Journal Pre-proofs

Validation of Deep Learning techniques for quality augmentation in diffusion MRI for clinical studies

Santiago Aja-Fernández, Carmen Martín-Martín, Álvaro Planchuelo-Gómez, Abrar Faiyaz, Md Nasir Uddin, Giovanni Schifitto, Abhishek Tiwari, Saurabh J. Shigwan, Rajeev Kumar Singh, Tianshu Zheng, Zuozhen Cao, Dan Wu, Stefano B. Blumberg, Snigdha Sen, Tobias Goodwin-Allcock, Paddy J. Slator, Mehmet Yigit Avci, Zihan Li, Berkin Bilgic, Qiyuan Tian, Xinyi Wang, Zihao Tang, Mariano Cabezas, Amelie Rauland, Dorit Merhof, Renata Manzano Maria, Vinícius Paraníba Campos, Tales Santini, Marcelo Andrade da Costa Vieira, SeyyedKazem HashemizadehKolowri, Edward DiBella, Chenxu Peng, Zhimin Shen, Zan Chen, Irfan Ullah, Merry Mani, Hesam Abdolmotalleby, Samuel Eckstrom, Steven H. Baete, Patryk Filipiak, Tanxin Dong, Qiuyun Fan, Rodrigo de Luis-García, Antonio Tristán-Vega, Tomasz Pieciak

PII: S2213-1582(23)00174-2
DOI: https://doi.org/10.1016/j.nicl.2023.103483
Reference: YNICL 103483

To appear in: *NeuroImage: Clinical*

Received Date: 2 March 2023
Revised Date: 24 July 2023
Accepted Date: 25 July 2023

Please cite this article as: S. Aja-Fernández, C. Martín-Martín, A. Planchuelo-Gómez, A. Faiyaz, M.N. Uddin, G. Schifitto, A. Tiwari, S.J. Shigwan, R. Kumar Singh, T. Zheng, Z. Cao, D. Wu, S.B. Blumberg, S. Sen, T. Goodwin-Allcock, P.J. Slator, M. Yigit Avci, Z. Li, B. Bilgic, Q. Tian, X. Wang, Z. Tang, M. Cabezas, A. Rauland, D. Merhof, R. Manzano Maria, V.P. Campos, T. Santini, M.A. da Costa Vieira, S. HashemizadehKolowri, E. DiBella, C. Peng, Z. Shen, Z. Chen, I. Ullah, M. Mani, H. Abdolmotalleby, S. Eckstrom, S.H. Baete, P. Filipiak, T. Dong, Q. Fan, R. de Luis-García, A. Tristán-Vega, T. Pieciak, Validation of Deep Learning techniques for quality augmentation in diffusion MRI for clinical studies, *NeuroImage: Clinical* (2023), doi: https://doi.org/10.1016/j.nicl.2023.103483

Validation of Deep Learning techniques for quality augmentation in diffusion MRI for clinical studies

Santiago Aja-Fernández [a,*], Carmen Martín-Martín[a], Álvaro Planchuelo-Gómez[a,b], Abrar Faiyaz[c], Md Nasir Uddin[c], Giovanni Schifitto[c], Abhishek Tiwari[d], Saurabh J. Shigwan[d], Rajeev Kumar Singh[d], Tianshu Zheng[e], Zuozhen Cao[e], Dan Wu[e], Stefano B. Blumberg[f], Snigdha Sen[f], Tobias Goodwin-Allcock[f], Paddy J. Slator[f], Mehmet Yigit Avci[g], Zihan Li[g], Berkin Bilgic[g], Qiyuan Tian[g], Xinyi Wang[h], Zihao Tang[h], Mariano Cabezas[h], Amelie Rauland[i], Dorit Merhof[j], Renata Manzano Maria[k], Vinícius Paraníba Campos[k], Tales Santini[l], Marcelo Andrade da Costa Vieira[k], SeyyedKazem HashemizadehKolowri[m], Edward DiBella[m], Chenxu Peng[n], Zhimin Shen[n], Zan Chen[n], Irfan Ullah[o], Merry Mani[o], Hesam Abdolmotalleby[o], Samuel Eckstrom[q], Steven H. Baete[q], Patryk Filipiak[q], Tanxin Dong[p], Qiuyun Fan[p], Rodrigo de Luis-García[a], Antonio Tristán-Vega[a], Tomasz Pieciak[a]

[a]*Laboratorio de Procesado de Imagen (LPI), ETSI Telecomunicación, Universidad de Valladolid, Spain*
[b]*Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, Cardiff, UK*
[c]*University of Rochester, USA*
[d]*Shiv Nadar Institution of Eminence, India*
[e]*Zhejiang University, China*
[f]*University College London, UK*
[g]*Athinoula A. Martinos Center for Biomedical Imaging, USA*
[h]*University of Sydney, Australia*
[i]*RWTH Aachen University, Germany* [j]*University of Regensburg, Germany* [k]*University of Sao Paulo, Brazil*
[l]*Western University, Canada*
[m]*University of Utah, USA*
[n]*Zhejiang University of Technology, China*
[o]*University of Iowa, USA*
[p]*Tianjin University, China*
[q]*New York University, USA*

Abstract

The objective of this study is to evaluate the efficacy of deep learning (DL) techniques in improving the quality of diffusion MRI (dMRI) data in clinical applications. The study aims to determine whether the use of artificial intelligence (AI) methods in medical images may result in the loss of critical clinical information and/or the appearance of false information. To assess this, the focus was on the angular resolution of dMRI and a clinical trial was conducted on migraine, specifically between episodic and chronic migraine patients. The number of gradient directions had an impact on white matter analysis results, with statistically significant differences between groups being drastically reduced when using 21 gradient directions instead of the original 61. Fourteen teams from different institutions were tasked to use DL to enhance three diffusion metrics (FA, AD and MD) calculated from data acquired with 21 gradient directions and a b-value of 1000 s/mm^2. The goal was to produce results that were comparable to those calculated from 61 gradient directions. The results were evaluated using both standard image quality metrics and Tract-Based Spatial Statistics (TBSS) to compare episodic and chronic migraine patients. The study results suggest that while most DL techniques improved the ability to detect statistical differences between groups, they also led to an increase in false positive. The results showed that there was a constant growth rate of false positives linearly proportional to the new true positives, which highlights the risk of generalization of AI-based tasks when assessing diverse clinical cohorts and training using data from a single group. The methods also showed divergent performance when replicating the original distribution of the data and some exhibited significant bias. In conclusion, extreme caution should be exercised when using AI methods for harmonization or synthesis in clinical studies when processing heterogeneous data in clinical studies, as important information may be altered, even when global metrics such as structural similarity or peak signal-to-noise ratio appear to suggest otherwise.

Keywords: Deep learning, machine learning, artificial intelligence, diffusion MRI, angular resolution, diffusion tensor.

* *Address correspondence to: Santiago Aja-Fernández, ETSI Telecomunicación, Universidad de Valladolid, Spain.
Email address: sanaja@tel.uva.es (Santiago Aja-Fernández)

1. Introduction

In the field of medical imaging, the application of Artificial Intelligence (AI) in general, and Deep Learning techniques in particular, has brought about a significant revolution, with an increasing number of new applications emerging every year. These techniques have shown to effectively improve image quality and generate new images from limited medical imaging data. However, it is important to note that the majority of validations for these DL approaches in medical images have been performed visually and/or qualitatively, rather than being rigorously assessed in clinical studies.

A crucial question arises regarding the impact of these techniques on the preservation of relevant quantitative clinical information in medical images: *Are we sacrificing critical data for the sake of high-quality images generated by artificial intelligence techniques?* The validity of traditional quality measures such as the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Root Mean Squared Error (RMSE) must be taken into consideration for this task. It is not sufficient for the AI-generated images to simply resemble an original one; they must also preserve all relevant clinical information while not generating false information.

In this work, we worked with a specific modality of medical imaging: diffusion Magnetic Resonance Imaging (dMRI). dMRI is a non-invasive imaging technique that provides information about the diffusion of water molecules in biological tissues, allowing for the assessment of microstructural properties of the brain and other organs (Basser, 2002; Westin et al., 2002). The use of DL techniques in this modality is a rapidly growing field that has shown great potential in various stages of the dMRI data pipeline. These techniques include the correction of phase errors in multishot dMRI acquisitions (Aggarwal et al., 2019), the automatic identification and removal of artifacts (Qiao and Shi, 2021; Ahmad et al., 2023), and noise filtering (Tian et al., 2020; Fadnavis et al., 2020).

In addition, DL has also been used in tissue segmentation (Zhang et al., 2021a), registration between datasets (Zhang et al., 2021b), simultaneous segmentation and registration of white matter tracts in longitudinal measurements (Li et al., 2021a), super-resolution of fiber orientation distributions (Zeng et al., 2022), and the parcellation of superficial white matter (Xue et al., 2023).

Artificial intelligence techniques have also been extensively used for the harmonization of data from different sources. This is particularly important in the case of MRI, where data acquired with different parameters or from different vendor scanners can vary significantly. To enable multisite studies, it is crucial to harmonize such databases. DL techniques have shown potential to integrate data from multiple sources, increasing the statistical power of the analysis and generalizing findings across different sites (Zhu et al., 2019; Tax et al., 2019; Moyer et al., 2020; Blumberg et al., 2019). The harmonization process using DL can normalize data acquired under varying conditions, and this is expected to have a significant impact.

DL has also established its effectiveness in quantifying scalar metrics for a range of dMRI representations and models, including diffusion tensor imaging (DTI) (Li et al., 2021b; Sabidussi et al., 2023), diffusion kurtosis imaging (DKI) (Golkov et al., 2016), ensemble average propagator measures (Ye et al., 2019), neurite orientation dispersion and density imaging (NODDI) (Golkov et al., 2016; Ye, 2017; Gibbons et al., 2019; Qin et al., 2021; Sedlar et al., 2021; Faiyaz et al., 2022a), and white matter tract integrity (WMTI) models (Diao and Jelescu, 2023). While these frameworks can be optimized through regularized solutions (Daducci et al., 2015) or simplified with model symmetries (Oeschger et al., 2023), they still require a densely sampled **q**-space to produce quantitatively reliable biomarkers. DL techniques have enabled the modeling of a non-linear relationship between the **q**-space and the parameter space (Golkov et al., 2016; Gibbons et al., 2019; HashemizadehKolowri et al., 2022). Other applications include managing incomplete or reduced data, such as sparse sampling and time-efficient acquisition protocols (Gibbons et al., 2019; Li et al., 2021b; Sabidussi et al., 2023; Mani et al., 2021, 2022) or to increase angular sampling of the **q**-space (Lyon et al., 2022; Zeng et al., 2022).

Given the potential benefits of using DL on incomplete data in dMRI, our study aims to determine whether DTI parameters generated from volumes acquired with a reduced number of data and processed by DL techniques can match the statistical results obtained from standard quality acquisitions. We intend to validate the usefulness of DL-based reconstruction techniques in real clinical studies.

For clarity, the study has focused on a single critical aspect of dMRI, specifically the angular resolution. This parameter, which is proportional to the inverse of diffusion sensitizing gradient directions, is a crucial design element in dMRI (de Figueiredo et al., 2011; Basser and Pierpaoli, 1996). The number of gradient directions required to fit the basis can vary depending on the method used to represent the dMRI signal, ranging from a minimum of six gradients in DTI to several dozens or hundreds in High Angular Resolution Diffusion Imaging (HARDI) techniques (Tuch et al., 2003; Tristán-Vega et al., 2009; Özarslan et al., 2006). In clinical settings, it is necessary to optimize the number of gradient directions to minimize the examination duration and ensure patient comfort. However, reducing the number of gradient directions may result in a loss of subtle changes in the angular characteristics of dMRI data (Jones, 2004), leading to inaccuracies in the quantitative measures derived from a fitted model. Consequently, clinical studies could provide different results with different number of gradient directions.

In order to increase the angular sampling, i.e., to augment the number of gradient directions, one can average the local angular neighborhood using the spherical radial basis functions representation (Tuch, 2004) or decompose the

dMRI signal into orthogonal basis and then reconstruct the angular information under a different gradient configuration (Descoteaux et al., 2007; Chen et al., 2018, 2019). While alternative techniques based on AI have also been proposed (Ren et al., 2021; Lyon et al., 2022; Zeng et al., 2022), their effectiveness was evaluated using numerical measures such as the aforementioned RMSE, PSNR or SSIM indexes. However, recent studies have shown that reducing the number of gradients can result in a loss of clinically relevant information and make it difficult to detect differences in various medical conditions (Landman et al., 2007; Barrio-Arranz et al., 2015). The number of diffusion gradient orientations has been recognized as a crucial factor that influences the values of diffusion and DTI descriptors and affects the results of their statistical comparisons between clinical groups.

In this study, the assessed clinical groups were based on migraine, specifically between episodic migraine (EM) and chronic migraine (CM) patients. This disorder was selected for the following reasons:

1. A comprehensive and unique database of migraine patients was available, with an appropriate acquisition scheme that allows for downsampling the number of gradient directions without losing the coverage of the **q**-space.
2. Previous studies using the same database have identified brain regions with statistically significant differences using fully-sampled data, making them the ideal reference for reduced acquisitions.
3. As migraine findings are subtle, reducing the sample size or the number of gradient directions could negatively impact the significance of the differences, making this database ideal for evaluating the relevance of DL methods.

It is important to note that the objective of the study was not the diagnosis of migraine using diffusion-based parameters or the potential optimization of the dMRI acquisition. Migraine was exclusively employed to assess the preservation of the statistical relationships between diverse clinical groups after the application of the DL methods to synthesize the volumes with unsampled diffusion gradient orientation. The objective was the analysis of the effect of increasing the diffusion angular resolution in the context of the clinical studies. Thus, the asked DL network architecture design was not focused on the best possible distinction between clinical groups from the dMRI data.

We used 160 dMRI volumes, all including a unique **q**-space coverage scheme that enables us to easily subsample the data by merely selecting appropriate 21 gradient directions out of 61 without the need of applying interpolation algorithms. It is worth noting that there was no aim related to the optimization of the dMRI acquisition. We wanted to recreate a realistic situation in a clinical context with an available reduced dMRI dataset. The whole proposal is surveyed in Fig. 1. Our analysis revealed that 60% of the statistically significant differences between EM and CM patients, identified in a white matter study utilizing 61 gradient directions, were no longer present when only 21 directions were used. Thus, we wanted to study if the downsampled data could be enhanced using DL techniques so that we achieve an outcome similar to the original data.

The study here presented was initiated as a challenge[2] hosted at the 2022 Computational Diffusion MRI Workshop of the MICCAI conference in Singapore. It is important to note that the evaluation of the methods in this study was conducted using a specific database focused on migraine. Therefore, it is crucial to acknowledge that the results presented here may not be directly generalizable to other databases or pathologies.

2. Materials

2.1. Datasets: subjects' selection

As previously stated, the purpose of this study was to evaluate the validity of DL-based reconstructed dMRI images in a real clinical setting for the pathology of migraine. Migraine is a prevalent primary disabling disorder that is characterized by recurrent episodes of headache and is more common among young and middle-aged women. Despite its high prevalence, the pathophysiological mechanisms of migraine are not well understood and there are no current biomarkers. There are two classifications of migraine, episodic migraine (EM) and chronic migraine (CM), which are differentiated based on the number of headache days per month (15 or more days with headache per month for chronic migraine patients) (Headache Classification Committee of the International Headache Society, 2018).

Migraine is advantageous for this kind of study since the findings related to dMRI are subtle compared to healthy controls, as noted in previous studies. This makes it challenging to appreciate techniques or parameters that can better define pathophysiological properties, as opposed to disorders such as Alzheimer's disease or schizophrenia, where it is relatively easy to find statistically significant results with classic methods based on DTI, T1-, and T2-weighted MR imaging.

There have been some dMRI studies assessing migraine, with DTI being the most used technique to evaluate

[2] QUaD22: '*Quality augmentation in diffusion MRI for clinical studies: Validation in migraine'*. CDMRI Workshop, MICCAI 2022, Singapore. https://www.lpi.tel.uva.es/quad22/

microstructural properties. These studies have found differences between healthy controls and migraine patients (Chong et al., 2019; Kattem Husøy et al., 2019; Planchuelo-Gómez et al., 2020b; Yu et al., 2013), as well as between EM and CM patients (Coppola et al., 2020; Planchuelo-Gómez et al., 2020b), for DTI-related scalars such as fractional anisotropy (FA), mean diffusion (MD) and Axial Diffusivity (AD). DTI has been the most employed technique to evaluate microstructural properties with differences found between controls and migraine patients (MP) and between EM and CM patients. The most reported result in these studies is lower FA in MP compared to controls. Regarding the AD, a similar number of studies have found both lower and higher values in the MP compared to controls. For the MD and also the radial diffusivity, as with AD, both trends with higher and lower values in MP compared to controls have been detected, but with a higher number of studies reporting the higher values in MP. A detailed description of these results and further comparisons with a higher number of references is available elsewhere (Rahimi et al., 2022).

The subjects used for this work were obtained from a previous migraine clinical study (Planchuelo-Gómez et al., 2020a). A dataset was built comprising healthy controls (HC) and patients with EM and CM.

HC were recruited by convenience sampling and snowball sampling. Controls with a history of migraine, other headache disorders different to infrequent tension-type-headache (less than one attack per month), or a history of other neurological or psychiatric disorders were excluded. Healthy controls were aged between 18 and 65 years. Additionally, a questionnaire was provided to the controls to assess whether they suffered from headaches with migraine features.

Migraine patients were recruited to a neurologist specialized in headache disorders at their first visit. Due to migraine, these patients had been referred to the Headache Unit at the Hospital Clínico Universitario de Valladolid (Valladolid, Spain). Patients were included after a definite diagnosis of episodic migraine or chronic migraine according to the third edition of the International Classification of Headache Disorders (ICHD-3).

All the participants were aged between 18 and 60 years and read and signed a written informed consent form before acquiring the MRI data. The local Ethics Committee of Hospital Clínico Universitario de Valladolid (Valladolid, Spain) approved the study regarding the MRI acquisitions (PI: 14-197). All participants read and signed a written consent form prior to their participation.

2.2. Datasets: Acquisition

For all the clinical groups included in the sample, the acquisition protocol was identical. All the subjects were scanned using a Philips Achieva 3T MRI unit (Philips Healthcare, Best, The Netherlands) equipped with a 32-channel head coil in the MRI facility at the Universidad de Valladolid (Valladolid, Spain).

The single-shell dMRI acquisition protocol was the following: repetition time TR = 9000 ms, echo time TE = 86 ms, flip angle = 90^o, 61 non-collinear diffusion-sensitizing gradient orientations, one baseline volume (b = 0), b-value = 1000 s/mm^2, volume size of 2×2×2 mm^3, 128×128 matrix size, and 66 axial slices covering the whole brain.

The acquisition time for this sequence was around 12 minutes. Regarding the diffusion gradient directions, they were acquired so they could be subsampled to an alternative scheme composed of 21 orientations, see Fig. 2. The design of the gradients is grounded in a regular icosahedron, which has 20 facets, 30 edges, and 12 vertices. It can be "refined" by inserting one new vertex at the mid-point of each of its edges, and projecting it onto the unit sphere. This splits each facet in 4 new triangles. The resulting convex polyhedron has 42 vertexes and 80 facets, coupled in 21 and 40 pairs of antipodes. The 21 unique vertexes and the 40 unique barycenters of the facets comprise the interleaved gradients scheme, each set uniformly covering the orientations space.

The training and test datasets were composed of different clinical groups with a specific number of diffusion gradient orientations.

Training: this dataset included 60 HC. All the diffusion-weighted volumes, i.e., 61 directions and one non-diffusion weighted volume (b = 0), were provided to the participants of the challenge. The sampling scheme allows the 61 gradient directions to be easily subsampled to 21 gradients.

Test: this dataset included 50 patients with EM and 50 patients with CM. The subsample composed of 21 diffusion gradient directions together with the baseline volume of each subject was provided to the participants of the challenge. Datasets were shuffled, so the participant could not know if a volume belonged to EM or CM. To validate the results, the organizers (but not the participants) also had the complete acquisition, with 61 gradient directions.

The training dataset included no patients with migraine because our objective was to assess the effects of the application of a general machine learning method for increasing the number of diffusion gradient orientations in the statistical relationships between clinical groups. The target was to avoid a development of a "migraine-specific" method or a migraine classifier.

2.3. Diffusion MRI preprocessing

The training and tests datasets provided in the challenge were preprocessed to avoid any bias caused by different preprocessing pipelines conducted by each participant group. The dMRI preprocessing was composed of the following

steps: denoising following the Marchenko-Pastur Principal Component Analysis method (Veraart et al., 2016), correction of eddy currents and motion (Andersson and Sotiropoulos, 2016), and B1 field inhomogeneity (Smith et al., 2004; Zhang et al., 2001). All these steps were applied with MRtrix3 software (Tournier et al., 2019).

A mask of the preprocessed dMRI volumes was extracted (Dhollander et al., 2016). The diffusion tensor (DT) was estimated at the voxels defined within the brain mask following the ordinary least squares method implemented in FSL software (Jenkinson et al., 2012). To avoid potential bias related to the extraction of the diffusion tensor or its descriptors compared to the original study comparing CM, EM and HC groups (Planchuelo-Gómez et al., 2020a), the FA, MD and AD maps of the training dataset were obtained with the FSL estimation. The maps of the test dataset were separately obtained for each method by each group and submitted to the organizers.

3. Methods

3.1 Task: quality enhancement

For the analysis carried out in this study, only three DTI-derived metrics are considered: FA, AD and MD. AD and MD were selected for being the ones detecting significant differences in preliminary clinical studies with migraine patients. FA was also calculated as a complementary metric. Participants were asked to estimate these three metrics from the migraine dataset acquired with 21 diffusion gradient directions at b=1000 s/mm^2, trying to achieve a quality similar to the parameters estimated from 61 gradient directions. To that end:

1. They used the training data set to train an AI-based system that could angularly augment the dMRI data from 21 to 61 gradient directions to provide the most faithful representation of the signal and consequently the quantitative parameters, including FA, MD and AD. DL methods were recommended here.
2. Then, the participants applied the enhancement method to the migraine dataset. Three volumes (FA, MD, AD) were submitted by each participant group for each of the 50 EM and 50 CM subjects.

3.2 Evaluation

The dataset for the statistical study consisted of maps from two migraine groups (CM and EM), which were compared using the tract-based spatial statistics (TBSS) pipeline (Smith et al., 2006) for three metrics (FA, MD, AD). To avoid bias, the participants were unaware of the distribution of patients between CM and EM.

For the TBSS analysis, the same steps were applied to the original maps and the maps provided by the participant groups. The extracted FA images were nonlinearly registered to the FMRIB-58 template in the Montreal Neurological Institute (MNI) space, composed of averaged FA maps, using the b-spline representation of the registration warp field with the FNIRT tool from FSL (Rueckert et al., 1999). After registration, the white matter skeleton was defined from the thinning of a generated mean FA image, using an FA threshold of 0.2 to distinguish white matter from gray matter. The aligned FA images from the subjects were projected onto the white matter skeleton. The MD and AD maps were registered to the MNI space using the same warp transformations employed to register the FA images and were projected onto the skeleton. To identify the regions with statistically significant differences, the Johns Hopkins University ICBM-DTI-81 White-Matter Labels Atlas was employed (Mori et al., 2005). The minimum volume per region to consider statistically significant results was 30 mm^3. To minimize the sources of variability, the registration was done using the fully sampled data (61 gradient directions) and then applied to the data provided by the groups. Likewise, the FA mask derived from the 61 gradients scheme was used for all teams.

The voxelwise differences in FA, MD and AD values of the CM and EM were assessed following the *randomise* permutation-based inference tool by non-parametric statistics implemented in FSL, considering the threshold-free cluster enhancement (TFCE) results (Nichols and Holmes, 2002; Smith and Nichols, 2009). This procedure was applied to the original maps and the maps submitted by the participant groups using 5000 permutations and considering a threshold for statistical significance of $p < 0.05$ after family wise error (FWE) correction for multiple comparisons.

For the sake of comparison, two measures are considered: True Positives (TP) and False Positives (FP). TP are those voxels found in the analysis of the 61-gradients reference, considered as the silver standard. Conversely, FP are those voxels found as having significant differences in the comparison carried out with the improved maps submitted by the participants but were not found to have significant differences in the comparison carried out with the 61-gradients reference. For a better comparison, we also used the following sensitivity and specificity parameters, based on the ratios of TPs and FPs:

- ***Sensitivity***, or ***true positive rate*** (TPR)

$$\mathrm{TPR}_i(\mathrm{Team}) = \frac{\mathrm{TP}_i(\mathrm{Team})}{\mathrm{TP}_i(\mathrm{Team})+\mathrm{FN}_i(\mathrm{Team})} = \frac{\mathrm{TP}_i(\mathrm{Team})}{\mathrm{TP}_i(\mathrm{T61g})} \times 100[\%], \qquad (1)$$

- ***Specificity***, or ***true negative rate*** (TNR)

$$\mathrm{TNR}_i(\mathrm{Team}) = \frac{\mathrm{TN}_i(\mathrm{Team})}{\mathrm{TN}_i(\mathrm{Team})+\mathrm{FP}_i(\mathrm{Team})} = \frac{\mathrm{TN}_i(\mathrm{Team})}{\mathrm{TN}_i(\mathrm{T61g})} \times 100[\%], \qquad (2)$$

- ***Precision***, or ***positive predicted*** value (PPV)

$$\mathrm{PPV}_i(\mathrm{Team}) = \frac{\mathrm{TP}_i(\mathrm{Team})}{\mathrm{TP}_i(\mathrm{Team})+\mathrm{FP}_i(\mathrm{Team})} \times 100[\%], \quad (3)$$

- ***False positive rate*** (FPR)

$$\mathrm{FPR}_i(\mathrm{Team}) = \frac{\mathrm{FP}_i(\mathrm{Team})}{\mathrm{FP}_i(\mathrm{Team})+\mathrm{TN}_i(\mathrm{Team})} = \frac{\mathrm{FP}_i(\mathrm{Team})}{\mathrm{TN}_i(\mathrm{T61g})} \times 100[\%], \quad (4)$$

- ***Accuracy*** (ACC)

$$\mathrm{ACC}_i(\mathrm{Team}) = \frac{\mathrm{TP}_i(\mathrm{Team})+\mathrm{TN}_i(\mathrm{Team})}{\mathrm{TP}_i(\mathrm{Team})+\mathrm{FP}_i(\mathrm{Team})+\mathrm{TN}_i(\mathrm{Team})+\mathrm{FN}_i(\mathrm{Team})} = \frac{\mathrm{TP}_i(\mathrm{Team})}{\mathrm{TP}_i(\mathrm{T61g})+\mathrm{TN}_i(\mathrm{T61g})} \times 100[\%], \quad (5)$$

where $\mathrm{TP}_i(\mathrm{T61g})$ denotes the number of TP found for metric *i* by the reference calculated by 61 gradient directions. TN stands for True Negative and FP for False Positive. In all cases, we consider $i = \{\mathrm{FA}, \mathrm{MD}, \mathrm{AD}\}$. For all the metrics, we also defined a global metric that merges all the individual values into a single one. For instance, for TPR is defined as:

$$\mathrm{TPR}_{\mathrm{Total}}(\mathrm{Team}) = \frac{\sum_i \mathrm{TP}_i(\mathrm{Team})}{\sum_i \mathrm{TP}_i(61\mathrm{g})} \times 100[\%] \quad (6)$$

Additionally, we have added a new metric that calculates the gain respect to the comparison carried out with the 21-gradients maps, without the use of any technique (AI or not) to enhance their quality:

$$\mathrm{Compar.\ 21\ (Team)} = \frac{\sum_i(\mathrm{TP}_i(\mathrm{Team})-\mathrm{FP}_i(\mathrm{Team}))-\sum_i(\mathrm{TP}_i(21\mathrm{g})-\mathrm{FP}_i(21\mathrm{g}))}{\sum_i \mathrm{TP}_i(61\mathrm{g})-[\sum_i(\mathrm{TP}_i(21\mathrm{g})-\mathrm{FP}_i(21\mathrm{g}))]} \times 100[\%] \quad (7)$$

where $\mathrm{TP}_i(\mathrm{T21g})$ and $\mathrm{FP}_i(\mathrm{T21g})$ denote respectively the number of TP and FP found for metric *i* by the reference calculated from 21 gradient directions.

3.3 Participants and methods

Thirteen different institutions participated in the study, providing results from 14 different methods (one institution provided two methods). The AI-schemes were different among them, with different designs, training, and validation procedures, see Tables 1 and 2. More detailed insight of each method can be found in the supplementary material where each team deeply describes their procedure. All teams provided enhanced FA/AD/MD volumes for all migraine patients calculated from 21 directions.

In general, we can divide the methods provided by the participants into three categories (see Table 1), i.e., techniques that operate in a) the magnitude domain of the data, b) spherical harmonics (SH) representation of the signal, or c) map the DTI measures between 21 and 61 gradients directly. Regarding the architecture, the most straightforward solutions handled the feed-forward network that translated the low-resolution diffusion weighted imaging (DWI) data into high-resolution DWIs, attaching local neighborhood information (Team 2) or naturally including the context of a voxel via the CNN-based architecture (Teams 1, 4, 10 and 11). Other approaches employed a U-Net-based architecture (Team 14) and its modification with extra dropout layers in the decoder part (Team 5) or a gated attention mechanism (Schlemper et al., 2019) (Team 9). Others used more advanced methods, such as denoising autoencoders (Teams 6 and 7) by adapting the qModeL (Mani et al., 2021) previously used to reconstruct the DWIs in parallel imaging from under-sampled raw data. The participants mostly handled the preprocessed data by the organizers, without additional handling. Nevertheless, two teams applied additional preprocessing to enhance the input data quality: Team 5 denoises again DWI data used for AD and MD generation with the DeepDTI technique (Tian et al., 2020) to achieve high fidelity image generation prior to DTI estimation; Team 8 repeats the whole preprocessing pipeline including MP-PCA denoising, Gibbs-ringing artifacts removal, motion, and distortion corrections and a nonuniform intensity normalization. Interestingly, one participant (Team 11) additionally handled tissue differences segmented into the WM, GM and CSF regions. It is worth noting that the participants were allowed to optimize the hyperparameters and training of the developed methods to find the optimal results and were not restricted to just apply a pre-trained network.

In Table 2, we show the training procedure carried out by the different teams. Note that most of them used error-related metrics for the loss function, such as MSE or mean absolute error (MAE). One group (Team 10) used the perceptual loss (Johnson et al., 2016) that takes into account the features from the pre-trained VGG16 network (Simonyan and Zisserman, 2015). Regarding the division of the training dataset (HC data), most groups incorporated the highest number of subjects in the training subset, including between 36 (Team 3) and 54 (Team 14) subjects out of 60 composing the whole dataset. Team 2 included only 13 out of 60 HCs to train their method, including three of these 13 subjects in the training subset. Team 13 trained the method with three different sets of 20 HCs (80% in the training subset). All the groups that used the *in vivo* data except Teams 9 and 11 used a validation subset that included between 5% (Team 14) and 20% (Team 1) of subjects. A similar proportion of subjects was included in the testing subset, being the number of subjects equal or higher than those included in the validation subset for most methods.

Two groups used two different divisions of the three subsets, using the first one to determine the best method, and the second one removing all subjects from the testing subset, increasing the size of the training and validation subsets, for cross-validation (Team 1) or improving the results of the chosen approach (Team 5). Two teams employed 5-fold cross-validation (Teams 1 and 9) and one group used 10-fold cross-validation (Team 12). Interestingly, Teams 6 and 7 did not make use of the *in vivo* data in the training and testing subsets. Instead, the samples were generated artificially using a three-compartment biophysical model (Behrens et al., 2003; Jelescu et al., 2016) with the physically-compatible parameters observed in the WM. Only one group (Team 11) used an augmentation technique (random flipping) to increase the training sample artificially.

4. Results

4.1 Quality metrics of image reconstruction

A preliminary visual assessment of the different methods was performed using one slice from a CM patient. The considered metrics (FA, AD and MD) were calculated from the original data and from the different AI-enhancement procedures and shown in Fig. 3. Upon visual inspection, the majority of the images appear similar in terms of their overall appearance. While there are slight variations in intensity levels, no discernible differences were observed with respect to structural features.

The quality assessment of reconstructed images in medical imaging commonly involves utilizing visual references and various error or noise metrics. This process enables us to determine the similarity between the reconstructed and original images. Our initial step is, therefore, to compute two image-based metrics, namely structural similarity index measure (Wang et al., 2004, SSIM) and peak signal to noise ratio (PSNR). In this regard, the 100 enhanced volumes for the three considered metrics (FA, MD, AD) have been considered for each team. The original data, reconstructed from 61 gradient directions, has been used as a reference for the calculation of the same metrics. To minimize the influence of the background on the SSIM calculation, the metric has been computed only on a white matter mask, determined for those points where FA $>$ 0.2 with the FA calculated with 61 gradient directions. The results of these metrics are presented in Table 3.

With respect to the SSIM metric, most of the methods exhibit improvement or are comparable to the reference. Only Team 9 shows results that indicate a significant discrepancy between the reconstructed signal and the reference. Additionally, Team 13 also performs slightly worse than the reference. Concerning the PSNR, Team 9 again exhibits significant deviation from the reference, while the other teams display results that are either slightly better or slightly worse than the reference. Overall, if we exclude Team 9, it can be concluded that the results of most teams demonstrate a high degree of similarity with the original data. In subsequent sections, we will examine whether these results align with the statistical results obtained from the clinical study.

4.2 TBSS results

Second, we show the results of the TBSS analysis of the data for each of the teams and every considered metric. Results are in Table 4. Table 4-(a) shows the raw numbers given by TBSS: the number of voxels with statistically significant differences in the skeleton of the FA, for FA, AD and MD, considering a total of 39,256 points over that skeleton. Two measures are considered, TPs and FPs. We also show the number of ROIs that presents statistically significant differences for each of the metrics. The first number represent the number of ROIS already present in the original study, i.e., TPs, and the second one, the number of ROIs that were not present in the original study, i.e., FPs. For a better understanding of the results, in Table 4-(b) we present an alternative version based on ratios, using Eqs. (1)-(7).

If we first focus on the TPR for 21 gradient direction scheme with no enhancement, we can see that only 41% of the differences are detected for AD and 40% for MD, which means that, with the reduction of the number of gradient directions, around 60% of the differences were lost. In addition, a small number of FPs also arises, see the FPR (3% for both metrics) This result was precisely the motivation behind the study. In addition, no differences were found for FA. Let us now check the performance of the enhanced sets.

According to the Accuracy (ACC) of the methods in Table 4-(b), three methods did not improve the 21 grads reference (Teams 4, 9, and 14) one method showed similar results (Team 13) and 7 of them show a performance over 80%. These results are similar to the comparison with 21-gradient directions see column 'Compar. 21 grads', whereonly one method improved by over 30% (Team 8). If we only consider the TPR, note that most methods performed better than the reference, and eight methods found over 60% of the original points with significant differences. However, there is a counterpart: the higher the number of TPs, the higher also the number of FPs. Note that one method shows a FPR of 77% but, at the same time, it shows a FPR of 15%. For all the methods with TPR over 65%, the global FRR is over 10%. This is an undesired effect that must be carefully analyzed in the next section.

4.3 Analysis of False Positives

False positives in clinical studies refer to those results that indicate the presence of statistically significant

differences when they do not actually exist. These errors can have serious implications and can lead to several negative consequences, such as misdiagnosis and reduced reliability of study results. Thus, it is of paramount importance to deeply analyze the presence of these results in the studies carried out with AI-enhanced volumes, since they can limit the applicability of these techniques.

We have decomposed the results in Table 4 in the bar diagrams in Fig. 4, where we show the number of voxels with significant differences detected by TBSS for the different methods. We have used a color coding to show: TPs detected by the 21-gradient reference (blue); the TPs detected by the 61-gradient reference but not by the 21 gradient reference (green) and the FPs (red). The values marked in green can be seen as the *new* TPs found by each method. For the sake of simplicity, we only show results for AD and MD, since FA detects no significant differences.

As illustrated in Fig. 4, it is evident that the cost of identifying new significant differences is the occurrence of additional FPs. Fig. 5 presents a histogram of these new TPs and FPs across the teams. For instance, the bar in position 4 indicates the number of voxels detected as statistically significant in four teams. We have calculated these histograms to check if the new TPs and the FPs occur in the same areas for all the teams or, on the contrary, the different groups found differences in distinct voxels. When examining the FP plots, it can be observed that the histograms are both skewed towards the left. This suggests that the FPs detection is not consistent across the groups, but differences are primarily identified by only one or two teams. This finding indicates that the presence of FPs is not a result of the data itself, but rather a consequence of the algorithms employed for reconstruction.

The question that arises is the existence of a correlation between the number of new TPs and the number of FPs produced by each team. To examine this relationship, we have fitted simple linear regression models for the AD and MD metrics using least squares, as shown in Fig. 6. Each marker represents the value of new TPs to FPs for each team. For the AD metric, the model accuracy is $R^2 = 0.907$ and the linear model has a slope of $\alpha = 2.11$ which means that, according to the model, for every two new TPs the methods produce one FP. The results are similar for the MD metric, with $R^2 = 0.749$ and a slope of $\alpha = 1.57$, where for every one and a half new TPs, the methods produce one FP.

It is important to note that the data used in this experiment was generated from various methods, in multiple laboratories, by distinct teams, and using diverse AI-based techniques. Despite the diversity in the data source, the results produced by each team align perfectly with the model.

4.4 Analysis of the variance

In this final section, we evaluate the results produced by different teams by analyzing the out values and their variability. We start by plotting histograms of the values of the three metrics (FA, AD, MD) over the region defined by the FA mask for each team. These results are shown in Fig. 7, with histograms for EM subjects only, for the sake of simplicity (similar results can be found for CM). Additionally, we calculate several statistics over the same FA mask, including mean, standard deviation (std), and coefficient of variation (CV), which are presented in Table 5.

Based on Figure 7, the histograms of the three metrics calculated with only 21 gradient directions (R21) are the closest to the reference calculated with 61 directions (R61). The std and CV of AD and MD are only slightly higher than the reference. On the other hand, Team 9 stands out as having the largest difference in its results. Its distribution has a clear bias towards smaller values compared to the reference, which suggests that its data has been scaled and does not have the same range as the original data. This effect can also be seen, albeit less pronounced, in Teams 10, 13, and 14. Meanwhile, Team 2, the only team that does not use DL, also shows a slight discrepancy in the three metrics, with reduced mean and standard deviation. In contrast, Teams 1, 3, 5, 8, and 12 display high similarity with the original data.

5. Discussion

In this study, we have focused on a critical aspect of dMRI: the sampling resolution of the **q**-space. It is well known that a reduction in the number of gradient directions may result in changes in the estimation of the DT and, subsequently, in the values of the scalar metrics derived. In the case selected for this study (EM vs. CM), when reducing the angular samples from 61 to 21, we detected a loss of around 60% of the voxels with statistically significant differences detected by TBSS. We proposed different research groups to *improve* the quality of the subsampled data using DL techniques, so that the loss of quality produced by the reduction in the number of available directions were compensated by their AI-based algorithms.

We have gathered the results and numerical data of the different methods in Table 6 to gain a deeper insight into the impact of each technique on performance.

One of the methods (Team 9) exhibits remarkably disparate results compared to the other methods in terms of SSIM, PSNR, and the histogram shape. It is evident that the data generated using this technique deviates significantly from the expected results. The method does not accurately reconstruct the original data but instead, it produces a scaled version, as demonstrated in Fig. 7. The results presented in Table 5 indicate that the mean and variance of the data have undergone substantial alteration. Consequently, the method fails to detect statistically significant differences. Interestingly, the selected AI method, inputs, and outputs are similar to those of other teams that produce much better

results. Therefore, Team 9 can be considered an outlier and excluded from further analysis.

The image-based quality metrics (SSIM and PSNR) show similar values for most of the methods but, however, they do not correlate with the results of the statistical test. Note, for instance, Team 5, which exhibits relatively low values in these metrics compared to the 21-gradients reference, but still demonstrates a good performance in the clinical study with a 28% improvement. Among the first four teams, three of them have lower SSIM and higher PSNR values compared to the reference but the best TBSS results. Conversely, Team 4 has improved quality metrics, but does not show improvement in TBSS. These results suggest that metrics based on visual features or global errors are not sufficient to evaluate the significant differences in clinical studies. The *global* improvement of visual quality in medical imaging data does not necessarily mean that the processed image is better for quantitative purposes. Therefore, alternative metrics should be used for those AI techniques that enhance medical imaging for numerical processing.

Upon analysis of the TBSS results, three methods did not outperform the reference, while two methods displayed marginal improvement and only one method demonstrated a significant improvement of over 30%. When considering the TPR, most methods performed better than the reference, with eight methods identifying over 60% of the original voxels with statistically significant differences. There was no method with a number of TPs close to the reference with 61 gradient directions, with at least 2000 unidentified voxels with TPs. However, this comes with a trade-off, as a higher number of TPs leads to a higher number of FPs. Note that one method achieved a TPR of 77%, but also generated an excessive FPR of 14.5%. Additionally, for all methods with TPR over 65%, the overall false positive rate exceeded 8%.

The generation of false positives is a critical issue in clinical studies. The errors produced by FPs can have far-reaching consequences for the validity and reliability of study findings, potentially undermining the usefulness of the study for informing clinical practice. Furthermore, if FPs are not adequately addressed, they can lead to misdiagnosis and patients receiving unnecessary treatment, resulting in a waste of resources and further testing. It is therefore crucial to strive for accurate results in clinical studies and to minimize the occurrence of FP through rigorous study design and analysis.

The occurrence of FPs in this study is driven by two factors: (1) increased variance in the data and (2) AI-based processing. The former is a result of the relationship between angular resolution and variance of the tensor metrics (Poonawalla and Zhou, 2004; Tristán-Vega et al., 2012): the variance of the estimation error in the DT increases when fewer gradients are considered. As a consequence, as shown in Table 5, the standard deviations for the AD and MD metrics are slightly larger for the 21 gradient case compared to the 61 gradient case. This small increase of the variance results in a significant decrease in findings in the TBSS analysis and the presence of FPs. However, some methods that produce more FPs do not increase the variance, such as Team 8, where the FPs are instead generated by the processing algorithm itself.

The results of the linear regression models in Fig 6, fitted for the AD and MD metrics (with accuracies of 0.907 and 0.749, respectively) are particularly illuminating in this regard. Despite the differences in performance, inputs, and outputs among the methods and teams, there appears to be a constant rate of growth in FPs. This raises questions about the generalization of the application of AI-based methods from a single group, as only healthy controls were included in the training dataset. The results of this study suggested that the more new TPs are found, the more FPs appear, potentially indicating that some methods are generating significant differences regardless of their existence. In this context, the lack of patients with migraine in the training dataset prevented the DL methods from learning the specific properties of these patients in relation to the differences based on the number of diffusion gradient orientations. Thus, our results suggest that DL-based techniques oriented for clinical studies must include a broad sample composed of the diverse groups of interest to learn their particular features. Generalization of machine learning methods trained only in one group or of pre-trained networks not including clinical cohorts of interest in the training process for a task should be applied with extreme caution, as the data from these cohorts may be altered generating spurious results. Some tasks included in this generalization include processes not directly oriented to classification such as increasing the angular resolution, as assessed in this study, or data harmonization. Previous studies reached similar DTI-derived metrics using reduced compared to large datasets (Tian et al., 2020) or were able to generalize the extraction of DTI maps from diverse datasets (Sabidussi et al., 2023). However, these studies included no clinical cohorts, except a person with white matter hyperintensities, but no specific clinical condition (Sabidussi et al., 2023). It is worth noting that the method by Tian et al., (2020) was used by Team 5 in the challenge, and showed similar volumes compared to the 61-gradient reference, and, following the general trend of our results, a high increase in the number of TP (28.1%) and FP (14.2%). Considering the use of clinical cohorts, a recent study has shown that the additional inclusion of DTI features to genetic data improved the prediction results related to the prognosis of patients with glioma using a DL network (Yan et al., 2021), raising the importance of including the clinical groups of interest to learn the features potentially present in the patients.

In this study, the type of AI method, inputs, and outputs used for augmenting the **q**-space sampling resolution showed no clear impact on the final results. Teams using different methods, such as the well-known U-Net, were found both in the top and bottom rankings. The use of a particular architecture does not ensure successful results. Similarly, the type of inputs and outputs used varied among teams with no clear correlation to performance. One team using scalar metrics as inputs (Team 4) had poor results, but no significant conclusion can be drawn as no other teams used

this approach. The use of SH coefficients as inputs (Team 7) could be worth exploring in dMRI-related applications, but its impact on results is once more not clear, since only one team used it in this study.

There was no observed influence of output format on results, as the bottom-performing methods considered scalar metrics but so did the top-performing method. Interestingly, Team 2, which did not use DL, showed great performance similar to DL methods but with a simpler configuration. The results suggest that the selection of methods and format of input and output is not a determining factor for success in the task. The training process may play a role, see, for instance, Teams 4, 9, 10 and 14, which use only few volumes for validation and testing and achieved the poorest results. On the other hand, Teams 8, 5 and 1 used a similar training procedure with opposite results. Teams 6 and 7 that did not train the model with the *in vivo* data but with a biophysical model, remains in the middle of the table. Note that these two cases did not use in-vivo training data at all, but the physics-based training, and they succeeded in generating DWI signal and the metrics for the healthy and unhealthy cases with good agreement. The improvement over the 21 gradients is moderate but, at the same time, they generate a low number of false positives. The lower number of TP of this method could be due to not using in-vivo data for training and it may indicate that in-vivo data may have richer features that are not captured by the biophysical models.

Finally, the ultimate goal of the study was to replicate the scalar metrics calculated using 61 directions so histograms of the reconstructed data should be comparable to the original histogram, see Fig. 7. It is interesting to notice that, despite the TBSS results, the histogram of the scalars calculated with only 21 gradient directions are very similar to the original one. Some teams (9, 10, 13, 14) showed clear differences with the reference histogram and poor results in TBSS. On the other hand, the methods of the top four (teams 8, 5, 1, 12) showed high similarity with the original data. The remaining teams had slight variations in shape, mean and/or variance.

This raises another crucial issue regarding the use of AI methods in data harmonization. In this study, only one source of difference was considered, i.e., the number of acquired gradient directions. However, in real-life studies, data from different facilities can also differ in terms of resolution, acquisition parameters, or vendor. In such cases, harmonization is essential to compare data sets from different sources on the same study. For harmonization of heterogeneous data, methods unassessed in this study such as transfer learning can be employed. Considering the results from the DL-based reconstructions included in this study, as shown in Fig. 7, the processed datasets may contain significant biases, even with decent results in metrics such as SSIM, PSNR, or MSE. Therefore, future studies must assess the effects of transfer learning or similar techniques for data harmonization in relation to the possible alteration of relevant features, differences between subjects or high rate of FPs.

Finally, it is essential to highlight that the assessment of the methods used in this study was performed utilizing a database focused solely on migraine. Therefore, the findings presented herein may not be directly generalizable to other databases or pathologies. The specificities of the migraine database may influence the given results and their relevance to broader medical research. Different pathologies could exhibit distinct characteristics, patient populations, and data quality, potentially impacting the performance and applicability of the different methods. Thus, while this study's findings hold valuable insights into AI, caution must be exercised when extrapolating these results to other databases.

6. Conclusions

For this particular dMRI study, the performance of different AI methods to generate DTI-based measurements from 21 diffusion gradient orientations to those equivalent from a 61-gradient protocol was vastly divergent, with some methods even producing results inferior to using the unenhanced data. Methods that increase the number of new TPs do so at the cost of also increasing the number of FPs. This is a critical issue in clinical studies, as it has the potential to negatively impact results. While the results of this study cannot be generalized to other problems, it is advisable to exercise caution when using DL techniques in MRI-based clinical studies. Some important conclusions that can be drawn from this study include:

1. Global image metrics based on visual quality and/or errors are inadequate for evaluating the quality of reconstructed datasets if they are to be used in statistical analysis. Such metrics could overlook false positives and false negatives, and alternative metrics that account for small, relevant differences should be adopted by the DL community. End-to-end methods are specifically designed with this goal in mind and may be considered in such situations.
2. Enhancing the quality of medical imaging data using AI methods could significantly increase the number of FPs. In some cases, it might be more appropriate not to process the data in order to avoid biases in data analysis.
3. The issues that may arise with the AI-based processing of dMRI are not limited to a particular scheme. The generalization of AI-based techniques to clinical groups using methods trained in a single group alters the information included in the clinical cohorts of interest. AI-based networks design options, the loss function and the hyperparameter search should be focused on a specific clinical application including data from all groups of interest in the training process.

Acknowledgments

Results here presented are derived from the MICCAI 2022 challenge QuaD22: "Quality augmentation in dMRI for clinical studies: Validation in migraine", held by the CDMRI Workshop. We want to thank all participants, those who succeed in sending the final data, but also those who only participated in the first steps. We also want to thank Dr. David García-Azorín, Dr. Ángel L. Guerrero and Dr. Margarita Rodríguez from Hospital Clínico Universitario (Valladolid, Spain) for the creation of the migraine database used in this work. We acknowledge MICCAI society for hosting the QuaD22 challenge. This work was supported by: Ministerio de Ciencia e Innovación (Spain) by research grant PID2021-124407NB-I00, funded by MCIN/AEI/10.13039/501100011033/FEDER, UE, and TED2021-130758B-I00, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU/PRTR"; Polish National Agency for Academic Exchange for the grant PPN/BEK/2019/1/00421 under the Bekker programme; EPSRC grants M020533 R006032 R014019; Microsoft scholarship; NIHR UCLH Biomedical Research Centre at UCLH NHS Foundation Trust and UCL, EPSRC EP/S021930/1, EPSRC grant EP/V034537/1; Australia Medical Research Future Fund under Grant (MRFFAI000085); German Research Foundation grants ME 3737/19-1 and 269953372/GRK2150; Coordenaçao de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; BrainsCAN and Canada Research Chairs; National Council for Scientific and Technological Development (CNPq) and National Institutes of Health (NIH, R01-EB028774, R01-NS082436, R01-EB031169, R01-AG054328 and R01-MH118020); Ministry of Science and Technology of the People's Republic of China (2021ZD0200202); National Natural Science Foundation of China (81971606, 82122032); and Science and Technology Department of Zhejiang Province (202006140, 2022C03057).

Conflict of interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

Data and code availability

The dataset used in this study is available upon the request of Santiago Aja-Fernández. The source code of all methods is available upon request to the authors of each team. Source code of Team 1 is available upon request to a.rauland@fz-juelich.de. The source code of Team 3 is available upon request to Prof. Edward DiBella (edward.dibella@hsc.utah.edu). Source code of Team 5 is available and can be downloaded from these links: DUnet codes (for FA estimation): http://github.com/myigitavci/ dropout_ISMRM, deepDTI codes (for AD,MD estimation): http://github.com/qiyuantian/DeepDTI. Source code of Team 9 is available upon request to Dr Cabezas (mariano.cabezas@sydney.edu.au). Source code of Team 10 is available upon request to Renata Manzano Maria (re.maria@usp.br). Source code of Team 11 is available under these premises: email to pcx0521@163.com and it can be downloaded from https://github.com/chaineypung. Team 12 uses a simplified version of the PROSUB method. The PROSUB source code is available under the Apache License, Version 2.0 and can be downloaded from https://github.com/sbb-gh/PROSUB. Source code of Team 13 is available upon request to Prof Saurabh J. Shigwan (saurabh.shigwan@snu.edu.in).

References

Aggarwal, H. K., Mani, M. P., Jacob, M., 2019. MoDL-MUSSELS: Model-based deep learning for multishot sensitivity-encoded diffusion MRI. IEEE transactions on medical imaging 39 (4), 1268–1277.

Ahmad, A., Parker, D., Dheer, S., Samani, Z. R., Verma, R., 2023. 3D-QCNet–A pipeline for automated artifact detection in diffusion MRI images. Computerized Medical Imaging and Graphics 103, 102151.

Andersson, J. L., Sotiropoulos, S. N., 2016. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage 125, 1063–1078.

Barrio-Arranz, G., de Luis-García, R., Tristán-Vega, A., Martín-Fernández, M., Aja-Fernández, S., 2015. Impact of MR acquisition parameters on DTI scalar indexes: a tractography based approach. PloS one 10 (10), e0137905.

Basser, P., Pierpaoli, C., 1996. Microstructural features measured using diffusion tensor imaging. J Magn Reson B 111 (3), 209–219.

Basser, P. J., 2002. Relationships between diffusion tensor and q-space MRI. Magnetic Resonance in Medicine 47 (2), 392–397.

Behrens, T. E., Woolrich, M. W., Jenkinson, M., Johansen-Berg, H., Nunes, R. G., Clare, S., Matthews, P. M., Brady, J. M., Smith, S. M., 2003. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 50 (5), 1077–1088.

Blumberg, S. B., Lin, H., Grussu, F., Zhou, Y., Figini, M., Alexander, D. C., 2022. Progressive subsampling for oversampled data-application to quantitative MRI. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI. Springer, pp. 421–431.

Blumberg, S. B., Palombo, M., Khoo, C. S., Tax, C. M. W., Tanno, R., Alexander, D. C., 2019. Multi-stage prediction networks for data harmonization. In: Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, pp. 411–419.

Chen, G., Dong, B., Zhang, Y., Lin, W., Shen, D., Yap, P.-T., 2018. Angular upsampling in infant diffusion MRI using neighborhood matching in x-q space. Frontiers in Neuroinformatics 12, 57.

Chen, G., Dong, B., Zhang, Y., Lin, W., Shen, D., Yap, P.-T., 2019. XQ-SR: joint xq space super-resolution with application to infant diffusion MRI. Medical image analysis 57, 44–55.

Chen, Z., Peng, C., Zhang, H., Zeng, Q., Feng, Y., 2021. Deep-based super-angular resolution for diffusion imaging. In: Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Proceedings, Part III 4. Springer, Beijing, China, pp. 513–523.

Chong, C. D., Peplinski, J., Berisha, V., Ross, K., Schwedt, T. J., 2019. Differences in fibertract profiles between patients with migraine and those with persistent post-traumatic headache. Cephalalgia 39 (9), 1121–1133.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. Springer, pp. 424–432.

Coppola, G., Di Renzo, A., Tinelli, E., Petolicchio, B., Di Lorenzo, C., Parisi, V., Serrao, M., Calistri, V., Tardioli, S., Cartocci, G., Caramia, F., Di Piero, V., Pierelli, F., 2020. Patients with chronic migraine without history of medication overuse are characterized by a peculiar white matter fiber bundle profile. The Journal of Headache and Pain 21 (1), 92.

Daducci, A., Canales-Rodríguez, E. J., Zhang, H., Dyrby, T. B., Alexander, D. C., Thiran, J.-P., 2015. Accelerated microstructure imaging via convex optimization (AMICO) from diffusion MRI data. Neuroimage 105, 32–44.

de Figueiredo, E. H., Borgonovi, A. F., Doring, T. M., 2011. Basic concepts of MR imaging, diffusion MR imaging, and diffusion tensor imaging. Magnetic Resonance Imaging Clinics 19 (1), 1–22.

Descoteaux, M., Angelino, E., Fitzgibbons, S., Deriche, R., 2007. Regularized, fast, and robust analytical Q-ball imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 58 (3), 497–510.

Dhollander, T., Raffelt, D., Connelly, A., 2016. Unsupervised 3-tissue response function estimation from single-shell or multi-shell diffusion MR data without a co-registered T1 image. In: ISMRM Workshop on Breaking the Barriers of Diffusion MRI. ISMRM Lisbon, Italy, p. 5.

Diao, Y., Jelescu, I., 2023. Parameter estimation for WMTI-Watson model of white matter using encoder–decoder recurrent neural network. Magnetic Resonance in Medicine 89 (3), 1193–1206.

Fadnavis, S., Batson, J., Garyfallidis, E., 2020. Patch2Self: Denoising diffusion MRI with self-supervised learning. Advances in Neural Information Processing Systems 33, 16293–16303.

Faiyaz, A., Doyley, M., Schifitto, G., Zhong, J., Uddin, M. N., 2022a. Single-shell NODDI using dictionary-learner-estimated isotropic volume fraction. NMR in Biomedicine 35 (2).

Faiyaz, A., Uddin, M. N., Schifitto, G., 2022b. Angular upsampling in diffusion MRI using contextual hemihex sub-sampling in q-space. arXiv preprint arXiv:2211.00240.

Gibbons, E. K., Hodgson, K. K., Chaudhari, A. S., Richards, L. G., Majersik, J. J., Adluru, G., DiBella, E. V., 2019. Simultaneous NODDI and GFA parameter map generation from subsampled q-space imaging using deep learning. Magnetic resonance in medicine 81 (4), 2399–2411.

Golkov, V., Dosovitskiy, A., Sperl, J. I., Menzel, M. I., Czisch, M., Sämann, P., Brox, T., Cremers, D., 2016. Q-space deep learning: twelve-fold shorter and model-free diffusion MRI scans. IEEE transactions on medical imaging 35 (5), 1344–1351.

HashemizadehKolowri, S., Chen, R.-R., Adluru, G., DiBella, E. V., 2022. Jointly estimating parametric maps of multiple diffusion models from undersampled q-space data: A comparison of three deep learning approaches. Magnetic Resonance in Medicine 87 (6), 2957–2971.

Headache Classification Committee of the International Headache Society, 2018. The International Classification of Headache Disorders, 3rd edition. Cephalalgia 38 (1), 1–211.

Jelescu, I. O., Veraart, J., Fieremans, E., Novikov, D. S., 2016. Degeneracy in model parameter estimation for multi- compartmental diffusion in neuronal tissue. NMR Biomed.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., Smith, S. M., 2012. FSL. Neuroimage 62(2), 782–790. Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer, pp. 694–711.

Jones, D. K., 2004. The effect of gradient sampling schemes on measures derived from diffusion tensor MRI: a Monte Carlo study. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 51 (4), 807–815.

Kattem Husøy, A., Eikenes, L., Haberg, A. K., Hagen, K., Stovner, L. J., 2019. Diffusion tensor imaging in middle-aged headache sufferers in the general population: a cross-sectional population-based imaging study in the Nord-Trøndelag health study (HUNT-MRI). The Journal of Headache and Pain 20 (1), 78.

Landman, B. A., Farrell, J. A., Jones, C. K., Smith, S. A., Prince, J. L., Mori, S., 2007. Effects of diffusion weighting schemes on the reproducibility of DTI-derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 T. Neuroimage 36 (4), 1123–1138.

Li, B., Niessen, W. J., Klein, S., de Groot, M., Ikram, M. A., Vernooij, M. W., Bron, E. E., 2021a. Longitudinal diffusion MRI analysis using Segis-Net: a single-step deep-learning framework for simultaneous segmentation and registration. NeuroImage 235, 118004.

Li, H., Liang, Z., Zhang, C., Liu, R., Li, J., Zhang, W., Liang, D., Shen, B., Zhang, X., Ge, Y., et al., 2021b. SuperDTI: Ultrafast DTI and fiber tractography with deep learning. Magnetic resonance in medicine 86 (6), 3334–3347.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022.

Lyon, M., Armitage, P., Álvarez, M. A., 2022. Angular super-resolution in diffusion MRI with a 3D recurrent convolutional autoencoder. In: International Conference on Medical Imaging with Deep Learning. PMLR, pp. 834–846.

Mani, M., Magnotta, V. A., Jacob, M., 2021. qModeL: A plug-and-play model-based reconstruction for highly accelerated multi-shot diffusion MRI using learned priors. Magnetic resonance in medicine 86 (2), 835–851.

Mani, M., Yang, B., Bathla, G., Magnotta, V., Jacob, M., Apr 2022. Multi-band- and in-plane-accelerated diffusion MRI enabled by model-based deep learning in q-space and its extension to learning in the spherical harmonic domain. Magn Reson Med 87 (4), 1799–1815.

Mori, S., Wakana, S., Van Zijl, P. C., Nagae-Poetscher, L., 2005. MRI atlas of human white matter. Elsevier.

Moyer, D., Ver Steeg, G., Tax, C. M., Thompson, P. M., 2020. Scanner invariant representations for diffusion MRI harmonization. Magnetic resonance in medicine 84 (4), 2174–2189.

Nichols, T., Holmes, A. P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapp 15 (1), 1–25.

Oeschger, J. M., Tabelow, K., Mohammadi, S., 2023. Axisymmetric diffusion kurtosis imaging with Rician bias correction: A simulation study. Magnetic Resonance in Medicine 89 (2), 787–799.

Özarslan, E., Sepherd, T. M., Vemuri, B. C., Blackband, S. J., Mareci, T. H., 2006. Resolution of complex tissue microarchitecture using the Diffusion Orientation Transform (DOT). NeuroImage 31, 1086–1103.

Planchuelo-Gómez, Á., García-Azorín, D., Guerrero, Á. L., Aja-Fernández, S., Rodríguez, M., de Luis-García, R., 2020a. Structural connectivity alterations in chronic and episodic migraine: A diffusion magnetic resonance imaging connectomics study. Cephalalgia 40 (4), 367–383.

Planchuelo-Gómez, Á., García-Azorín, D., Guerrero, Á. L., Aja-Fernández, S., Rodríguez, M., de Luis-García, R., 2020b. White matter changes in chronic and episodic migraine: a diffusion tensor imaging study. The journal of headache and pain 21 (1), 1–15.

Poonawalla, A. H., Zhou, X. J., 2004. Analytical error propagation in diffusion anisotropy calculations. Journal of Magnetic Resonance Imaging, 19 (4), 489–498.

Qiao, Y., Shi, Y., 2021. Unsupervised deep learning for FOD-based susceptibility distortion correction in diffusion MRI. IEEE Transactions on Medical Imaging 41 (5), 1165–1175.

Qin, Y., Liu, Z., Liu, C., Li, Y., Zeng, X., Ye, C., 2021. Super-Resolved q-Space deep learning with uncertainty quantification. Medical Image Analysis 67, 101885.

Rahimi, R., Dolatshahi, M., Abbasi-Feijani, F., Momtazmanesh, S., Cattarinussi, G., Aarabi, M. H., Pini, L., Oct 2022. Microstructural white matter alterations associated with migraine headaches: a systematic review of diffusion tensor imaging studies. Brain Imaging and Behavior 16 (5), 2375–2401.

Ren, M., Kim, H., Dey, N., Gerig, G., 2021. Q-space conditioned translation networks for directional synthesis of diffusion weighted images from multi-modal structural MRI. In: Medical Image Computing and Computer Assisted Intervention– MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. Springer, pp. 530–540.

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., Hawkes, D. J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE transactions on medical imaging 18 (8), 712–721.

Sabidussi, E., Klein, S., Jeurissen, B., Poot, D., 2023. dtiRIM: A generalisable deep learning method for diffusion tensor imaging. NeuroImage, 119900.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis 53, 197–207.

Sedlar, S., Alimi, A., Papadopoulo, T., Deriche, R., Deslauriers-Gauthier, S., 2021. A spherical convolutional neural network for white matter structure imaging via dMRI. In: Medical Image Computing and Computer Assisted Intervention– MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer, pp. 529–539.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: The 3rd International Conference on Learning Representations (ICLR2015). San Diego, CA, USA.

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M., et al., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage 31 (4), 1487–1505.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., et al., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23, S208–S219.

Smith, S. M., Nichols, T. E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44 (1), 83–98.

Tang, Z., Cabezas, M., Liu, D., Barnett, M., Cai, W., Wang, C., 2021. LG-Net: lesion gate network for multiple sclerosis lesion inpainting. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. Springer, pp. 660–669.

Tang, Z., Wang, X., Cabezas, M., D'Souza, A., Calamante, F., Liu, D., Barnett, M., Wang, C., Cai, W., 2022. Diffusion MRI fibre orientation distribution inpainting. In: Computational Diffusion MRI: 13th International Workshop, CDMRI 2022, Held in Conjunction with MICCAI 2022, Singapore, Singapore, September 22, 2022, Proceedings. Springer, pp. 65–76.

Tax, C. M., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., Evans, C. J., St-Jean, S., Leemans, A., Koppers, S., Merhof, D., et al., 2019. Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms. NeuroImage 195, 285–299.

Tian, Q., Bilgic, B., Fan, Q., Liao, C., Ngamsombat, C., Hu, Y., Witzel, T., Setsompop, K., Polimeni, J. R., Huang, S. Y., 2020. DeepDTI: High-fidelity six-direction diffusion tensor imaging using deep learning. NeuroImage 219, 117017.

Tournier, J.-D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., Christiaens, D., Jeurissen, B., Yeh, C.-H., Connelly, A., 2019. MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. Neuroimage 202, 116137.

Tristán-Vega, A., Aja-Fernández, S., Westin, C.-F., 2012. Least squares for diffusion tensor estimation revisited: Propagation of uncertainty with Rician and non-Rician signals. Neuroimage 59 (4), 4032–4043.

Tristán-Vega, A., Westin, C.-F., Aja-Fernández, S., 2009. Estimation of fiber orientation probability density functions in high angular resolution diffusion imaging. NeuroImage 47 (2), 638–650.

Tuch, D. S., 2004. Q-ball imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 52 (6), 1358–1372.

Tuch, D. S., Reese, T. G., Wiegell, M. R., Wedeen, V. J., 2003. Diffusion MRI of complex neural architecture. Neuron 40, 885–895.

Veraart, J., Novikov, D. S., Christiaens, D., Ades-Aron, B., Sijbers, J., Fieremans, E., 2016. Denoising of diffusion MRI using random matrix theory. Neuroimage 142, 394–406.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13 (4), 600–612.

Westin, C.-F., Maier, S. E., Mamata, H., Nabavi, A., Jolesz, F. A., Kikinis, R., 2002. Processing and visualization for diffusion tensor MRI. Medical image analysis 6 (2), 93–108.

Xue, T., Zhang, F., Zhang, C., Chen, Y., Song, Y., Golby, A. J., Makris, N., Rathi, Y., Cai, W., O'Donnell, L. J., 2023. Superficial white matter analysis: An efficient point-cloud-based deep learning framework with supervised contrastive learning for consistent tractography parcellation across populations and dMRI acquisitions. Medical Image Analysis, 102759.

Yan, J., Zhao, Y., Chen, Y., Wang, W., Duan, W., Wang, L., Zhang, S., Ding, T., Liu, L., Sun, Q., Pei, D., Zhan, Y., Zhao, H., Sun, T., Sun, C., Wang, W., Liu, Z., Hong, X., Wang, X., Guo, Y., Li, W., Cheng, J., Liu, X., Lv, X., Li, Z.-C., Zhang, Z., Oct 2021. Deep learning features from diffusion tensor imaging improve glioma stratification and identify risk groups with distinct molecular pathway activities. eBioMedicine 72.

Ye, C., 2017. Tissue microstructure estimation using a deep network inspired by a dictionary-based framework. Medical image analysis 42, 288–299.

Ye, C., Li, X., Chen, J., 2019. A deep network for tissue microstructure estimation using modified LSTM units. Medical image analysis 55, 49–64.

Ye, C., Li, Y., Zeng, X., 2020. An improved deep network for tissue microstructure estimation with uncertainty quantification. Medical image analysis 61, 101650.

Yu, D., Yuan, K., Zhao, L., Dong, M., Liu, P., Yang, X., Liu, J., Sun, J., Zhou, G., Xue, T., Zhao, L., Cheng, P., Dong, T., von Deneen, K. M., Qin, W., Tian, J., 2013. White matter integrity affected by depressive symptoms in migraine without aura: a tract-based spatial statistics study. NMR in Biomedicine 26 (9), 1103–1112.

Zeng, R., Lv, J., Wang, H., Zhou, L., Barnett, M., Calamante, F., Wang, C., 2022. FOD-Net: A deep learning method for fiber orientation distribution angular super resolution. Medical Image Analysis 79, 102431.

Zhang, F., Breger, A., Cho, K. I. K., Ning, L., Westin, C.-F., O'Donnell, L. J., Pasternak, O., 2021a. Deep learning based segmentation of brain tissue from diffusion MRI. NeuroImage 233, 117934.

Zhang, F., Wells, W. M., O'Donnell, L. J., 2021b. Deep diffusion MRI registration (DDMReg): a deep learning method for diffusion MRI registration. IEEE Transactions on Medical Imaging 41 (6), 1454–1467.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE transactions on medical imaging 20 (1), 45–57.

Zheng, T., Zheng, W., Sun, Y., Zhang, Y., Ye, C., Wu, D., 2022. An adaptive network with extragradient for diffusion MRI-based microstructure estimation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I. Springer, pp. 153–162.

Zhu, A. H., Moyer, D. C., Nir, T. M., Thompson, P. M., Jahanshad, N., 2019. Challenges and opportunities in dMRI data harmonization. In: Computational Diffusion MRI: International MICCAI Workshop, Granada, Spain, September 2018 Springer, pp. 157–172.

Figures and Tables

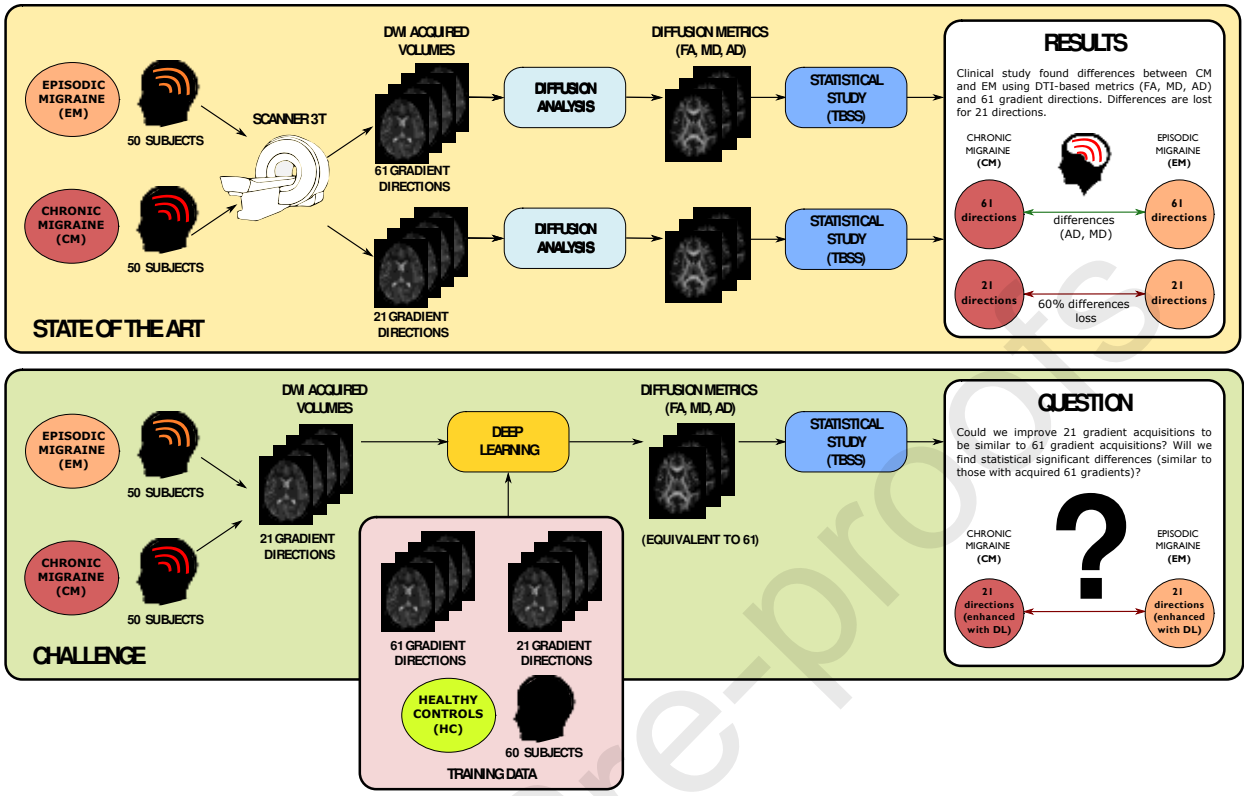


Figure 1: Overview of the study carried out in this work. TOP: State of the art process of DTI data using 61 and 21 gradient directions EM is compared to CM using data acquired with 61 gradient directions. BOTTOM: Overview of the task proposed in the challenge: a Deep Learning network is trained using healthy controls. The trained network is used to estimate parameters from patients acquired with 21 gradient directions.

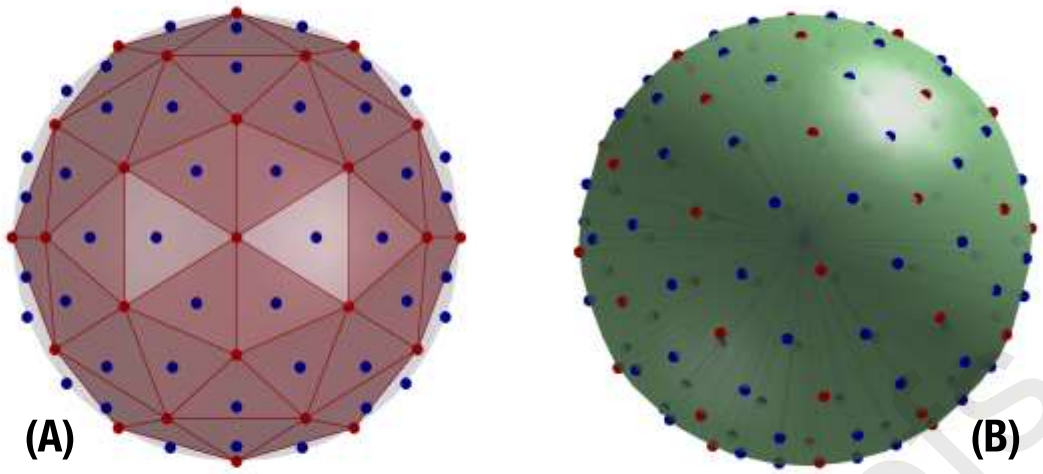


Figure 2: Sampling scheme of the sphere for the definition of the gradient directions in the dMRI acquisition protocol. Red points indicate the samples for 21 gradient directions. Blue points indicate a sampling of 40 directions. The complete sampling (61 directions) requires the blue and red points. (a) Samples over the icosahedron. (b) Sampling directions over the surface of the sphere (podal and antipodal directions are shown).

	Architecture	Input data	Output data	Description
Conventional machine learning				
Team 2	FCN	DWIs	DWIs	FCN interpolation of DWI data using HemiHex (Faiyaz et al., 2022b) subsampling in q-space.
Convolutional neural networks				
Team 1	CNN+ MLP	DWIs	FA/AD/MD	Direct mapping the DWIs into the enhanced metrics through convolutions and multilayer perceptron.
Team 4	CNN	FA/AD/MD	FA/AD/MD	Mapping between the diffusion metrics estimated from 21 gradients and enhanced metrics via convolutions and residual learning.
Other deep learning architectures				
Team 3	UN	DWIs	FA/AD/MD	Mapping sparse spatial-angular resolution into diffusion parameters by fully connected layers.
Team 5	U-Net	FA + DWIs	FA + DWIs	FA: Mapping the DWIs and FA estimated from 21 gradients into the corresponding enhanced FA through U-net with extra dropout layers. AD, MD: Denoising the DWIs and DTI estimation.
Team 6	DAE	DWIs	DWIs	Modeling dMRI signal with a three-compartment model and mapping 21 DWIs into enhanced equivalents via the auto-encoder network.
Team 7	DAE	SH coeffi- cients	DWIs	Modeling dMRI signal with a three-compartment model and mapping the spherical harmonics coefficients between DWIs using the auto-encoder network.
Team 8	AEME	DWIs	FA/AD/MD	Direct mapping the DWIs from 21 gradients into the corresponding metric through iteration blocks for sparse representation of dMRI signals with extragradient.
Team 9	U-Net	DWIs	FA/AD/MD	Direct mapping the DWIs estimated from 21 gradients into the corresponding measure through Encoder-Decoder framework.
Team 10	CNN+ residue learning	DWIs	DWIs	Mapping DWIs to obtain diffusion metrics by use of synthetic DWIs generated by video frame interpolation in the polar space.
Team 11	CNN+ SARDI-Net	DWIs	FA/AD/MD	Mapping the DWIs into the enhanced metrics through q-space sampling.
Team 12	DNSR	DWIs	DWIs	Mapping the DWIs into the corresponding enhanced metrics by progressive subsampling and reconstruction through convolution.
Team 13	Swin-Transformer	DWIs	FA/AD/MD	Mapping the DWIs into diffusion metrics by attention and perception mechanisms and adjusting the data to each value range.
Team 14	U-Net	DWIs	FA/AD/MD (difference 21-61)	Mapping the DWIs into the difference between 21 and 61 scenario by max pooling and up sampling with skip layers.

Table 1: Employed AI-reconstruction methods to synthesize FA, AD and MD from 21 gradient directions. FCN: Fully Connected Network; CNN: Convolutional neural network; MLP: Multilayer perceptron; UN: Unrolled network (Ye et al., 2020; HashemizadehKolowri et al., 2022); U-Net (C¸ i¸cek et al., 2016; Tang et al., 2021, 2022); DAE: Denoising autoencoder (Faiyaz et al., 2022b); AEME: Adaptive network with extragradient for dMRI based microstructure estimation (Zheng et al., 2022); SARDI-Net: Super-angular Resolution Diffusion Imaging Network (Chen et al., 2021); DNSR: Dual network scoring and reconstruction (Blumberg et al., 2022); Swin-Transformer (Liu et al., 2021). The term *enhanced metrics* used in the description of the methods refers to the 3 considered metrics (FA, MA, AD) calculated by the different methods to achieve a quality similar to the parameters estimated from 61 gradient directions.

	Number of subjects				
	Training	Validation	Testing	Loss function	Cross-validation
Team 1*	45	5	10	MSE	No
	48	12	0	MSE	5-fold
Team 2†	3	5	5	MSE	No
Team 3	36	9	15	MSE	No
Team 4	44	10	6	MSE	No
Team 5*	40	10	10	FA: MSE/ AD, MD: MAE	No
	50	10	0	FA: MSE/ AD, MD: MAE	No
Team 6‡	0	5	0	MSE	No
Team 7‡	0	5	0	MSE	No
Team 8	42	6	12	MSE	No
Team 9	48	0	12	MAE	5-fold
Team 10	50	5	5	VGG loss	No
Team 11	40	0	20	MSE	No
Team 12	48	6	6	MSE	10-fold
Team 13◇	16	2	2	MAE	No
Team 14	54	3	3	MSE	No

Table 2: Training procedure for the different AI methods considered: Number of subjects used for training, validation, and testing; loss function used for training. *Teams 1 and 5 trained the method with two different divisions of subjects to improve the results after selecting the best method. †Only 13 subjects were used to train the method, and the remaining subjects were unused. ‡Artificially generated samples using a three-compartment biophysical model for training and testing subsets. ◇The method was trained with three different sets of 20 subjects.

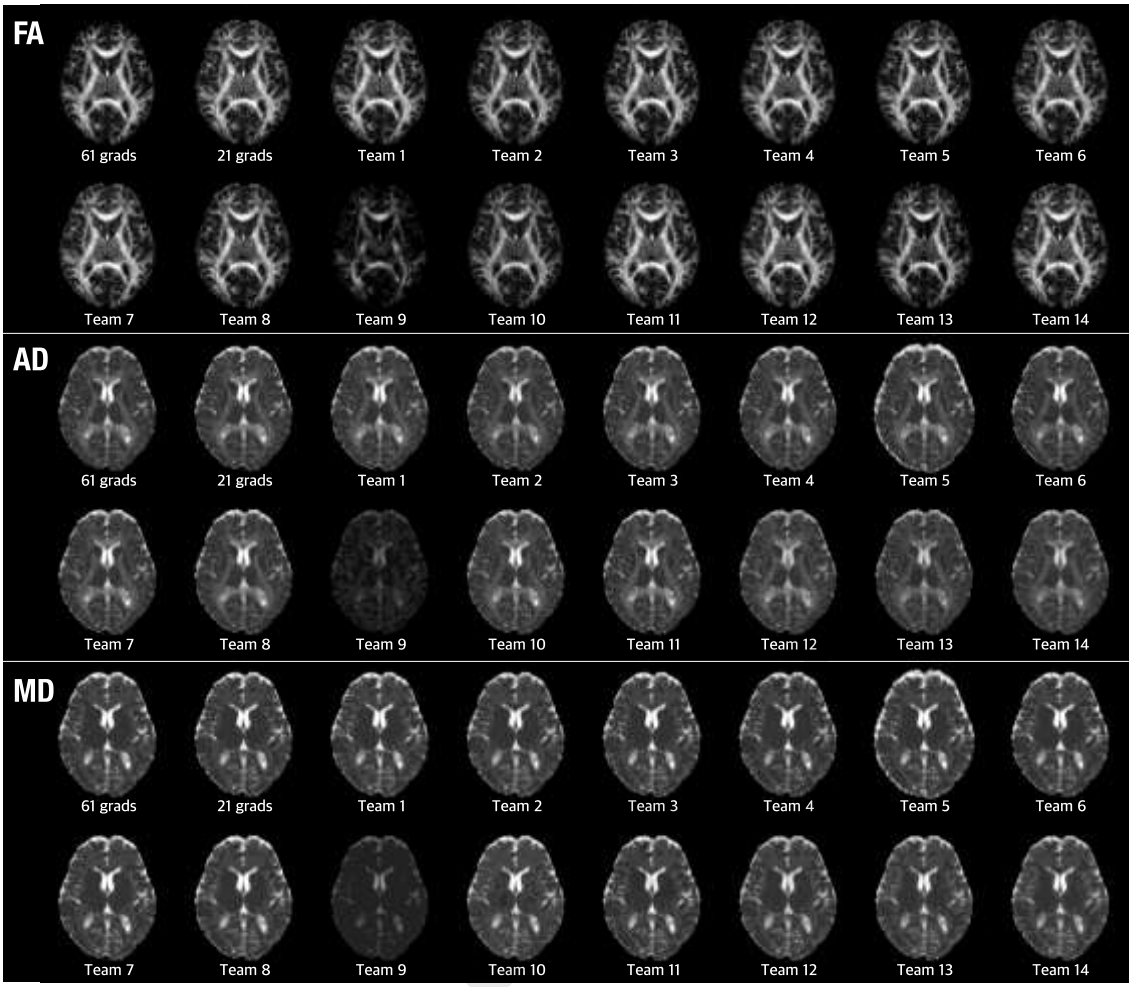


Figure 3: Slide 81 from one CM patient. Three metrics are considered (FA, AD, MD), estimated from the original data ("*61 grads*"), the original data with only 21 directions ("*21 grads*") and the different AI-enhanced methods.

	SSIM			PSNR		
	FA	AD	MD	FA	AD	MD
REF (21 grad)	**0.90**	**0.91**	**0.94**	**30.0**	**80.2**	**83.3**
Team 1	0.92	0.92	0.94	31.5	81.0	83.3
Team 2	0.91	0.91	0.94	30.8	80.6	**82.7**
Team 3	0.92	0.92	0.94	31.4	80.8	83.3
Team 4	0.92	0.92	0.94	31.5	81.1	83.4
Team 5	0.92	**0.84**	**0.88**	31.5	**68.8**	**71.1**
Team 6	0.90	0.91	0.94	30.4	**79.9**	81.6
Team 7	0.90	0.91	0.94	**29.1**	80.2	**82.8**
Team 8	**0.89**	**0.90**	**0.92**	**29.8**	**76.2**	**78.6**
Team 9	**0.48**	**0.42**	**0.53**	**22.4**	**68.9**	**70.7**
Team 10	**0.87**	**0.90**	**0.93**	**29.1**	**79.3**	**80.8**
Team 11	0.92	0.92	0.94	31.4	80.7	83.2
Team 12	0.91	**0.90**	**0.93**	30.7	**78.3**	**79.3**
Team 13	**0.84**	**0.86**	**0.90**	**28.4**	**78.4**	**80.1**
Team 14	0.90	0.91	0.94	30.7	**78.3**	**79.3**

Table 3: Quality metrics between AI-enhanced scalars (FA, AD, MD) using 21 gradients compared to the original scalars calculated with 61 gradients. Structural similarity index measure (SSIM) and Peak Signal to noise ratio (PSNR) are calculated. "REF (21 grad)" stands for the metrics calculated directly from 21 gradient directions without using any AI algorithm to process the data. Results show the average for the 100 reconstructed volumes. In red, those cases that did not improve the reference. In amber, those cases that are slightly worse than the reference.

	FA		AD		MD		ROIS					
							TP			FP		
	FP	TP	FP	TP	FP	TP	FA	AD	MD	FA	AD	MD
61 GRADS	**0**	**0**	**0**	**17029**	**0**	**14981**	**0**	**40**	**40**	**0**	**0**	**0**
21 GRADS	**0**	**0**	**626**	**6981**	**797**	**5941**	**0**	**29**	**27**	**0**	**0**	**0**
Team 1	0	0	1516	10721	2572	10266	0	32	38	0	0	**2**
Team 2	0	0	1553	10594	1880	9592	0	34	38	0	0	**1**
Team 3	0	0	1345	10049	2330	9887	0	34	37	0	0	**2**
Team 4	0	0	373	4797	1859	8563	0	21	35	0	0	**4**
Team 5	0	0	1938	11766	2594	10022	0	36	38	0	**2**	**2**
Team 6	0	0	1312	9005	818	6313	0	34	30	0	**1**	0
Team 7	0	0	978	8845	1680	8433	0	33	35	0	0	**1**
Team 8	0	0	3381	13402	3384	11375	0	37	40	0	**2**	1
Team 9	0	0	0	0	0	0	0	0	0	0	0	0
Team 10	0	0	2386	10332	3150	9586	0	35	38	0	**1**	**2**
Team 11	0	0	1957	11354	2540	9841	0	34	38	0	**1**	**2**
Team 12	0	0	2117	11799	2785	9999	0	35	37	0	**1**	**2**
Team 13	0	0	1197	7069	1879	7802	0	33	36	0	0	**1**
Team 14	0	0	222	3736	1710	7016	0	19	31	0	0	**1**

(a) Absolute numbers

	AD					MD					TOTAL					
	ACC	TPR	TNR	PPV	FPR	ACC	TPR	TNR	PPV	FPR	ACC	TPR	TNR	PPV	FPR	Comp 21 grads
21 GRADS	**73%**	**41%**	**97%**	**92%**	**3%**	**75%**	**40%**	**97%**	**88%**	**3%**	**74%**	**40%**	**97%**	**90%**	**3%**	**0,0%**
Team 1	80%	63%	93%	88%	7%	81%	69%	89%	80%	11%	81%	66%	91%	84%	9%	**26,3%**
Team 2	80%	62%	93%	87%	7%	81%	64%	92%	84%	8%	81%	63%	93%	85%	7%	**25,6%**
Team 3	79%	59%	94%	88%	6%	81%	66%	90%	81%	10%	80%	62%	92%	84%	8%	**23,2%**
Team 4	68%	28%	98%	93%	2%	79%	57%	92%	82%	8%	73%	42%	95%	86%	5%	**-1,8%**
Team 5	82%	69%	91%	86%	9%	81%	67%	89%	79%	11%	81%	68%	90%	83%	10%	**28,1%**
Team 6	76%	53%	94%	87%	6%	76%	42%	97%	89%	3%	76%	48%	95%	88%	5%	8,2%
Team 7	77%	52%	96%	90%	4%	79%	56%	93%	83%	7%	78%	54%	94%	87%	6%	15,2%
Team 8	82%	79%	85%	80%	15%	82%	76%	86%	77%	14%	82%	77%	85%	79%	15%	**31,8%**
Team 9	57%	0%	100%	0%	0%	62%	0%	100%	0%	0%	59%	0%	100%	0%	0%	**-56,1%**
Team 10	77%	61%	89%	81%	11%	78%	64%	87%	75%	13%	78%	62%	88%	78%	12%	14,1%
Team 11	81%	67%	91%	85%	9%	80%	66%	90%	79%	10%	80%	66%	90%	82%	10%	**25,3%**
Team 12	81%	69%	90%	85%	10%	80%	67%	89%	78%	11%	81%	68%	89%	82%	11%	**26,3%**
Team 13	72%	42%	95%	86%	5%	77%	52%	92%	81%	8%	74%	46%	93%	83%	7%	1,4%
Team 14	66%	22%	99%	94%	1%	75%	47%	93%	80%	7%	70%	34%	96%	85%	4%	**-13,1%**

(b) Ratios

Table 4: Statistically significant differences for the comparison between CM and EM using TBSS in terms of True Positives (TP) and False Positives (FP) for the 3 considered metrics (FA, AD, MD) and total TPs and FPs over all metrics. "61 grads" and "21 grads" stand for the metrics calculated directly from 61 and 21 gradient directions without using any AI algorithm to process the data. A total number of 39256 points is considered. (a)

Absolute number for each Team. (b) Ratios of the metrics in the previous table. “Compar. 21” stands for the global improvement with respect to the reference (21 gradients).

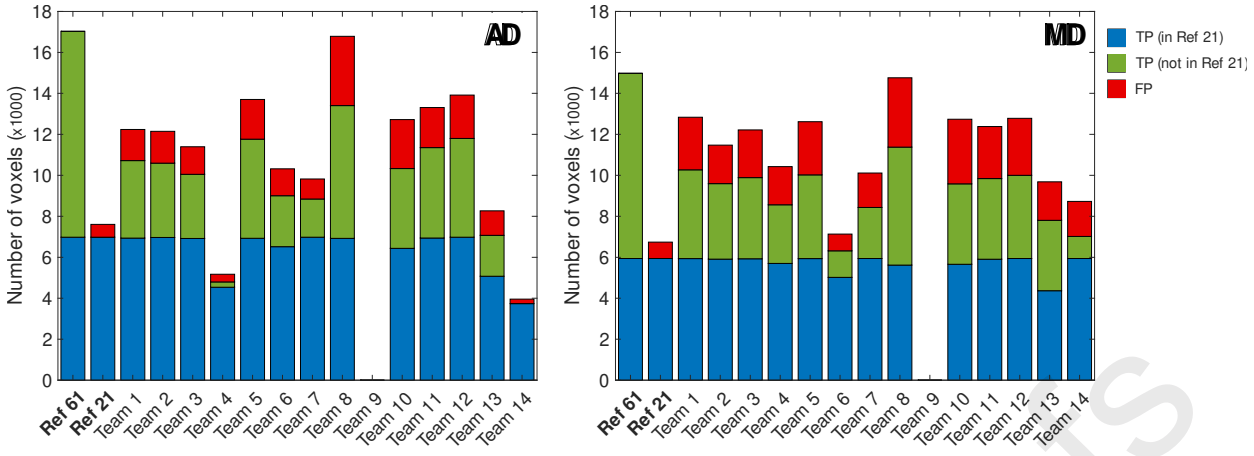


Figure 4: Number of voxels with significant differences detected by TBSS for the different methods. Blue: true positives detected by Ref21. Green: True positives detected by Ref61 but not by Ref21. Red: False positives.

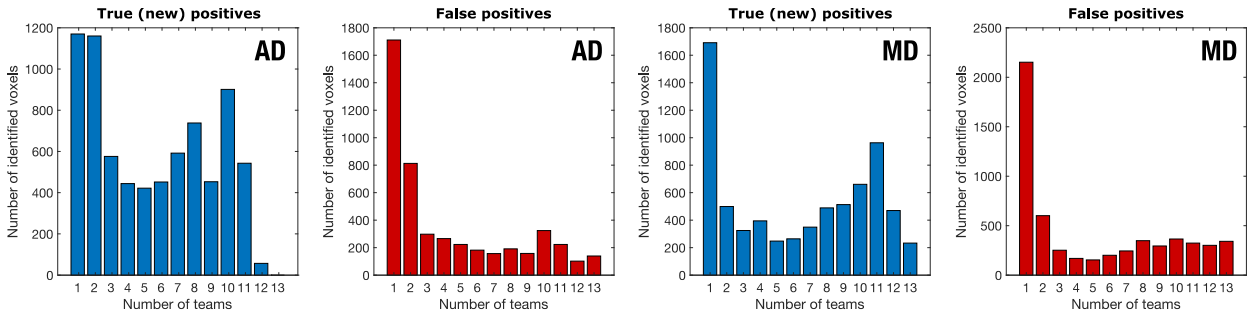


Figure 5: Histograms of (new) TPs (blue) and FPs (red) found for AD and MD. The number of coincidences '*n*' in the abscissa indicates the number of voxels detected as statistically significant in '*n*' teams at the same time.

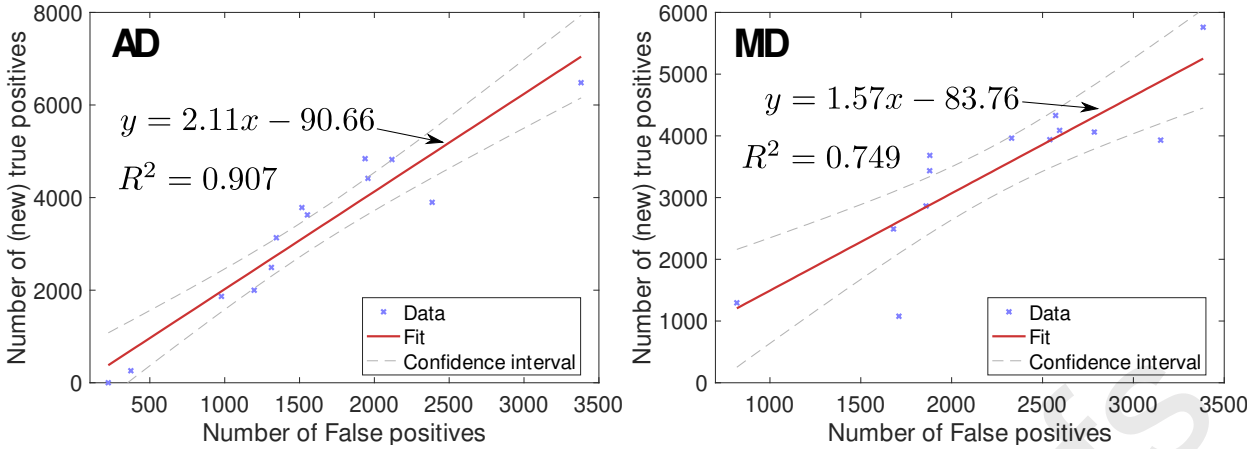


Figure 6: Relation between the (new) true positives and the false positives for each method computed by a linear regression. Each marker denotes the value for a different Team using values from Table 4. The red lines indicate the linear model fitted to the data. The dashed lines indicate the 5th and 95th percentiles. The regions between them were shaded for visualization purposes. The goodness-of-fit R^2 is indicated in each plot with the parameters of the linear fitting.

		Ref 61	**Ref 21**	**T.1**	**T.2**	**T.3**	**T.4**	**T.5**	**T.6**
FA	Mean	**0.5962**	0.5967	0.5932	0.5508	0.5948	0.5853	0.5859	0.5647
	STD	**0.1428**	0.1416	0.1412	0.1367	0.1428	0.1360	0.1405	0.1427
	CV	**0.2395**	0.2373	0.2381	0.2481	0.2400	0.2324	0.2399	0.2527
AD	Mean ($\times 10^{-3}$)	**13.086**	13.123	13.065	12.369	13.053	12.792	13.003	12.581
	STD ($\times 10^{-3}$)	**0.2365**	0.2392	0.2360	0.2023	0.2356	0.2207	0.2420	0.2230
	CV	**0.1807**	0.1823	0.1806	0.1635	0.1805	0.1725	0.1861	0.1773
MD	Mean ($\times 10^{-3}$)	**0.7325**	0.7347	0.7322	0.7277	0.7312	0.7346	0.7305	0.7284
	STD ($\times 10^{-3}$)	**0.7720**	0.7784	0.7631	0.7576	0.7680	0.7228	0.7772	0.7706
	CV	**0.1054**	0.1059	0.1042	0.1041	0.1050	0.0984	0.1064	0.1058
		T.7	**T.8**	**T.9**	**T.10**	**T.11**	**T.12**	**T.13**	**T.14**
FA	Mean	0.5937	0.5776	**0.3048**	**0.4909**	0.5919	0.5866	**0.5112**	0.5877
	STD	0.1421	0.1392	**0.1252**	0.1318	0.1409	0.1409	0.1368	0.1408
	CV	0.2393	0.2411	**0.4108**	**0.2686**	0.2380	0.2402	**0.2677**	0.2396
AD	Mean ($\times 10^{-3}$)	13.066	12.924	**0.5606**	12.618	13.061	12.924	12.400	**12.076**
	STD ($\times 10^{-3}$)	0.2379	0.2352	**0.0867**	0.2114	0.2392	0.2304	**0.1936**	0.2390
	CV	0.1821	0.1820	**0.1546**	0.1675	0.1831	0.1783	**0.1562**	0.1979
MD	Mean ($\times 10^{-3}$)	0.7337	0.7353	**0.4177**	**0.7853**	0.7296	0.7301	0.7306	0.7334
	STD ($\times 10^{-3}$)	0.7773	0.7465	**0.3321**	**0.8333**	0.7642	0.7538	**0.5086**	0.7794
	CV	0.1060	0.1015	**0.0795**	0.1061	0.1047	0.1032	**0.0696**	0.1063

Table 5: Statistics (Mean, standard deviation, and coefficient of variation) of metrics FA, AD and MD, measured over the skeleton of the FA for all the EM subjects. Similar results can be found for CM subjects. We have highlighted those values that differ the most with respect the original data (61-direction reference).

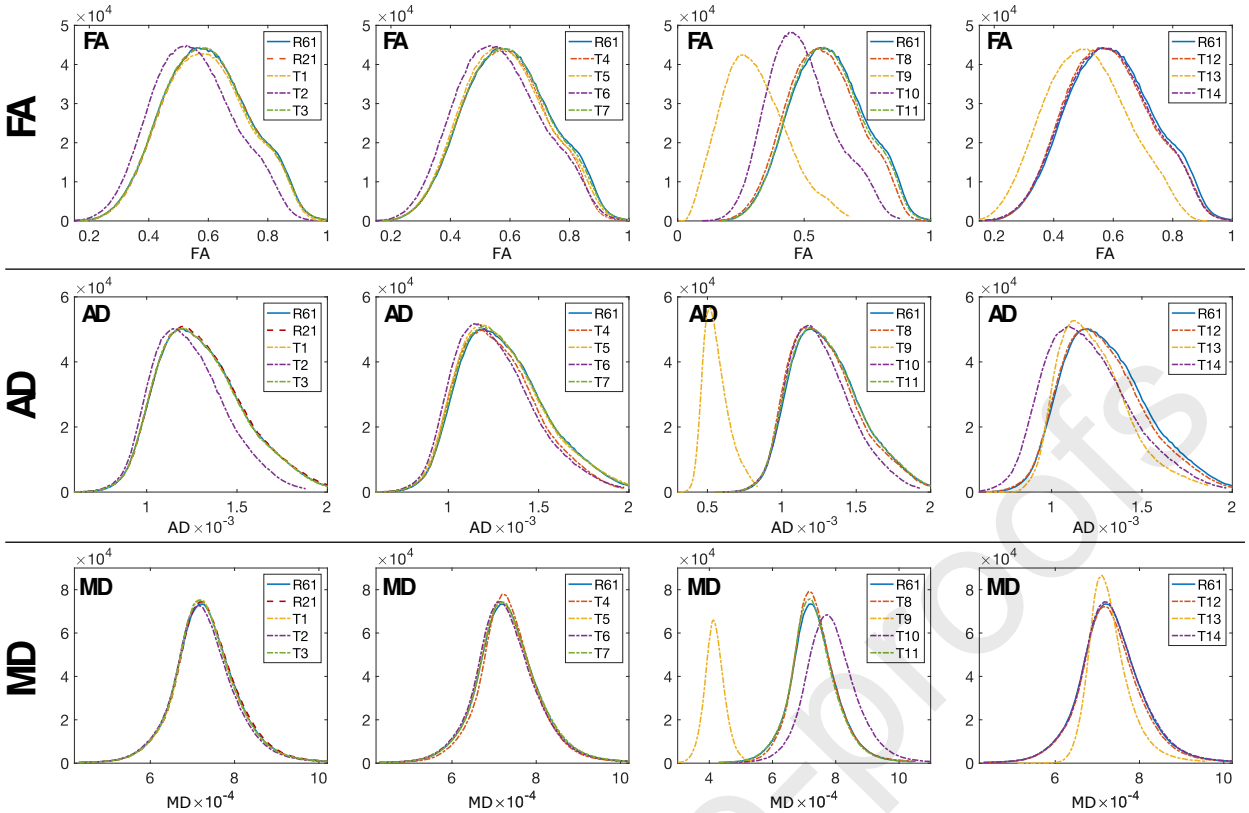


Figure 7: Histograms of the values of the three considered metrics (FA, AD, MD) for the EM processed volumes provided by the different teams. The histograms are calculated over the area given by the FA mask. R61: Histogram of the original data calculated with 61 gradient directions; R21: histogram of the original data calculated with 21 gradient directions; T*n*: histogram of the data provided by Team *n*. R61 (continuous blue line) is included to all the figures for the sake of reference.

	TBSS ORDER	Compar. 21 grads	ACC	TPR	FPR	SSIM	PSNR	TRAIN	VALID.	TEST	LOSS FUNC	SCHEME	INPUT	OUTPUT	
Team 8	1	31.8%	82.2%	77.4%	14.5%	0.91	61.5	42	6	12	MSE	AEME	DWIs	FA/AD/MD	IMPROVEMENT
Team 5	2	28.1%	81.2%	68.1%	9.7%	0.88	57.1	40	10	10	MSE / MAE	U-Net	DWIs + FA	DWIs + FA	
Team 1	3	26.3%	80.8%	65.6%	8.8%	0.93	65.2	45	5	10	MSE	CNN + MLP	DWIs	FA/AD/MD	
Team 12	4	26.3%	80.8%	68.1%	10.5%	0.91	62.8	48	6	6	MSE	DNSR	DWIs	DWIs	
Team 2	5	25.6%	80.6%	63.1%	7.4%	0.92	64.7	3	5	5	RMSE	FCN	DWIs	DWIs	
Team 11	6	25.3%	80.5%	66.2%	9.7%	0.93	65.1	40	0	20	PSNR	CNN + SARDI-Net	DWIs	FA/AD/MD	
Team 3	7	23.2%	79.9%	62.3%	7.9%	0.93	65.2	36	9	15	MSE	UN	DWIs	FA/AD/MD	
Team 7	8	15.2%	77.9%	54.0%	5.7%	0.92	64.0	0	5	0	MSE	DAE	SH coeffs	DWIs	SMALL IMPROV
Team 10	9	14.1%	77.5%	62.2%	11.9%	0.90	63.0	50	5	5	VGG loss	CNN + RL	DWIs	DWIs	
Team 6	10	8.2%	76.0%	47.9%	4.6%	0.92	64.0	0	5	0	MSE	DAE	DWIs	DWIs	
Team 13	11	1.4%	74.3%	46.5%	6.6%	0.87	62.3	16	2	2	MAE	Swin-Transformer	DWIs	FA/AD/MD	
21 GRADS	12	0.0%	73.9%	40.4%	3.1%	0.92	64.5	-	-	-	-	-	-	-	Ref
Team 4	13	-1.8%	73.4%	41.7%	4.8%	0.93	65.3	44	10	6	MSE	CNN	FA/AD/MD	FA/AD/MD	NO IMPROV
Team 14	14	-13.1%	70.5%	33.6%	4.2%	0.92	62.8	54	3	3	MSE	U-Net	DWIs	FA/AD/MD	
Team 9	15	-56.1%	59%	0.0%	0%	0.48	54.0	48	0	12	MAE	U-Net	DWIs	FA/AD/MD	

Table 6: Overview of all the results. The Methods have been ordered following results of the comparison with 21 gradient directions. SSIM and PSNR shows the average of the values in Table 3. For Teams 1 and 5, only the numbers for the first training methods are shown.

EPISODIC MIGRAINE (EM)

50 SUBJECTS

CHRONIC MIGRAINE (CM)

50 SUBJECTS

SCANNER 3T

DWI ACQUIRED VOLUMES

61 GRADIENT DIRECTIONS

21 GRADIENT DIRECTIONS

DIFFUSION ANALYSIS

DIFFUSION ANALYSIS

DIFFUSION METRICS (FA, MD, AD)

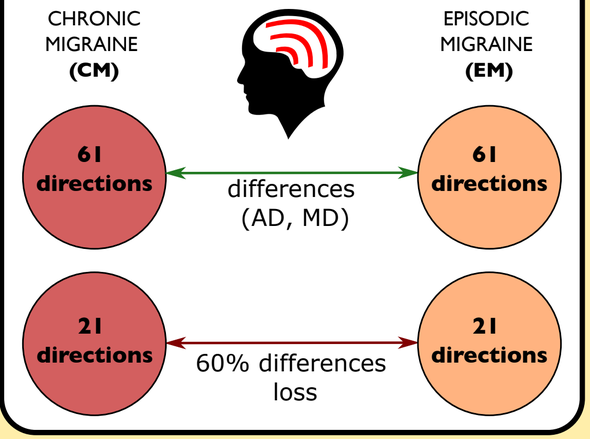
STATISTICAL STUDY (TBSS)

STATISTICAL STUDY (TBSS)

STATE OF THE ART

RESULTS

Clinical study found differences between CM and EM using DTI-based metrics (FA, MD, AD) and 61 gradient directions. Differences are lost for 21 directions.



EPISODIC MIGRAINE (EM)

50 SUBJECTS

CHRONIC MIGRAINE (CM)

50 SUBJECTS

DWI ACQUIRED VOLUMES

21 GRADIENT DIRECTIONS

DEEP LEARNING

DIFFUSION METRICS (FA, MD, AD)

(EQUIVALENT TO 61)

STATISTICAL STUDY (TBSS)

61 GRADIENT DIRECTIONS

21 GRADIENT DIRECTIONS

HEALTHY CONTROLS (HC)

60 SUBJECTS

TRAINING DATA

CHALLENGE

QUESTION

Could we improve 21 gradient acquisitions to be similar to 61 gradient acquisitions? Will we find statistical significant differences (similar to those with acquired 61 gradients)?