (93%). Approximately half of the images in class 0 were assigned to class 1 while half of the images in class 1 were assigned to class 2 further indicating conservative image classification of the network.

As a first use case we applied our trained networks on an independent large database (see Sec. 2.2. neurofeedback database). Out of a total of 181 T1-weighted images, five were classified as images showing more pronounced motion artifacts. Visual inspection of these images confirmed the network classification, indicating the feasibility of network-assisted motion artifact detection.

3.4. *Impact of age and gender on motion artifacts*

A weak positive correlation emerged between age and manual rated RHM prominence in our dataset. The Spearman correlation coefficient of age and motion artifact prominence divided into three classes yields $r = 0.268$ with $p < 0.001$.

Considering the black-box nature of CNNs it is important to investigate performance on expected variations within our data. A comparison of network performance between male and female samples revealed comparable accuracy estimates for both groups (female: 74.9%, male 75.5%; three-class classification). This indicates applicability of our algorithm irrespective of gender.

4. Discussion

In this study, we applied a CNN to classify the prominence of RHM, reflecting a “red flag” approach. Consequently, our primary goal was the identification of images showing pronounced motion artifacts that may require further inspection by clinicians, study personnel or data analysts. In a second step, we integrated the classification of artifacts according to three classes: no visible artifacts, moderate artifacts, and pronounced artifacts. Our results indicate a high accuracy for the detection of images with pronounced artifact load (95%). The addition of an intermediate class resulted in an accuracy of 76%. Hence, performance was clearly above chance, however, the network is optimal for an automatic artifact detection that marks images with pronounced artifacts. Results indicate similar performance of the approach across gender.

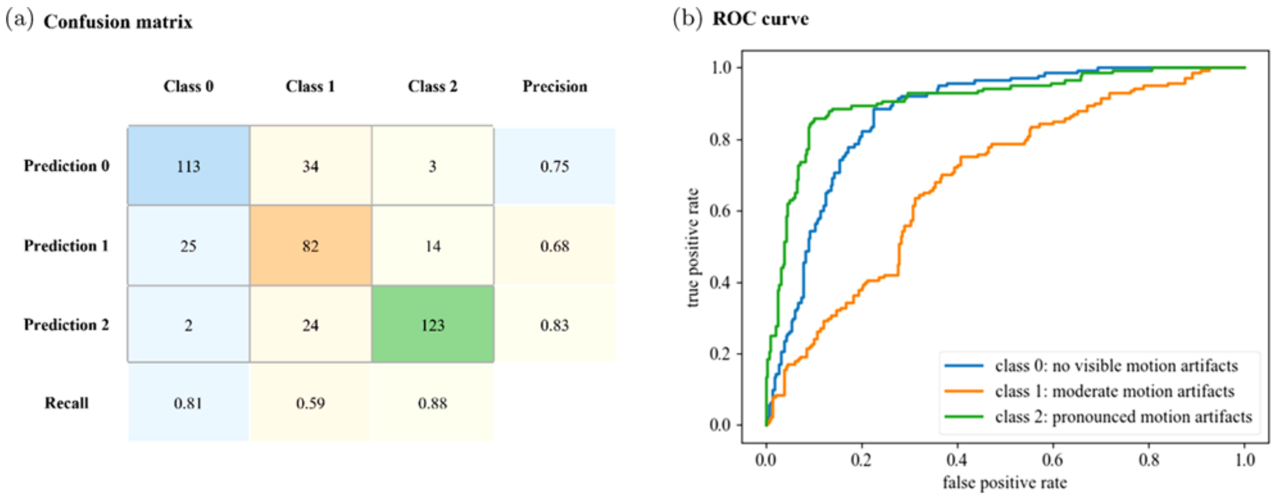
Artifacts in imaging data can have a substantial influence on the analysis and subsequent results, including over- and underestimation of effects, as many image segmentation approaches and probabilistic brain-area matching algorithms rely on voxel intensity. In daily clinical and research data acquisition routines a fast and reliable on-site motion artifact detection may thus facilitate data quality assurance and hence data utilization. An accuracy of 95% suggests that the current approach is suitable to support identification of images with pronounced visual motion artifacts. The identified images may require additional inspections due to possible impact on further analyses (e.g. Refs. 5 and 35). Our model yields comparable results to Küstner and colleagues.[23] However, the authors applied their algorithm to a small dataset consisting solely of data with volitionally induced motion artifacts ($N = 16$).[23] The authors used a clearly defined assignment of images due to rest (motion-free) versus motion conditions. It remains unclear if the algorithm performs similarly for artifacts caused by naturally occurring and potentially more subtle motions. We extend previous research by applying a machine-learning approach to a larger naturalistic dataset. While a subset of our data included volitionally induced motion artifacts ($N = 19$), most images contained natural motion artifacts and hence presumably more subtle variations. All included images underwent expert ratings before training. In addition, generalizability indicated suitable performance for application on clinical data.

ML-based approaches have a high potential to increase the efficiency of data quality evaluation by automatic labeling of potentially critical datasets. For anatomical 3D volume images, a large number of slices is generated by the MR scanner. Hence, stack-by-stack autoclassification of artifacts such as provided by our algorithm can considerably increase detailed evaluations while minimizing time constraints on the medical or study personnel. As a first use case, we integrated the autoclassification of artifacts as part of a newly established real-time fMRI neurofeedback database. The approach was used to detect images with insufficient quality and “flag” the respective measurements to prompt researchers working with the database to carefully consider inclusion of the respective T1-weighted images in their data analyses.

To generalize network performance for artifact detection, it is necessary to validate the network on independent data. Here, it is important to highlight two aspects of the present approach: (1) we used data acquired at two different scanners for the training dataset, (2) we tested the generalizability of our network on datasets that were acquired on our inhouse MRI device in an independent study using identical imaging protocols, (3) we applied the algorithm to a local database containing images with different imaging sequences. Our results in both two-class and three-class classification are substantially above theoretical chance level of 50.00% and 33.33% classification accuracy, respectively. However, due to the limited amount of training examples, we also investigated empirical chance level thresholds as reported by Combrisson and Jerbi.[36] Using the provided equation, we arrived at a chance level of 59.62% for the two-class classification and 40.48% for the three-class classification with $p = 0.001$. Thus, the accuracy of both, the two-class stack network and the three-class ensemble network approach showed that our neural networks performed above chance which renders them a viable option for computer-assisted QA.

Importantly, the introduction of our two-network approach (stack network and volume network) increased classification performance notably compared with a single network trained directly on full volume images while simultaneously increasing generalizability. Further, our volume network recognized an underlying pattern in the prediction results generated by our stack networks, increasing classification performance by 2% (two-class classification) and 9% (three-class classification) when compared with a simple statistical average across all stack predictions. A visual inspection of stack predictions did not reveal an obvious pattern in predictions. Lastly, our ensemble approach using three stack networks for the three-class classification increased classification accuracy compared to a volume network trained on stack network predictions of a single network.

In addition to the overall average network performance, we investigated how our trained networks perform across different classes to gain a better understanding of the decision-making process of the network. In specific, we were interested in whether classification is conservative or rather liberal. Conservative in this context refers to a high recall value for images exhibiting pronounced motion artifacts, while liberal would indicate that images showing pronounced artifacts are subjected to less pronounced artifact classes. A conservative approach appears favorable in the context of a "red flag" approach since it prompts the clinician or researcher to further inspect image quality and to decide on the subsequent procedure (e.g. repetition of scans or artifact correction). To identify whether the network performed conservatively or liberally, we investigated the confusion matrix and the receiver operating characteristic (ROC) curves (see Fig. 4). The confusion matrix shows the assigned label and the predicted class for all validation images by the ensemble network. The majority of images in each class were correctly classified. Incorrectly classified images belonging to the no visible motion artifacts and pronounced motion artifacts classes were mostly mapped to the neighboring moderate motion artifact class. Precision and recall of the pronounced motion artifact class is high with precision $= 0.83$ and recall $= 0.88$, denoting solid detection of images containing motion artifacts. This is confirmed by the ROC curves in Fig. 4. However, the ROC curve for the intermediate class indicates less discriminatory power. Whilst most images in this class are still mapped correctly by the ensemble network, precision and recall show overall poorer performance when compared to the other two classes. It is important to note, that Cohen's kappa between rater A and rater B prior to supervisory decision of $\kappa = 0.52$ represents only moderate interrater reliability for the three-class classification. In comparison Cohen's kappa of $\kappa = 0.72$ for the two-class classification represents



(a) **Confusion matrix**

	Class 0	Class 1	Class 2	Precision
Prediction 0	113	34	3	0.75
Prediction 1	25	82	14	0.68
Prediction 2	2	24	123	0.83
Recall	0.81	0.59	0.88	

(b) **ROC curve**

Fig. 4. The figure illustrates the difficulty to classify images belonging to the moderate motion artifact class. (a) on the left shows the confusion matrix for the three-class classification of all validation images by the ensemble network. Precision and recall by class are denoted in the last column and the bottom row, respectively. (b) on the right shows the ROC curves by class. A perfect prediction is characterized by a curve through the upper left corner of the ROC space, while a random classifier is represented by a point on the diagonal line. Class 0 with no visible motion artifacts and class 2 with pronounced motion artifacts show a higher sensitivity and specificity as compared to the intermediate class.

substantial agreement. For reference, Landis and Koch regarded a kappa of $\kappa > 0.81$ as almost perfect agreement.[37] Accordingly, while two-class classification yielded substantial agreement, our data indicates the difficulty for human medical experts to unambiguously classify motion artifacts in marginal cases, which was also reflected by our network performance. Notably, Bottani *et al.* reported a comparable Cohen's kappa of 0.68 for their two-rater, three-class motion artifact annotation.[21]

It is important to consider the age distribution in investigations focusing on motion artifacts. Previous studies indicated an acceleration of motion artifacts in children and adolescents[5,38] possibly related to difficulties in inhibiting unwanted behavior as well as older individuals that may experience more difficulties to remain still for the duration of the measurement.[23,35] While we cannot confirm higher proneness for movement in young children and adolescents in this study due to the inclusion of data from adults only, our data indicates a slight positive association between age and the prominence of motion artifacts as determined with the Spearman Correlation. This finding is consistent with previous investigations.[35,39] To confirm generalizability across sexes, we explored whether sex had an influence on prediction accuracy of our algorithm. Our findings indicate that the algorithm performance was similar for images acquired for males and females. Accordingly, our results suggest the sex-independent validity of the presented CNN model.

5. Limitations

Our developed CNNs were trained and evaluated on T1-weighted images acquired using identical MP-RAGE (Magnetization Prepared — RApid Gradient Echo) sequences with isovolumetric voxel. The generalizability of our networks was tested on independent datasets, however, further testing our approach on different equipment and T1-weighted sequences is required to confirm universal applicability. Furthermore, due to our slice-stack approach, predictions were based on stacks of slices, while our images were rated on a volume basis. A manual rating of individual slices might improve network performance. However, while this approach would certainly be desirable, it would require more extensive expert ratings that are highly time-consuming and thus, costly.

Lastly, RHM prominence was treated as a classification task. This is desirable for the applicability of our approach and easier integration into a clinical routine; however, artifact prominence is by definition a continuous scale and the cutoff between classes had to be chosen by our human raters. While research indicates that motion artifacts impact subsequent data analysis, the extent of motion artifacts acceptable needs to be evaluated by human experts depending on individual use cases. Future investigations need to consider user-adjustable cutoffs between motion artifact classes. Further, it is important to note that ratings of image quality depend on several aspects including, but not limited to, motion artifacts. Hence, while the current focus on a "red flag" approach supports identification of critical images further decisions on repetition, exclusion of data, or artifact correction have to be made depending on the specific use cases.

6. Conclusion

The present artifact classification CNN yields promising accuracy for the detection of images with pronounced motion artifacts. Training and validation of the algorithm was performed on naturalistic data and generalizability was demonstrated on an independent dataset. Our approach may support the efficiency of QAs in large datasets by enabling fast detection of images with motion artifacts. This may also prompt decisions on immediate scan repetitions. Prospective development needs to address the sensitivity and specificity of predictions. Adjustable thresholds for the individual classes support adaptation to different applications. Further refinement of our networks based on additional training data, i.e. containing anatomical irregularities, is desirable.

Acknowledgments

This study was supported by the German Research Foundation [DFG; IRTG 2150, MA 2631/6-1]; the German Ministry for Education and Research [BMBF; APIC: 01EE1405A, B, and C]; and the Interdisciplinary Center for Clinical Research (IZKF) of the University Hospital Aachen. The authors report no further financial disclosures or potential conflicts of interest.

ORCID

Erik Roecher https://orcid.org/0000-0002-5622-7410
Lucas Mösch https://orcid.org/0000-0002-0041-2879
Jana Zweerings https://orcid.org/0000-0002-5724-1054
Frank O. Thiele https://orcid.org/0000-0002-4232-8878
Svenja Caspers https://orcid.org/0000-0001-5083-4669
Arnim Johannes Gaebler https://orcid.org/0000-0002-8816-3415
Pegah Sarkheil https://orcid.org/0000-0002-6903-7974
Klaus Mathiak https://orcid.org/0000-0002-2276-7726

References

1. M. C. Fu *et al.*, Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions, *Spine J.* **14** (2014) 2442–2448.
2. J. S. Siegel *et al.*, Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points, *Hum. Brain Mapp.* **35** (2014) 1981–1996.
3. J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar and S. E. Petersen, Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion, *Neuroimage* **59** (2012) 2142–2154.
4. M. Reuter *et al.*, Head motion during MRI acquisition reduces gray matter volume and thickness estimates, *Neuroimage* **107** (2015) 107–115.
5. J. D. Blumenthal, A. Zijdenbos, A. E. Molloy and J. N. Giedd, Motion artifact in magnetic resonance imaging: Implications for automated analysis, *Neuroimage* **16** (2002) 89–92.
6. B. A. Duffy *et al.*, Retrospective motion artifact correction of structural MRI images using deep learning improves the quality of cortical surface reconstructions.*Neuroimage* **230** (2021) 117756.
7. G. Mirzaei and H. Adeli, Segmentation and clustering in brain MRI imaging, *Rev. Neurosci.* **30** (2018) 31–44.
8. G. Mirzaei and H. Adeli, Resting state functional magnetic resonance imaging processing techniques in stroke studies, *Rev. Neurosci.* **27** (2016) 871–885.
9. C. Dold *et al.*, Advantages and limitations of prospective head motion compensation for MRI using an optical motion tracking device, *Acad. Radiol.* **13** (2006) 1093–1103.
10. M. B. Ooi, S. Krueger, W. J. Thomas, S. V. Swaminathan and T. R. Brown, Prospective real-time correction for arbitrary head motion using active markers, *Magn. Reson. Med.* **62** (2009) 943–954.
11. M. Zaitsev, C. Dold, G. Sakas, J. Hennig and O. Speck, Magnetic resonance imaging of freely moving objects: prospective real-time motion correction using an external optical motion tracking system, *Neuroimage* **31** (2006) 1038–1050.
12. G. Mirzaei, A. Adeli and H. Adeli, Imaging and machine learning techniques for diagnosis of Alzheimer's disease, *Rev. Neurosci.* **27** (2016) 857–870.
13. A. V. Perez-Sanchez, J. P. Amezquita-Sanchez, M. Valtierra-Rodriguez and H. Adeli, A new epileptic seizure prediction model based on maximal overlap discrete wavelet packet transform, homogeneity index, and machine learning using ECG signals, *Biomed. Signal Process. Control* **88** (2024) 105659.
14. M. H. Rafiei, L. V. Gauthier, H. Adeli and D. Takabi, Self-supervised learning for electroencephalography, *IEEE Trans, Neural Netw. Learn. Syst.* **35** (2024) 1457–1471.
15. A. S. Lundervold and A. Lundervold, An overview of deep learning in medical imaging focusing on MRI, *Z. Med. Phys.* **29** (2019) 102–127.
16. H. Selcuk Nogay and H. Adeli, Diagnostic of autism spectrum disorder based on structural brain MRI images using, grid search optimization, and convolutional neural networks, *Biomed. Signal Process. Control* **79** (2023) 104234.
17. H. S. Nogay and H. Adeli, Multiple classification of brain MRI autism spectrum disorder by age and gender using deep learning, *J. Med. Syst.* **48** (2024) 15.
18. N. J. Herzog and G. D. Magoulas, Convolutional neural networks-based framework for early identification of dementia using MRI of brain asymmetry, *Int. J. Neural Syst.* **32** (2022) 2250053.
19. W. Feng *et al.*, Automated MRI-based deep learning model for detection of Alzheimer's disease process, *Int. J. Neural Syst.* **30** (2020) 2050032.
20. M. Leming, J. M. Górriz and J. Suckling, Ensemble deep learning on large, mixed-site fMRI datasets in autism and other tasks, *Int. J. Neural Syst.* **30** (2020) 2050012.
21. S. Bottani *et al.*, Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse, *Med. Image Anal.* **75** (2022) 102219.
22. A. Keshavan, J. D. Yeatman and A. Rokem, Combining citizen science and deep learning to amplify expertise in neuroimaging, *Front. Neuroinform.* **13** (2019), https://doi.org/10.3389/fninf.2019.00029.
23. T. Küstner *et al.*, Automated reference-free detection of motion artifacts in magnetic resonance images, *Magn. Reson. Mater. Phys. Biol. Med.* **31** (2017) 243–256.
24. T. Küstner *et al.*, Retrospective correction of motion-affected MR images using deep learning frameworks, *Magn. Reson. Med.* **82** (2019) 1527–1540.

25. M. A. Al-masni *et al.*, Stacked U-Nets with self-assisted priors towards robust correction of rigid motion artifact in brain MRI, *Neuroimage* **259** (2022) 119411.
26. S. Caspers *et al.*, Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS, *Front. Aging Neurosci.* **6** (2014) 149.
27. M. Bauer *et al.*, Das deutsche forschungsnetz zu psychischen erkrankungen, *Nervenarzt* **87** (2016) 989–1010.
28. J. Cohen, A coefficient of agreement for nominal scales, *Educ. Psychol. Meas.* **20** (1960) 37–46.
29. K. Friston, J. Ashburner, S. Kiebel, T. Nichols and W. Penny, *Statistical Parametric Mapping* (Elsevier, 2007), doi: 10.1016/B978-0-12-372560-8.X5000-1.
30. G. Van Rossum and F. L. Drake, *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).
31. M. Abadi *et al.*, TensorFlow: Large-scale machine learning on heterogeneous distributed systems, Preprint at http://download.tensorflow.org/paper/whitepaper2015.pdf (2015).
32. K. He, X. Zhang, S. Ren and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, *2015 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2015), pp. 1026–1034.
33. V. Nair and G. E. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines, in *Proc. 27th Int. Conf. Machine Learning* (Omnipress, Madison, WI, USA, 2010), pp. 807–814.
34. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* **15** (2014) 1929–1958.
35. N. K. Savalia *et al.*, Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion, *Hum. Brain Mapp.* **38** (2017) 472–492.
36. E. Combrisson and K. Jerbi, Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy, *J. Neurosci. Methods* **250** (2015) 126–136.
37. J. R. Landi and G. G. Koch, The measurement of observer agreement for categorical data, *Biometrics* **33** (1977) 159.
38. O. Afacan *et al.*, Evaluation of motion and its effect on brain magnetic resonance image quality in children, *Pediatr. Radiol.* **46** (2016) 1728–1735.
39. A. F. G. Rosen *et al.*, Quantitative assessment of structural image quality, *Neuroimage* **169** (2018) 407–418.