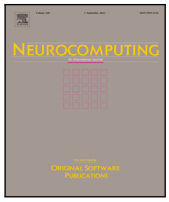
Neurocomputing 602 (2024) 128292

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



RAD-IQMRI: A benchmark for MRI image quality assessment

Yueran Ma [a], Jianxun Lou [b,a,*], Jean-Yves Tanguy [c], Padraig Corcoran [a], Hantao Liu [a]

[a] School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AG, United Kingdom
[b] School of Computer Science, Northeast Electric Power University, Jilin, China
[c] Department of Radiology, Angers University Hospital, Angers, 49933, France

ARTICLE INFO

Communicated by S. He

Keywords:
MRI
Artifacts
Radiology
Subjective experiment
Image quality assessment

ABSTRACT

Magnetic resonance imaging (MRI) is susceptible to visual artifacts that can degrade the perceptual image quality, potentially leading to inaccurate or inefficient diagnoses in clinical practice. It is critical to evaluate the perceptual image quality and build this technique into clinical solutions. In a previous study, an MRI database was created for image quality assessment (IQA), where various types of MRI artifacts with different degrees of degradation were simulated. Application specialists assessed the image quality; however, radiologists' perception of MRI image quality remains unknown. To make IQA clinically relevant, in this paper we conduct a new subjective experiment where 13 radiologists rated the quality of images contained in the MRI database. Based on this subjective IQA benchmark named RAD-IQMRI, we evaluate the performance of state-of-the-art objective IQA models, providing insights into their application for MRI image quality assessment in clinical settings.

1. Introduction

Magnetic resonance imaging (MRI) represents an advanced, non-invasive imaging technique, capable of revealing detailed tissue structures within the body. It provides invaluable biochemical information on the body's metabolism, reflecting cellular activity [1]. The advantage of MRI over other imaging techniques, such as X-ray radiography and Computed Tomography (CT), is not only the avoidance of hazardous ionizing radiation but also the ability to produce images that better represent the structure of soft tissues [2]. However, MRI suffers from the disadvantage of extended acquisition times and complex scanning protocols and parameters. During acquisition and processing, various sources of interference such as hardware imperfections (e.g., equipment noise, and electromagnetic interference), technician errors, patient motion, and underlying physiological processes can cause visual artifacts in MRI images [3]. These artifacts degrade perceptual image quality, which can potentially lead to misinterpretations, erroneous diagnoses, and substandard patient care [4]. Accurately measuring perceptual image quality is crucial for ensuring the accuracy and reliability of MRI-based diagnoses, as well as the efficiency of the clinical workflow.

Perceptual image quality assessment (IQA) provides numerous advantages for medical imaging, such as enhancing diagnostic performance, improving patient outcomes, and facilitating the development of advanced imaging technologies [5]. For example, various methods have been developed to reduce noise including both structural and non-structural noise in MRI images [6–8]. The ability to assess the perceptual quality of output images is crucial for quality monitoring and assurance. In real clinical practice, a large volume of MRI images is generated. These images are initially screened by qualified inspectors to exclude unusable and low-quality images, creating a substantial workload and leading to variable results due to differing review criteria amongst inspectors [9]. In this case, automatic image quality assessment is highly beneficial for providing fast and consistent solutions for quality screening.

The development of image quality assessment (IQA) models is grounded in perception studies where human participants rate image quality within a fully controlled experimental environment [10]. For IQA of natural images, several widely recognized databases have been established, such as LIVE [11], TID2013 [12], and CSIQ [13]. These IQA databases enhance our understanding of how human viewers perceive image quality and provide ground truth data essential for developing IQA models that can automatically predict image quality. However, the quality perception for medical images significantly differs from natural images [14]. For instance the perceived quality of natural images is mainly determined by the visibility of artifacts. In contrast, for medical images, both the diagnostic task and the presence of artifacts play a crucial role in determining the perceptual image quality [14]. Unfortunately, limited research has been undertaken in the area of perceptual image quality assessment of medical images, particularly there is a paucity of studies involving a sufficient number

* Corresponding author.
E-mail address: louj2@cardiff.ac.uk (J. Lou).

https://doi.org/10.1016/j.neucom.2024.128292
Received 24 March 2024; Received in revised form 28 June 2024; Accepted 28 July 2024
Available online 31 July 2024

Table 1
MRI parameters of original MRI images used in the RAD-IQMRI database.

Anatomical area	Sequence	Sequence type	TR (ms)	TE (ms)	Voxel size (mm)	FA (°)	ETL	Coil	NSA
Brain	T1	Plain Spin-Echo	650	15	0.72*0.72*5	69	N/A	SENSE-head-8 coil	2
	T2	N/A	4877	100	0.47*0.47*5	N/A	15	SENSE-head-8 coil	3
Liver	4D	Field echo	117	4.6	1.3*1.3*5	80	N/A	SENSE-torso-XL coil	2
Breast	T2	N/A	6107	120	0.74*0.74*3	N/A	25	SENSE-body coil	2
Fetus	PD	Single shot	N/A	140	0.9*0.9*4	N/A	N/A	SENSE-cardiac coil	N/A
Hip	T2	N/A	2760	60	0.31*0.31*3.5	20	20	SENSE-body coil	4
Knee	PD	N/A	5000	30	0.3*0.3*2.5	N/A	11	SENSE-knee coil	2
Spine	T2	N/A	3255	120	0.52*0.52*4	N/A	22	SENSE-spine coil	6

PD = Proton density; T1 = T1-weighted; T2 = T2-weighted; TR = Repetition Time; TE = Echo Time;
FA = Flip Angle; ETL = Echo Train Length; NSA = Number of Signal Averages.
4D = Four-Dimension, indicating an imaging technique where the fourth dimension is time, capturing the spatial structure as it evolves over time.

of radiologists in IQA research. In this paper, we advance the understanding of radiologists' perception of image quality by conducting a fully controlled psychovisual experiment. The quality of MRI images including eight pristine undistorted images and 112 distorted images was assessed by 13 radiologists. This study resulted in the creation of a new IQA database of MRI images, named **RAD-IQMRI**. In addition, we perform a comprehensive comparative study to evaluate the feasibility of popular IQA models designed for natural images in the medical image domain.

2. Related work

2.1. Medical image quality assessment databases

In recent years, significant progress has been made in establishing medical image quality assessment databases. For example, a Magnetic Resonance Imaging Quality Assessment (MRIQA) database was introduced by Chen et al. which encompasses 3809 images categorized into two distinct classes including high and low quality [15]. A quality assessment database based on MRI images of the brain was created by Narai et al. which includes images of 148 patients, categorized into three levels of quality, i.e., good, medium, bad [16]. An MRI image quality assessment database containing 635 images of six anatomical areas was developed by Lei et al. [17]. A breast-based MRI image quality assessment database including 2618 dichotomous images labeled by the presence or absence of artifacts was constructed by Kapsner et al. [18]. A 3D-MRI database based on the similar strategy is established by Pizarro et al. which includes 1457 images labeled according to their diagnostic usability [19]. Beyond MRI, other imaging modalities have also benefited from the development of specialized IQA databases. The "Chest-X-ray8" database introduced by Wang et al. serves as a hospital-scale chest X-ray database, providing a benchmark on classification and localization of common thorax diseases [20]. Furthermore, Zeng et al. developed a simple low-dose X-ray CT simulation method from high-dose scans, which aids in understanding and evaluating CT image quality [21]. Chen et al. introduced the Muiqa database for medical ultrasound images, which assesses image quality using a specialized algorithm [22]. The majority of these IQA databases used the Absolute Category Rating (ACR) scale for rating quality, which often fails to capture subtle differences in image quality. The ordinal nature of the ACR scale makes it challenging to apply certain statistical analyses; and treating ordinal data as internal data can lead to incorrect conclusions and reduced statistical power. Also, these IQA databases often involve a limited number of radiologists, reducing the clinical relevance of IQA ratings.

2.2. Image quality assessment algorithms

Over the past few decades, many image quality assessment (IQA) algorithms have been developed, which are tailored for predicting the quality of natural images as perceived by human viewers. These IQA algorithms are mainly classified into two categories including full-reference (FR) method and no-reference (NR) method. The FR IQA method requires both the original reference image and the distorted image for the evaluation of image quality. Methods such as Peak Signal-to-Noise Ratio (PSNR) [23] and Structural Similarity Index (SSIM) [24] are commonly used in the FR IQA framework, where the reference image provides a benchmark for identifying and quantifying distortions. These methods have been proven effective in accurately assessing image quality by comparing the test/distorted image against its undistorted counterpart. In contrast, the NR IQA method relies solely on the distorted image for quality prediction. Methods such as Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) and Natural Image Quality Evaluator (NIQE) are widely used in the NR IQA framework, as they assess the quality based on statistical features extracted from the test/distorted image itself without needing a reference. Research has shown that these methods can effectively predict perceived image quality by modeling natural scene statistics or learning from large datasets of distorted images. Compared to the FR IQA method, the NR IQA method has a wider range of applications, mainly due to the fact that a reference is often unavailable in many real-world scenarios. For example, in medical imaging distortions often stem from factors such as equipment limitations and patient movements [4]. Typically, the image data often consists of a single distorted image without an accompanying reference image.

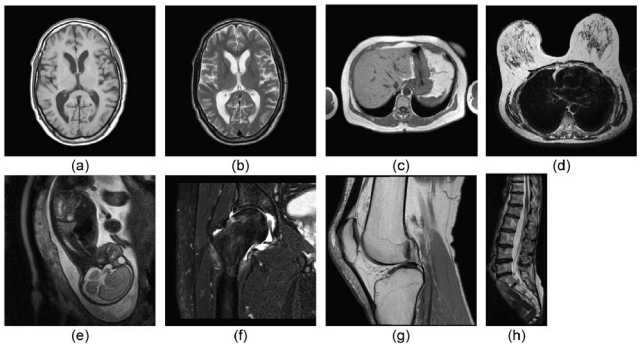


Fig. 1. Original MRI images used in our RAD-IQMRI database. The images are referred to (a) Brain_T1, (b) Brain_T2, (c) Liver, (d) Breast, (e) Fetus, (f) Hip, (g) Knee, and (h) Spine.

3. RAD-IQMRI: Subjective IQA database

3.1. Stimuli

Eight original MRI images, as shown in Fig. 1, were used in our study. Each image was acquired using a Philips Achieva 1.5T MRI system and is of high quality in terms of artifacts, signal-to-noise ratio and resolution. All source images were taken from patients without significant pathological conditions. The MRI parameters of these original images are shown in Table 1. Based on the eight original MR images,

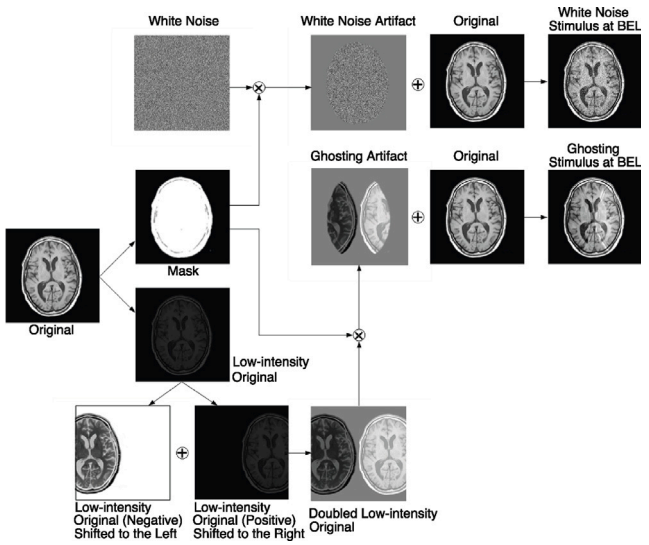


Fig. 2. Illustration of the simulation of ghosting and white noise in an MRI image [29].

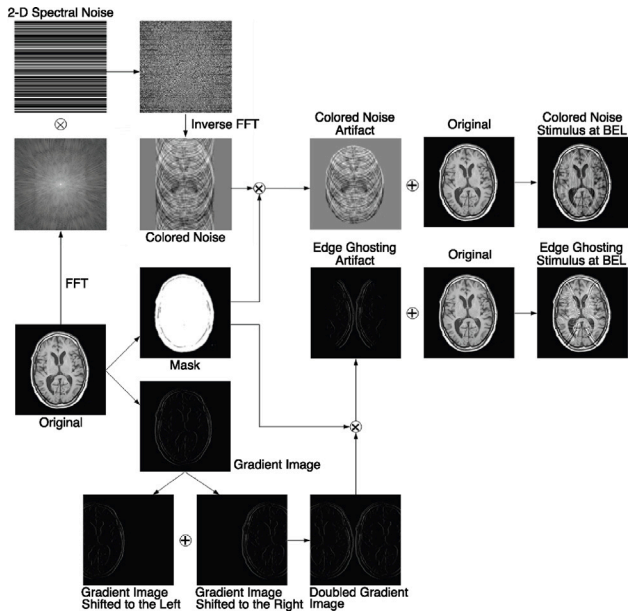


Fig. 3. Illustration of the simulation of edge ghosting and colored noise in an MRI image [29].

various distorted images were generated. More specifically, four types of artifacts [25–27] were simulated at two levels (low and high) of energy and linearly added to the original image content. The types of artifacts includes structured colored artifacts (plain ghosting), structured white artifacts (edge ghosting), unstructured colored artifacts (colored noise), and unstructured white artifacts (white noise). They represent common distortions encountered in clinical settings [28]. A benchmark energy level (BEL), i.e., the high-level energy, was defined by the amount of energy in a typical ghosting artifact for each original image and calculated as:

$$BEL = \sum_{i=1}^{M}\sum_{j=1}^{N} I_g(i,j)^2, \tag{1}$$

where M and N denote the pixel size of an original image (height and width), $I_g(i,j)$ denotes the intensity of the simulated ghosting artifact at pixel (i,j) ($i \in [1,M]$, $j \in [1,N]$). Based on the BEL, the low-level energy was determined by reducing the BEL with 80%. Note, all artifacts were applied to anatomical object areas rather than the background.

Ghosting: As shown in Fig. 2, for a given original image, a corresponding binary mask image (i.e., Mask) representing the anatomical object area, and a low-intensity image with 20% of the original image's intensity, were generated. Two displaced images were generated by shifting the low-intensity image by $1/3$ of its width: once leftward with negative intensity values and once rightward with positive intensity values, both relative to the original position of the anatomical object area. A new low-intensity image featuring a double copy of the anatomical area was obtained by superimposing these two displaced images. A ghosting artifact image (I_g) was generated by multiplying the new low-intensity image with the Mask pixel by pixel. A test stimulus with a high energy level of ghosting was produced by adding the ghosting artifact image to the original image.

White Noise: As shown in Fig. 2, a white noise artifact image was generated by multiplying an image containing additive white Gaussian noise, of the same size as the original image, with the Mask pixel by pixel and then scaling the intensity to achieve a total energy equal to the BEL. By combining the white noise artifact image with the original image, a test stimulus with a high energy level of white noise was generated.

Edge ghosting: The simulation of edge ghosting was similar to that of ghosting. As shown in Fig. 3, two images were generated based on the original image, including a Mask and a gradient image (GI). GI can be calculated as:

$$GI(i,j) = |I(i,j+1) - I(i,j)|, \quad j \in [1, N-1] \tag{2}$$

where i and j represent the row and column indices of a pixel within the image matrix I, traversing the image height and spanning from 1 to $N-1$ columns respectively, and N denotes the total number of columns in the image. The gradient image was shifted by $1/3$ of its width leftward and rightward, separately, based on the original position of the anatomical object area and then superimposed to yield a doubled gradient image. The edge ghosting artifact image was generated by first multiplying the doubled gradient image with the Mask, pixel by pixel, and then performing intensity scaling to achieve a total energy equal to the BEL. This artifact image was subsequently combined with the original image to create a test stimulus characterized by a high energy level of edge ghosting.

Colored noise: A Fourier transform was applied to the original image in the left–right direction, based on the original position of the anatomical object area (with constant values in the horizontal direction), and multiplied by a 2D spectrum with random values in the vertical direction. An inverse Fourier transform was subsequently applied to the result to generate an image of "colored noise". The colored noise artifact image was generated by multiplying the colored noise pattern and the Mask pixel by pixel and scaling the intensity to achieve a total energy equal to the BEL. A test stimulus with a high energy level of colored noise was produced by adding the colored noise artifact image to the original image.

Due to the effect of randomization of the 2D spectrum with random values in the vertical direction on the simulation of the colored noise image, four different versions of the 2D spectrum were used to yield four colored noise images. Therefore, this experiment contains 112 stimuli (i.e., 8 original images ×7 distortion versions ×2 energy levels) in total.

3.2. Psychovisual experiment

The subjective experiment used a simultaneous-double-stimulus (SDS) method, where subjects scored each test image (i.e., distorted image) on a continuous scale from 0 to 100 in the presence of a

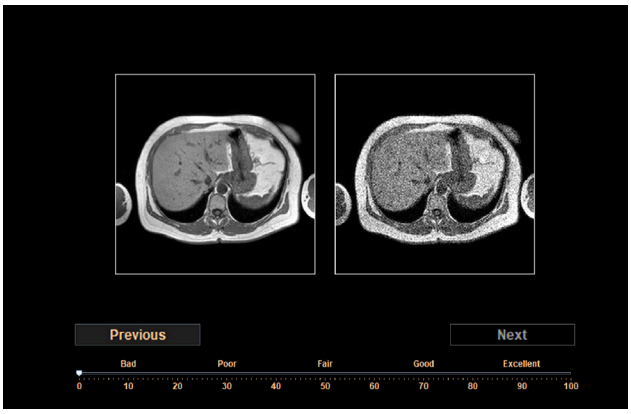


Fig. 4. Illustration of the scoring interface used in the psychovisual experiment. The interface presents two stimuli side-by-side, with the reference image on the left and the test image on the right.

reference image simultaneously [30]. The SDS method enables direct comparisons between reference and test images, reducing subjective variability and highlighting subtle differences in quality [30]. Thirteen radiologists from Angers University Hospital in France participated in this experiment. Before starting the experiment, each subject was given a written description of the procedure along with training instructions. First, a set of 10 images containing the same types of artifacts as those used in the actual experiment was shown to the subjects to familiarize them with the stimuli. Subsequently, six representative stimuli were presented one by one and each subject was asked to score them familiarize themselves with the scoring procedure. The images from the training phase were not included in the formal experiment. In the formal experiment, each test image was presented only once and in a random order. The subjective experiments were performed in a typical radiological reading room at the Angers University Hospital in France, with a consistent viewing environment ensured for all subjects. The images were displayed on a 24" wide-screen liquid-crystal monitor with a resolution of 1920 × 1200 pixels, calibrated to the Digital Imaging and Communications in Medicine (DICOM): Grayscale Standard Display Function (GSDF) standard [31–33]. The viewing distance was maintained at around 60 cm. No image adjustment (zoom, window level) was allowed. The scoring interface is shown in Fig. 4, with the reference image on the left and the test image on the right. The rating scale ranges from 0 to 100 and includes five semantic labels ('Bad', 'Poor', 'Fair', 'Good', 'Excellent') to assist in scoring. No time limit was imposed on the subjects for completing the experiment.

3.3. Processing of raw data

An outlier detection and subject rejection procedure was applied to the raw data prior to data analysis. An individual score was considered an outlier if it was more than two standard deviations from the mean score for that image [34]. For each subject, if twenty percent of all scores were outliers, the subject was excluded. Overall, none of the 13 subjects was excluded from the subsequent analysis and less than three percent of all image scores were excluded as outliers. In order to account for the differences in the use of the scoring scale between subjects, the raw scores were normalized using z-scores after applying the outlier removal and the subject rejection procedure:

$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}, \tag{3}$$

where μ_i and σ_i represent the mean and the standard deviation of all images scored by radiologist i, respectively. r_{ij} and z_{ij} represent the raw score and z score provided by the radiologist i for the image j, respectively. These z-scores were then linearly mapped to the interval [1, 10] for ease of interpretation. Next, the Mean Opinion Score (MOS) was calculated for each image:

$$\text{MOS}_j = \frac{1}{N}\sum_{i=1}^{N} z_{ij}, \tag{4}$$

where N represents the number of remaining radiologists in the group. MOS has been regarded as the benchmark of perceptual image quality measurement (IQA) [34]. This results in the creation of a new radiologist-rated subjective IQA database for MRI, named **RAD-IQMRI**.

To assess the validity of the obtained MOS, we quantify the differences in scores between individual subjects using the Pearson linear correlation coefficient (PLCC), i.e., calculating the PLCC between the MOS and each subject's scores. Fig. 5 shows the PLCC values for all subjects as well as the average PLCC value. The results indicate a high degree of consistency (i.e., with the majority of PLCC values around 0.8) in the subjects' scores for image quality.

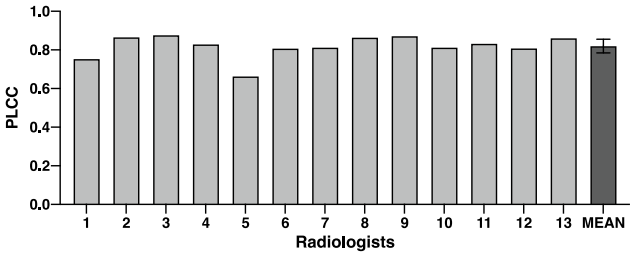


Fig. 5. Illustration of the correlation (i.e., PLCC) between MOS and each individual subject's scores. The right-most bar shows the mean correlation with a 95% confidence interval.

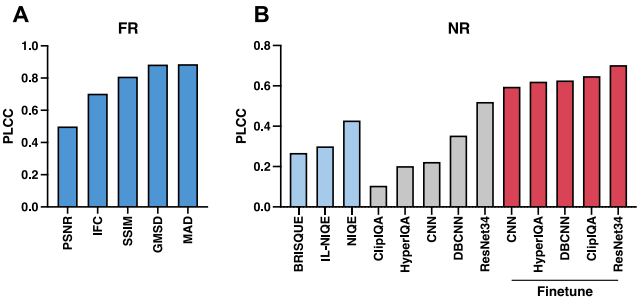


Fig. 6. Performance comparison of FR and NR IQA algorithms on our proposed RAD-IQMRI database. (A) Performance of FR methods. (B) Performance of NR methods including deep learning-based method without and with fine-tuning.

4. Benchmarking and evaluation of IQA algorithms on the RAD-IQMRI database

4.1. IQA algorithms

Based on our RAD-IQMRI database, we conduct thorough evaluations of several typical IQA algorithms to quantify their applicability and performance in assessing the perceptual quality of MRI images. Despite their widespread use in assessing natural image quality, these algorithms have not yet been applied to the medical imaging domain. To bridge this gap, we specifically benchmark the following popular IQA algorithms, including both FR and NR IQA methods, on our newly established RAD-IQMRI database.

The Full-reference (FR) IQA methods used in our study are:

- **Peak Signal-to-Noise Ratio** (PSNR) [35] is a measure of the difference between the reference and test images that is based on the Mean Squared Error (MSE) and it sets a baseline for the performance of objective IQA algorithms.

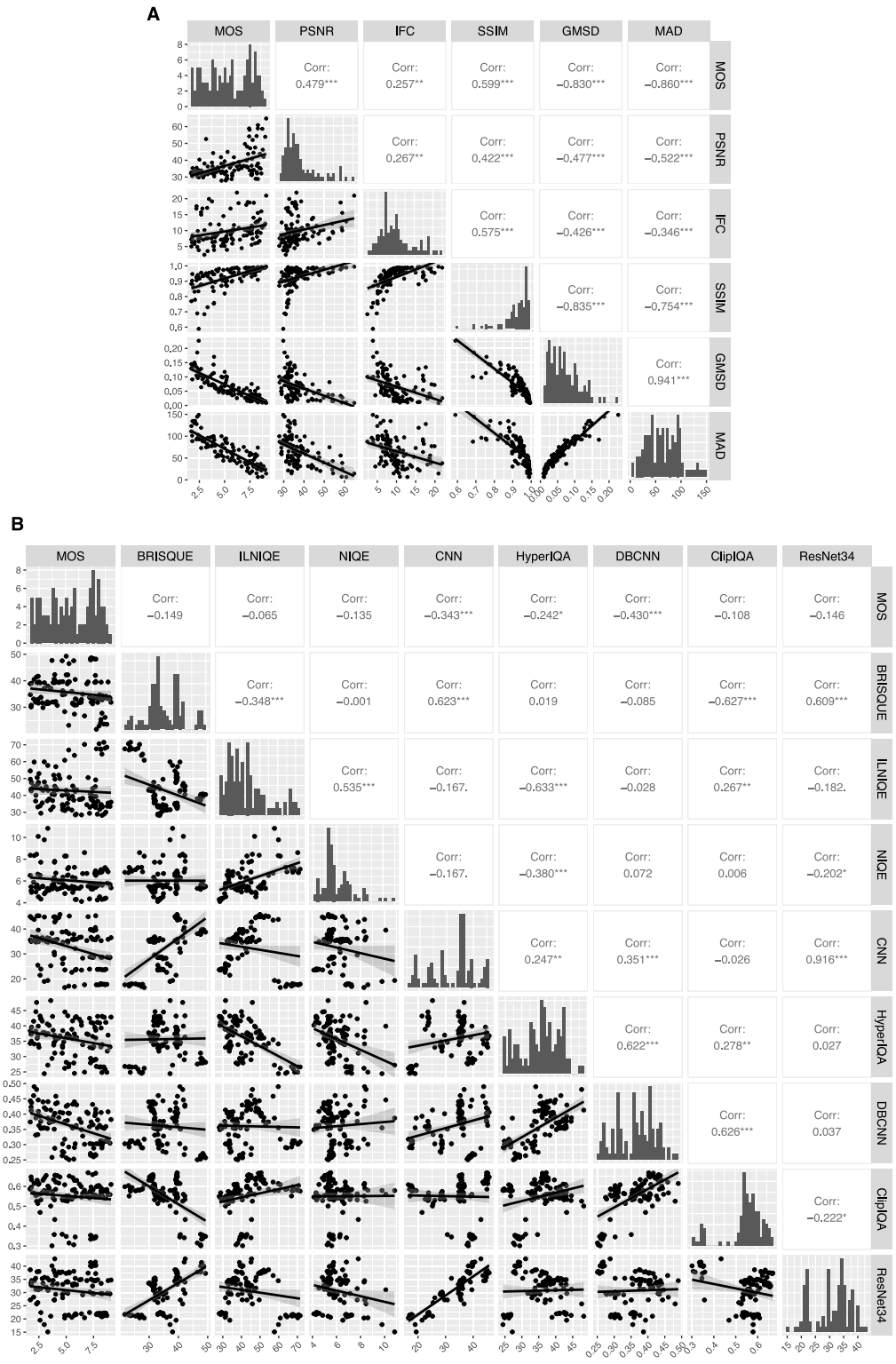


Fig. 7. Distribution of discrepancies between FR/NR IQA method predicted values and MOS. (A) Results of FR methods. (B) Results of NR methods. The histogram show the data distribution characteristics of MOS or an IQA method. The scatter plot show the relationship between two methods (i.e., MOS or an IQA method). Correlation coefficients and their significance levels (marked with an asterisk: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$) are displayed.

- **Information Fidelity Criterion** (IFC) [36] is an information-theoretic based FR IQA method, which is based on the Gaussian Scale Mixtures (GSM) model by quantifying the mutual information between the local wavelet coefficients of the reference image and the distorted image. The assessment of the image quality is achieved by summing the mutual information over all sub-bands.
- **Structural Similarity** (SSIM) [37] is a method for evaluating image quality through variations in image structural information, simulating the ability of the human visual system (HVS) to extract structural information from images. Firstly, using the calculation process of the Universal image Quality Index (UQI) [23] method, the local similarity score of the image is obtained by comparing

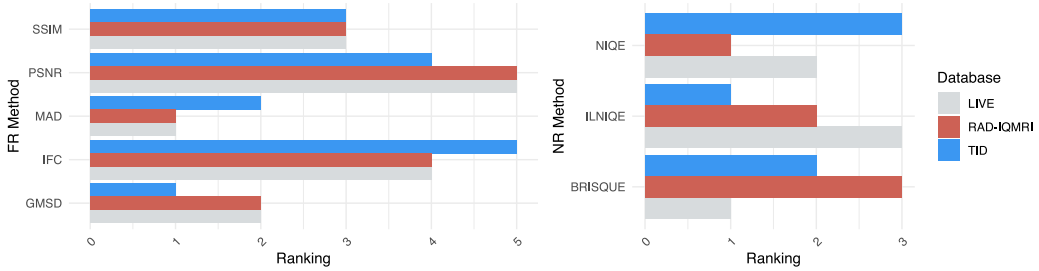


Fig. 8. Cross-database comparison of the PLCC Rankings of traditional IQA methods on our proposed RAD-IQMRI database and two widely used IQA databases for natual images (i.e., LIVE and TID).

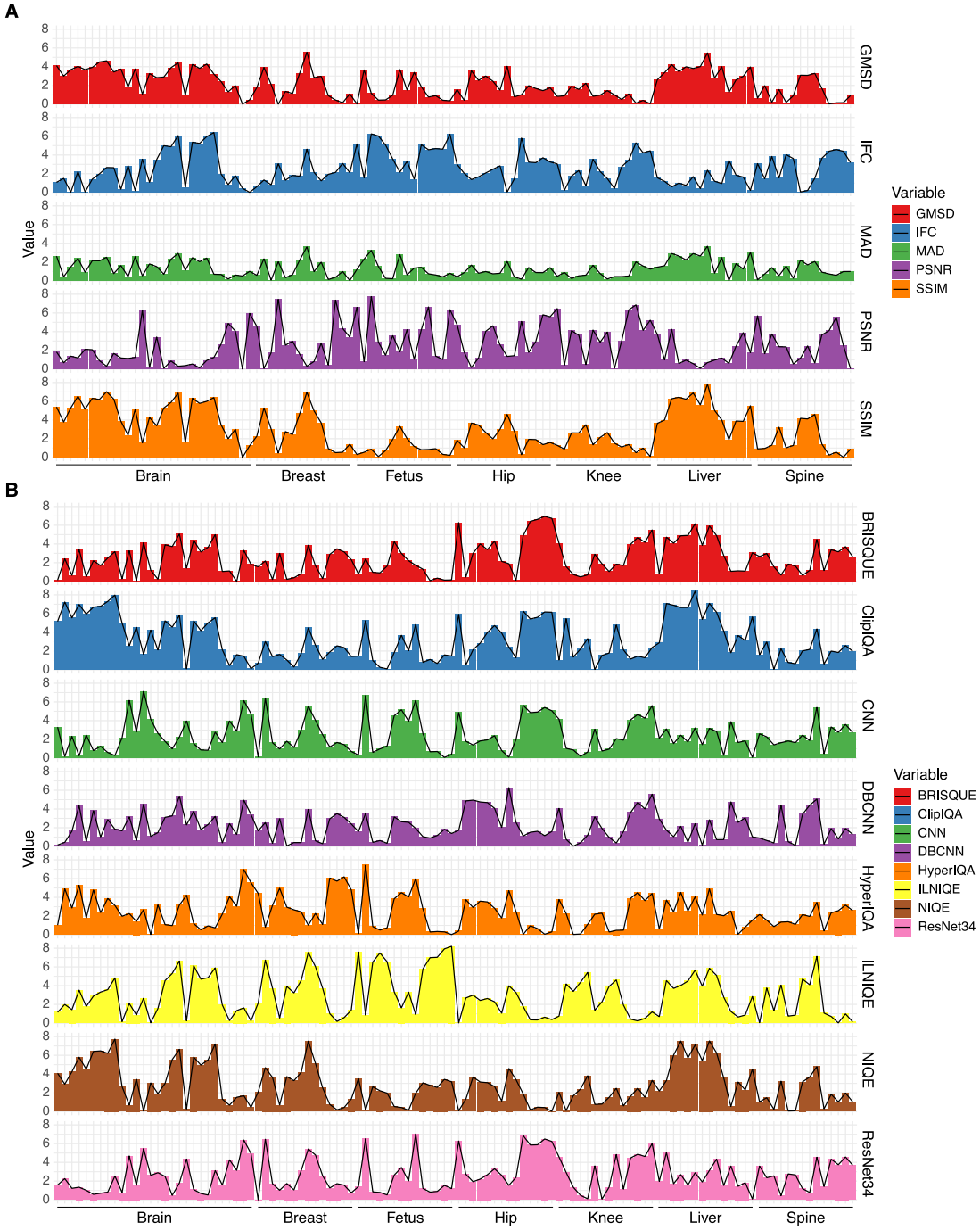


Fig. 9. Performance comparison of IQA algorithms on different anatomical sites (i.e., brain, breast, fetus, hip, knee, liver and spine), based on the residuals (i.e., absolute differences) between MOS and an FR/NR method predicted values. (A) Results of FR methods. (B) Results of NR methods.

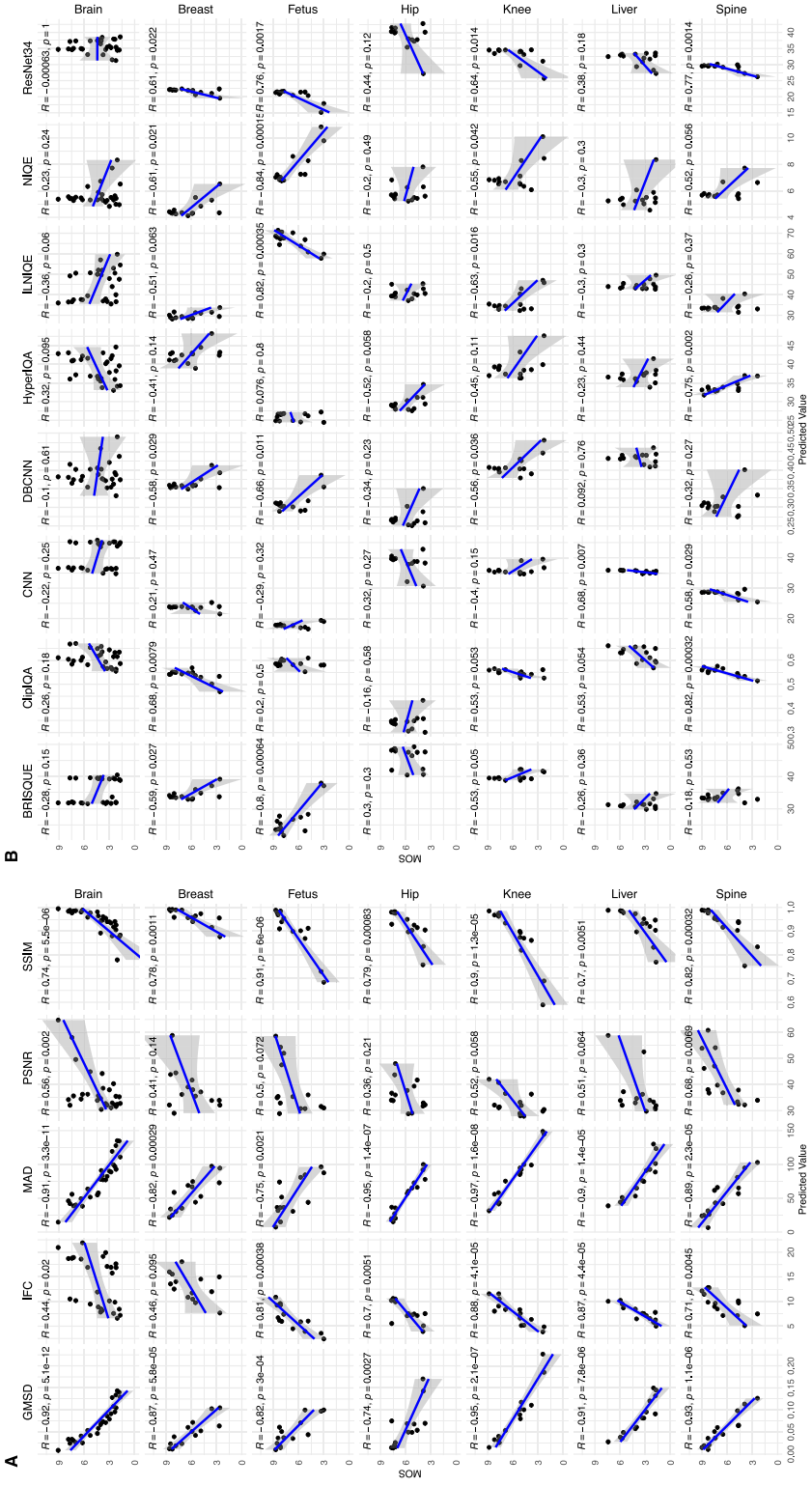


Fig. 10. Performance comparison of IQA algorithms for different anatomical sites (i.e., brain, breast, fetus, hip, knee, liver and spine), based on scatter plot of MOS versus predictions of an FR/NR method. (A) Results of FR methods. (B) Results of NR methods.

the local brightness information, local contrast information and local structure information between the reference image and the distorted image according to the multi-channel characteristic of HVS, and then the overall quality score of the distorted image is obtained using the Minkowski model.

- **Gradient Magnitude Similarity Deviation** (GMSD) [38] is also a structural similarity based FR IQA method, which discards the extra information and calculates the gradient similarity in different local structures to obtain local quality maps. It uses the standard deviation of the local quality maps as a pooling strategy to predict the overall image quality.
- **Most Apparent Distortion** (MAD) [24] is a mixed strategy based FR IQA method, which considers that the HVS places different emphasis under varying image quality conditions. It classifies

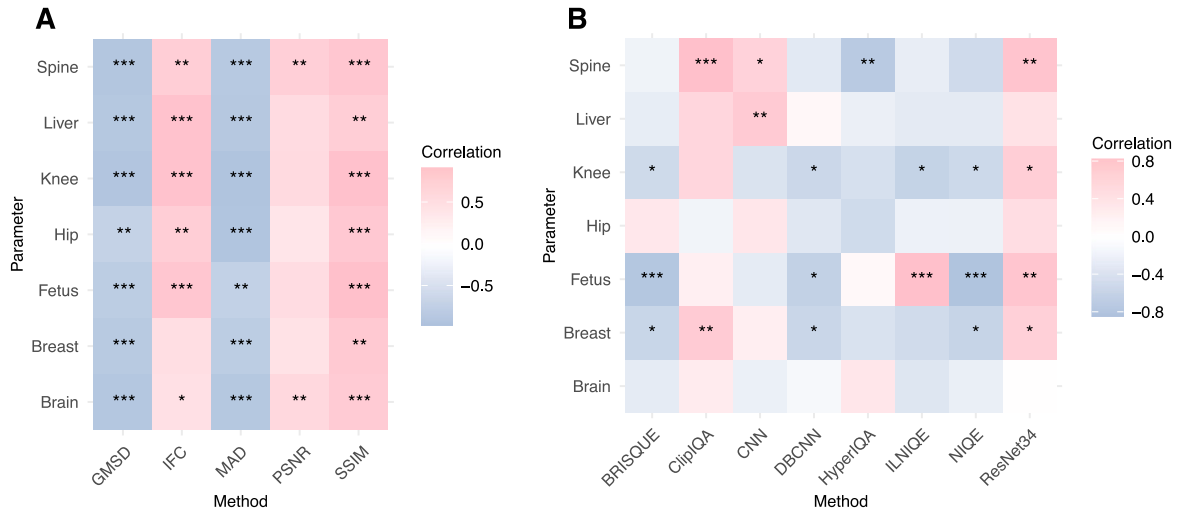


Fig. 11. Illustration of correlation strength (using a heatmap) between MOS and predictions of an FR/NR method across different for different anatomical image sites including brain, breast, fetus, hip, knee, liver and spine. (A) Results of FR methods. (B) Results of NR methods.

images into two categories: high quality and low quality. For high quality images, perceptual distortion is evaluated by considering contrast sensitivity, local luminance and contrast masking, while for low quality images, perceptual distortion is evaluated by the change in local statistics between the sub-bands of the reference and distorted images.

The No-reference (NR) IQA methods used in our study are:

- **Blind/Referenceless Image Spatial Quality Evaluator** (BRISQUE) [39] is a natural scene statistics (NSS) based NR IQA method that operates in the spatial domain. BRISQUE does not compute distortion-specific features. Instead, it uses the statistics of the locally normalized luminance coefficients of the scene to assess the potential loss of "naturalness" in the image caused by the presence of distortions, resulting in an image quality score. Other approaches based on the similar concept are CORNIA [40] and NIQE [41].
- Deep learning-based NR IQA algorithms: **Convolutional Neural Network** (CNN) [42] is used for NR IQA, which combines feature learning and regression in an end-to-end optimization process. In addition, we could replace the backbone network with **ResNet34** to produce another NR IQA method. Other deep learning based approaches are **DBCNN** [43]: a deep bilinear CNN-based NR model; **HyperIQA** [44]: a NR model that adaptively establishes perceptual rules; **ClipIQA** [45]: a text-image pair NR model.

In this paper, we distinguish between the IQA methods based on their underlying technologies, depending on how they leverage image features. Hence, we refer to IQA algorithms that do not utilize deep learning as traditional methods, while those that employ deep learning are referred to as learning-based methods.

4.2. Evaluation metrics

The metric that is used to quantify the prediction accuracy of an objective IQA algorithm is Pearson linear correlation coefficient (PLCC):

$$PLCC = \frac{\sum_{i=1}^{N}(p_i - \bar{p})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{N}(p_i - \bar{p})^2(s_i - \bar{s})^2}} \tag{5}$$

where p_i and s_i are values of subjective and objective measures, respectively, and $\bar{p}$ and $\bar{s}$ are the mean values, while N is the number of images in the test database.

The metric that is used to quantify the prediction monotonicity of an objective IQA algorithm is Spearman rank-order correlation coefficient (SROCC):

$$SROCC(Q, S) = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{6}$$

where d_i is the difference between the ith image's ranks in the objective (Q) and subjective (S) scores, while N is the number of images in the test database.

4.3. Performance evaluation

4.3.1. Overall performance

The performance results of all IQA algorithms, including FR and NR methods, are shown in Fig. 6. Since deep learning models typically require substantial datasets for training, and medical image data is often not as abundant as natural image data, we adopt a two-stage evaluation process for deep learning-based IQA algorithms. In the #1 phase, we pre-train a deep learning-based IQA model on a large number of natural images, and test the resulting model on our RAD-IQMRI database. In the #2 phase, the pre-trained IQA model is fine-tuned on our RAD-IQMRI database via transfer learning, and is tested on the same database.

From the results, it is observed that FR IQA methods generally outperform NR methods. This is attributed to the utilization of reference in the FR framework, which provides additional context for assessing image quality. Among the FR IQA methods, SSIM and GMSD based on structural similarity and MAD based on hybrid strategies show better performance compared to PSNR based on errors and IFC based on information theory. The deep learning-based IQA methods, which are also part of the NR approach, show improved performance after fine-tuning on medical images.

Furthermore, we found some key trends and patterns of these IQA algorithms' performance, as shown in Fig. 7. First, the histogram of the MOS data appears to be closer to a uniform distribution. The distribution of images in many existing IQA databases tends to be normal [2], which has certain drawbacks. Specifically, there is a higher concentration of images with mid-range quality, while fewer images fall into the low- and high-quality extremes. This imbalance can affect the accuracy as well as the robustness of the assessment methods. In contrast, our RAD-IQMRI database features an even distribution of images across low to high quality. This balanced distribution is more conducive to the development of robust and accurate IQA algorithms.

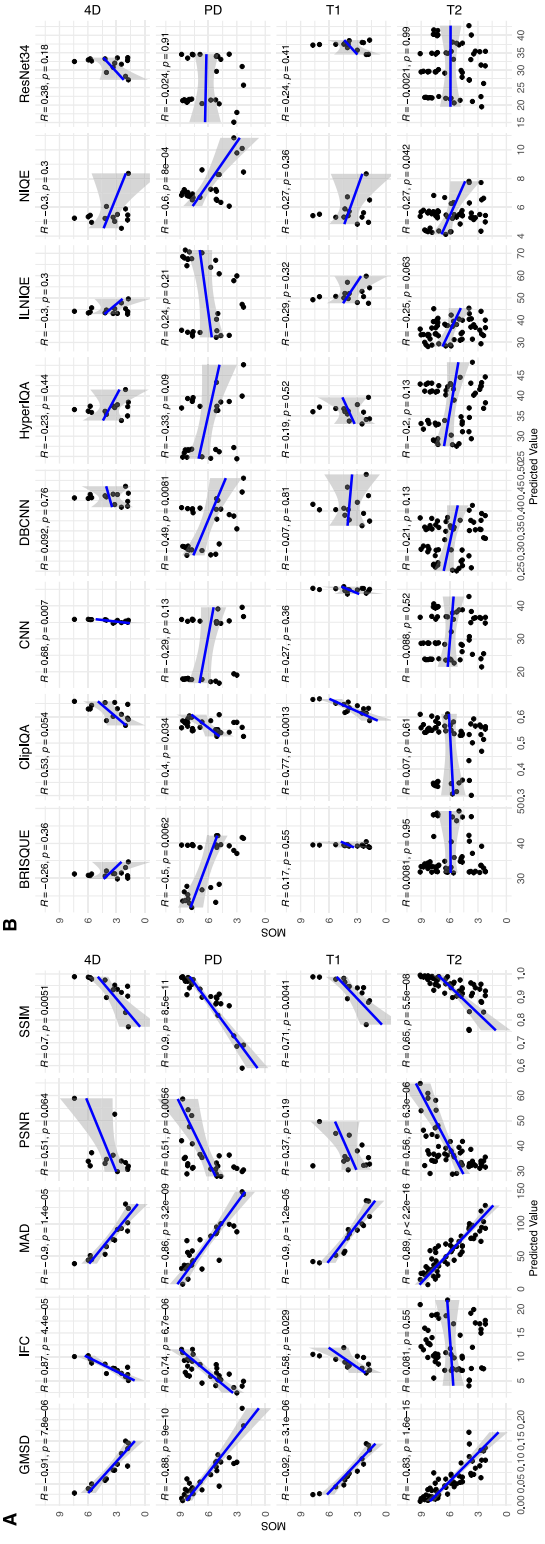


Fig. 12. Performance comparison of IQA algorithms for different imaging parameters (i.e., 4D, PD, T1 and T2), based on scatter plot of MOS versus predictions of an FR/NR method. (A) Results of FR methods. (B) Results of NR methods.

Fig. 7(A) shows that the distribution of MAD closely resembles the distribution of MOS. This similarity in shape suggests that the MAD scores align well with the subjective scores in terms of statistical properties. As illustrated in the scatter plots in Fig. 7(A), both MAD and GMSD demonstrate a significant linear relationship with the subjective scores (MOS). This indicates their sensitivity and ability to accurately reflect changes in image quality, showing a high degree of consistency with the subjective scores. Among all the IQA algorithms examined, MAD exhibits the most significant correlation with MOS (Corr: −0.860***). This strong correlation confirms the reliability and validity of MAD in image quality assessment. For the NR IQA algorithms, as illustrated in Fig. 7(B), we found a lack of a strong linear correlation with MOS. *Note the #1 phase of NR IQA evaluation is used hereafter to ensure a fair comparison between the FR and NR methods, where the entire RAD-IQMRI database is used as the test set.* This indicates that NR methods, in their current form, are limited in their ability to directly predict image quality of medical images and require further improvement and development. These results provide an empirical basis for further optimization and development of more suitable IQA models for medical imaging.

4.3.2. Cross-database comparison

We conduct a cross-database comparison of the performance of FR and NR IQA methods to investigate whether these methods perform consistently on our RAD-IQMRI database compared to other widely used natural IQA databases, such as LIVE [11] and TID [12]. As shown in Fig. 8, the performance rankings of the FR IQA methods show consistency cross the three IQA databases including LIVE, TID, and the proposed RAD-IQMRI. In particular, MAD and GMSD consistently rank as the top models across all three databases. This consistency implies the robustness of these methods in aligning with subjective quality assessment. In contrast, the rankings of NR IQA methods exhibit greater variation across databases. This variability suggests a heightened sensitivity of NR methods to particular content types and distortion categories. Certain NR methods may excel in assessing specific image types or distortions but under-perform with others. This performance dependency underscores the influence of content specificity and highlights potential constraints in algorithm adaptability. These findings emphasize the importance of considering a comprehensive assessment of algorithm adaptation in different IQA application contexts.

4.3.3. Performance on different anatomical sites

Ensuring consistent quality assessment results across imaging of diverse anatomical sites is a critical performance indicator for IQA algorithms. This reliability and applicability are essential for their effective use in various medical diagnostic contexts. To evaluate the performance of IQA algorithms across various anatomical sites, including brain, breast, fetus, hip, knee, liver, and spine, we calculate the residuals (i.e., absolute differences) between the MOS and the predicted scores of each IQA algorithm. The results are illustrated in Fig. 9. It can be seen that MAD and GMSD demonstrate a high degree of consistency with the subjective scores (MOS) across different anatomical regions. Specifically, MAD exhibits the highest agreement, showing minimized variability from MOS across different anatomical regions, thus demonstrating robust scoring stability. This consistency highlights MAD's capability to reliably assess image quality across diverse content types, highlighting its significant practical value. The NR IQA methods consistently under-perform across different anatomical regions Compared to the FR IQA methods. This may be due to the inherent difficulty of accurately capturing comprehensive information about image quality degradation in the absence of a reference image. Moreover, certain methods show high reliability in the prediction of specific anatomical regions. For example, GMSD shows consistent performance in assessing quality of fetal, hip, knee, and spine images; IFC exhibits strong consistency in evaluating quality of liver images; PSNR proves reliable in assessing quality of brain and liver images, despite some instability in other anatomical regions; and SSIM demonstrates consistency in predicting quality of fetal, hip, and knee images.

In addition, as illustrated in Figs. 10 and 11, FR IQA methods not only score each anatomical site consistently, but also maintain a

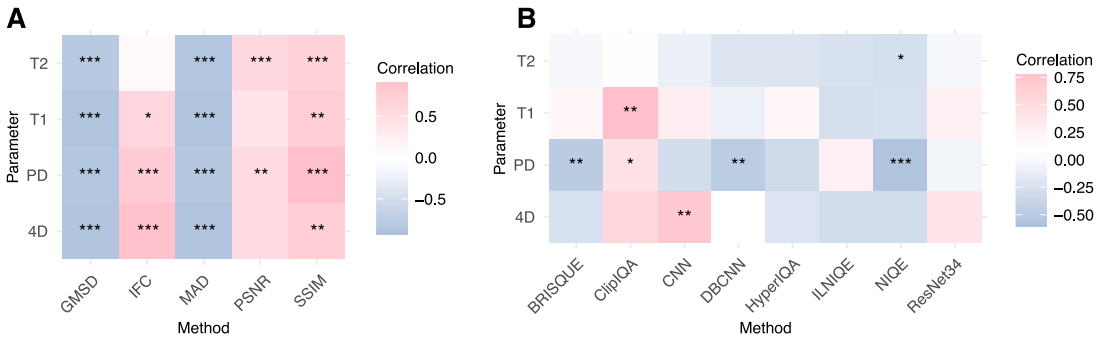


Fig. 13. Illustration of correlation strength (using a heatmap) between MOS and predictions of an FR/NR method for different imaging parameters including 4D, PD, T1 and T2. (A) Results of FR methods. (B) Results of NR methods.

consistent scoring pattern across the entire dataset. This suggests that FR methods can provide stable quality scores for similar image content, regardless of minor differences in the images. In contrast, NR IQA methods exhibit significant variability in scores for different anatomical sites, suggesting higher sensitivity to varying image characteristics. For example, while some NR methods may consistently score the quality of spine images, they may show considerable fluctuations when assessing the quality of brain or liver images. These variations may be attributed to the fact that NR methods rely on the intrinsic properties of the image, leading to inconsistent results due to visual differences across different anatomical regions.

Our results emphasize the importance of selecting IQA methods that are appropriate for specific anatomical image sites. For example, certain FR methods may deliver more reliable scores for fetal or hip images, while a different set of methods might be better suited for brain or liver images. This targeted approach allows us to more accurately model the image quality assessment process in a clinical setting, hereby enhancing the precision of image quality control in the field of medical imaging.

4.3.4. Performance on different imaging parameters

Our analysis of image quality assessment (IQA) algorithms reveals distinct performance characteristics when evaluating different imaging parameters, including 4D, PD, T1 and T2. As shown in Figs. 12 and 13, FR IQA methods exhibit a uniform range of scores for each parameter, suggesting consistent application regardless of parameter-specific image characteristics. This consistency in scoring suggests a degree of robustness of these methods, which may be attributed to their utilization of reference images for evaluation. In contrast, NR IQA methods show greater variability in scores when applied to different parameters, highlighting their potential sensitivity to the unique features of each parameter. This increased variability may reflect the underlying models of NR methods, which are potentially more finely tuned or responsive to artifacts specific to certain image quality aspects or unique to particular imaging parameters.

4.3.5. Discussion

Overall, the results show that extending IQA methods from the natural image domain to the medical image domain is feasible and has a certain room for improvement. While traditional FR methods generally outperform traditional NR methods and deep learning-based methods without fine-tuning, deep learning-based methods that have been fine-tuned on medical images can achieve performance on par with FR methods. This demonstrates the technical feasibility and potential of deep learning-based IQA models in the medical imaging field. In a real clinical setting, the integration of IQA methods can offer several clinical benefits. For example, high-quality medical images are critical for accurate diagnosis; and IQA models can ensure superior image quality, leading to more reliable diagnostic outcomes. Automated IQA can streamline the workflow in radiology departments, reducing the time radiologists spend on assessing image quality, leading to faster diagnosis and treatment planning. The integration of automated IQA in healthcare facilities can reduce the need for repeat scans due to poor image quality, optimizing resource utilization and reducing patient exposure to additional radiation, lowering operational costs.

5. Conclusions

In this study, we established a novel benchmark for assessing the quality of MRI images. A fully-controlled psychovisual experiment was undertaken to construct the RAD-IQMRI database, in which 13 radiologists assessed the quality of MRI images of varying quality. The proposed database comprises eight distinct undistorted MRI images, along with 112 associated distorted images that span a range of perceived quality levels. We conducted a comprehensive evaluation to benchmark the effectiveness of a series of IQA methods using the RAD-IQMRI database, thereby demonstrating the viability of adapting IQA methods from the natural image domain to the medical domain. The RAD-IQMRI database not only establishes a baseline for quality assessment in the radiological context but also serves as a foundation for further refinement and application of these methods in a clinical setting. Recognizing the limitations of our current database, which relies on simulated rather than real clinical artifacts and is constrained by the unique challenges of medical image data resulting in a smaller scale, future work will aim to develop a more expansive and clinically representative database that includes pathological information to accommodate diverse research objectives and improve diagnostic processes.

CRediT authorship contribution statement

Yueran Ma: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Jianxun Lou:** Writing – original draft, Validation, Software, Methodology. **Jean-Yves Tanguy:** Resources, Investigation, Data curation. **Padraig Corcoran:** Writing – review & editing, Validation, Supervision, Formal analysis. **Hantao Liu:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Y. Wang, Principles of Magnetic Resonance Imaging: Physics Concepts, Pulse Sequences, & Biomedical Applications, CreateSpace Independent Publishing, 2014.

[2] H. Khalid, M. Hussain, M.A. Al Ghamdi, T. Khalid, K. Khalid, M.A. Khan, K. Fatima, K. Masood, S.H. Almotiri, M.S. Farooq, et al., A comparative systematic literature review on knee bone reports from mri, X-rays and CT scans using deep learning and machine learning methodologies, Diagnostics 10 (8) (2020) 518.

[3] J.J. Ma, U. Nakarmi, C.Y.S. Kin, C.M. Sandino, J.Y. Cheng, A.B. Syed, P. Wei, J.M. Pauly, S.S. Vasanawala, Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges, in: 2020 IEEE 17th International Symposium on Biomedical Imaging, ISBI, IEEE, 2020, pp. 337–340.

[4] E.A. Krupinski, Current perspectives in medical image perception, Attention Percept. Psychophys. 72 (5) (2010) 1205–1217.

[5] P. Suetens, Fundamentals of Medical Imaging, Cambridge University Press, 2017.

[6] S. Vaishali, K.K. Rao, G.S. Rao, A review on noise reduction methods for brain MRI images, in: 2015 International Conference on Signal Processing and Communication Engineering Systems, IEEE, 2015, pp. 363–365.

[7] M. Zaitsev, J. Maclaren, M. Herbst, Motion artifacts in MRI: A complex problem with many partial solutions, J. Magn. Reson. Imaging 42 (4) (2015) 887–901.

[8] Y. Hirokawa, H. Isoda, Y.S. Maetani, S. Arizono, K. Shimada, K. Togashi, MRI artifact reduction and quality improvement in the upper abdomen with PROPELLER and prospective acquisition correction (PACE) technique, Am. J. Roentgenol. 191 (4) (2008) 1154–1158.

[9] M.D. Abràmoff, M.K. Garvin, M. Sonka, Retinal imaging and image analysis, IEEE Rev. Biomed. Eng. 3 (2010) 169–208.

[10] K. Ma, Y. Fang, Image quality assessment in the modern age, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5664–5666.

[11] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (2006) 3440–3451.

[12] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al., Image database TID2013: Peculiarities, results and perspectives, Signal Process.: Image Commun. 30 (2015) 57–77.

[13] X. Liu, M. Pedersen, J.Y. Hardeberg, CID: IQ–A new image quality database, in: Image and Signal Processing: 6th International Conference, ICISP 2014, Cherbourg, France, June 30–July 2, 2014. Proceedings 6, Springer, 2014, pp. 193–202.

[14] M. Outtas, L. Zhang, O. Deforges, W. Hammidouche, A. Serir, C. Cavaro-Ménard, A study on the usability of opinion-unaware no-reference natural image quality metrics in the context of medical images, in: 2016 International Symposium on Signal, Image, Video and Communications, ISIVC, IEEE, 2016, pp. 308–313.

[15] Q. Chen, F. Liu, H. Duan, Y. Wang, X. Min, Y. Zhou, G. Zhai, MRIQA: Subjective method and objective model for magnetic resonance image quality assessment, in: 2022 IEEE International Conference on Visual Communications and Image Processing, VCIP, IEEE, 2022, pp. 1–5.

[16] Á. Nárai, P. Hermann, T. Auer, P. Kemenczky, J. Szalma, I. Homolya, E. Somogyi, P. Vakli, B. Weiss, Z. Vidnyánszky, Movement-related artefacts (MR-ART) dataset of matched motion-corrupted and clean structural MRI brain scans, Sci. Data 9 (1) (2022) 630.

[17] K. Lei, A.B. Syed, X. Zhu, J.M. Pauly, S.S. Vasanawala, Artifact-and content-specific quality assessment for MRI with image rulers, Med. Image Anal. 77 (2022) 102344.

[18] L.A. Kapsner, E.L. Balbach, L. Folle, F.B. Laun, A.M. Nagel, A. Liebert, J. Emons, S. Ohlmeyer, M. Uder, E. Wenkel, et al., Image quality assessment using deep learning in high b-value diffusion-weighted breast MRI, Sci. Rep. 13 (1) (2023) 10549.

[19] R.A. Pizarro, X. Cheng, A. Barnett, H. Lemaitre, B.A. Verchinski, A.L. Goldman, E. Xiao, Q. Luo, K.F. Berman, J.H. Callicott, et al., Automated quality assessment of structural magnetic resonance brain images based on a supervised machine learning algorithm, Front. Neuroinform. 10 (2016) 52.

[20] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.

[21] D. Zeng, J. Huang, Z. Bian, S. Niu, H. Zhang, Q. Feng, Z. Liang, J. Ma, A simple low-dose X-ray CT simulation from high-dose scan, IEEE Trans. Nucl. Sci. 62 (5) (2015) 2226–2233.

[22] Q. Chen, X. Min, H. Duan, Y. Zhu, G. Zhai, Muiqa: Image quality assessment database and algorithm for medical ultrasound images, in: 2021 IEEE International Conference on Image Processing, ICIP, IEEE, 2021, pp. 2958–2962.

[23] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (3) (2002) 81–84.

[24] E.C. Larson, D.M. Chandler, Most apparent distortion: Full-reference image quality assessment and the role of strategy, J. Electron. Imaging 19 (1) (2010) 011006.

[25] E. Pusey, R.B. Lufkin, R. Brown, M.A. Solomon, D.D. Stark, R. Tarr, W. Hanafee, Magnetic resonance imaging artifacts: Mechanism and clinical significance, Radiographics 6 (5) (1986) 891–911.

[26] J.A. Clark II, W.M. Kelly, Common artifacts encountered in magnetic resonance imaging, Radiol. Clin. North Am. 26 (5) (1988) 893–920.

[27] C.E. Willis, S.K. Thompson, S.J. Shepard, Artifacts and misadventures in digital radiography, Appl. Radiol. 33 (1) (2004) 11.

[28] E.M. Bellon, E.M. Haacke, P.E. Coleman, D.C. Sacco, D.A. Steiger, R.E. Gangarosa, MR artifacts: A review, Am. J. Roentgenol. 147 (6) (1986) 1271–1281.

[29] H. Liu, J. Koonen, M. Fuderer, I. Heynderickx, The relative impact of ghosting and noise on the perceived quality of MR images, IEEE Trans. Image Process. 25 (7) (2016) 3087–3098.

[30] R. BT, Methodology for the subjective assessment of the quality of television pictures, Int. Telecommun. Union 4 (2002).

[31] E. Samei, A. Badano, D. Chakraborty, K. Compton, C. Cornelius, K. Corrigan, M.J. Flynn, B. Hemminger, N. Hangiandreou, J. Johnson, et al., Assessment of display performance for medical imaging systems: executive summary of AAPM TG18 report, Med. Phys. 32 (4) (2005) 1205–1225.

[32] B.M. Hemminger, R.E. Johnston, J.P. Rolland, K.E. Muller, Introduction to perceptual linearization of video display systems for medical image presentation, J. Digit. Imaging 8 (1) (1995) 21–34.

[33] V. Rosslyn, Digital imaging and communications in medicine (DICOM) part 14: Gray scale standard display function, Medicine (Baltimore) 10 (S1) (2004) 3–4.

[34] Z. Wang, A.C. Bovik, Modern Image Quality Assessment (Ph.D. thesis), Springer, 2006.

[35] Z. Wang, A.C. Bovik, Mean squared error: Love it or leave it? A new look at signal fidelity measures, IEEE Signal Process. Mag. 26 (1) (2009) 98–117.

[36] H.R. Sheikh, A.C. Bovik, G. De Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (12) (2005) 2117–2128.

[37] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[38] W. Xue, L. Zhang, X. Mou, A.C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE Trans. Image Process. 23 (2) (2013) 684–695.

[39] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.

[40] P. Ye, J. Kumar, L. Kang, D. Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1098–1105.

[41] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Process. 24 (8) (2015) 2579–2591.

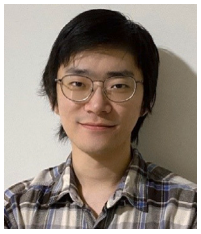
[42] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1733–1740.

[43] W. Zhang, K. Ma, J. Yan, D. Deng, Z. Wang, Blind image quality assessment using a deep bilinear convolutional neural network, IEEE Trans. Circuits Syst. Video Technol. 30 (1) (2018) 36–47.

[44] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, Y. Zhang, Blindly assess image quality in the wild guided by a self-adaptive hyper network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3667–3676.

[45] J. Wang, K.C. Chan, C.C. Loy, Exploring clip for assessing the look and feel of images, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, (no. 2) 2023, pp. 2555–2563.

Yueran Ma received the B.Eng. from Beijing Jiaotong University, in 2016 and the M.S. degree from Southern Methodist University in 2018. He is now pursuing his Ph.D. degree at the School of Computer Science and Informatics, Cardiff University, Cardiff, UK. His interests are Image Processing, Biomedical Image Processing, Image Quality Assessment and Saliency Prediction.



Jianxun Lou received the B.Eng. degree from Central South University, Changsha, China, in 2018, and the M.S. degree in 2020 from Cardiff University, Cardiff, U.K., where he is currently working toward the Ph.D. degree with the School of Computer Science and Informatics.

Jean-Yves Tanguy, M.D., is a Neuroradiologist and Head and Neck imaging specialist at the University Hospital Center in Angers, France. He has given lessons on technical aspects of medical imaging to medicine students, and future radiologists, technicians, and engineers in Angers Faculty of Medicine, and ESEO since the beginning of his career.

Padraig Corcoran is a Reader and the Director of Research in the School of Computer Science and Informatics at Cardiff University, UK. His research interests are in the fields of network science and operations research.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently a Professor at the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. His research interests sit at the intersection of Image Processing, Machine Learning, Computer Vision, Applied Perception, and Medical Imaging.