frontiers in Neuroscience

ORIGINAL RESEARCH
published: 22 January 2020
doi: 10.3389/fnins.2019.01456

QC-Automator: Deep Learning-Based Automated Quality Control for Diffusion MR Images

Zahra Riahi Samani, Jacob Antony Alappatt, Drew Parker, Abdol Aziz Ould Ismail and Ragini Verma*

Diffusion and Connectomics in Precision Healthcare Research Lab, Department of Radiology, University of Pennsylvania, Philadelphia, PA, United States

OPEN ACCESS

Edited by:
Andrew L. Alexander,
University of Wisconsin–Madison,
United States

Reviewed by:
Nan-kuei Chen,
University of Arizona, United States
Matthew Brett,
University of Birmingham,
United Kingdom

***Correspondence:**
Zahra Riahi Samani
Zahra.RiahiSamani@
PennMedicine.upenn.edu;
zahra.riahy@gmail.com

Specialty section:
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

Received: *01 May 2019*
Accepted: *31 December 2019*
Published: *22 January 2020*

Citation:
Samani ZR, Alappatt JA, Parker D, Ismail AAO and Verma R (2020) QC-Automator: Deep Learning-Based Automated Quality Control for Diffusion MR Images. Front. Neurosci. 13:1456. doi: 10.3389/fnins.2019.01456

Quality assessment of diffusion MRI (dMRI) data is essential prior to any analysis, so that appropriate pre-processing can be used to improve data quality and ensure that the presence of MRI artifacts do not affect the results of subsequent image analysis. Manual quality assessment of the data is subjective, possibly error-prone, and infeasible, especially considering the growing number of consortium-like studies, underlining the need for automation of the process. In this paper, we have developed a deep-learning-based automated quality control (QC) tool, *QC-Automator*, for dMRI data, that can handle a variety of artifacts such as motion, multiband interleaving, ghosting, susceptibility, herringbone, and chemical shifts. QC-Automator uses convolutional neural networks along with transfer learning to train the automated artifact detection on a labeled dataset of ∼332,000 slices of dMRI data, from 155 unique subjects and 5 scanners with different dMRI acquisitions, achieving a 98% accuracy in detecting artifacts. The method is fast and paves the way for efficient and effective artifact detection in large datasets. It is also demonstrated to be replicable on other datasets with different acquisition parameters.

Keywords: MRI, artifacts, diffusion MRI, quality control, convolutional neural networks

INTRODUCTION

Diffusion MRI (dMRI) (Basser and Jones, 2002; Assaf and Pasternak, 2008) is now widely used to probe the microstructural properties of biological tissues, as well as the structural connectivity of the brain. dMRI is prone to different kinds of artifacts including motion, multiband interleaving, ghosting, susceptibility, herringbone, and chemical shift (Wood and Henkelman, 1985; Smith et al., 1991; Simmons et al., 1994; Schenck, 1996; Heiland, 2008; Moratal et al., 2008; Krupa and Bekiesińska-Figatowska, 2015). If these artifacts remain undetected or insufficiently corrected, it could bias the results of subsequent analyses, weakening their interpretability (Bammer et al., 2003; Van Dijk et al., 2012; Reuter et al., 2015). Thus, quality control (QC) is an essential step before dMRI goes into further processing like motion correction and tensor estimation (Bastiani et al., 2019).

Quality control is undertaken mostly by visual inspection prior to any processing or analysis, in order to assess the quality of the data. Based on this QC, appropriate corrections can be applied, or a decision can be made to exclude affected slices or volumes. This is very time consuming and challenging to undertake manually, especially in large datasets of dMRI data, a modality with inherently low signal to noise ratio. Furthermore, manual visual QC is subjective based on the level of sensitivity, expertise, or even tolerance to fatigue of the QC expert, leading to high inter-rater variability (Victoroff et al., 1994). This warrants the need for automated QC methods, to limit the work of the QC expert to the inspection of slices that have been flagged by an automated algorithm. In this paper, we propose to design such an automated QC method to detect a wide range of artifacts that may occur individually or in combination, flagging affected slices for subsequent inspection. This can be applied prior to processing, as well as at any stage when the results of an analysis step need to be tested. This will help the user determine the presence of an artifact, and whether corrective steps need to be employed or the slices need to be excluded.

Some form of QC is present in the different artifact correction tools such as FSL EDDY (Andersson et al., 2016; Bastiani et al., 2019), DTI studio (Jiang et al., 2006), DTIPrep (Oguz et al., 2014), and TORTOISE (Pierpaoli et al., 2010). Such tools are usually limited to detecting and correcting the specific artifact that they have been designed for, mostly motion, and eddy current induced distortions (Liu et al., 2015; Kelly et al., 2016; Iglesias et al., 2017; Alfaro-Almagro et al., 2018; Graham et al., 2018). The results of these correction packages also need subsequent inspection to detect the presence of any remaining artifacts, making QC essential before and after these correction methods. However, these methods do not detect or correct for other prominent artifacts like ghosting, herringbone, and chemical shifts, further underlining the need for a comprehensive QC paradigm, outside of these artifact correction packages.

While traditional feature-based machine learning methods can be considered as a natural choice for training artifact detection, these require careful feature selection, which can present a challenge considering the variety of artifacts, and noise, in dMRI data. This is further compounded by the fact that the same artifact may present differently across scanners/sites, making the feature-based learners' site and scanner specific. Human QC experts rely on the brain's ability to identify and integrate patterns specific to artifacts in dMRI data to detect them. Deep learning tools, especially convolutional neural networks (CNNs), that emulate human visual feature extraction in an automated manner, can be a very powerful tool for training an automated QC detector. The superior performance of CNN in many computer vision tasks, and in medical imaging, motivated us to use it to train an automated QC method for dMRI data.

In order to train a CNN that emulates human behavior, a large set of parameters need to be optimized during the training process, which in turn necessitates a high volume of training data (slices with artifacts, and slices of good brain tissue), and increased computational cost. Providing this huge volume of data is a challenging task, especially in medical imaging. In order to fulfill the requirement of a large amount of labeled data for training a deep CNN, transfer learning (Mazurowski et al., 2018) is used. Transfer learning involves taking a pre-trained CNN and re-training a subset of its parameters using a smaller amount of data to perform well on a new task (Mazurowski et al., 2018). As a result of this vast reduction in the number of parameters, transfer learning has the advantage of requiring less training time and computational cost. Pre-trained models have been applied successfully to various computer vision and medical imaging tasks, such as breast cancer diagnosis in digital breast tomosynthesis from mammography data (Samala et al., 2018a,b), classification of radiographs to identify hip osteoarthritis (Xue et al., 2017), or diagnosis of retinal diseases in retinal tomography images (Rampasek and Goldenberg, 2018). As our sample size was limited due to the difficulty of manual QC labeling of dMRI data, we adopted a transfer learning approach in this paper.

A significant problem in artifact detection is that the same artifact may present differently across sites and scanners. In order to make the CNN insensitive to scanner and site differences, we use manually labeled datasets from different sites and scanners. In addition to this, we apply data augmentation techniques that led to demonstrably improved results of CNN classifiers (Wang and Perez, 2017). In this manner, classical image transformations, including rotating, cropping, zooming, and shearing, are applied on the original images to increase the heterogeneity of the sample, by providing a simulated variation of the original data. In the process, both heterogeneity and size of the sample are increased.

In summary, we present a CNN-based automated QC paradigm, called QC-Automator, to detect various artifacts in dMRI data, including motion, multiband interleaving, ghosting, susceptibility, herringbone, and chemical shift. We will use transfer learning and data augmentation. The method will be trained and cross-validated on a large sample of expert-labeled images that combine dMRI data from multiple scanners.

MATERIALS AND METHODS

Proposed method contains two CNN-based classifiers, one for artifacts that manifest clearly in axial slices (e.g., ghosting), and one for artifacts that manifest in sagittal slices (e.g., motion). An input dMRI volume is converted into axial and sagittal slices and the slices are sent to the axial or sagittal classifier correspondingly. Finally, the slices in which artifacts are detected, by either of the two classifiers described above, are flagged and a slice-wise report is created.

We first describe the datasets that are used for training and testing, and describe the different artifacts in the section "Database for Training and Testing QC-Automator." QC-Automator is described next. The performance of a number of different CNN architectures is compared for their suitability to the problem of artifact detection and compared to traditional machine learning approaches using texture features. Additionally, we report the performance of the detectors on data of different acquisition protocols, that are not a part of the training set.

Database for Training and Testing QC-Automator

Our database for all the following experiments included data from 155 unique subjects across 5 different scanners and dMRI acquisition schemes. The details are reported in **Table 1**. The ground truth labels in this paper were provided by manual visual inspection. In order to reduce the manual labeling errors, QC was done by two experts with 2–8 years of experience. The labels were binarized, in order to create a classifier which categorizes images as “artifact free” or “artifactual.”

The artifacts labeled from these datasets were divided into six categories, motion, multiband interleaving, ghosting, susceptibility, herringbone, and chemical shifts. These six categories of artifacts manifested differently in the images; thus, the QC experts inspected axial slices for herringbone, chemical shift, susceptibility, and ghosting artifacts, while they inspected sagittal slices for motion and multiband interleaving artifacts.

To exclude slices that capture the periphery of the brain which contain mostly background voxels, we excluded sagittal slices which were entirely outside of the brain and five sagittal slices starting from the left and right edges of the brain. We excluded five axial slices inferior to the superior surface of the skull, as well as slices superior to the skull and inferior to the cerebellum, as they represent non-brain tissue. Overall, ∼132,000 axial slices and ∼200,000 sagittal slices were annotated as either artifactual or artifact-free. Details are reported in **Table 2**. **Figure 1** shows representative examples of the artifacts that were annotated, based on the view used.

Convolutional Neural Networks

An Overview

Convolutional neural networks are a special kind of artificial neural network that are composed of a set of convolutional and pooling layers in their architectures (**Figure 2**). Convolutional layers are designed to detect certain local features throughout the input image; they perform a convolution operation to the input image and pass the result to the next layer, a pooling layer, which reduces the dimensionality of the data by combining the outputs of a set of neurons into a single one, via a max or average operation. A sequence of convolutional and pooling layers is followed by some successive fully connected layers, in which all the neurons in a prior layer are connected to all the neurons in the next layer. Finally, a softmax, or regression layer, tags the data with the desired output label (Krizhevsky et al., 2012).

Various CNN architectures have been proposed in the literature. The VGG (Simonyan and Zisserman, 2014) networks, along with the earlier AlexNet (Krizhevsky et al., 2012), are the most basic architectures which follow the traditional layout of CNNs as shown in **Figure 2**. ResNet (He et al., 2016), Inception (Szegedy et al., 2015), and Xception (Chollet, 2017) are newer architectures. While ResNet introduces residual networks that make some connections between non-consecutive layers in very deep networks, Inception uses a module that performs different transformations over the same input in parallel and concatenates their results. Xception, on the other hand, is based on separating cross-channel and spatial correlations. Each of these architectures convey their own unique advantages and pitfalls, warranting a comparison of the performance of different CNN architectures.

Transfer Learning

To train CNNs, a large number of parameters need to be optimized, which in turn requires a large amount of computational power and labeled training data (in our case, the database of artifactual and artifact-free slices). Manually labeling data, however, are a time-consuming process, and with the limited number of datasets in medical imaging, it may not be possible to create a large and heterogeneous enough database to train a CNN from scratch for a given task. To overcome these issues, transfer learning methods have been proposed in which, an existing CNN network/architecture, pre-trained on a certain task, is adapted to a new task and CNN parameters are adjusted for a few layers of the network.

In general, the early layers of a CNN learn low-level features, which are applicable to most computer vision tasks, while the subsequent layers learn high-level features that are mostly application-specific. Therefore, adjusting the last few layers of an existing CNN architecture is usually sufficient for transfer learning (Tajbakhsh et al., 2016).

The efficiency of the transfer learning method depends on the similarity between the images of the database that the selected CNN architecture was trained on, and the database that the CNN is transferred to. Although the heterogeneity between the images used in the pre-trained CNNs (see the section “An Overview”) and medical imaging databases is considerable, an extensive study on medical imaging data has demonstrated that adjusting the parameters of an existing, pre-trained CNN, is as effective as training a CNN from scratch while being more robust to the size of training data (Tajbakhsh et al., 2016) and requiring significantly less computational power.

TABLE 1 | The acquisition parameters across our datasets.

Datasets	Number of subjects	*b*-values (s/mm^2)	Number of repeated acquisitions	Number of *b* = 0 images	Number of weighted gradients	TR (ms)	TE (ms)
Dataset-1	30	1000	2	1	32	8000	51
Dataset-2	32	1000	2	7	30	6500	84
Dataset-3	17	300, 800, 2000	1	9	108	4300	75
Dataset-4	31	1000	1	7	64	8100	82
Dataset-5	57	1000	1	1	30	11,000	76.4

TABLE 2 | Distribution of different types of artifacts in our dataset.

Artifact type	Slice view	Total samples
Herringbone	Axial	120
Chemical shift	Axial	1054
Susceptibility	Axial	442
Ghosting	Axial	11,619
Motion	Sagittal	21,436
Multiband interleaving	Sagittal	4017
Total-artifact	Axial	13,235
Total-artifact	Sagittal	25,453
Total-artifact-free	Axial	118,641
Total-artifact-free	Sagittal	179,911

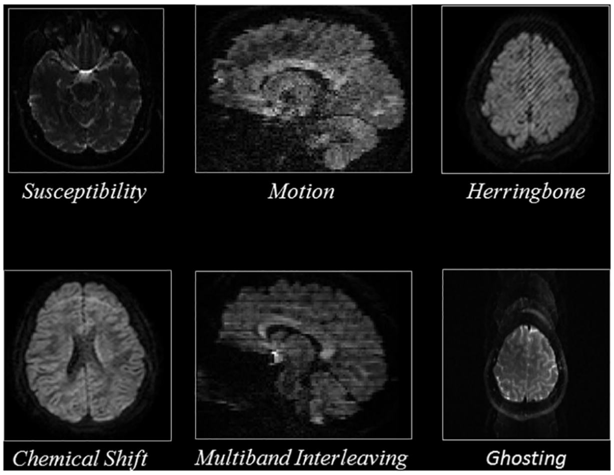

FIGURE 1 | Representative slices of the different artifacts that the QC-Automator was trained to detect.

As providing the manual labels for dMRI data is a time consuming and laborious task, our sample size was limited and insufficient to train a CNN from scratch. In this paper, we used transfer learning, to create QC-Automator, described in detail in the following section.

Creation of QC-Automator

Figure 3 shows the pipeline of the proposed approach. The artifacts detected by QC-Automator are motion, multiband interleaving, ghosting, susceptibility, herringbone, and chemical shifts. As different artifacts manifested more clearly either in the axial or sagittal view, QC-Automator consisted of two detectors: the "axial detector" which detected artifacts that presented better axially (herringbone, chemical shift, susceptibility, and ghosting) and the "sagittal detector" which detected artifacts that presented in the sagittal plane (motion and multiband interleaving artifacts).

For creating training samples, every dMRI volume was converted to axial and sagittal slices, and was assigned manual labels (see the section "Database for Training and Testing QC-Automator" for details). The two detectors were fed a new dMRI volume, in order to determine the slices that manifest with artifacts.

Transfer Learning and Data Augmentation for QC-Automator

In the proposed architecture, the detectors were CNN based and applied transfer learning to adapt existing knowledge, obtained from a large database of labeled training samples in other domains, to our problem of artifact detection (see the section "Transfer Learning" for details). The transfer learning process consists of two main steps: selection of the pre-trained model, and applying the pre-trained model to the new domain. To select an optimal pre-trained model for our axial and sagittal detectors, we compared the performance of four different pre-trained CNN architectures namely, VGGNet, ResNet, Inception, and Xception. These CNN architectures were pre-trained on ImageNet (Russakovsky et al., 2015), which is the most popular public dataset with a very large amount of labeled images across various number of classes. This makes these architectures capable of learning generic features from images, making them good feature extractors for a variety of classification tasks.

To implement transfer learning, we removed the top layer of a pre-trained CNN and replaced it with a fully connected layer with 256 neurons, followed by a softmax layer which performs the classification between two classes (artifact-present vs. artifact-free). All parameters of CNN architectures were fixed except those in the newly added layer, which were re-trained with the augmented manually labeled artifactual and non-artifactual data described above.

Data augmentation techniques are strategies that enable a significant increase in the diversity and size of data available for training, without collecting new data. They perform different image transformations to provide a simulated variation of the original data for training. We performed extensive data augmentation of the manually labeled data by applying horizontal and vertical translations, rotations, zooming, shearing, and flipping of the original slices. This was undertaken to increase the sample size of the labeled dataset, as well as to increase the heterogeneity of the data.

Training QC-Automator

The two classifiers were trained using the first three datasets, by passing axial and sagittal slices of the brain along with ground-truth labels (artifactual, or artifact-free). For both classifiers, the intensity values for each slice were normalized to have zero center and unit variance as calculated by the value subtracted by the mean and divided by the standard deviation. Training was done for 20 epochs using the RMSprop optimizer with a learning rate of 2×10^{-4} and a cross entropy loss function. The network structure was implemented in Python, using Keras with Tensorflow as the backend (Python 2.7, Keras 2.0.8, Tensorflow 1.3.0).

Slice-Based and Volume-Based Reports

Quality control Automator was designed to produce a report of the presence or lack of artifacts in individual slices in a diffusion-weighted image. However, an alternate way of reporting such

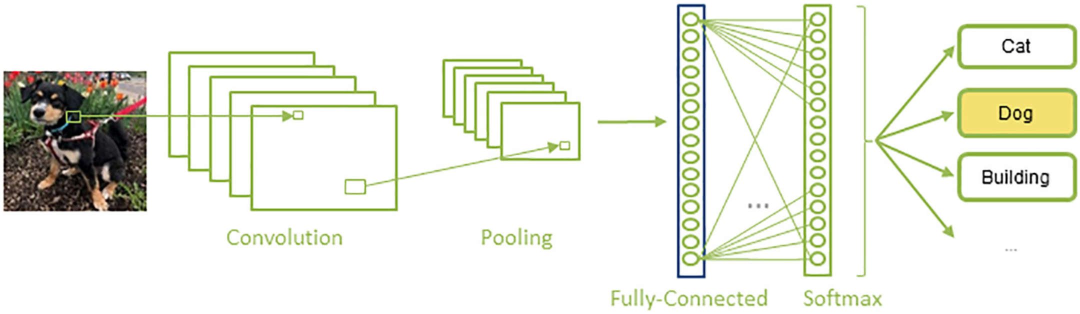

FIGURE 2 | A typical architecture of a CNN: A set of convolution and pooling layers with successive fully connected and softmax layer.

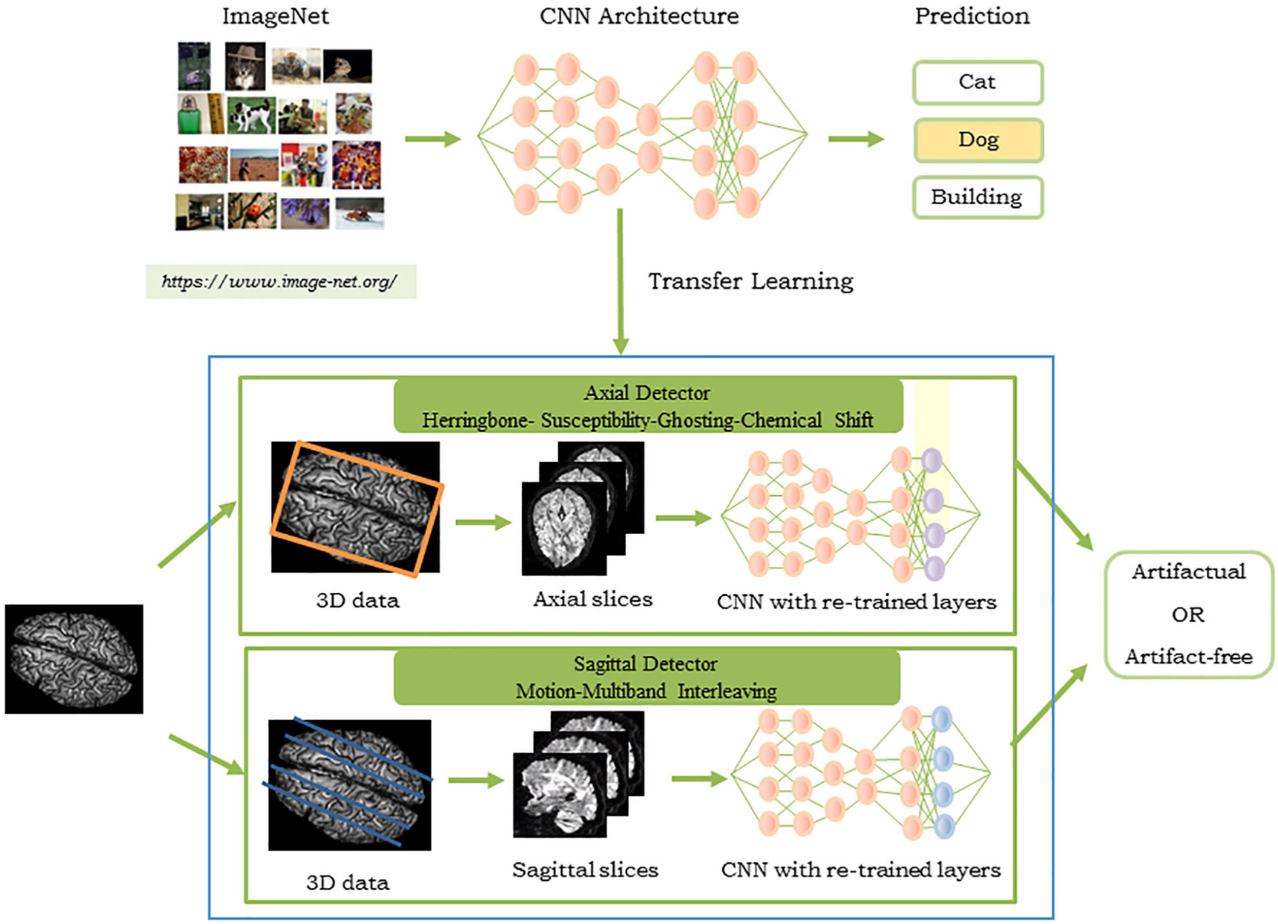

FIGURE 3 | Pipeline of the proposed approach for the QC-Automator: **(Top)** CNN pre-trained on ImageNet to obtain parameters used for transfer learning, where the last layer of the network was re-trained with our dataset of manually labeled artifactual and artifact-free data. The process was replicated to create the axial **(Middle)** and the sagittal detector **(Bottom)**. The blue box represents the QC-Automator. Given an input image **(Left)**, both the axial and sagittal detectors are applied to it and the status of each slice as artifact-free or artifactual is predicted.

information is based on the presence or absence of artifacts in an entire volume. To this end, we used a slice-count threshold to label a volume as "artifactual." If QC-Automator found that a given volume contained more artifactual slices than the slice-count threshold, it would flag this volume as artifactual. While choosing low values of slice-count threshold could lead to over detection, choosing high values for threshold could lead to higher chances of missing artifacts by not flagging a volume. We chose different slice-count values for the threshold from 1 to 10 in order to find an optimal threshold.

To summarize the pipeline of the QC-Automator, an input dMRI volume was sliced axially and sagittally, and the respective slices were sent to the axial classifier, or the sagittal classifier. The presence of an artifact was detected by either of the two classifiers, and the artifactual slices were flagged in a slice-wise report.

Evaluating the Performance of QC-Automator

The following measures were used to evaluate the performance of QC-Automator:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where true positive (TP) represents the number of cases correctly recognized as artifactual, false positive (FP) represents the number of cases incorrectly recognized as artifactual, true negative (TN) represents the number of cases correctly recognized as artifact-free and false negative (FN) represents the number of cases incorrectly recognized as artifact-free.

Comparing CNN Architectures to Traditional Methods

The performance of the four different architectures (VGGNet, ResNet, Inception, and Xception) was evaluated on the first three datasets in a fivefold cross-validation setting. The slices from the first three datasets were shuffled and randomly partitioned into five equally sized subsamples. For each run, a single subsample was retained as test data, while the remaining four subsamples were used as training data, and this process was repeated for all five subsamples. This cross-validation process was run on the "axial" and "sagittal" detectors separately.

We compared our approach with three traditional feature extraction and learning methods: Gabor features, Zernike moments, and local binary patterns. Gabor features are constructed from the responses of applying Gabor filters made on several frequencies (scales) and orientations (Manjunath and Ma, 1996). We applied Gabor filters with four directions and four scales. Zernike moments are a global image feature constructed by projecting the image onto Zernike Polynomials, which are a set of orthogonal basis functions mapped to the unit circle in different orders and repetitions (Khotanzad and Hong, 1990; Revaud et al., 2009). We applied Zernike moments with order 4 and repetition 2. Local binary patterns are a non-parametric method to detect local structures of images by comparing each pixel with its neighboring pixels (Ojala et al., 2002; Huang et al., 2011). We applied the aforementioned features in combination with random forest classifiers using the same cross-validation scheme described above.

In order to investigate the CNN classifiers and the traditional filters individually, we designed two more experiments. In the first experiment, the outputs of Gabor filters on images were fed into a fully connected layer with 256 neurons followed by a dropout and a softmax layer. In the second experiment, we applied principal component analysis (PCA) on the output of final convolutional layer of the CNN. We kept enough principal components to cover 98% of the variation in the data, and fed them into an SVM classifier.

Evaluating Performance on New Datasets

We performed the following experiments to evaluate whether the artifact detection is replicable to other dMRI datasets with different acquisition protocols. We tested the applicability of the QC-Automator on two new datasets, Dataset 4 and Dataset 5 (details are in **Table 1**). These datasets contained a variety of artifacts, encompassing all those that the detector was trained to detect, but were acquired with different scanning parameters. The training was done using the first three datasets, while the performance was evaluated using data from the fourth and fifth datasets.

In addition, we investigated whether generalizability in performance across datasets was improved by retraining QC-Automator after adding a small subsample (10%) of the new datasets to the training set, to see if incorporating samples could improve the accuracy, precision, and recall of the classifier versus application to a hitherto unseen dataset. We performed two different experiments by adding data from Dataset 4 and Dataset 5 to the original training set, separately.

RESULTS

Comparison Across CNN Architectures and Traditional Methods

Tables 3, **4** show the performance of our artifact detection method, using different architectures. As VGGNet outperformed other architectures, it was selected as the architecture of choice for QC-Automator. Using VGGNet, we obtained 98% accuracy for all artifacts in both the axial and sagittal detectors. Precision and recall values are reported accordingly. Representative instances of the true and false detections for QC-Automator are shown in **Figures 4**, **5**.

Tables 5, **6** compare our method with traditional pattern recognition approaches including Gabor filters, Zernike moments, and local binary patterns in combination with random forest classifiers. As seen, VGGNet outperformed the traditional methods. **Table 7** shows the result of applying Gabor filters to a fully connected layer and **Table 8** shows the results of performing SVM on top of VGGNet final convolutional layer features after PCA. Although Gabor filters and SVM classifiers could achieve high accuracy (87 and 91% for axial detector), the value of precision and recall was poor compared to our method using CNNs, showing that our transfer learning approach outperformed traditional SVM classifiers and Gabor filters for this task.

Volume-wise results for VGGNet are reported in **Tables 9**, **10**. As seen, we obtained 96% accuracy for our axial detector at a slice-count threshold of three slices, and 98% accuracy for our sagittal detector at a slice-count threshold of seven slices. Correspondingly, recall values were 98 and 95% for the volume-wise axial and sagittal detectors. This means we only missed 2% of volumes that contain artifacts manifesting in axial view and 5% of sagittal ones.

TABLE 3 | The result of different CNN architectures in detecting artifact type 1 (Axial Detector).

	Accuracy	Precision	Recall
VGG 16	0.98	0.97	0.91
Resnet 50	0.89	0.82	065
Inception V3	0.96	0.89	0.82
Xception	0.96	0.88	0.82

TABLE 4 | The result of different CNN architectures in detecting artifact type 2 (Sagittal Detector).

	Accuracy	Precision	Recall
VGG 16	0.98	0.92	0.91
Resnet 50	0.98	0.91	0.78
Inception V3	0.98	0.90	0.67
Xception	0.99	0.92	0.82

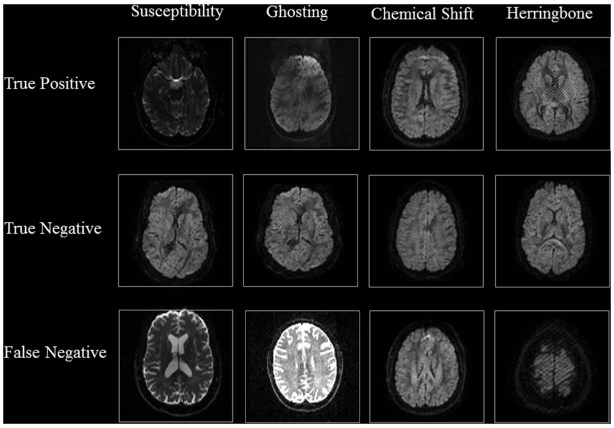

FIGURE 4 | Results of Axial Detector: Representative slices of correctly and incorrectly classified slices are presented.

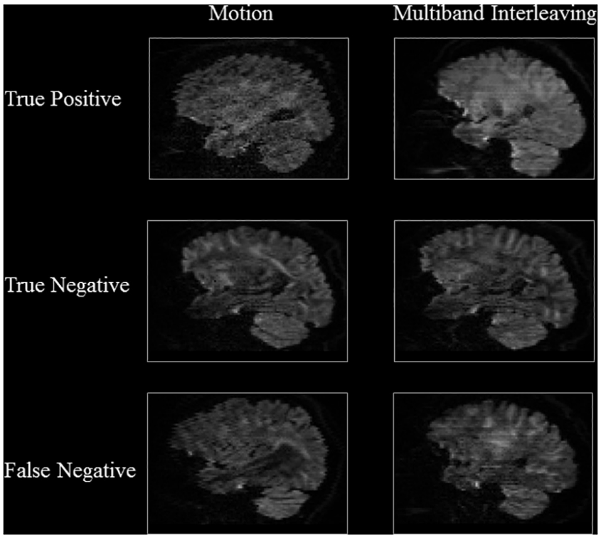

FIGURE 5 | Results of Sagittal Detector Representative slices of correctly and incorrectly classified artifactual slices.

TABLE 5 | Results of different texture features in detecting artifact type 1 (Axial Detector).

	Accuracy	Precision	Recall
Gabor 32	0.91	0.89	0.87
Zernike moments	0.87	0.58	0.19
Local binary patterns	0.83	0.85	0.12

TABLE 6 | Results of different texture features in detecting artifact type 2 (Sagittal Detector).

	Accuracy	Precision	Recall
Gabor 32	0.98	0.96	0.48
Zernike moments	0.97	0.45	0.55
Local binary patterns	0.97	0.40	0.52

TABLE 7 | The result of Gabor filter combined with fully connected layers.

Gabor filters – fully connected	Accuracy	Precision	Recall
Axial detector	0.87	0.37	0.35
Sagittal detector	0.90	0.30	0.46

TABLE 8 | The result of feeding CNN features to support vector machines.

CNN–SVM	Accuracy	Precision	Recall
Axial detector	0.91	0.94	0.85
Sagittal detector	0.87	0.93	086

Performance on New Datasets

These experiments were undertaken in order to evaluate whether the artifact detection is replicable to other datasets acquired through different imaging protocols. The detectors were trained on the first three datasets and tested on the fourth and fifth datasets. Their performance is reported in **Tables 11**, **12**. For Dataset 4, the accuracy of detecting artifacts through the axial detector decreased by 7% comparing to the previous results in **Table 3**. There was also a 14% decrease in the accuracy of the sagittal artifact detector, compared to **Table 4**. For Dataset 5, the value of accuracy dropped by 7 and 11% for the axial and sagittal detectors, respectively.

In order to see if these results could be improved, we evaluated the results of adding a small percentage of the new datasets to the original training data, to acclimatize the deep learner to new scanning parameters. We added a small subset (10% of each whole dataset) from the fourth and fifth datasets to the original training set, the results of which are displayed in **Tables 13**, **14**. It can be seen that we attained a higher accuracy, recall, and precision than those of the previous experiment (**Tables 11**, **12**). Results were in the range of 90% recall for both new datasets, demonstrating that we missed <10% of artifacts. We provided an example of a FP case for this experiment in **Figure 6**.

TABLE 9 | QC-Automator volume-wise result – Axial Detector.

Threshold	Accuracy	Precision	Recall
Threshold = 1	0.92	0.86	0.99
Threshold = 3	0.96	0.94	0.98
Threshold = 5	0.94	0.97	0.90
Threshold = 7	0.0.87	0.97	0.74
Threshold = 10	0.84	0.69	0.67

TABLE 10 | QC-Automator volume-wise result – Sagittal Detector.

Threshold	Accuracy	Precision	Recall
Threshold = 1	0.74	0.64	0.97
Threshold = 3	0.90	0.79	0.96
Threshold = 5	0.97	0.87	0.95
Threshold = 7	0.98	0.94	0.95
Threshold = 10	0.98	0.97	0.95

TABLE 11 | Results of applying the QC-Automator to the fourth dataset.

	Accuracy	Precision	Recall
Artifact type-1 axial	0.91	0.75	0.81
Artifact type-2 sagittal	0.84	0.70	0.79

TABLE 12 | Results of applying the QC-Automator to the fifth dataset.

	Accuracy	Precision	Recall
Artifact type-1 axial	0.91	0.91	0.71
Artifact type-2 sagittal	0.87	0.75	0.69

TABLE 13 | Results of applying the QC-Automator on the fourth dataset, after adding small subsample (10%) data from the fourth dataset to the training set.

	Accuracy	Precision	Recall
Artifact type-1 axial	0.94	0.87	0.91
Artifact type-2 sagittal	0.95	0.84	0.90

TABLE 14 | Results of applying the QC-Automator on the fifth dataset, after adding small subsample (10%) from the fifth dataset to the training set.

	Accuracy	Precision	Recall
Artifact type-1 axial	0.89	0.82	0.91
Artifact type-2 sagittal	0.94	0.84	0.94

DISCUSSION

In this paper, we created an automated QC method, QC-Automator, using CNN and transfer learning, via data augmentation on a manually labeled dataset encompassing several scanners and dMRI acquisition parameters. We demonstrated the ability of QC-Automator to distinguish between artifactual slices from artifact-free ones, as well as its performance across different acquisitions from multiple sites. Given a diffusion MRI volume, the QC-Automator was able to flag slices based on the presence of several artifacts, including motion, multiband interleaving, ghosting, susceptibility, herringbone, and chemical shift. The flagged slices can be manually inspected to determine if the corresponding volume would be safe to use for further analysis for a given study.

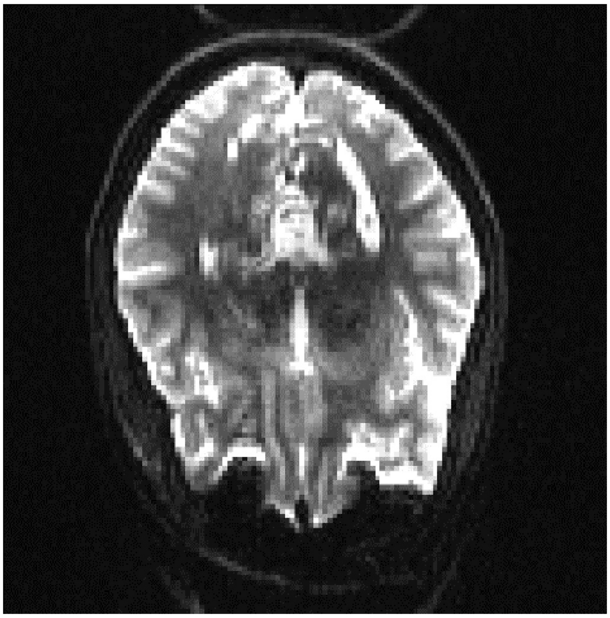

FIGURE 6 | A sample of false positive slice for Dataset 4: The slice contains aliasing artifact. Our expert labeled it as artifact-free one. But our QC-Automator caught it as it contained a similar pattern to ghosting artifact.

The QC-Automator consisted of two classifiers: one for all artifacts that manifest in the axial view (namely herringbone, chemical shift, susceptibility, and ghosting), and one for artifacts that manifest in the sagittal view (namely motion and multiband interleaving). For both the classifiers, VGGNet performed better than Inception, ResNet, and Xception, based on the comparison of transfer learning results for various architectures (**Tables 3**, **4**). This might be because of the uniform structure of VGGNet, which uses consecutive layers of 3 × 3 filters and max pooling, with each successive layer detecting features at a more abstract, semantic level than the layer before. Residual networks introduce connections between layers at different resolutions, which results in a jump in the semantic abstraction. Inception and Xception networks compute and concatenate multiple different transformations over the same input. These architectures are more complex and did not perform as well on our data.

The representative slices in **Figures 4**, **5** demonstrate that our method correctly classified artifacts in different slices of the brain. Despite having correctly classified most artifacts, the QC-Automator also incorrectly flagged some artifactual slices as artifact-free, and we inspected some of these examples. We hypothesize that the FN case for ghosting (**Figure 4**) happened because the pattern of ghosting was particularly faint in this specific slice. For herringbone, chemical shift, and susceptibility artifacts, our classifier successfully labeled multiple slices of the given volume as artifactual, but sometimes failed to label slices where the artifact was less prominent (**Figure 4**). Thus, although our classifier failed to correctly label some artifactual slices, it was able to capture adjacent slices where the artifact was more prominent. As the pattern of artifact is more visible in susceptibility, herringbone, and chemical shift, we believe we can get better performance by adding more training data for other artifacts in the future.

The transfer-learning-based approach presented in this paper performed better than Gabor filters, Zernike moments, and local binary patterns in combination with random forest classifiers (**Tables 5**, **6**). Gabor filters performed better than Zernike moments and local binary patterns. The fact that Gabor filters analyze the presence of specific frequencies in specific directions of localized regions in the input image might explain this result. These specific frequencies can capture the edge patterns in motion and multi-band interleaving, the checkerboard pattern in herringbone and chemical shift, and the curves visible in the background of ghosting artifacts. While Gabor filters had the best performance of the three, the precision and recall were still poor compared to VGGNet. Zernike and local binary patterns on the other hand look for patterns of intensity. This is enough for detecting high intensities but fail to find patterns of edges and curves. However, the performance of these methods is bound to the quality of the features, which need human experts to hand-craft them manually. The fact that the Gabor features did well lends support to the notion that most of our features were discriminated in the early layers of the CNN, and thus the transfer learning approach, which consists of adjusting only last layer, performed well. The proposed approach also performed better than Gabor filters in combination with fully connected layer neural networks (**Table 7**) and it performed better than Support Vector Machine classifiers (**Table 8**). This indicates that VGGNet is a good choice both as a feature extractor and as a classifier.

As an alternative to the slice-wise report, we also measured the performance of QC-Automator when reporting the presence of artifact in an entire volume (**Tables 9**, **10**). This way of reporting is easier for the human analyst to interpret, than a flat list of bad slices. In this manner, a volume was labeled as artifactual if it has more artifactual slices than a certain threshold. For the axial detector, we observed 99% recall with a lower threshold, meaning we detected 99% of volumes containing artifacts. However, the precision was 86% at this point, implying that we over-detected in 14% of cases. As we increased the threshold, the precision improved and the recall decreased as the detector missed some artifactual volumes. Optimal results appeared at threshold of three slices, with precision of 95% and recall of 98%. For our sagittal detector, however, the optimal threshold was higher. We got 97% recall at slice-count threshold of one slice, while precision was poor at this point (64%). As we increased the threshold, precision improved and recall dropped. The optimal point was at a slice-count threshold of 10 slices, as it had the highest values of recall and precision. This difference in the optimal thresholds between the detectors might be because of the nature of motion artifacts, which are generally visible in more than one sagittal slices in a volume. Overall, with the thresholds of 3 and 10 for axial and sagittal detector, respectively, QC-Automator only missed 2% of artifactual volumes which contain artifact in axial view and 5% of volumes with artifact present in sagittal view. However, these thresholds might not be ideal for all cases. Data analysts are encouraged to use their own slice-count thresholds for flagging volumes based on the data quality requirement of their given study.

Furthermore, the framework was tested on how well it performs on acquisitions from different scanners. Evaluation was performed by training on three datasets and testing on the two remaining datasets, which had different acquisition parameters compared to the three training datasets. The corresponding results for Datasets 4 and 5 are shown in **Tables 11**, **12**. It indicates that, we could achieve high accuracy (90% approximate average), however, the values of precision and recall decreased. To inspect the decrease in precision and recall, we added small subsample of the fourth and fifth datasets to the training set which covers 10% of the dataset (see the section "Performance on New Datasets"). In this experiment, we achieved higher accuracy, close to the intra-dataset experiment (**Tables 3**, **4**). The value of precision and recall also increased substantially for both detectors in both datasets (∼10% for Dataset 4 and 20% for Dataset 5). This suggests that adding a small subsample of the new datasets to the original training set could decrease the false detection. As seen, in this experiment we achieved a higher value of recall, around 90% for both datasets, showing that QC-Automator had low chance of missing an artifact, while staying in the range of 85% for precision. This indicates that there were some artifact-free slices that are detected as artifactual representing that our method over-detected in some cases. However, considering the nature and purpose of QC, a FP is favorable to FN, as we do not want to miss an artifact. To summarize, by adding a subsample of the new datasets to the original training set, a drastic increase in recall was observed, giving us reason to believe that the classifier could be gradually improved to reach the same level of precision and recall as that of the intra-dataset experiment. This means that with a little effort we can apply our classifier to a new dataset.

Moreover, we inspected the FPs of the cross-dataset experiment which uncovered another potential cause of FP; an error in labeling of the data. As it can be seen in **Figure 6**, there was an aliasing artifact inside the slice, despite the fact that our QC expert had labeled that slice as artifact-free. However, our classifier detected them as artifactual slices possibly due to the fact that this slice had similar patterns to the ghosting artifact. The fact that QC-Automator was able to detect such artifacts, despite potential mislabeling in the training dataset, indicates the high performance of the detectors.

Comprehensively, QC-Automator is able to detect artifacts in a fraction of time comparing to manual labeling, which is more prone to errors introduced by subjectivity and fatigue on part of the data analyst. Considering the constantly increasing size of datasets, we believe that this contribution is a valuable framework, and can save a tremendous amount of time and effort. QC Automator is the first tool that can detect the wide range of artifacts presented in this paper.

With respect to the practical usage of QC-Automator, a few specifics need to be highlighted. QC-Automator is trained to detect the presence of artifacts, by performing slice-wise detection in both the axial, and sagittal planes, and does not alter or correct the data in any way. It creates a report, documenting the slice-wise presence of artifacts, that an analyst can use to zone in on scans that need inspection. Alternatively, a user can set slice-count thresholds, to create a report that flags volumes instead of individual slices as described in the section "Slice-Based and Volume-Based Reports." This reduces the number of images that an analyst needs to go through, to capture most of the artifacts present in a dataset. As an example, for manual QC, our analysts reported spending a total of 69.38 h, to inspect 4163 volumes, finding 557 volumes with axial artifacts and 138 volumes with sagittal artifacts. Comparing manual QC with our method (**Tables 9**, **10**), at $T = 3$ axial threshold, we note a FN rate of 2%. If the data analyst decides to go through the flagged volumes reported at this threshold, this leads to a reduction of volumes to be inspected by 86.05%. Similarly, at $T = 10$ sagittal threshold, we note a FN rate of 5%, coupled with a 96.75% decrease in the number of volumes that analyst has to go through. QC-Automator is meant to be used as a tool for speedy and reliable detection of artifacts in large datasets where manual QC is extremely time-consuming.

Despite the impressive results of QC Automator, there is still room for improvement, such as by adding more training data. We trained our classifier on three datasets acquired on different scanners with varying fields of view and gradient sampling schemes and tested our classifier on two other datasets, again of different acquisition sequences. We observed a high accuracy in our cross-dataset experiment; however, there was a decrease in the precision and recall implying higher rates of false detection. We believe that this issue can be solved by adding small subsamples of the target dataset so the training set so classifier can gradually get improved over-time with seeing more data.

Quality control Automator was trained to detect artifacts across multiple datasets with varying acquisition parameters and it was trained and tested on brain images of healthy participants. Efficacy of QC-Automator was not tested in the presence of signs of pathology such as brain tumors or micro-bleeds. Additionally, future automated artifact detection methods would do well to test the efficacy of their algorithms in detecting artifacts on poorly acquired scans, such as those with a partial field of view.

The ground truth labels in this paper were provided manually. Artifacts manifest differently in different slices, from very subtle to clearly visible patterns. The subjectivity of manual visual inspection in our case was lowered by labeling using two experts, with varying degrees of expertise. The labels were binarized into two classes to create a classifier to categorize images as "artifact-free" or "artifactual." If an objective "artifact severity" threshold can be determined through characterization of artifacts, it might provide a better alternative to the use of binary labels.

Overall, the QC-Automator can gain from large training samples, limited by the effort and quality of manually labeling data on different artifacts. Given the recent progress in deep networks, and further advances in GPU hardware, the accuracy of convolutional neural nets is expected to further improve in the future. That provides the potential for better QC tools.

CONCLUSION

In summary, QC-Automator is a deep learning-based method for QC of diffusion MRI data that are able to detect a variety of artifacts. QC is a well-suited task for CNNs. The difficulty in obtaining huge amounts of expert-labeled dMRI data to train a CNN is alleviated by using transfer learning, and data augmentation. The proposed approach achieves superior performance with respect to pattern recognition methods and is considerably faster and less computationally expensive in comparison to purely learning-based approaches with neural networks. We demonstrated that our method achieves high accuracy and generalizes well to other datasets, different from those used for training. This artifact detector enhances analyses of dMRI data by flagging artifactual slices. This substantially reduces the effort and time of human experts and allows for an almost instantaneous access to clean dMRI data.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This study was approved by the Institutional Review Board of the University of Pennsylvania.

AUTHOR CONTRIBUTIONS

ZS: methodology, writing – original draft, and visualization. JA: data labeling, writing – review and editing, and visualization. DP: conceptualization, data labeling, and writing – review and editing. AI: visualization and writing – review. RV: conceptualization, writing – review and editing, and supervision.

FUNDING

This work was supported by funding from the National Institute of Health (NIH R01MH117807 and NIH R01HD089390, PI: RV).

REFERENCES

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., and Douaud, G. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* 166, 400–424. doi: 10.1016/j.neuroimage.2017.10.034

Andersson, J. L., Graham, M. S., Zsoldos, E., and Sotiropoulos, S. N. (2016). Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. *NeuroImage* 141, 556–572. doi: 10.1016/j.neuroimage.2016.06.058

Assaf, Y., and Pasternak, O. (2008). Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review. *J. Mol. Neurosci.* 34, 51–61. doi: 10.1007/s12031-007-0029-0

Bammer, R., Markl, M., Barnett, A., Acar, B., Alley, M., Pelc, N., et al. (2003). Analysis and generalized correction of the effect of spatial gradient field distortions in diffusion-weighted imaging. *Magn. Resonance Med.* 50, 560–569. doi: 10.1002/mrm.10545

Basser, P. J., and Jones, D. K. (2002). Diffusion-tensor MRI: theory, experimental design and data analysis–a technical review. *NMR Biomed.* 15, 456–467. doi: 10.1002/nbm.783

Bastiani, M., Cottaar, M., Fitzgibbon, S. P., Suri, S., Alfaro-Almagro, F., Sotiropoulos, S. N., et al. (2019). Automated quality control for within and between studies diffusion MRI data using a non-para-metric framework for movement and distortion correction. *NeuroImage* 184, 801–812. doi: 10.1016/j.neuroimage.2018.09.073

Chollet, F. (2017). Xception: deep learning with depthwise separable convolutions. *arXiv [preprint]*

Graham, M. S., Drobnjak, I., and Zhang, H. (2018). A supervised learning approach for diffusion MRI quality control with minimal training data. *NeuroImage* 178, 668–676. doi: 10.1016/j.neuroimage.2018.05.077

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ: IEEE).

Heiland, S. (2008). From A as in Aliasing to Z as in Zipper: artifacts in MRI. *Clin. Neuroradiol.* 18, 25–36. doi: 10.1007/s00062-008-8003-y

Huang, D., Shan, C., Ardabilian, M., Wang, Y., and Chen, L. (2011). Local binary patterns and its application to facial image analysis: a survey. *IEEE Trans. Syst. Man Cybernet. Part C* 41, 765–781. doi: 10.1109/tsmcc.2011.2118750

Iglesias, J. E., Lerma-Usabiaga, G., Garcia-Peraza-Herrera, L. C., Martinez, S., and Paz-Alonso, P. M. (2017). "Retrospective head motion estimation in structural brain MRI with 3D CNNs," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, eds M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. Collins, and S. Duchesne, (Cham: Springer).

Jiang, H., Van Zijl, P. C., Kim, J., Pearlson, G. D., and Mori, S. (2006). DtiStudio: resource program for diffusion tensor computation and fiber bundle tracking. *Comput. Methods Programs* 81, 106–116. doi: 10.1016/j.cmpb.2005.08.004

Kelly, C., Pietsch, M., Counsell, S., and Tournier, J. D. (2016). "Transfer learning and convolutional neural net fusion for motion artefact detection," in *Proceedings of the Conference ISMRM*, Honolulu, HI.

Khotanzad, A., and Hong, Y. H. (1990). Invariant image recognition by Zernike moments. *IEEE Trans. Pattern Anal. Mach. Intell.* 12, 489–497. doi: 10.1109/34.55109

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 60, 1097–1105.

Krupa, K., and Bekiesińska-Figatowska, M. (2015). Artifacts in magnetic resonance imaging. *Pol. J. Radiol.* 80, 93–106. doi: 10.12659/PJR.892628

Liu, B., Zhu, T., and Zhong, J. (2015). Comparison of quality control software tools for diffusion tensor imaging. *Magn. Reson. Imaging* 33, 276–285. doi: 10.1016/j.mri.2014.10.011

Manjunath, B. S., and Ma, W. Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.* 18, 837–842. doi: 10.1109/34.531803

Mazurowski, M. A., Buda, M., Saha, A., and Bashir, M. R. (2018). Deep learning in radiology: an overview of the concepts and a survey of the state of the art. *arXiv [Preprint]*

Moratal, D., Vallés-Luch, A., Martí-Bonmatí, L., and Brummer, M. E. (2008). k-Space tutorial: an MRI educational tool for a better understanding of k-space. *Biomed. Imaging Interv. J.* 4:e15. doi: 10.2349/biij.4.1.e15

Oguz, I., Farzinfar, M., Matsui, J., Budin, F., Liu, Z., Gerig, G., et al. (2014). DTIPrep: quality control of diffusion-weighted images. *Front. Neuroinform.* 8:4. doi: 10.3389/fninf.2014.00004

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 971–987. doi: 10.1109/tpami.2002.1017623

Pierpaoli, C., Walker, L., Irfanoglu, M., Barnett, A., Basser, P., Chang, L., et al. (2010). TORTOISE: an integrated software package for processing of diffusion MRI data. in *Proceedings of the Book TORTOISE: an Integrated Software Package for Processing of Diffusion MRI Data ISMRM 18th Annual Meeting*, Stockholm.

Rampasek, L., and Goldenberg, A. (2018). Learning from everyday images enables expert-like diagnosis of retinal diseases. *Cell* 172, 893–895. doi: 10.1016/j.cell.2018.02.013

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J., and Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage* 107, 107–115. doi: 10.1016/j.neuroimage.2014.12.006

Revaud, J., Lavoué, G., and Baskurt, A. (2009). Improving Zernike moments comparison for optimal similarity and rotation angle retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 627–636. doi: 10.1109/TPAMI.2008.115

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Samala, R. K., Chan, H. P., Hadjiiski, L., Helvie, M. A., Richter, C., and Cha, K. (2018a). Cross-domain and multi-task transfer learning of deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Med. Imaging* 10575:105750Q7.

Samala, R. K., Chan, H. P., Hadjiiski, L. M., Helvie, M. A., Richter, C., and Cha, K. (2018b). Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Phys. Med. Biol.* 63:095005. doi: 10.1088/1361-6560/aabb5b

Schenck, J. F. (1996). The role of magnetic susceptibility in magnetic resonance imaging: MRI magnetic compatibility of the first and second kinds. *Med. Phys.* 23, 815–850. doi: 10.1118/1.597854

Simmons, A., Tofts, P. S., Barker, G. J., and Arridge, S. R. (1994). Sources of intensity nonuniformity in spin echo images at 1.5 T. *Magn. Reson. Med.* 32, 121–128. doi: 10.1002/mrm.1910320117

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv [preprint]*

Smith, R., Lange, R., and McCarthy, S. (1991). Chemical shift artifact: dependence on shape and orientation of the lipid-water interface. *Radiology* 181, 225–229. doi: 10.1148/radiology.181.1.1887036

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Silver Spring, MD: IEEE).

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., et al. (2016). Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans. Med. Imaging* 35, 1299–1312. doi: 10.1109/TMI.2016.2535302

Van Dijk, K. R., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* 59, 431–438. doi: 10.1016/j.neuroimage.2011.07.044

Victoroff, J., Mack, W., Grafton, S., Schreiber, S., and Chui, H. (1994). A method to improve interrater reliability of visual inspection of brain MRI scans in dementia. *Neurology* 44, 2267–2267. doi: 10.1212/wnl.44.12.2267

Wang, J., and Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convol. Neural Netw. Vis. Recogn. arXiv* [Preprint]. arXiv:1712.04621.

Wood, M. L., and Henkelman, R. M. (1985). MR image artifacts from periodic motion. *Med. Phys.* 12, 143–151. doi: 10.1118/1.595782

Xue, Y., Zhang, R., Deng, Y., Chen, K., and Jiang, T. (2017). A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. *PLoS One* 12:e0178992. doi: 10.1371/journal.pone.0178992

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.