2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Automatic Volumetric Quality Assessment of Diffusion MR Images via Convolutional Neural Network Classifiers

Nabil Ettehadi*, Xuzhe Zhang, Yun Wang, David Semanek, Jia Guo, Jonathan Posner, and Andrew F. Laine

Abstract— **Diffusion Tensor Imaging (DTI) is widely used to find brain biomarkers for various stages of brain structural and neuronal development. Processing DTI data requires a detailed Quality Assessment (QA) to detect artifactual volumes amongst a large pool of data. Since large cohorts of brain DTI data are often used in different studies, manual QA of such images is very labor-intensive. In this paper, a deep learning-based tool is developed for quick automatic QA of 3D raw diffusion MR images. We propose a 2-step framework to automate the process of binary (*i.e.,* 'good' vs 'poor') quality classification of diffusion MR images. In the first step, using two separately trained 3D convolutional neural networks with different input sizes, quality labels for individual Regions of Interest (ROIs) sampled from whole DTI volumes are predicted. In the second step, two distinct novel voting systems are designed and fine-tuned to predict the quality label of whole brain DTI volumes using the individual ROI labels predicted in the previous step. Our results demonstrate the validity and practicality of our tool. Specifically, using a balanced dataset of 6,940 manually-labeled 3D DTI volumes from 85 unique subjects for training, validation, and testing, our model achieves 100% accuracy via one voting system, and 98% accuracy via another voting system on the same test set.**

I. Introduction

Diffusion Tensor Imaging (DTI) [1] [2] is a relatively novel technique of Magnetic Resonance Imaging (MRI). DTI is widely used to study the neural and structural development of the human brain. For example, DTI is used to study white matter development over time [3-6], as well as other non-normal conditions such as Alzheimer's disease [7] [8], and Schizophrenia [9] [10]. Additionally, popular large cohort studies such as the Human Connectome Project [11] and the Adolescent Brain Cognitive Development (ABCD [12]) also use DTI data for different biomarker measurements. However, raw DTI data are often prone to various MRI artifacts such as motion, Eddy currents, ghosting, gradient distortions, etc. that could potentially influence the reliability of results. In order to achieve reliable and reproducible results in studies using DTI, the data needs to be checked for such artifacts. Hence, Quality Assessment (QA) of DTI volumes is an essential step prior to carrying any further analysis.

Traditional approaches for QA of DTI data are often performed via the opinion of an MRI expert (or sometimes a consensual opinion of a group of MRI experts) by visually inspecting all volumes (sometimes all individual slices) and manually detecting any patterns of artifacts. Large cohorts of data, such as the ABCD dataset, make this task extremely challenging, and labor-intensive. Moreover, because this approach to QA is subjective, experts may not necessarily agree on the quality of an image. Hence, there is a need for fast automatic robust QA of large DTI datasets.

Computerized QA of DTI data provides an automated solution to artifact detection and quality control of diffusion MR images. Several works have attempted to include some level of automatic QA of DTI data in their artifact correction tools such as FSL [13-15], DTIPrep [16], and DTI studio [17]. Such tools mostly work based on various statistical measures to detect and estimate the underlying noise characteristics [18] or Artificial Intelligence (AI) approaches to detect a specific class of artifact patterns (e.g., motion, and Eddy current) [19-21]. Several articles such as [22] and [23] have evaluated the performance of these tools to QA DTI data and have reported that such tools are mostly limited to detecting the specific type of artifact they have been designed for. Moreover, other AI-based approaches such as [24] that take into account a wider range of artifacts, often require slice-wise labels (as opposed to volume-wise labels) in their training set which is usually highly labor-intensive.

In this paper, we propose a Deep Learning-based (DL) tool for automatic and fast QA of 3D raw DTI volumes into 'good' and 'poor' quality classes regardless of the type of potentially existing artifacts. Unlike previous works, our framework only uses volume-wise labels instead of slice-wise labels which significantly reduces the workload for creating a manually labeled dataset for training and evaluation purposes. The proposed QA tool operates in two steps. In the first step, using 2 separately trained 3D Convolutional Neural Networks (CNNs), Regions of Interest (ROIs) of different sizes randomly sampled from the whole DTI volume are classified into 'good' or 'poor' quality classes. In the second step, the ROI-level predicted labels are utilized in a voting system with two different proposed tailored score functions (each can be used separately) to predict the quality label of the whole DTI volume. Our results demonstrate the practicality of our tool for fast automatic QA of 3D raw DTI volumes regardless of the class of existing artifacts.

The rest of this paper is organized as follows. The proposed DTI QA framework is fully detailed in Section II. Section III presents the data used for this project and the pre-processing

*Corresponding author: ne2289@columbia.edu

Research was supported by the National Institute of Mental Health award number NIMH R01MH110445.

Nabil Ettehadi, Xuzhe Zhang, and Andrew F. Laine are with the Department of Biomedical Engineering, Columbia University, New York, NY, 10027 USA.

Jia Guo is with the Department of Psychiatry at Columbia University Medical Center, and the Zuckerman Institute at Columbia University, New York, NY, 10032 USA.

Yun Wang, David Semanek, and Jonathan Posner are with Department of Psychiatry, Columbia University Medical Center, New York, NY, 10032 USA.

procedures. The performance of our model is then demonstrated in Section IV. Finally, the conclusion and future works are addressed in Section V.

II. METHOD

As mentioned in the previous section, our framework operates in 2 steps. The first step is to predict the labels of individual 3D ROIs sampled from whole DTI volumes. In the second step, the individual predicted labels are used to assign the final quality label for each whole DTI volume using a voting system. In this section, we detail the two major steps of our tool. Figure 1 demonstrates the schematic of our overall framework.

A. ROI-level label prediction using CNNs

In order to capture diverse artifacts locally and globally across the 3D diffusion MR image volume, we use two different CNNs with distinct 3D ROI sizes. In particular, for each DTI whole volume, we use 2 sets of ROIs: 1) a single center-cropped large ROI, as well as 2) a number of small overlapping ROIs. Using multi-size ROIs allows the overall framework to search for both local and global noise patterns across the image. For each ROI type (*i.e.,* large or small), quality labels are predicted independently using their corresponding CNN. An overview of the architectures of CNN 1 and 2 is depicted in Fig. 2. Both architectures consist of multiple blocks of Convolutional layers; each followed by a batch normalization layer. At the end of each block, a Max-pooling layer is utilized. Throughout both networks, the image dimension is eventually reduced while the number of channels is increased. For each layer, the kernel sizes and channels are presented in Fig. 2. After the 3rd block, the features (128 elements) are unrolled and a fully connected layer is used with a SoftMax classifier to perform ROI-level binary quality classification (*i.e.,* class 0: poor quality vs class 1: good quality). A dropout unit is placed after the fully connected layer to minimize overfitting.

B. Whole DTI volume label prediction using two tailored voting systems

After predicting the ROI-level labels, we utilize these labels to predict the quality label of the whole 3D DTI volume via the voting system block shown in Fig. 1. Specifically, we propose two separate tailored score functions for the voting block, each can be used individually to perform the final classification. The score function one, defined by equation (1), gives a total score S_{tot} to each whole DTI volume using a linear combination of the predicted label of the single center-cropped large ROI $S_{center\text{-}cropped\text{-}ROI}$, and the summation of the predicted

$$S_{tot} = \alpha\, S_{center\_cropped\_ROI} + \beta \sum_{i=1}^{N} S_i \qquad (1)$$

If:

$$S_{tot} \geq threshold \;\; \longrightarrow Label = 1\ (Good\ volume)$$

$else$:

$$\longrightarrow Label = 0\ (Poor\ volume)$$

labels S_i's of N overlapping small ROIs sampled randomly from the whole DTI volume. If the total score S_{tot} is above some threshold then the whole DTI volume is considered of good quality. Otherwise, we classify the whole DTI volume as a poor-quality volume. There are 4 hyperparameters in this voting system, namely constant numbers α, β, N, and the choice of threshold value. The only conditions on α and β are α and β ≥ 0 and α + β = 1. On the other hand, voting system 2 uses a two-layer voting process as in equations (2) and (3). In the first layer, an aggregate label is assigned to the average score of N overlapping small ROIs (2) using a thresholding condition. Equation (3) then assigns a total score S_{tot} to the whole DTI volume by adding the value of the predicted label of the single center-cropped large ROI $S_{center\text{-}cropped\text{-}ROI}$ and the single aggregate label $S_{N\text{-}ROIs}$ of the small ROIs. In the 2nd layer, a decision is made on the quality label of the whole DTI

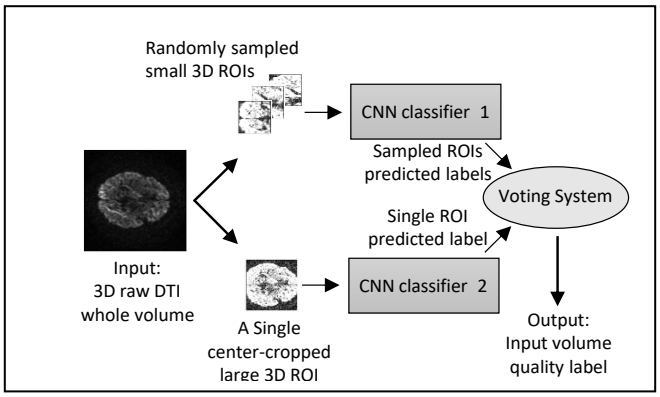

Figure 1. An overall view of the whole DTI volume automatic quality assessment framework

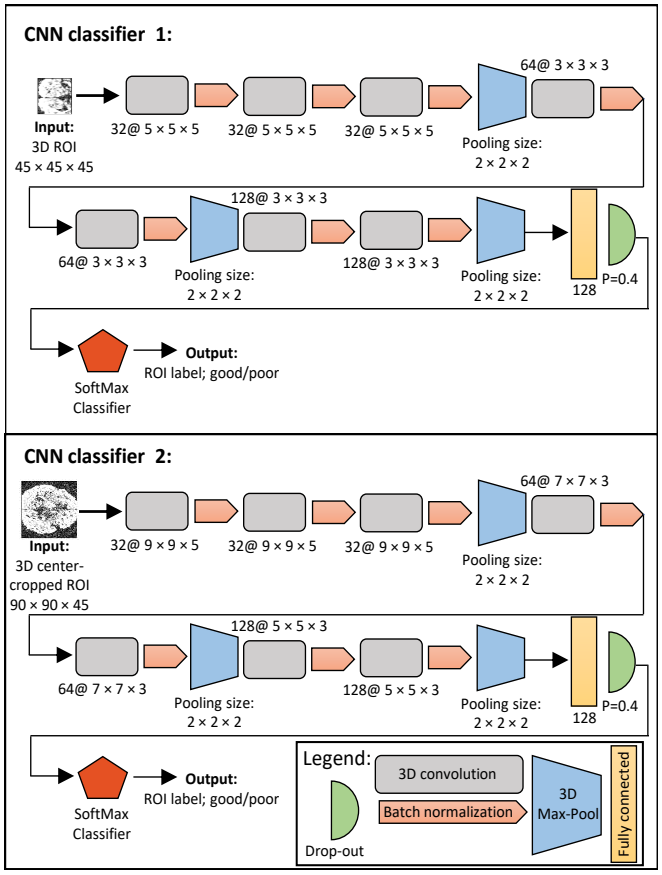

Figure 2. The architectures of CNN 1 (top) and CNN 2 (bottom). After the 3rd block, the input is reduced to a 128-element feature vector. The SoftMax unit performs the classification task using this feature vector.

volume using $S_{N\text{-}ROIs}$ and $S_{center\text{-}cropped\text{-}ROI}$. Particularly, if both labels $S_{N\text{-}ROIs}$ and $S_{center\text{-}cropped\text{-}ROI}$ are good (*i.e.,* equal to 1) then the whole DTI volume is considered a good volume otherwise it is assigned a poor-quality label.

$$S_{AVG\_ROIs} = \frac{1}{N}(\sum_{i=1}^{N} S_i) \quad (2)$$

Layer 1:

If:
$S_{AVG\_ROIs} \geq threshold \quad \longrightarrow S_{N\_ROIs} = 1\ (Good\ aggregate\ ROI)$
$else$:
$\longrightarrow S_{N\_ROIs} = 0\ (Poor\ aggregate\ ROI)$

$$S_{tot} = S_{center\_cropped\_ROI} + S_{N\_ROIs} \quad (3)$$

Layer 2:

$If\ S_{tot} \geq 2$: $\longrightarrow Label = 1\ (Good\ volume)$
$else$: $\longrightarrow Label = 0\ (Poor\ volume)$

Voting system 1 has 4 hyperparameters introducing more freedom to the overall framework and allowing it to perform more challenging final classifications at the cost of being a more complex model. On the other hand, voting system 2 has only two hyperparameters (*i.e.,* the threshold and N) which reduces the fine-tuning process load but gives less freedom to the overall framework to perform the final classification.

III. Data and Pre-processing

For training and testing of both CNN models, we utilized a small portion of the ABCD study, which recruited over 10,000 9-10 years old participants to track human brain development from childhood through adolescence [12]. Institutional review boards at each site approved the study procedures. Written consent was obtained from all parents, and children gave verbal assent. Multi-shell (b=0, 500, 1000, 2000, 3000 s/mm^2) DTI scans from eighty-five subjects in the ABCD study were used in the following experiments. All the scans have isotropic resolution (1.7×1.7×1.7 mm^3), same matrix size (140×140×81), and identical diffusion directions (96). Because the scans were acquired across twenty-one sites, the acquisition parameters vary slightly and are reported in Table 1.

To implement volume-wise classification and reduce labor intensity, a total of 8,530 DTI magnitude volumes from 85 subjects were labeled by an expert with 12 years of experience in MRI and DTI analysis. Among this cohort, around 5,060 (59.32%) volumes were classified as good quality. For training, we further balanced the two classes by randomly excluding 1,590 good quality volumes. Compared with slice-wise labeling, volume-wise QA is more practical and efficient since most volumes containing bad quality slices will often be excluded from any analysis. Volume-wise labeling makes manual annotation much easier to achieve, with only ~15 minutes to label one subject (~100 volumes).

The labeled volumes are then randomly shuffled and split into training, validation, and testing datasets, with a ratio of 6:2:2. The data intensity of each volume is normalized to the range of [0, 1]. A brain mask is then extracted using the function **dwi2mask** provided in the toolbox MRtrix3 [25]. The mask is used in the following ROI sampling procedure to ensure that most of the ROI contains brain regions, instead of background.

TABLE 1. The detailed DTI scans acquisition parameters across our datasets

Number of subjects	Number of b=0 volumes	Number of acquisitions	TE (ms)	TR (ms)
3	8	1	81.9	4100
4	7	1	81.9	4100
7	6	2	89	5300
1	3	1	89	5300
2	8	2	96	6050
1	3	1	96	6050
1	10	2	88	4100
66	7	1	88	4100

The QA performance of our tool was validated using both accuracy and confusion matrices of the predicted labels compared to the ground truth.

IV. Experiments and Results

In this section, we detail the choices of hyperparameters for the two CNNs' and their training history. Moreover, we present the results of individual ROI-level quality assessments. Next, we detail the choice of hyperparameters for both voting systems 1 and 2. Finally, we demonstrate the quality assessment accuracy of our tool on the whole DTI volumes in the test set using accuracy measures and confusion matrices.

A. ROI sampling per each DTI volume

As mentioned earlier, for each DTI whole volume we have two sets of ROIs: a single center-cropped ROI and N overlapping small ROIs. In particular, for each DTI whole volume, we randomly sampled 10 overlapping small ROIs of size [45, 45, 45] voxels and a single-center-cropped large ROI of size [90, 90, 45]. The maximum percent overlap allowed between any two small ROIs was set to 52%. Moreover, in order to take advantage of patterns of artifacts that manifest themselves better on the border between the background and the brain (such as motion), the small ROIs were allowed to include a maximum of 16% background voxels.

B. CNN 1 and 2 parameters and training history

The architectures of CNN 1 and 2 are presented in Fig. 2. For all convolutional layers, RELU activation function was used and the kernels' parameters were initialized using the random Glorot initialization technique [26]. The stride size for all convolutional layers was set to [1, 1, 1] and no paddings were used. For the Max-pooling layers the stride size was set to [2, 2, 2]. We used the RELU activation function for the fully connected layer with 128 elements. The dropout probability was set to 0.4. Finally, we used a SoftMax activation function for the SoftMax layer with two quality classes: 0 (poor), and 1 (good).

Both CNN architectures were implemented in Python, using Keras with Tensorflow as backend. The two models

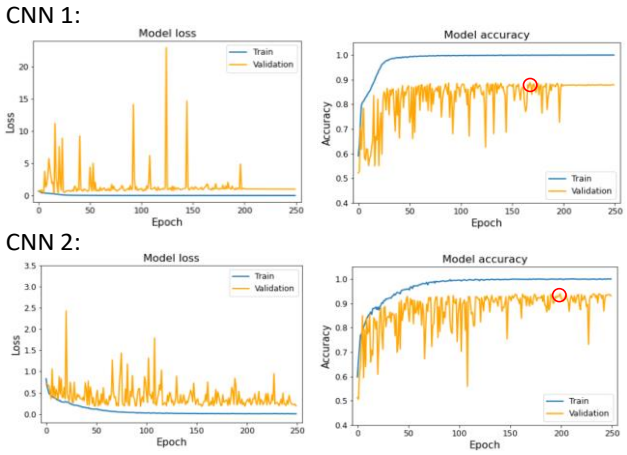

Figure 3. Training loss (left) and classification performance (right) of CNN 1 (top) and CNN 2 (bottom). The best performing models are shown using red circles.

were trained separately to minimize the Cross-entropy loss function using the manually labeled training dataset. Due to the differences in the kernel sizes of the two CNNs, we trained them with different hyperparameters. In particular, CNN 1 was trained for 250 epochs using a batch size of 128. For the first 200 epochs, Adam optimizer with default values was used, and then for the last 50 epochs the model was stabilized using the SGD with Nesterov optimizer [27] (initial learning rate = 0.0001, Momentum = 0.5, and decay rate = 10^{-6}). On the other hand, CNN 2 was trained for 250 epochs using a batch size of 32 and the SGD with Nesterov optimizer (with the same parameter values as in CNN 1). The training history is depicted in Fig. 3. The best model for each CNN was chosen as the best performing model (*i.e.,* highest classification accuracy) on the validation set. The test set's confusion matrices for the best performing models of CNN1 and 2 are presented in Fig. 4. The classification accuracies for CNN 1 on the training, validation, and test sets were 99.85%, 88.72%, and 89.15%, respectively. For CNN 2 the accuracies were 99.99%, 93.98%, and 94.49% on the training, validation, and test sets, respectively. Note that there is no need for achieving higher accuracies on the individual ROI-level classification since in the next step, the framework uses the labels of multiple ROIs per DTI volume to predict the quality label of the whole DTI volume.

C. Voting systems for Whole DTI volume label prediction

The equations for the two proposed voting systems are in (1-3). For voting system 1, we have 4 hyperparameters. In our implementation we fixed the value of N = 10 for both voting

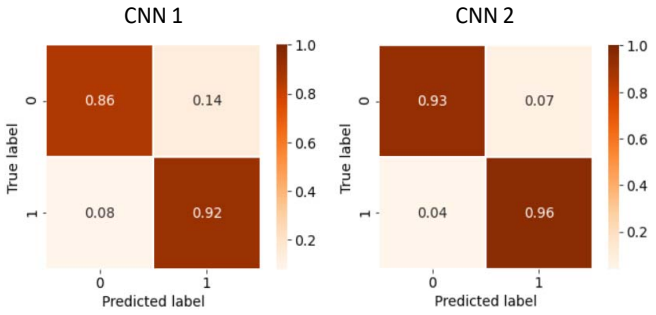

Figure 4. Confusion matrices for CNN 1 (right) and CNN 2 (left) on the test set. CNN 1 achieves an accuracy of 89.15% and CNN 2 achieves an accuracy of 94.49% on the test set.

systems. Since the performance of CNN 2 is significantly better than CNN 1, we put more weight on the scores associated with CNN 2 (*i.e.,* $S_{center\text{-}cropped\text{-}ROI}$). Hence, we chose $\alpha = 0.6$ and $\beta = 0.4$. Since (1) takes finite discrete values as inputs, S_{tot} will also have finite discrete values. Hence, in order to find the best value for a selected threshold, we simply vary the threshold value over all possible values of S_{tot} and choose the value leading to the best performance on the validation set. Fig. 5a depicts this process by demonstrating the model's total accuracy as well as true positive and true negative rates. As shown in Fig. 5a. the framework achieves the best performance on the validation set over a range of threshold values depicted by the black arrow. We chose a threshold value of 2.5. Similarly, for voting system 2, we varied the threshold value over all the possible values of $S_{AVG\text{-}ROIs}$ and chose the one leading to the highest validation accuracy. Fig. 5b shows the results of this fine-tuning process. The ideal model is the one performing similarly on both 'good' and 'poor/bad' classes. As depicted in Fig. 5b, this voting system achieved high accuracies (slightly less than voting system 1) over a wider range of thresholding values compared to voting system 1. We chose the value of 0.1. The selected threshold values were used to predict the quality labels on the test set. The classification accuracies were 100% and 98% for voting frameworks 1 and 2, respectively (99.85±0.15%, and 97.95±0.76% Wilson adjusted accuracy and confidence interval [28]). Precision, recall, and F1-score were 1, 1, 1, and, 1, 0.962, 0.981, for voting frameworks 1 and 2, respectively. One possible reason for the small drop in the performance of voting system 2 lies in situations where there is a high false-positive rate in individual ROI scores predicted by the two CNNs. In such situations voting system

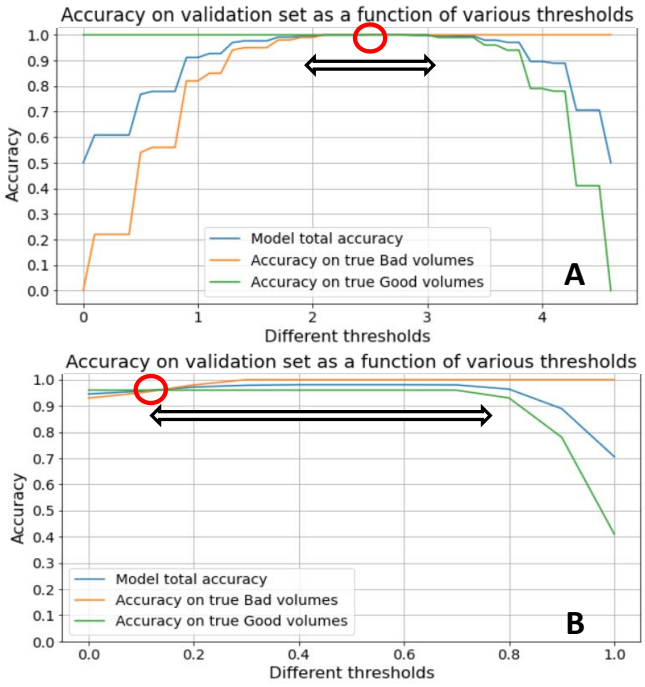

Figure 5. Hyperparameter fine-tuning for a) voting system 1, and b) voting system 2. Voting system 1 results in higher accuracy over a narrower range of thresholds while voting system 2 achieves a lower accuracy over a wider range of threshold values.

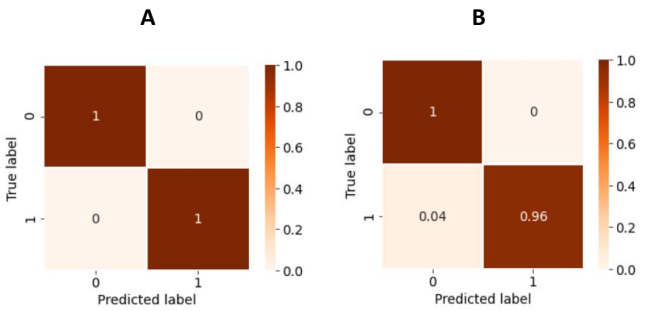

Figure 6. Confusion matrices for voting system 1 (a), and 2 (b) on the test set. Voting system 1 achieves an accuracy of 100% and voting system 2 achieves an accuracy of 98% on the test set.

1 performs better since, unlike voting system 2, it does not necessarily require $S_{\text{center-cropped-ROI}}$ to be equal to 1 (*i.e.,* good ROI) to declare the whole DTI volume a good one. Figure 6 shows the confusion matrices of the whole DTI volume quality assessments using the two voting systems. Regardless of the choice for the voting system, high accuracies were achieved on quality assessment of DTI whole volumes demonstrating the practicality of our developed automatic quality control tool.

V. Conclusion and Future Works

In this work, an artificial intelligence-based framework for automatic fast quality assessment of raw 3D DTI volumes was presented. The proposed framework utilized a voting approach to assess the quality of 3D DTI whole volumes using CNNs in conjunction with two distinct tailored score functions. The performance of the framework was tested using a subset of the ABCD dataset resulting in 100% and 98% accuracies for voting systems with score function 1 and 2, respectively. In future work, we plan to further enhance our framework to predict the types of the existing artifacts if the image is classified as 'poor' quality.

References

[1] Basser, Peter J., and Derek K. Jones. "Diffusion-tensor MRI: theory, experimental design and data analysis–a technical review." *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 15.7-8 (2002): 456-467.

[2] Alexander, Andrew L., et al. "Diffusion tensor imaging of the brain." *Neurotherapeutics* 4.3 (2007): 316-329.

[3] Hüppi, Petra S., and Jessica Dubois. "Diffusion tensor imaging of brain development." *Seminars in Fetal and Neonatal Medicine*. Vol. 11. No. 6. WB Saunders, 2006.

[4] Ladouceur, Cecile D., et al. "White matter development in adolescence: the influence of puberty and implications for affective disorders." *Developmental cognitive neuroscience* 2.1 (2012): 36-54.

[5] Simmonds, Daniel J., et al. "Developmental stages and sex differences of white matter and behavioral development through adolescence: a longitudinal diffusion tensor imaging (DTI) study." *Neuroimage* 92 (2014): 356-368.

[6] Krogsrud, Stine K., et al. "Changes in white matter microstructure in the developing brain—A longitudinal diffusion tensor imaging study of children from 4 to 11 years of age." *Neuroimage* 124 (2016): 473-486.

[7] Hoy, Andrew R., et al. "Microstructural white matter alterations in preclinical Alzheimer's disease detected using free water elimination diffusion tensor imaging." *PloS one* 12.3 (2017): e0173982.

[8] Lo Buono, Viviana, et al. "Diffusion tensor imaging of white matter degeneration in early stage of Alzheimer's disease: a review." *International Journal of Neuroscience* 130.3 (2020): 243-250.

[9] Roalf, David R et al. "White matter microstructure in schizophrenia: associations to neurocognition and clinical symptomatology." *Schizophrenia research* vol. 161,1 (2015): 42-9. doi:10.1016/j.schres.2014.09.026

[10] Tønnesen, Siren, et al. "White matter aberrations and age-related trajectories in patients with schizophrenia and bipolar disorder revealed by diffusion tensor imaging." *Scientific reports* 8.1 (2018): 1-14.

[11] Van Essen, D C et al. "The Human Connectome Project: a data acquisition perspective." *NeuroImage* vol. 62,4 (2012): 2222-31. doi:10.1016/j.neuroimage.2012.02.018

[12] Casey, B. J., et al. "The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites." *Developmental cognitive neuroscience* 32 (2018): 43-54.

[13] M. Jenkinson, C.F. Beckmann, T.E. Behrens, M.W. Woolrich, S.M. Smith. FSL. NeuroImage, 62:782-90, 2012

[14] Andersson, Jesper L R et al. "Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images." *NeuroImage* vol. 141 (2016): 556-572. doi:10.1016/j.neuroimage.2016.06.058

[15] Bastiani, Matteo, et al. "Automated quality control for within and between studies diffusion MRI data using a non-parametric framework for movement and distortion correction." *Neuroimage* 184 (2019): 801-812.

[16] Oguz, Ipek, et al. "DTIPrep: quality control of diffusion-weighted images." *Frontiers in neuroinformatics* 8 (2014): 4.

[17] Jiang, Hangyi, et al. "DtiStudio: resource program for diffusion tensor computation and fiber bundle tracking." *Computer methods and programs in biomedicine* 81.2 (2006): 106-116.

[18] Roalf, David R., et al. "The impact of quality assurance assessment on diffusion tensor imaging outcomes in a large-scale population-based cohort." *Neuroimage* 125 (2016): 903-919.

[19] Iglesias, Juan Eugenio, et al. "Retrospective head motion estimation in structural brain MRI with 3D CNNs." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2017.

[20] Alfaro-Almagro, Fidel, et al. "Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank." *Neuroimage* 166 (2018): 400-424.

[21] Graham, Mark S., Ivana Drobnjak, and Hui Zhang. "A supervised learning approach for diffusion MRI quality control with minimal training data." *NeuroImage* 178 (2018): 668-676.

[22] Liu, Bilan, Tong Zhu, and Jianhui Zhong. "Comparison of quality control software tools for diffusion tensor imaging." *Magnetic resonance imaging* 33.3 (2015): 276-285.

[23] Haddad, Seyyed MH, et al. "Comparison of quality control methods for automated diffusion tensor imaging analysis pipelines." *PloS one* 14.12 (2019): e0226715.

[24] Samani, Zahra Riahi, et al. "QC-Automator: Deep learning-based automated quality control for diffusion mr images." *Frontiers in neuroscience* 13 (2020): 1456.

[25] Tournier, J-Donald, et al. "MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation." *NeuroImage* 202 (2019): 116137.

[26] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010. W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[27] Sutskever, Ilya, et al. "On the importance of initialization and momentum in deep learning." *International conference on machine learning*. PMLR, 2013. B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.

[28] Wilson, Edwin B. "Probable inference, the law of succession, and statistical inference." *Journal of the American Statistical Association* 22.158 (1927): 209-212.