Medical Image Analysis 99 (2025) 103343



Contents lists available at ScienceDirect

Medical Image Analysis

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





Low-dose computed tomography perceptual image quality assessment

Wonkyeong Lee [a], Fabian Wagner [b], Adrian Galdran [c], Yongyi Shi [d], Wenjun Xia [d], Ge Wang [d], Xuanqin Mou [e], Md. Atik Ahamed [f], Abdullah Al Zubaer Imran [f], Ji Eun Oh [g], Kyungsang Kim [h], Jong Tak Baek [g], Dongheon Lee [g], Boohwi Hong [g], Philip Tempelman [i], Donghang Lyu [j], Adrian Kuiper [i], Lars van Blokland [i], Maria Baldeon Calisto [k], Scott Hsieh [l], Minah Han [m], Jongduk Baek [m], Andreas Maier [b], Adam Wang [n], Garry Evan Gold [n], Jang-Hwan Choi [a,o,*]

[a] *Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Republic of Korea*
[b] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Schloßplatz 4, Erlangen 91054, Germany*
[c] *Universitat Pompeu Fabra, Plaça de la Mercè, 12, Ciutat Vella, Barcelona 08002, Spain*
[d] *Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180, USA*
[e] *Xi'an Jiaotong University, 28, Xianning West Road, Xi'an City, Shaanxi Province 710049, People's Republic of China*
[f] *Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA*
[g] *Chungnam National University College of Medicine, 266 Munghwa-ro, Daejeon 35015, Republic of Korea*
[h] *MGH and Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA*
[i] *Delft University of Technology, Mekelweg 5, CD Delft 2628, Netherlands*
[j] *Leiden University, Rapenburg 70, EZ Leiden 2311, Netherlands*
[k] *Universidad San Francisco de Quito, Campus Cumbayá, Diego de Robles s/n, Quito 170901, Ecuador*
[l] *Mayo Clinic, 200 First St., SW Rochester, MN 55905, USA*
[m] *Yonsei University, A50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea*
[n] *Stanford University, 450 Jane Stanford Way, Stanford, CA, 94305, USA*
[o] *Computational Medicine, Graduate Program in System Health Science and Engineering, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Republic of Korea*

ARTICLE INFO

Dataset link: Low-dose Computed Tomography Perceptual Image Quality Assessment Grand Challenge Dataset (MICCAI 2023) (Original data)

Keywords:
Image quality assessment (IQA)
Medical IQA challenge
Computed tomography (CT) Imaging
No-reference Image quality metric
Artifacts in CT imaging
Open-access benchmark dataset

ABSTRACT

In computed tomography (CT) imaging, optimizing the balance between radiation dose and image quality is crucial due to the potentially harmful effects of radiation on patients. Although subjective assessments by radiologists are considered the gold standard in medical imaging, these evaluations can be time-consuming and costly. Thus, objective methods, such as the peak signal-to-noise ratio and structural similarity index measure, are often employed as alternatives. However, these metrics, initially developed for natural images, may not fully encapsulate the radiologists' assessment process. Consequently, interest in developing deep learning-based image quality assessment (IQA) methods that more closely align with radiologists' perceptions is growing. A significant barrier to this development has been the absence of open-source datasets and benchmark models specific to CT IQA. Addressing these challenges, we organized the Low-dose Computed Tomography Perceptual Image Quality Assessment Challenge in conjunction with the Medical Image Computing and Computer Assisted Intervention 2023. This event introduced the first open-source CT IQA dataset, consisting of 1,000 CT images of various quality, annotated with radiologists' assessment scores. As a benchmark, this challenge offers a comprehensive analysis of six submitted methods, providing valuable insight into their performance. This paper presents a summary of these methods and insights. This challenge underscores the potential for developing no-reference IQA methods that could exceed the capabilities of full-reference IQA methods, making a significant contribution to the research community with this novel dataset. The dataset is accessible at https://zenodo.org/records/7833096.

* Corresponding author at: Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Republic of Korea.
E-mail addresses: ewonkyong@ewhain.net (W. Lee), choij@ewha.ac.kr (J.-H. Choi).

https://doi.org/10.1016/j.media.2024.103343
Received 14 February 2024; Received in revised form 30 August 2024; Accepted 3 September 2024
Available online 6 September 2024
1361-8415/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Computed tomography (CT) imaging is a fundamental diagnostic tool in patient care. In 2012, approximately 50% of the average radiation exposure to the population resulted from medical procedures, with CT examinations contributing around 25% of this exposure (Boone et al., 2012). This scenario highlights a growing concern regarding the radiation doses patients receive, as such doses of ionizing radiation can have deleterious effects, including the potential to cause cancer. Therefore, to align with the aim of minimizing radiation to the lowest level reasonably achievable in CT imaging (Sodickson et al., 2009), researchers are interested in developing image reconstruction techniques, such as denoising algorithms (Chen et al., 2017; Li et al., 2020; Kim et al., 2023; Wu et al., 2023). The goal is to produce images of acceptable quality with the least possible radiation.

The increased focus on image quality in medical imaging has made adopting an effective image quality assessment (IQA) metric for CT images imperative. Currently, the gold standard for assessing the image quality of medical images is subjective assessment by radiologists (Chow et al., 2016). This approach is based on the premise that the quality of medical images is related to their diagnostic contribution. Consequently, as the primary users of these images for patient diagnosis, radiologists are best positioned to evaluate their quality based on this criterion. However, radiologists' evaluations can be costly, time-consuming, and subject to individual bias, leading to variability in evaluations between practitioners (Chow and Paramesran, 2016). In response to these limitations, objective assessment metrics, such as the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) (Wang et al., 2004), have been proposed as alternatives. These metrics are considered to evaluate image quality, providing a more standardized approach. In addition, the need for no-reference (NR) IQA methods in real-world applications, where image quality must be assessed without access to clean reference images, has led to the development of NR-IQA techniques such as Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) (Mittal et al., 2012a). This method is based on natural scene statistics that quantify the naturalness of a given image without modeling specific distortion features. It has been demonstrated that BRISQUE correlates well with human perception and statistically outperforms traditional full-reference (FR) IQA methods like PSNR and SSIM, despite its simplicity. However, they were initially developed for natural images; thus, their ability to represent and capture subtle differences for diagnostic quality accurately is limited (Lee et al., 2022; Mason et al., 2019).

Research efforts have also been conducted in the medical domain with the aim of identifying the relationship between subjective and objective assessments, which could further lead to the development of appropriate IQA metrics for medical imaging. For example, Chow et al. generated an IQA database for magnetic resonance images with six types of distortions, including Gaussian Blur, Gaussian White Noise, Rician Noise, DCT compression, JPEG compression, and JPEG2000 compression, along with difference mean opinion score values obtained from human subjects. Additionally, they analyzed the impact of different distortion types on image quality and their correlation with human perception (Chow et al., 2016). Similarly, Schuppert et al. examined the applicability of quantitative image quality parameters in guiding decisions on preferred acquisitions by human readers (Schuppert et al., 2023).

Furthermore, particularly in CT imaging, there have been studies focusing on developing novel deep learning-based IQA metrics specifically for CT images. Gong et al. developed a new deep learning-based model observer (DL-MO) using a pretrained convolutional neural network (CNN), a partial least squares discriminant analysis (PLS-DA) model, and an internal noise component (Gong et al., 2019). First, the images were fed into the CNN to extract image features. These features were then input into the PLS-DA to create initial test statistics. Finally, the internal noise component added a weighted normal random variable to the initial test statistics to generate the final test statistics. Similarly, Han et al. proposed an anthropomorphic model observer using a simple three-layer CNN consisting of two convolutional layers and a fully connected layer for signal-known-statistically and background-known-statistically detection tasks in breast anatomical CT images (Han and Baek, 2020). Their method includes a labeling strategy that reflects human observer inefficiency by artificially assigning incorrect labels of signal present or absent. Inspired by the success of these deep learning model observers, Lee et al. proposed Deep-detector IQA (D2IQA) that utilizes a deep learning detection model to convert detectability into image quality scores for CT images (Lee et al., 2022). By inserting artificial lesion-like virtual objects into CT images and allowing D2IQA to detect these objects, the detection accuracy is then converted into image quality scores.

However, three significant limitations have hindered the advancement of IQA research in CT imaging, affecting the establishment of a standard IQA metric. First, the need to minimize patient radiation exposure makes acquiring high-quality reference images in medical imaging particularly challenging. This limitation impedes the development of FR-IQA methods, necessitating a more challenging approach to NR-IQA metrics. Second, the costly and labor-intensive process of obtaining subjective radiologists' assessments has led to a dearth of publicly available datasets for CT IQA. This absence hampers the development of deep learning-based CT IQA models, requiring substantial data volumes for effective training. Finally and most importantly, data scarcity has resulted in most medical IQA research relying on private datasets, leading to incomparable metrics. This approach complicates establishing standardized IQA metrics for medical images.

Therefore, the Low-dose Computed Tomography Perceptual Image Quality Assessment Grand Challenge (LDCTIQAC) 2023 was organized to overcome these limitations and develop an IQA method that closely correlates with radiologists' assessments, considering the diagnostic contributions of the images. The NR-IQA is essential in real clinical settings, where the ability to obtain high-quality images is often limited; therefore, this challenge emphasizes the development of deep learning-based NR-IQA algorithms. By providing the first open-source low-dose CT (LDCT) IQA dataset along with the challenge, researchers can develop novel algorithms using this dataset. Furthermore, the comparison and analysis of various approaches in developing these algorithms could significantly contribute to future research on establishing a standardized metric for CT images.

1.1. Challenge format and task

The LDCTIQAC 2023 was conducted in conjunction with the Medical Image Computing and Computer Assisted Intervention 2023 conference. The structured challenge proposal (Lee et al., 2023) underwent a peer review process by three reviewers. Participants received training data, including an axial view of abdominal images in vivo, along with the corresponding scores rated by five radiologists. These data were distributed about three months prior to the final submission deadline, allowing participants ample time to develop their algorithms. Supplementary materials, such as scoring criteria, image preprocessing details, and an artifact generation pipeline, were also provided to enhance participants' understanding of the dataset. Algorithm testing was automated via the grand-challenge.org platform, and the testing dataset remained undisclosed to participants. Therefore, a reference Docker container was made available to ensure that the participants submitted their algorithms in a format compatible with the challenge infrastructure. A sample test dataset and instructions on how to test and package the Docker file were provided to facilitate the local testing of these reference algorithms before submission.

The aim of the challenge was to develop IQA methods that correlate well with radiologists' assessments. Accordingly, 1500 images were divided into training, validation, and testing sets, comprising 1000, 200, and 300 images, respectively. The validation and testing

sets were kept confidential from participants, whereas the training images were provided along with the average subjective scores from five radiologists, serving as ground-truth annotations. Images were sourced from two institutions to ensure racial diversity in the dataset: the Mayo Clinic in the USA and the National Cancer Center (NCC) in South Korea. These datasets were used to generate artifacts that degrade image quality using a physics-based noise and streak artifact generation pipeline (Gholizadeh-Ansari et al., 2020; Kim et al., 2021). This process resulted in images with four dose levels, three streak levels, and combinations of both types of artifacts. The compositions of the validation and testing sets remained undisclosed to participants and were used to evaluate the algorithms, conducted in two phases through the Grand Challenge platform.

The challenge consisted of two phases: training and testing. Training was held for approximately three months, allowing participants to develop their algorithms and become familiar with the submission system by submitting their algorithms up to 15 times. Following the training phase, testing lasted for 14 days, during which participants had the opportunity to submit their final algorithm three times, with only the first submission being considered final. Automated evaluations were performed for both phases using undisclosed validation and testing sets. However, only the results from the testing phase using the testing set were applied for the final rankings. These rankings were determined based on an overall score that summed three representative correlation coefficient metrics: Pearson's linear correlation coefficient (PLCC), Spearman's rank-order correlation coefficient (SROCC), and Kendall's rank-order correlation coefficient (KROCC). When participants submitted their algorithms using the Docker container, the platform automatically performed the evaluation and displayed each participant's overall score on the leaderboard. For model training, external datasets, pretrained models, and weights were permitted. For the final submission, participants were required to submit their algorithms through the Grand Challenge platform and provide a short paper describing their algorithm, methodology, and contributions.

On the Grand Challenge platform, 122 participants joined the challenge. Members of the organizers' institution were not allowed to rank on the leaderboard. During the training phase, 23 participants submitted at least one entry, and six teams submitted their final algorithms in the testing phase. A single peer review was conducted for all submitted algorithms. Teams who submitted their algorithms were invited to be coauthors of this paper, with a brief description of their algorithms provided in the following section.

2. Material and methods

We applied the following principles while acquiring the dataset:

1. We addressed the principal artifacts caused by LDCT images rather than focusing on a single artifact type, ensuring general applicability to LDCT imaging.
2. We assembled a dataset representing diverse racial backgrounds, capturing a broad spectrum of diagnostic scenarios.
3. We secured images with subjective scores that consider diagnostic quality.

Accordingly, as LDCT images are typically obtained by reducing the number of projections per rotation and lowering the x-ray current, this challenge addresses two primary artifacts: sparse view streaks and noise, which are generated by these methods. An accurate physics-based artifact insertion pipeline was applied to simulate these effects. Additionally, image datasets were sourced from two institutions with distinct racial distributions: the Mayo Clinic in the USA and the NCC in South Korea. Last, a Likert scale (Lee et al., 2022) specifically designed to include diagnostic properties was employed to ensure that the diagnostic quality is accurately reflected in the scores, incorporating input from radiologists.

2.1. Challenge cohort

The challenge dataset was generated using source datasets acquired from the 2016 LDCT Grand Challenge dataset (AAPM, 2017) of the Mayo Clinic, USA, and a dataset from the NCC, South Korea. The 2016 LDCT Grand Challenge dataset comprises contrast-enhanced abdominal scans performed with a Siemens SOMATOM Flash scanner during the portal venous phase at two dose levels: full and quarter dose. Full-dose images were acquired using 120 kV with a 200 mAs quality reference, whereas quarter-dose images correspond to settings of 120 kV with a 50 mAs quality reference. For this challenge, images from 10 patients at the Mayo Clinic were selected. In addition, a dataset from NCC was used to ensure the representation of a more diverse population, encompassing North Americans and Asians. This dataset comprises actual clinical images from axial abdominal scans conducted using a Siemens SOMATOM force scanner, exclusively using full-dose settings. The full-dose settings varied, with tube voltages of either 100 or 90 kV and tube current-exposure time products of 219, 128, 249, or 89.5 mAs. From the NCC dataset, images from five patients were selected for this challenge.

For the training dataset, 50 slice images from each dataset (5 patients from the Mayo Clinic and 2 patients from NCC) were used as the source data. These slice images were employed to generate images of various quality by introducing noise and streak artifacts, resulting in 1000 images. For the validation dataset, 10 slice images from 2 patients at the Mayo Clinic and 10 slice images from one patient at NCC were used. Similarly, for the test dataset, 15 slice images from 3 patients at the Mayo Clinic and 15 slice images from 2 patients at NCC were used. This approach resulted in the creation of 200 images for the validation dataset and 300 images for the test dataset. To avoid data leakage between the training, validation, and test datasets, each dataset consisted of different patients from each other, ensuring independence among the datasets.

2.2. Low-dose CT noise artifact generation

Common methods employed in LDCT imaging to reduce the radiation dose include lowering the x-ray current and reducing the number of projections per rotation. Consequently, two primary artifacts, noise and streak artifacts, were addressed in the projection data using a physics-based artifact insertion pipeline.

First, streak artifacts were generated by forward-projecting LDCT images using stack sizes of 720, 360, and 180 projections per rotation, with corresponding angular increments of $0.5°$, $1°$, and $2°$, respectively. Thus, as the stack size decreases, distortion due to view aliasing increases. For noise artifact handling, Poisson noise levels ranging from 0 to 3 were added. In CT imaging, image noise is inversely proportional to the square root of the dose (Fahrig et al., 2006).

$$\text{Noise} \propto \frac{1}{\sqrt{\text{Dose}}} \tag{1}$$

Therefore, by adding different levels of Poisson noise to the original full-dose image, images with approximately 100%, 50%, 25%, and 10% of full-dose exposure were created due to the nonlinear relationship between noise and dose. Subsequently, the reconstruction of the projection data was executed using the filtered backprojection algorithm, specifically with the cone-beam in radiology framework system.[1] (Maier et al., 2013) This reconstruction process involved a three-step sequential pipeline, including applying a cosine weighting filter (Kak and Slaney, 2001), a Shepp–Logan ramp filter (Kak and Slaney, 2001), and backprojection accelerated by a compute unified device architecture-enabled graphics processing unit (Scherl et al., 2007). The reconstructed three-dimensional images were converted

[1] https://github.com/akmaier/CONRAD.

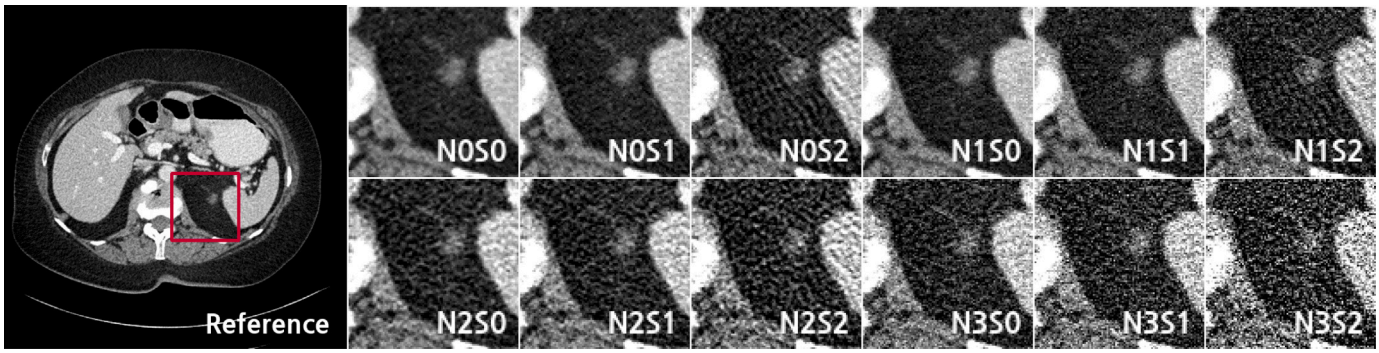


Fig. 1. Image examples with 12 distortion types generated by combining noise and streak artifacts. The numbers next to N and S represent the levels of noise and streak artifacts, respectively. Images are in abdominal soft-tissue windows (width: 350/level: 40).

Table 1
Image quality scoring criteria for computed tomography data.

Numeric score	Verbal descriptive scale	Diagnostic quality criteria
0	Bad quality	Desired features are not shown
1	Poor quality	Diagnostic interpretation is impossible
2	Fair quality	Images are suitable for limited clinical interpretation
3	Good quality	Images are suitable for diagnostic interpretation
4	Excellent quality	The anatomical structure is evident

into Hounsfield units through linear scaling, with the mean pixel values of the fat and air regions serving as reference points. The values were normalized within a range of 0 to 1, using a window width and level of 1400 and −300, respectively. Consequently, 12 distortion types were generated, resulting from a combination of four noise levels and three streak levels. Fig. 1 illustrates a sample of these 12 distortion types, indicating the distortion types where the numbers next to N and S denote the noise and streak artifact levels, respectively.

2.3. Image quality labeling

For the training and validation datasets, score labels were obtained from five radiologists, each with more than 10 years of experience in reading CT scans, from two institutions in South Korea (Seoul Metropolitan Government Seoul National University Boramae Medical Center and Veteran Health Service Medical Center). The images were presented to the radiologists in abdominal soft-tissue windows (width: 350/level: 40). Table 1 presents a five-point Likert scale that was used to capture subtle differences in diagnostic quality between images accurately based on radiologists' opinions, focusing on image noise, anatomical structure, and diagnostic interpretation (Lee et al., 2022). This scale was used with the consensus of radiologists from both institutions. The scores provided by each radiologist were averaged to minimize potential bias introduced during the assessment stage and represent the final subjective assessment score. In consideration of potential inter-rater bias, an additional radiologist from Seoul National University Hospital was included in the scoring process for the ground-truth label of the final testing dataset. The ground-truth label was obtained by averaging the scores from six radiologists.

2.4. Dataset statistics

Fig. 2 illustrates the dataset statistics, including the distribution of distortion types and subjective scores for the training, validation, and testing datasets. Data from the Mayo Clinic and NCC are marked in dark and light gray, respectively. As depicted in the figure, the distortion types are evenly distributed by selecting 10 random types from the 12 available for each slice image. Additionally, the average subjective scores range from 0 to 4.

The intraclass correlation coefficient (ICC (2,k) with two-way random effects, absolute agreement, and multiple raters/measurements) scores for the radiologists' subjective assessments were analyzed to evaluate the score reliability using the Pingouin statistical library (Vallat, 2018). The ICC score, where five radiologists assessed the training and validation datasets, was 0.9489, with a 95% confidence interval of [0.94, 0.96]. For the testing dataset, where six radiologists averaged the scores, the ICC score was 0.9618, with a 95% confidence interval of [0.95, 0.97]. These results indicate excellent reliability and substantial agreement among the radiologists in accordance with the guidelines reported by Koo and Li (2016). Additionally, the distributions of the standard deviation of subjective radiologists' scores per image are shown in Fig. 3. As seen in the figure, although radiologists sometimes reach consensus, slight variations in their assessments are common, with standard deviation values per image ranging from approximately 0 to 1.6 points on the Likert scale for the training and validation data, and from 0 to 1 point on the Likert scale for the testing data.

2.5. Baseline models

For the baseline models, we implemented task-based anthropomorphic model observers. In this approach, image quality is assessed based on the observer's performance on specific tasks. Consequently, these model observers employ mathematical predictions to estimate human performance in medical contexts, particularly in lesion detection (Barrett et al., 1993). A pivotal metric used in this methodology is the detectability index. This index quantifies the signal-to-noise ratio (SNR) associated with the test statistic, which the model observers calculate based on the image data (Abbey and Bochud, 2000; Solomon and Samei, 2016). While model observers are not yet suitable for complex clinical practice and are designed for simple tasks (Verdun et al., 2015), we include these model observers since they align with the aim of our challenge, which is to develop an IQA method that correlates with radiologists' assessments by considering task-based diagnostic quality in CT images. Among these model observers, we selected two conventional models: the channelized hotelling observer (CHO) (Yao and Barrett, 1992) with anthropomorphic channels of dense difference-of-Gaussian channels (Abbey and Barrett, 2001) and nonprewhitening observer with an eye filter (NPWE) that simulates the human visual system (Burgess, 1994). The dense difference-of-Gaussian CHO generates a channelized image v by applying multiple channels T to the given image g:

$$v = Tg, \tag{2}$$

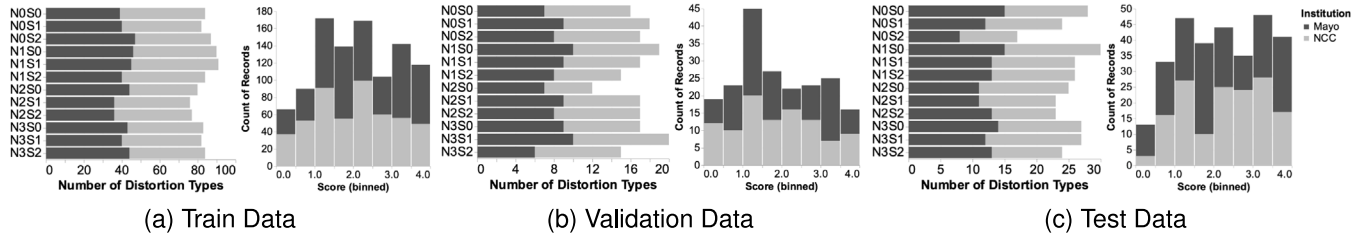


Fig. 2. Data distribution of training, validation, and testing datasets by distortion type and radiologist scores. The numbers next to N and S denote the levels of noise and streak artifacts, respectively. Data from the Mayo Clinic and NCC are represented in dark and light gray, respectively.

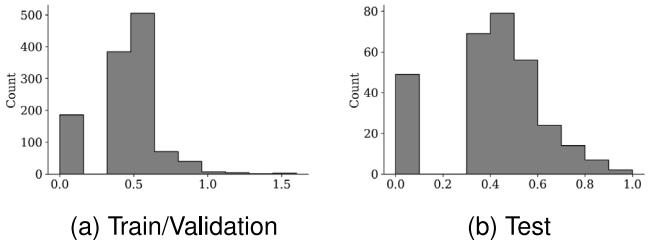


Fig. 3. Distribution of standard deviation of radiologists' scores per image.

where the channel matrix T is formed from discrete samples obtained via the inverse discrete Fourier transform (DFT) of Eq. (3).

$$C_j(\rho) = e^{-0.5\left(\frac{\rho}{Q\sigma_j}\right)^2} - e^{-0.5\left(\frac{\rho}{\sigma_j}\right)^2}. \tag{3}$$

In this equation, ρ, Q, and σ_j denote the radial frequency, a multiplicative factor, and the standard deviation of the jth channel, respectively. For the NPWE model, the template is given by:

$$w = F^{-1}\{E \cdot F\{\Delta g\}\}, \tag{4}$$

where F and F^{-1} correspond to a DFT and an inverse DFT operation, respectively. Additionally, Δg represents the difference between signal-present and signal-absent images, $E(f) = f^{1.3}\exp(-cf^2)$ represents the eye filter, and c is the eye filter parameter (Burgess, 1994; Abbey and Barrett, 2001).

Each slice image was segmented into 64 × 64 patches with a stride of 4, and lesions were introduced in half of these patches to calculate the SNR using these model observers. The lesions were modeled using a Gaussian signal characterized by a standard deviation of 3 and a maximum intensity of 0.2. For the development of the observer template, 70% of these patches were allocated to the training set. The remaining 30% of the patches constituted the testing set used to calculate the final SNR.

In addition to the model observers, we include the SNR, calculated by dividing the average signal of the fat by the standard deviation of the air, to directly measure noise. Moreover, BRISQUE and the Natural Image Quality Evaluator (NIQE) (Mittal et al., 2012b) are also included as baseline models to compare the effectiveness of established IQA metrics in the natural domain against the submitted methods specifically developed for CT images. Lastly, the frequently used evaluation metrics, PSNR and SSIM, are also applied as benchmarks to compare the effectiveness of the developed algorithms with these standard metrics in the medical domain. The IQA scores from PSNR and SSIM are calculated using full-dose images as references for each slice image to facilitate this comparison. Then, the scores acquired from SSIM, ranging from 0 to 1, and PSNR, which can range from 0 to infinity, are directly compared to the radiologists' scores by calculating the correlation coefficients. Subsequently, the performance of each IQA method is evaluated using images that are not full-dose to ensure a fair comparison with these FR-IQA metrics.

2.6. Evaluation methods

Participants were required to submit their algorithms as Docker containers, enabling their testing on the validation and testing datasets for the training and testing phases, respectively, on the Grand Challenge platform. Both the validation and testing datasets were concealed from participants to prevent manual fitting of the prediction results, ensuring that the algorithm performance was evaluated solely on its merit. The evaluation results for both datasets were accessible through the leaderboard. However, the performance scores from the validation dataset during the training phase were not considered final. The primary purpose of the training phase was to acquaint participants with the submission system and foster algorithm development by allowing them to compare their results with others on the leaderboard.

The overall performance was evaluated by summing three correlation coefficient metrics commonly used in IQA: PLCC, SROCC, and KROCC. In IQA, PLCC and SROCC are the most commonly used evaluation metrics for measuring the performance of IQA models (Bosse et al., 2018; Wu et al., 2020; Gu et al., 2022). PLCC is used to measure the linear relationship between the radiologists' assessments and each method, while SROCC is used to measure the monotonic relationship. Although not every IQA method may show a linear correlation with radiologists' assessments, we use PLCC as one of our evaluation metrics because linear correlation offers more straightforward interpretation and predictability. In addition to these metrics, KROCC, a measure of rank correlation similar to SROCC, was also used for evaluation due to its frequent use in IQA and its robust statistical performance under extreme conditions (Arndt et al., 1999).

Therefore, these three distinct metrics were employed to account for linear and nonlinear correlations with radiologists' assessments. First, the linear correlation was assessed using the PLCC, calculated as follows:

$$\text{PLCC} = \frac{\sum(\hat{s}_i - \hat{\mu})(s_i - \mu)}{\sqrt{\Sigma(\hat{s}_i - \hat{\mu})^2 \Sigma(s_i - \mu)^2}}, \tag{5}$$

where $\hat{s}_i$ represents the ith predicted image quality score, and s_i denotes the ith averaged score provided by the radiologists. Similarly, $\hat{\mu}$ and μ correspond to the mean values of the predicted and radiologists' scores, respectively.

In addition, two nonlinear correlation coefficients, SROCC and KROCC, were applied in this challenge. These metrics calculate the monotonic relationship between two methods:

$$\text{SROCC} = 1 - \frac{6\Sigma(\hat{S}_i - S_i)^2}{n(n^2-1)}. \tag{6}$$

In this equation, $\hat{S}_i$ and S_i represent the ranks of the ith element in the predicted image quality scores $\hat{S}$ and the radiologists' assessment scores S, respectively. The variable n denotes the total number of images. Additionally, KROCC is calculated as follows:

$$\text{KROCC} = \frac{M_c - M_d}{\sqrt{(M_c + M_d + T_{\hat{s}})(M_c + M_d + T_s)}}, \tag{7}$$

where M_c and M_d correspond to the number of concordant and discordant pairs, respectively. Moreover, $T_{\hat{s}}$ and T_s denote the numbers

Table 2
Summary of submitted methods detailing techniques, augmentation strategies, and pretrained weights.

Team	Method	Data augmentation	Pretrained
agaldran	Employed a multiloss, multihead classifier (Galdran et al., 2023) with swin transformer (Liu et al., 2021) and BiT ResNeXt-50 (Kolesnikov et al., 2020) backbones, each trained using a five-fold ensemble. The linear head of each backbone was augmented with four heads, each supervised using cross-entropy loss. Averaged predictions from each output minimized the ranked probability loss (Epstein, 1969; Galdran, 2023). Test-time augmentation was applied during inference.	Flips, rotations, and intensity scaling	ImageNet (Russakovsky et al., 2015)
RPI_AXIS	Applied a multidimension attention network for no-reference image quality assessment (Yang et al., 2022) with three core components for multidimensional feature extraction: transposed attention block, scale swin transformer block, and a dual-branch structure for patch-weighted quality prediction with mean squared error loss. Testing involved center cropping images to 448 × 448, followed by resizing to 224 × 224.	N/A	ImageNet
CHILL@UK	Adapted EfficientNetV2-L (Tan and Le, 2021) for regression tasks with single-channel input, modifying the first and final layers. Trained using the mean squared error loss and a five-fold cross-validation process. An ensemble of the four best-performing models was used for the final prediction.	Flips and rotations	ImageNet
FeatureNet	Employed a novel architecture comprising three components: a convolutional neural network module with wavelet transform, a vision transformer (Dosovitskiy et al., 2020) -based visual feature extraction module, and a gray-level co-occurrence matrix computation and Haralick texture feature extraction block (Löfstedt et al., 2019). Concatenated features from each module pass through a linear layer with the rectified linear unit activation function for prediction. Training involved three stages using the means squared error loss.	N/A	ImageNet
Team Epoch	Implemented a novel two-stage multitask model featuring two edge enhancement-based densely connected convolutional neural network (EDCNN) (Liang et al., 2020) models focused on noise features. The first stage extracts a noise map using a frozen pretrained EDCNN, and this noise map feeds into the second EDCNN, augmented with additional layers leading to dual output branches for regression and classification tasks. Employed Huber loss for regression and cross-entropy loss for classification, with final predictions derived by averaging outputs from both branches. An ensemble of three models trained under various parameters was used during inference.	Mosaic augmentation (Bochkovskiy et al., 2020), shuffling and mixing to disrupt image spacial structure, flips, and rotations	AAPM (Moen et al., 2021)
gabybaldeon	Used a vision transformer (Dosovitskiy et al., 2020)-based network with a convolutional stem and knowledge distillation, implementing two-stage training using the joint loss of the mean squared and mean absolute errors. A teacher model comprising an ensemble of five ViT-based networks was trained using five-fold cross validation. Subsequently, a single ViT-based network student model was trained, focusing on the loss between the teacher and student models and between the ground truth and student model.	Flips, rotations, and scaling	ImageNet

of ties found solely within the predicted $\hat{S}$ and radiologists' S scores, respectively. The final overall correlation score was determined by aggregating the values of all correlation coefficients:

$$\text{Score} = |\text{PLCC}| + |\text{SROCC}| + |\text{KROCC}|. \tag{8}$$

2.7. Generalization test

The final phase of the evaluation concentrated on examining the generalizability of the algorithms. These models were not explicitly trained on the datasets used in this phase, and generalization was not the primary focus of the challenge. However, for image quality metrics to replace standards established in the medical imaging field effectively, such as SSIM and PSNR, they must be robust across diverse medical images. Thus, we evaluated the generalizability of each algorithm to gauge their potential as universal metrics for CT IQA. This step was pivotal in determining their broader applicability and effectiveness across imaging scenarios.

To conduct the generalization test, we employed a diverse set of datasets, including the CQ500 dataset (Chilamkurthy et al., 2018) and three chest phantom datasets sourced from various scanner vendors: Siemens, GE, and Toshiba. Each of these datasets contains images characterized by varying noise levels. Because we do not have subjective scores for these datasets, we calculated each algorithm's performance based on the correlation between the noise levels in the images and the scores predicted by each algorithm.

3. Results

This section summarizes the methodologies and results employed by the six teams in this challenge. For a more comprehensive overview of each team's methodology, Table 2 succinctly summarizes the submitted methods. The Appendix provides detailed explanations of each methodology.

3.1. Overall performance of the submitted algorithms

Table 3 presents the assessment results for image quality, quantified using PLCC, SROCC, and KROCC. Fig. 4 further illustrates the comparative performance of each algorithm. Six algorithms, each employing distinct methodologies, were evaluated. Among these, the algorithm proposed by *agaldran* achieved the highest performance, registering an aggregate score of 2.7427, with individual PLCC, SROCC, and KROCC scores of 0.9491, 0.9495, and 0.8440, respectively. Next, *RPI_AXIS* ranked second with an overall score of 2.6843, comprising PLCC, SROCC, and KROCC scores of 0.9434, 0.9414, and 0.7995. *CHILL@UK* secured the third position with a composite score of 2.6719, including

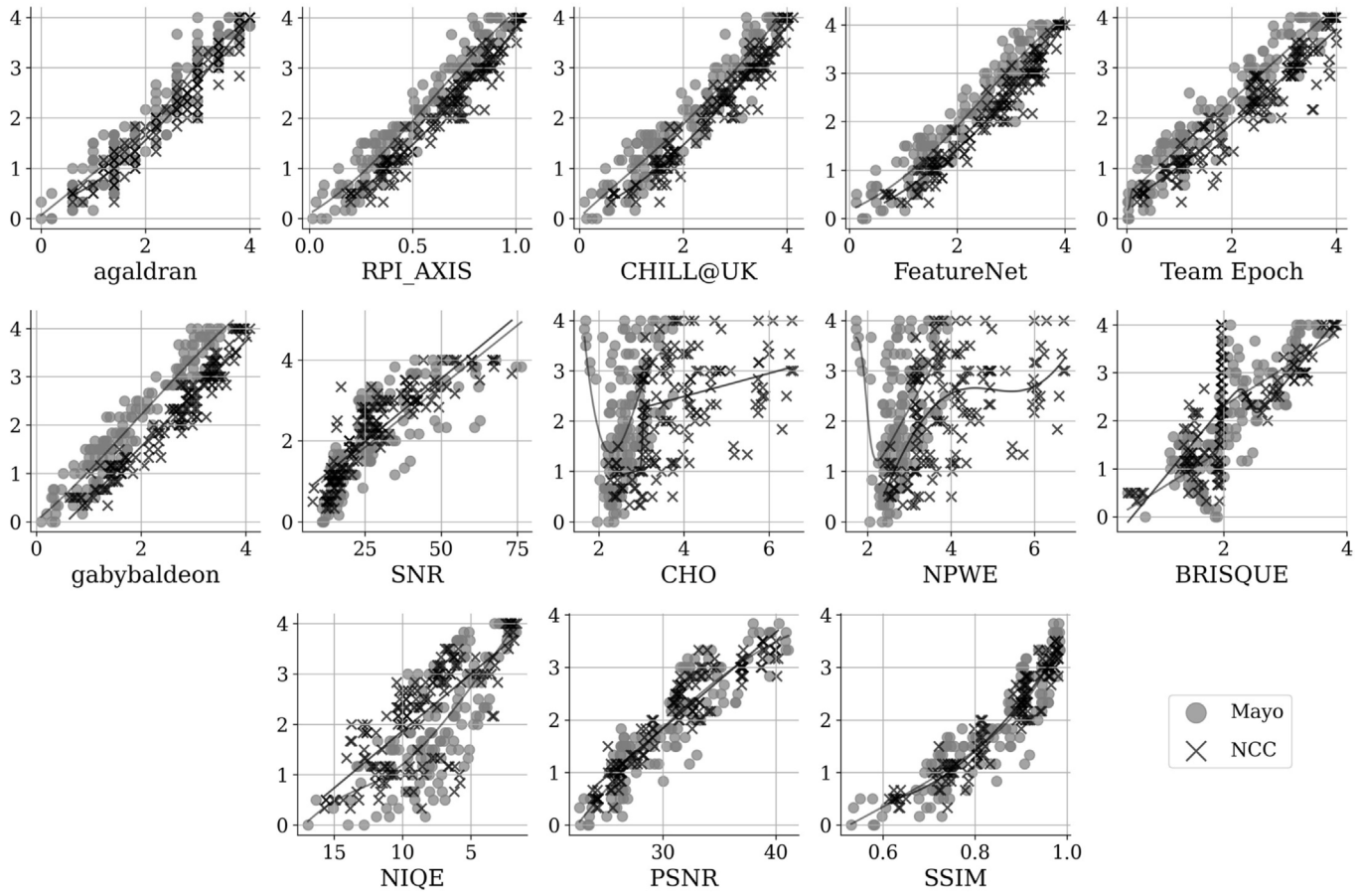


Fig. 4. Visualization of the correlation between each method and the radiologists' subjective scores, with data from the Mayo Clinic and NCC marked in gray and black, respectively. Regression lines illustrate the relationship between the scores of each dataset and the radiologists' scores for each method, calculated using a five-parameter logistic function described by Sheikh et al. (2006).

PLCC, SROCC, and KROCC scores of 0.9402, 0.9387, and 0.7930. All submitted algorithms surpassed the baseline model observers, CHO and NPWE, which had overall scores of 1.1331 and 1.0714, respectively, as well as the commonly used statistic SNR, which had an overall score of 2.4026, and the BRISQUE and NIQE models, which scored 2.1219 and 1.9760, respectively.

For a more equitable comparison with the conventional evaluation metrics of PSNR and SSIM, the results for each algorithm were recalculated, excluding 28 full-dose images from the testing set (Table 4). The revised analysis indicates that the leading three algorithms surpassed the baseline PSNR performance, with the second- and first-place algorithms also exceeding both PSNR and SSIM benchmarks. However, while PSNR and SSIM do not exhibit a perfect correlation trend with radiologists' scores, displaying distinctive prediction scores in the NCC dataset, they nearly perfectly align with the regression lines for the Mayo Clinic and NCC datasets (Fig. 4). Similarly, the high-ranked algorithms demonstrate good generalizability between the Mayo Clinic and NCC datasets; however, there was a gap in the alignment of the regression lines between the two datasets, increasing in the lower-ranked algorithms.

Table 3
Correlation scores between radiologists' scores and each submitted method, as well as the baseline scores, using the challenge testing dataset. The overall score represents the summation of three correlation coefficient scores: PLCC, SROCC, and KROCC. The column *n* indicates the number of submissions during the training phase.

Team	*n*	Overall	PLCC	SROCC	KROCC
agaldran	10	2.7427	0.9491	0.9495	0.8440
RPI_AXIS	2	2.6843	0.9434	0.9414	0.7995
CHILL@UK	5	2.6719	0.9402	0.9387	0.7930
FeatureNet	26	2.6550	0.9362	0.9338	0.7851
Team Epoch	58	2.6202	0.9278	0.9232	0.7691
gabybaldeon	14	2.5671	0.9143	0.9096	0.7432
SNR		2.4026	0.8226	0.8748	0.7052
CHO		1.1331	0.3778	0.4395	0.3159
NPWE		1.0714	0.3566	0.4165	0.2982
BRISQUE		2.1219	0.7500	0.7863	0.5856
NIQE		1.9760	−0.7278	−0.7143	−0.5339

3.2. Summary of submitted methods

The participating teams employed commonly used neural network architectures to predict image quality, except for *Team Epoch*. For instance, *RPI_AXIS* implemented the IQA model multidimension attention network for no-reference IQA (MANIQA) (Yang et al., 2022), integrating the vision transformer (Dosovitskiy et al., 2020) and swin transformer (Liu et al., 2021). Similarly, *CHILL@UK* adapted the EfficientNetV2 (Tan and Le, 2021) architecture for regression tasks. Furthermore, *agaldran*, *FeatureNet*, and *gabybaldeon* incorporated these state-of-the-art networks as foundational elements within their unique methodologies. Specifically, *agaldran* enhanced the swin transformer and BiT ResNeXt-50 (Kolesnikov et al., 2020) backbones with a multiloss, multihead strategy (Galdran et al., 2023) and ranked probability loss (RPL) (Epstein, 1969; Galdran, 2023). *FeatureNet* combined features derived from the ViT with those obtained via the wavelet transform and gray-level co-occurrence matrix (GLCM) and Haralick

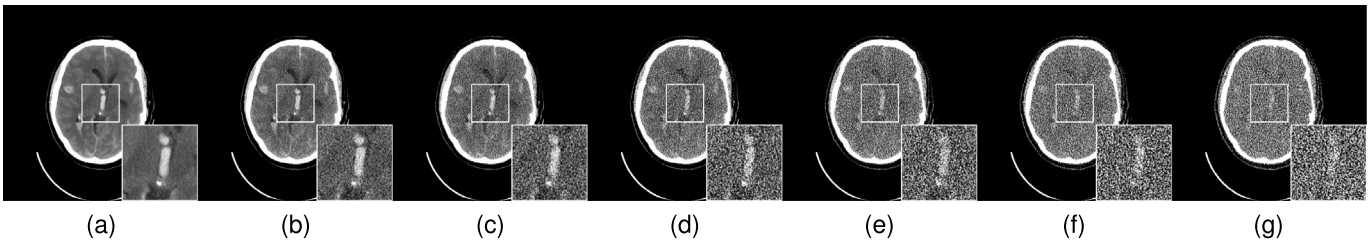


Fig. 5. Head computed tomography data sample image (Lee et al., 2022) generated using the CQ500 dataset with seven noise levels, increasing from left to right, visualized with a range of [0, 80] Hounsfield units.

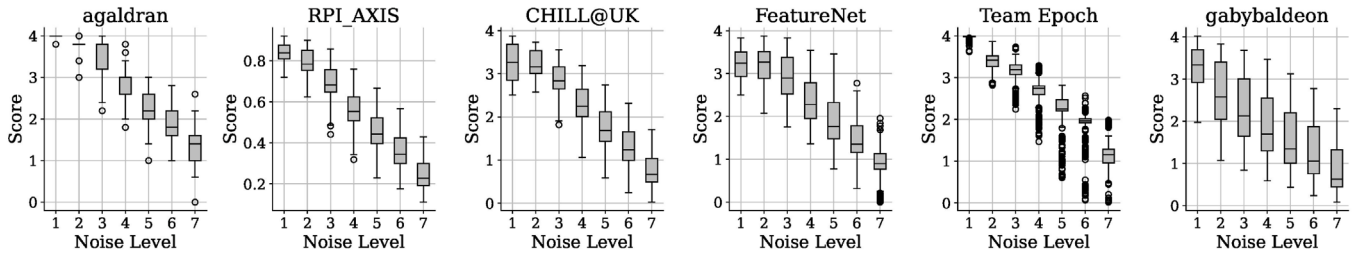


Fig. 6. Box plots representing the performance of each algorithm in the generalization test with the CQ500 dataset. The *x*-axis represents the noise level of the image; the *y*-axis denotes the image quality assessment scores predicted by each algorithm. Teams are plotted from left to right according to their rankings in the challenge.

Table 4
Correlation scores between radiologists and each submitted method, and the baseline metrics PSNR and SSIM. Correlation scores were calculated after excluding the full-dose images from the challenge testing dataset.

Team	Overall	PLCC	SROCC	KROCC
agaldran	2.6966	0.9376	0.9373	0.8216
RPI_AXIS	2.6323	0.9314	0.9279	0.7731
CHILL@UK	2.6143	0.9279	0.9230	0.7634
FeatureNet	2.5881	0.9211	0.9161	0.7509
Team Epoch	2.5554	0.9105	0.9058	0.7390
gabybaldeon	2.5225	0.9028	0.8967	0.7229
PSNR	2.6100	0.9149	0.9231	0.7720
SSIM	2.6252	0.9168	0.9284	0.7800

texture feature extraction (Löfstedt et al., 2019) modules. Team *gabybaldeon*'s approach involved knowledge distillation, transferring knowledge from an ensemble of ViT-based models to a singular ViT-based student model. Conversely, *Team Epoch* constructed a two-stage IQA network grounded on an efficient denoising model composed of two edge enhancement-based densely connected convolutional neural networks (EDCNNs) (Liang et al., 2020). Initially, this network delineates a noise map by subtracting the denoised image procured via a frozen EDCNN from the original image. This noise map is passed into the second EDCNN-based network to predict the image quality.

The majority of submitted algorithms were predicated on regression-based models, with the objective of computing quantitative scores for image quality. Nevertheless, *agaldran* and *Team Epoch* diverged by incorporating a classification paradigm into the IQA process. They each implemented a classification schema comprising 21 discrete classes, ranging from 0.0 to 4.0, in increments of 0.2. Team *agaldran* exclusively adopted classification tasks using cross-entropy loss and RPL. In contrast, *Team Epoch* pursued a multitask methodology using classification and regression tasks, calculating the mean of the outputs from each task using cross-entropy loss for classification and the Huber loss for regression.

Additionally, except for *Team Epoch*, all teams used pretrained weights from the ImageNet database (Russakovsky et al., 2015) to enhance their models. In contrast, *Team Epoch* used pretrained weights for EDCNN using the AAPM dataset (Moen et al., 2021). Furthermore, aside from *RPI_AXIS* and *FeatureNet*, the remaining teams adopted ensemble methods to improve model robustness. Team *agaldran* additionally applied a test-time augmentation technique. Regarding data augmentation, all participating teams, except *FeatureNet* and *RPI_AXIS*, implemented conventional techniques, including flipping and rotation, to enhance the variability of the training data. *Team Epoch* employed standard methods and introduced mosaic augmentation (Bochkovskiy et al., 2020), expanding the image dataset from 1000 to 5250 images, with 5000 of these designated for training. Furthermore, they used shuffling and mixing strategies in conjunction with mosaic augmentation to further disrupt the spatial structure of the images. Conversely, *RPI_AXIS* and *FeatureNet* abstained from employing any form of data augmentation.

3.3. Generalizability

The generalizability of each algorithm was tested to assess its performance on images with varying anatomical structures and acquisition parameters. Additional datasets, including CQ500 and anthropomorphic phantoms from three different scanners, were introduced to facilitate testing. However, we did not have subjective evaluations for these datasets; thus, the evaluation was performed by calculating the correlation scores between the noise level and the quality scores predicted by each algorithm. While the results from this phase do not provide a direct correlation between each algorithm and radiologists' scores, they offer insight into the broader applicability and effectiveness of the algorithms as potential universal metrics for CT IQA.

3.3.1. Correlation between the radiologists' score and the noise level

Before using the noise level as a proxy for radiologists' subjective scores, we investigated the correlation between the artifact level and radiologists' assessments. Table 5 displays the average subjective scores of radiologists for each distortion type in descending order. As shown in the table, when the level of streak artifact reaches 2, the subjective scores are markedly lower compared to those for other distortion types, regardless of the radiation dose. For other distortion types, however, the average subjective score decreases as the radiation dose decreases.

To quantitatively measure the correlation between radiologists' scores and the radiation dose, we calculated correlation scores between the two. The resulting PLCC, SROCC, and KROCC were 0.7956, 0.8303, and 0.6945, respectively. The correlation becomes more pronounced

Table 5
Distortion types, their corresponding radiation doses, and average subjective scores from radiologists. Numbers next to *N* and *S* represent the levels of noise and streak artifacts, respectively. Radiologists' scores were acquired using the Likert scale shown in Table 1.

Distortion type	Radiation dose (%)	Average subjective score
N0S0	100	3.8750
N1S0	50	3.3833
N0S1	50	2.9028
N1S1	25	2.7115
N2S0	25	2.5600
N2S1	12.5	2.0870
N3S0	10	2.0494
N3S1	5	1.5494
N0S2	25	1.2451
N2S2	12.5	1.1538
N2S2	6.25	0.8261
N3S2	2.5	0.4097

Table 6
Evaluation results of the generalization test conducted using the CQ500 dataset comprising seven distinct noise levels. The teams are listed in accordance with their respective rankings in the challenge.

Team	Overall	PLCC	SROCC	KROCC
agaldran	2.7238	0.9245	0.9423	0.8571
RPI_AXIS	2.6926	0.9405	0.9389	0.8132
CHILL@UK	2.5535	0.9018	0.9005	0.7513
FeatureNet	2.4684	0.8784	0.8757	0.7142
Team Epoch	2.7719	0.9435	0.9621	0.8663
gabybaldeon	2.1589	0.7746	0.7730	0.6113

when considering only one type of distortion. Specifically, the correlation scores between the noise level of images with only noise artifacts and the subjective scores were −0.8708, −0.9057, and −0.7969 for PLCC, SROCC, and KROCC, respectively. These results suggest that while not perfect, the noise level can serve as a reasonable proxy for radiologists' assessments.

3.3.2. Clinical head CT dataset, CQ500

A noncontrast head dataset, CQ500, was employed to assess the generalizability of the algorithms across anatomical parts. For this purpose, noise levels were generated in the dataset images, applying the same noise insertion pipeline used for creating the challenge dataset, albeit with different noise levels (Lee et al., 2022). This process yielded 6867 images, encompassing seven distinct noise levels (Fig. 5).

Table 6 and Fig. 6 present the results of the performance of each algorithm on this dataset. Table 6 indicates that the algorithms' estimations demonstrated a good extent of correlation with the noise levels. This outcome suggests that these models successfully learned to identify image quality artifacts, distinguishing them from anatomical structures.

A notable finding was observed with the algorithm proposed by *Team Epoch*. While maintaining the same correlation score ranking with the CQ500 dataset as in the challenge ranking, *Team Epoch*'s algorithm demonstrated exceptional correlation scores with the noise levels: 0.9435, 0.9621, and 0.8663 for PLCC, SROCC, and KROCC, respectively. Consequently, *Team Epoch* achieved the highest overall score for the CQ500 dataset among the six algorithms, at 2.7719. However, while these correlation scores are calculated based on the noise level, they do not necessarily imply that *Team Epoch*'s algorithm aligns best with radiologists' perception of image quality on the CQ500 dataset. Instead, these results indicate the algorithm's proficiency in predicting image quality scores that correlate with the noise level.

3.3.3. Cross-vendor anthropomorphic phantom CT dataset with real noise

Additional generalization tests were conducted using anthropomorphic chest phantom datasets to evaluate each model's performance on authentic images, as opposed to synthesized datasets. Images acquired from three scanner vendors (Siemens, GE, and Toshiba) were used. Each dataset from these scanners included noise levels at 100%, 50%, 25%, and 10% of the routine clinical dose, with the GE dataset additionally including a 5% dose level.

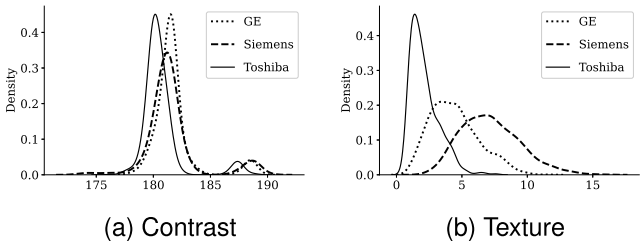


Fig. 7. Visualization of domain-dependent features (contrast and texture) across three datasets acquired from Siemens, GE, and Toshiba. Contrast was measured by average pixel intensity, while texture was represented by GLCM contrast values within randomly selected uniform ROIs, chosen separately for each feature.

Using these images, we first analyzed the domain shifts between the three datasets, which can potentially affect the performance of deep learning models. In medical imaging, domain-dependent features such as contrast and texture can vary based on the imaging protocols used to acquire images. Deep learning networks are susceptible to these variations, which can result in performance degradation (Kamnitsas et al., 2017; Dou et al., 2018; Pooch et al., 2020; Ouyang et al., 2023). Therefore, we further explored the variations in domain-dependent features among the three datasets by extracting and comparing contrast and texture features.

First, we identified uniform regions of interest (ROIs) corresponding to the [−100, 50] Hounsfield units, including fat and soft tissue, by calculating the standard deviation using sliding windows across full-dose images in each dataset. Within these ROIs, 1000 patches were randomly selected separately for contrast and texture features in each dataset to represent uniform areas devoid of anatomical structures, making them suitable for feature extraction. To improve computational efficiency, the images were then rescaled to a 0–255 range. For each patch, the average pixel intensity was used to quantify the contrast feature, while the GLCM (Haralick et al., 1973) contrast value was utilized to represent the texture feature. Fig. 7 shows the kernel density estimate plot of the 1000 contrast and texture values for each dataset. As illustrated in the figure, the contrast and texture distributions differ across the three datasets, with a particularly significant difference observed in the texture feature of the Toshiba dataset compared to the other two datasets. Therefore, by comparing the performance of each algorithm across these three different datasets, we can observe not only the model's performance under real noise conditions but also its generalization across various vendors, which are experiencing domain shifts due to different imaging sources and protocols.

To compare the performance of each algorithm, correlation scores were calculated based on the relationship between the noise levels and the image quality scores predicted by each method. Table 7 details the results for these three datasets. In the GE dataset, although the correlation scores were lower than those observed for the challenge dataset, most methods achieved substantial correlation. Notably, the method introduced by *Team Epoch* achieved the highest scores, recording 0.9152, 0.9222, and 0.8027 for PLCC, SROCC, and KROCC, respectively, culminating in an overall score of 2.6402.

In the Siemens dataset, the overall performance of the algorithms was lower than that in the GE dataset. The highest score was achieved by *agaldran*'s algorithm, which won first place in the challenge with an overall score of 1.8502. This algorithm was closely followed by the algorithm from *RPI_AXIS*, the second-place winner of the challenge, scoring 1.8082. *Team Epoch*'s algorithm ranked third among the six algorithms with a score of 1.7731, demonstrating a performance that was not significantly different from the top-ranked methods.

Table 7
Evaluation results of the generalization test comprising five noise levels using a phantom dataset from GE, Siemens, and Toshiba scanners. Teams are listed in the order of their rankings in the challenge.

Team	Siemens				GE				Toshiba			
	Overall	PLCC	SROCC	KROCC	Overall	PLCC	SROCC	KROCC	Overall	PLCC	SROCC	KROCC
agaldran	1.8502	0.5726	0.6818	0.5958	2.3173	0.7197	0.8494	0.7483	0.5318	0.1826	0.1826	0.1667
RPI_AXIS	1.8082	0.6358	0.6631	0.5093	2.3639	0.8327	0.8453	0.6858	0.1608	0.0236	0.0775	0.0598
CHILL@UK	1.7697	0.6374	0.6343	0.4981	2.2490	0.7938	0.8119	0.6434	0.3087	0.0333	0.1513	0.1241
FeatureNet	1.0078	0.3759	0.3594	0.2725	2.3141	0.8042	0.8328	0.6771	0.3237	0.0943	0.1283	0.0979
Team Epoch	1.7731	0.5663	0.6692	0.5376	2.6402	0.9152	0.9222	0.8027	2.5196	0.8668	0.8907	0.7621
gabybaldeon	0.8522	0.2890	0.3184	0.2448	1.8014	0.6476	0.6515	0.5023	0.0010	0.0060	0.0045	0.0024

Table 8
Average inference time and memory consumption for testing a single slice image using an Nvidia GeForce RTX 2080 Ti. The time is reported in milliseconds (ms), and memory consumption is reported as the peak memory usage during inference in megabytes (MB).

Team	Time (ms)	Memory consumption (MB)
agaldran	424.07	1309.89
RPI_AXIS	44.87	638.51
CHILL@UK	138.46	503.63
FeatureNet	51.08	572.31
Team Epoch	72.11	564.94
gabybaldeon	29.69	942.13

For the Toshiba dataset, the correlation scores were notably lower than those for the other two datasets, except for the results from *Team Epoch*'s algorithm. *Team Epoch*'s algorithm displayed robust performance on this dataset, achieving scores of 0.8668, 0.8907, and 0.7621 for PLCC, SROCC, and KROCC, respectively, resulting in an overall score of 2.5196. This score was significantly higher than the next highest score of 0.5318, achieved by *agaldran*.

4. Discussion

This paper introduces the first IQA challenge designed explicitly for CT images. In the absence of a standard evaluation metric for medical images, this challenge guides future research directions for CT IQA tasks, offers valuable insight, and provides the research community in medical IQA with the first open-source dataset for this purpose. The algorithms submitted to this challenge outperformed baseline models such as SNR, CHO, NPWE, BRISQUE, and NIQE. They achieved results comparable to established FR-IQA metrics like PSNR and SSIM, notably without relying on reference images. These findings not only demonstrate that the algorithms outperformed commonly used metrics like SNR, BRISQUE, NIQE, and model observers relevant to clinical tasks, but also suggest the potential of these new methods to replace traditional FR-IQA metrics. Participants in the challenge employed diverse approaches: some adapted state-of-the-art algorithms for CT IQA, whereas others used knowledge distillation or integrated various features, including those more suitable for medical images. Furthermore, although most algorithms approached the task through regression, some employed classification and multitasking methods.

Ensemble methods were common, though their applicability in real-world scenarios warrants further investigation. For example, as seen in Table 8, *agaldran*, *CHILL@UK*, and *Team Epoch* exhibited higher inference times compared to *RPI_AXIS* and *FeatureNet*, which did not use ensemble methods. On the other hand, *gabybaldeon's* results show that this can be overcome through techniques such as knowledge distillation. Specifically, *gabybaldeon* also utilized ensemble methods but demonstrated a comparably shorter inference time than other methods. This reduction in inference time is due to the use of knowledge distillation. In this approach, ensemble methods are employed during the training of the teacher model. However, through knowledge distillation, the student model learns the teacher model's knowledge while being much smaller in size.

The performance of the submitted algorithms was comparable with that of established metrics, such as PSNR and SSIM. However, in contrast to PSNR and SSIM, which did not exhibit significant score variations between the Mayo Clinic and NCC datasets, the submitted algorithms displayed noticeable differences in the score distribution between these datasets despite both being included in the training data. This discrepancy might be attributed to the fact that the PSNR and SSIM are FR-IQA methods that rely on reference images to assess statistical discrepancies in distorted images. In contrast, the submitted algorithms are NR-IQA methods that function without reference images. In this challenge, the ability to generalize across the Mayo Clinic and NCC datasets emerged as a crucial factor in determining the performance of each algorithm, with top-performing algorithms generally demonstrating more robust generalizability compared to their lower-ranked counterparts.

Contrastingly, all submitted algorithms exhibited superior performance compared to the model observers CHO and NPWE. This discrepancy can be attributed to the differing methodologies of model observers in comparison to those of both the radiologists and submitted algorithms. Although radiologists evaluate the overall diagnostic accuracy of medical images, and the submitted algorithms are directly trained based on radiologists' assessments, the predictions made by the model observers are exclusively focused on lesion detection accuracy within specific configurations, particularly Gaussian signals. Consequently, variations in the task or lesion configurations might lead to changes in the outcomes predicted by model observers. Furthermore, while radiologists may not perceive differences in image quality due to variations in image slices, the detection accuracy of model observers can be significantly influenced by such factors. This observation underscores the potential limitations of the model observers CHO and NPWE as primary metrics to assess image quality in a real-world clinical setting, where comparing quality across diverse images is crucial.

Additionally, the generalizability of the submitted methods was evaluated using in vivo head CT datasets generated by CQ500 and anthropomorphic chest phantom images acquired from three scanners. The algorithms performed well on the head CT dataset, which had noise artifact statistics similar to those of the challenge dataset. This outcome suggests that the models effectively learned texture information distinct from anatomical structures. However, their performance declined when assessed on authentic phantom images from various scanners, indicating the need for improved generalizability for these methods to become standard in CT imaging. Notably, *Team Epoch*'s algorithm performed consistently in the generalization tests. While the correlation scores in the generalization study were based on artifact levels rather than radiologists' scores, *Team Epoch*'s method consistently exhibited a strong correlation with these artifact levels. This result may be attributed to their unique approach of extracting noise maps for IQA score calculation and employing heavy data augmentation to disrupt the spatial structure. Although these techniques might have limited the algorithm's correlation with the radiologists' detailed scores, they potentially enhanced the ability to generalize by learning texture information independent of anatomical shapes.

While this challenge provided valuable insights into medical IQA research, it also highlighted several limitations and points that warrant further consideration in future studies. First, this study compared the

algorithm with radiologists' subjective assessments of image quality. In medical IQA, radiologist assessments are considered the gold standard. However, it is important to note that these assessments are subjective and could potentially include inherent biases, which may mislead the evaluation of IQA models by either overestimating or underestimating their performance. Therefore, to minimize potential bias in radiologist assessments for future studies, we could gather image scores from more radiologists and create a mean opinion score or different mean opinion scores to serve as the database for medical IQA, similar to practices currently used in natural IQA (Sheikh et al., 2006; Ponomarenko et al., 2015; Hosu et al., 2020; Jinjin et al., 2020).

Second, the dataset in this challenge lacked the breadth necessary to evaluate its efficacy as a standard evaluation metric fully. New metrics must demonstrate consistent performance across a wide range of datasets, irrespective of the imaging scanner, parameters, and depicted anatomical region, to supplant established metrics effectively, such as PSNR and SSIM. However, this challenge was confined to abdominal images sourced exclusively from the Mayo Clinic, USA and NCC, South Korea. Future challenges would benefit from incorporating a more diverse array of datasets, providing deeper insight into the suitability of each algorithm as a universally applicable standard metric with varying imaging statistics.

Third, this study did not analyze the perceptual sensitivity of radiologists to variations in model performance. There may be a threshold of correlation with radiologist assessments beyond which differences between algorithms are not perceptibly significant. Investigating this perceptual threshold could yield insightful findings, further elucidating the relative efficacy of each algorithm.

Fourth, accounting for the uncertainty of image quality scores for each image is an important factor in evaluating model performance. Specifically, image samples that exhibit high disagreement among radiologists regarding quality scores should not be weighted as heavily as those with unanimous agreement. Consequently, we have included additional weighted correlation scores in Appendix A.2, derived by considering the uncertainties of each sample. In future studies, further investigation of these uncertainties could yield deeper insights into model analysis, particularly by revealing the model's robustness with samples of low uncertainty and the challenges posed by samples with high uncertainty.

Fifth, while this study used PLCC, SROCC, and KROCC as evaluation metrics due to their common usage in IQA, it is important to question their applicability before selecting the appropriate metric for each specific task and context. For example, PLCC is commonly employed to assess the strength of a linear relationship, even though one of its key assumptions is that the two variables are jointly normally distributed—an assumption that may not always hold in real-world scenarios (Schober et al., 2018). Despite its widespread acceptance, the use of PLCC without careful consideration of its assumptions could lead to misleading conclusions. There are conflicting perspectives on whether this assumption must be strictly adhered to Nefzger and Drasgow (1957), Kowalski (1972) and Schober et al. (2018). In our study, the ranking trends among the six algorithms were consistent across all metrics: when an algorithm's correlation coefficient was higher in one metric, it was similarly higher in the others. Therefore, we believe the results of PLCC are reliable in this context. Nevertheless, this consistency might not apply universally, underscoring the importance of reporting individual correlation coefficients. For example, a new method might perform poorly on PLCC but excel on other metrics. In such cases, non-parametric methods like SROCC and KROCC might provide more reliable results.

Lastly, this study focused on predicting the overall diagnostic performance of medical images. Specifically, the assessment by radiologists was to judge the image quality, following a common approach in existing studies (Mason et al., 2019; Ma et al., 2020; Huh et al., 2023). That is, they determined whether the image quality was sufficient to reliably visualize any present pathology, without specific consideration for diagnostic tasks and findings confirmed by biopsy or other non-imaging tests. However, excluding direct diagnostic tasks may lead to an inaccurate assessment of diagnostic accuracy. On the other hand, including such tasks would make it more justifiable to claim that an image is of accurate diagnostic quality. For instance, the simple and frequent approach of assessing diagnostic accuracy concerning small lesions — which represents the ability to detect and characterize minute lesions accurately — is crucial for evaluating medical image quality. Therefore, future studies could incorporate scores derived from diagnostic tasks, such as sensitivity to small details within images, which are essential for precise diagnosis and treatment planning.

5. Conclusion

As the first challenge in the field of medical IQA, this event has notably contributed to and advanced medical IQA research. The introduction of the first open-source dataset specifically designed for CT IQA tasks was a critical achievement. This study highlights the feasibility of employing NR-IQA algorithms tailored for CT images as practical alternatives to the prevalent standard FR-IQA methods, such as PSNR and SSIM. Although the algorithms presented in this challenge display promising potential to supersede these standard IQA metrics, the necessity for improved generalization across diverse datasets and imaging conditions is evident. This enhancement is vital for their eventual adoption as standard metrics in the field. Thus, future research in CT IQA, alongside other promising research avenues, should focus on developing algorithms with an emphasis on generalization. Such a focus can ensure their applicability and efficacy in a broad spectrum of clinical scenarios.

CRediT authorship contribution statement

Wonkyeong Lee: Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Fabian Wagner:** Writing – review & editing. **Adrian Galdran:** Writing – original draft, Methodology. **Yongyi Shi:** Writing – original draft, Methodology. **Wenjun Xia:** Writing – original draft, Methodology. **Ge Wang:** Writing – original draft, Methodology. **Xuanqin Mou:** Writing – original draft, Methodology. **Md. Atik Ahamed:** Writing – original draft, Methodology. **Abdullah Al Zubaer Imran:** Writing – original draft, Methodology. **Ji Eun Oh:** Writing – original draft, Methodology. **Kyungsang Kim:** Writing – original draft, Methodology. **Jong Tak Baek:** Writing – original draft, Methodology. **Dongheon Lee:** Writing – original draft, Methodology. **Boohwi Hong:** Writing – original draft, Methodology. **Philip Tempelman:** Writing – original draft, Methodology. **Donghang Lyu:** Writing – original draft, Methodology. **Adrian Kuiper:** Writing – original draft, Methodology. **Lars van Blokland:** Writing – original draft, Methodology. **Maria Baldeon Calisto:** Writing – original draft, Methodology. **Scott Hsieh:** Writing – review & editing. **Minah Han:** Methodology. **Jongduk Baek:** Writing – review & editing. **Andreas Maier:** Writing – review & editing. **Adam Wang:** Writing – review & editing. **Garry Evan Gold:** Writing – review & editing. **Jang-Hwan Choi:** Writing – review & editing, Writing – original draft, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jang-Hwan Choi reports financial support was provided by the National Research Foundation of Korea. Jang-Hwan Choi reports was provided by the Korea Medical Device Development Fund grant funded by the Korea government. Adrian Galdran reports financial support was provided by the Marie Sklodowska-Curie Fellowship. If there are other authors, they 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

I have shared the link to my data.

Low-dose Computed Tomography Perceptual Image Quality Assessment Grand Challenge Dataset (MICCAI 2023) (Original data) (Zenodo).

Acknowledgments

This research was partly supported by the National Research Foundation of Korea, South Korea (NRF-2022R1A2C1092072, NRF-2022M3A9I2017587), and by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711174276, RS-2020-KD000016), and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00155966, Artificial Intelligence Convergence Innovation Human Resources Development (Ewha Womans University)). Adrian Galdran was funded by a Marie SkłodowskaCurie Fellowship (No. 892297).

Appendix

A.1. Methodology of the submitted algorithms

This section details the methodologies employed by each participating team, offering insight into the diverse approaches and techniques employed in this challenge. The source code for team agaldran is available at https://github.com/agaldran/ldct_iqa, whereas the source code for the rest of the teams is available at https://github.com/Ewha-AI/LDCTIQAC2023.

A.1.1. agaldran

Team *agaldran* won first place in the challenge. As illustrated in Fig. A.8, the team opted for a multiloss, multihead classifier (Galdran et al., 2023) to yield more calibrated models instead of direct regression to predict image quality. Their model integrates a CNN and transformer-based architecture, combining a swin transformer (Liu et al., 2021) and a BiT-ResNeXt50 (Kolesnikov et al., 2020). Each model's linear classifier at the network end was augmented with four heads. These heads are supervised with diverse cross-entropy losses, each weighted differently, to promote variety in their predictions. Additionally, an RPL (Epstein, 1969; Galdran, 2023) was implemented to train the model further. The predictions from each classification output were averaged and used to minimize the RPL, calculated as the squared L2 distance between the cumulative distributions Y and $\hat{Y}$ corresponding to the ground truth score y and the probabilistic prediction $\hat{y}$ in a K-class problem:

$$RPS(\hat{y}, y) = \frac{1}{K-1} \sum_{i=1}^{K-1} \left[\sum_{j=1}^{i} (\hat{y}_j - y_j) \right]^2 . \tag{A.1}$$

This method accounts for the actual category's distance, making it appropriate for ordinal challenges, such as medical image grading (Galdran, 2023). The number of classes K was set to 21, corresponding to the grading steps from 0 to 4 in increments of 0.2. The final loss is a sum of the weighted cross-entropy loss for each head and the RPL. The output class was scaled down by a factor of 5 to align with the radiologists' score range.

For image quality category predictions, separate probability calculations were performed for the swin transformer and BiT-ResNeXt50 models using five-fold ensembles, with each model undergoing five-fold cross validation using 80% of the data for training in each fold. The models initialized with ImageNet pretrained weights employed standard data augmentation techniques, including flips, rotations, and intensity scaling. An Adam optimizer with a learning rate of $1.3e{-}4$ and a cyclical schedule was used. During inference, horizontal and vertical flips were used as test-time augmentation.

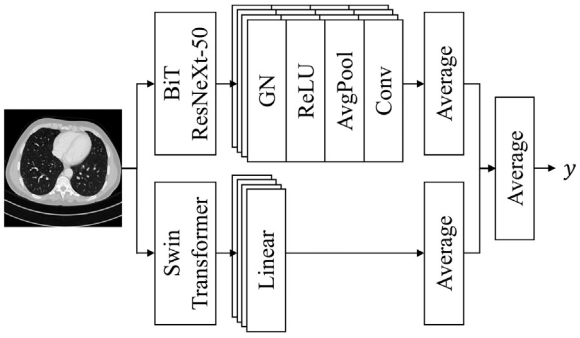


Fig. A.8. Architecture of the method submitted by *agaldran*; GN: group normalization (Wu and He, 2018), ReLU: rectified linear units, AvgPool: average pooling, Conv: convolutional.

A.1.2. RPI_AXIS

Team *RPI_AXIS*, which secured second place in the challenge, implemented a cutting-edge IQA model known as the MANIQA (Yang et al., 2022). This model is acclaimed for winning first place in the NTIRE 2022 Perceptual Image Quality Assessment Challenge Track 2: No-Ref. Gu et al. (2022). The MANIQA model employs three core components to extract multidimensional features: the transposed attention block, scale swin transformer block, and dual-branch structure for patch-weighted quality prediction (see Fig. A.9).

The process begins with the extraction and concatenation of features from four semantic levels, corresponding to the seventh, eighth, ninth, and tenth layers of the 12-layer ViT (Dosovitskiy et al., 2020). The transposed attention block calculates cross-covariance across channels via self-attention, effectively capturing critical information among channels and rearranging the weights to encode the global context. Subsequently, local feature information is processed through the scale swin transformer block, comprising swin transformer layers, a convolutional layer, a scale factor, and a residual connection applied to the final features. The scale factor (set at 0.8) stabilizes the training through a residual connection. The quality score is computed using a dual-branch structure, combining the scores and weights of each patch. This approach enables the model to evaluate quality across image regions.

For training, the model was initialized using ImageNet weights and trained using the Adam optimizer, with the weight decay set at $1e{-}5$. The initial learning rate was $1e{-}5$, and a minibatch size of 8 was used. The means squared error (MSE) loss was the chosen metric for the training procedure, implemented without augmentation strategies. During testing, images were center-cropped to 448 × 448 pixels and resized to 224 × 224 pixels. This cropping method excludes peripheral regions containing air, allowing the model to focus on significant parts corresponding to patients' anatomical sections. No additional ensemble strategies were employed.

A.1.3. CHILL@UK

CHILL@UK won third place in the challenge by developing an end-to-end quality evaluation model for CT images using a state-of-the-art deep learning network. This approach yielded a robust, large-scale model capable of discerning various image details for quality assessment (see Fig. A.10).

The team adapted the EfficientNetV2-L (Tan and Le, 2021) model for regression tasks, making simple modifications to its initial and final layers so that it can process grayscale images and output an IQA score. These changes enabled the model to directly predict image

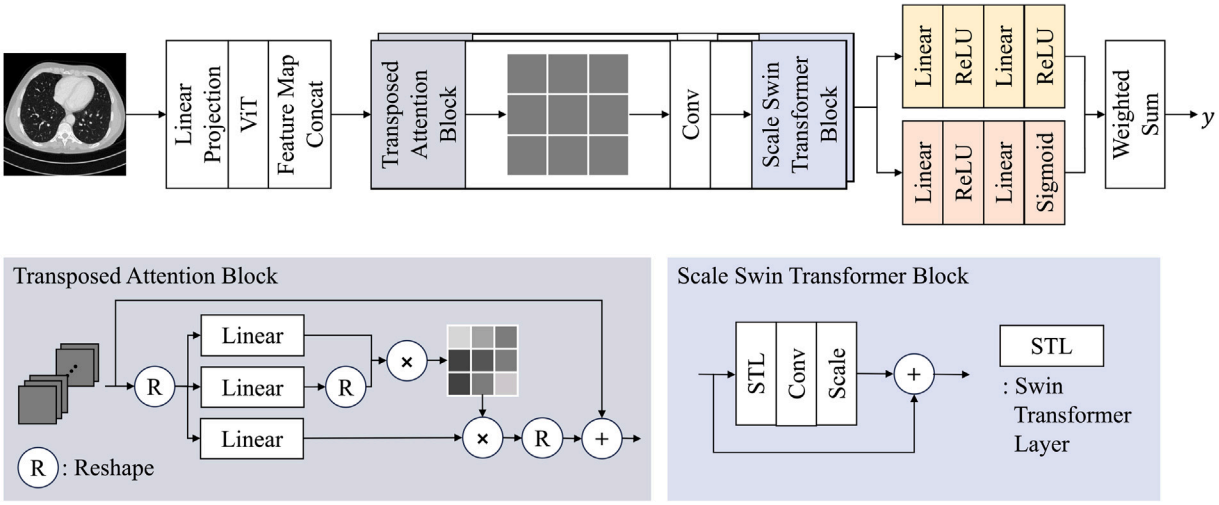


Fig. A.9. Architecture of the MANIQA (Yang et al., 2022) model adopted by *RPI_AXIS*.

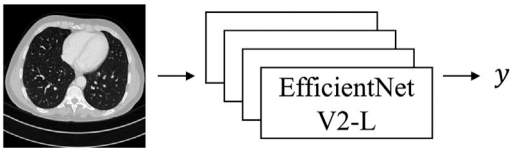


Fig. A.10. Architecture of the EfficientNetV2-L (Tan and Le, 2021) model adopted by *CHILL@UK*, with modifications to the first and last layers. The first layer is adapted for the input of grayscale images, and the last layer is modified to output a single numeric quality score.

quality scores, creating an efficient end-to-end pipeline. The model's first convolutional layer was adjusted to manage single-channel input, and the final layer was redesigned to predict a single image quality score ranging from 0 to 4.

Training was performed for 100 epochs, starting with ImageNet pretrained weights. A batch size of 4 and an initial learning rate of $1e{-}4$ were selected with a cosine annealing scheduler with $T_{max} = 100$ and $eta_{min} = 0$. The Adam optimizer with a weight decay of $1e{-}5$ was used. The MSE loss function was applied for training. During this phase, they applied standard augmentation techniques, such as horizontal and vertical flips and random 90° rotation, each with a 0.5 probability. Training involved a five-fold cross-validation process, with 80% of the data used for training in each fold. For inference, the team ensembled the four best-performing models out of the five EfficientNetV2-L models by averaging their quality scores.

A.1.4. FeatureNet

FeatureNet's development for medical IQA centered on two critical characteristics. First, addressing the data scarcity in training IQA models for LDCT images, they employed transfer learning with a shallow network. Second, recognizing that macro- and microstructures and textural information influence image quality in CT images, they introduced a novel architecture integrating a ViT-based module with a GLCM module to capture texture information effectively for quality assessment.

As depicted in Fig. A.11, the proposed architecture consists of three primary components: a CNN module using the wavelet transform, a visual feature extraction module based on the ViT, and a texture feature extraction module derived from the GLCM. In this setup, a CT image undergoes simultaneous processing by the wavelet-transform-based CNN, ViT, and GLCM modules. The features extracted from these modules are concatenated and passed through a linear layer, culminating in the calculation of the quality score using a rectified linear unit (ReLU) activation function to maintain the model's nonlinearity.

The training of each module is conducted in three stages. Initially, the ViT module, preinitialized with ImageNet weights, is fine-tuned with CT images. A preparatory convolutional layer, followed by batch normalization and the ReLU activation function, adapts these grayscale images into a three-channel format suitable for ViT processing. The second stage involves training a CNN module based on the wavelet transform. After performing the wavelet transform and reducing the images to a quarter of their original size, this module extracts patterns distinct from the original grayscale images using a three-layer CNN. The final stage trains the remaining modules, including the GLCM-based texture feature extraction module, comprising a GLCM computation block, Haralick texture feature (Löfstedt et al., 2019) extraction block, and two fully connected layers.

As a second-order statistical method, GLCM analyzes pixel pair relationships. For each CT image, GLCMs are computed with varying offsets ($\theta = [0, \pi/4, \pi/2,$ and $3\pi/4]$) and distances (1, 3), and five Haralick texture features are extracted: contrast, homogeneity, dissimilarity, correlation, and energy. These features provide quantitative measurements of local variations, pixel value similarities, intensity distribution differences, linear interdependence between pixels, and image regularity. The extracted features from the three modules are concatenated and further processed through additional linear layers to predict the final quality score after a ReLU activation function.

Across the three-stage training process, consistent training options were maintained, with the learning rate as the primary variable. The network employed a batch size of 10 and the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The learning rate was initially set at 3×10^{-5}, decaying by 95% every 10 epochs in the first stage. In the second stage, the learning rate started at 2×10^{-5} and followed a similar decay pattern. The MSE loss was used throughout to regress the quality score, with the best model selected based on the MSE, PLCC, SROCC, and KROCC metrics. Instead of relying on augmentation strategies or ensemble methods, the team focused on extracting and fusing multiple features to predict the final score.

A.1.5. Team epoch

Team Epoch developed a two-stage model comprising two EDCNN models (Liang et al., 2020), specifically targeting noise features. The choice of EDCNN was due to its simple structure and proven efficiency in the denoising task with the AAPM dataset (Moen et al., 2021). Furthermore, the team embraced a multitask design due to its efficiency and aimed to build a parameter-efficient model by freezing the first EDCNN model. To enhance robustness and generalization, they

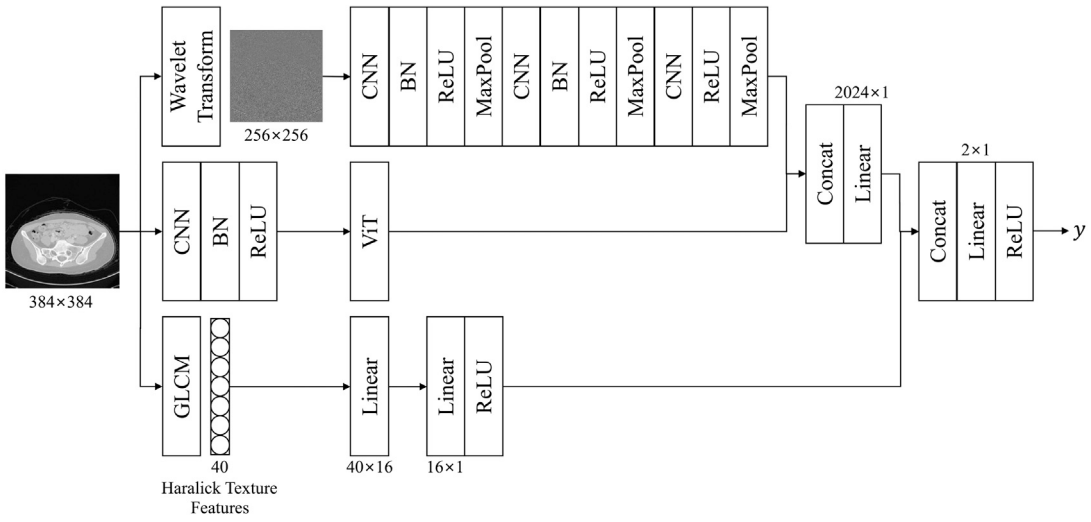


Fig. A.11. Architecture of the model from *FeatureNet*; BN: batch normalization (Ioffe and Szegedy, 2015).

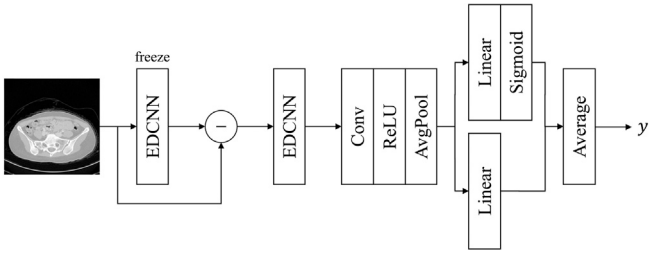


Fig. A.12. Architecture from Team Epoch, built on the edge enhancement-based densely connected convolutional neural network (EDCNN) (Liang et al., 2020).

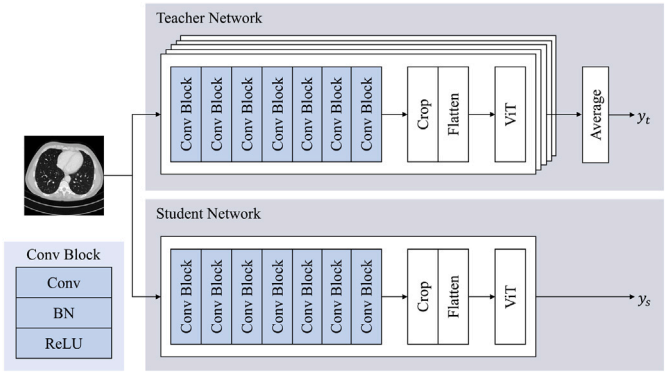


Fig. A.13. Architecture from *gabybaldeon* employing knowledge distillation; BN: batch normalization (Ioffe and Szegedy, 2015).

implemented heavy data augmentation techniques, such as the mosaic method (Bochkovskiy et al., 2020) (see Fig. A.12).

In the first stage, the team used a pretrained EDCNN model with the AAPM dataset to extract noise and artifact features. The model weights were frozen during training. The output of this stage, when subtracted from the original input, yielded an artifact map rich in noise features. This artifact map was processed in the second stage, another EDCNN model followed by additional layers: a convolutional layer to increase the channel number from 1 to 32, adaptive average pooling, and two output branches for regression and classification. The regression branch used a sigmoid activation function to constrain the output to between 0 and 1 and later scaled up to a maximum of 4. The classification branch did not use an activation function due to the subsequent application of cross-entropy loss, and the class number was set to 21, comprising a list of numbers from 0 to 4 in increments of 0.2. The final image quality score was determined by averaging the outputs of these two branches.

Team Epoch's approach included using the mosaic augmentation method to create 5250 data samples from the original 1000, assembling images with identical scores to form new samples. Of these, 5000 original and mosaic-generated images were used for training, with the remainder for testing. Additional augmentation strategies included shuffling and mixing to disrupt the spatial structure of the image, enhancing the model's ability to extract noise features, along with standard methods, such as random flipping and rotation. Notably, resizing and cropping were avoided to preserve image information.

The training employed the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay $1e{-}3$), with an initial learning rate of $2e{-}4$ and a cosine annealing scheduler including a warm-up period following the formula $2/(1-\beta_2)$, and a minimum learning rate of $1e{-}6$. The Huber and cross-entropy losses were used for the regression and classification branches, respectively. The model underwent 600 epochs of training with a batch size of 4, selecting the model weights with the highest PLCC, SROCC, and KROCC values for testing. For inference, three models trained under various settings were ensembled for quality calculation.

A.1.6. gabybaldeon

Team *gabybaldeon* introduced a ViT-based network specifically for IQA of CT images, employing knowledge distillation from an ensemble of ViT networks. This process is divided into two main steps (see Fig. A.13).

In the initial step, an ensemble of five separate ViT networks, collectively known as the teacher network, is trained on a CT imaging dataset. This training is accomplished using a five-fold division technique to ensure comprehensive learning. In the next step, a single ViT network, designated as the student network, undergoes training. It acquires knowledge from the teacher network, where the teacher's predictions are used as new labels alongside the original labeled dataset. During evaluations, only the student network is deployed for inferences.

The proposed ViT architecture comprises two principal components: a convolutional stem and the standard ViT network. The process begins with the input image passing through the convolutional stem, containing seven convolutional blocks. Each block consists of a convolutional layer, batch normalization layer, and ReLU activation function. The first six blocks use 3×3 kernels with a stride of 1, and the number of filters progressively doubles from 8 to 256. Differing slightly, the

Table A.9
Weighted correlation scores for each algorithm. The weights, ranging from 0 to 1, were calculated by subtracting the standard deviation of radiologists' assessment scores for each sample from 1.

Team	Overall	PLCC	SROCC	KROCC
agaldran	2.7676	0.9547	0.9551	0.8578
RPI_AXIS	2.7065	0.9482	0.9466	0.8117
CHILL@UK	2.7005	0.9457	0.9461	0.8086
FeatureNet	2.6789	0.9415	0.9395	0.7979
Team Epoch	2.6498	0.9363	0.9302	0.7833
gabybaldeon	2.5874	0.9205	0.9134	0.7535

final block employs a 2×2 kernel and a stride of 2 and produces three output channels, halving the feature map size. Next, a center crop to (3, 224, 224) aligns the feature map with the pretrained ViT, which is flattened and input into the traditional ViT network. The architecture culminates in a linear function with a single output node for predicting image quality.

Both the teacher and student networks follow identical training protocols. They used the stochastic gradient descent optimization with a momentum of 0.9 and an initial learning rate of $1e{-}3$, adjusted via the cosine annealing scheduler. The training extends for 1000 epochs with a batch size of 10. Data augmentation techniques, including horizontal and vertical flipping, random rotation, transposition, and scaling, are employed during training. The teacher networks are trained using a joint L1 and L2 loss function between the model prediction y_t and the ground-truth score y_g:

$$L_{reg}(y_t, y_g) = (y_t - y_g)^2 + |y_t - y_g|. \tag{A.2}$$

In the distillation phase, the training of the student network involves the same loss function applied between its predictions (y_s) and the teacher network predictions (y_t) and between its predictions and the ground-truth scores with a weight of 0.5:

$$L_{total}(y_s, y_t, y_g) = L_{reg}(y_s, y_g) + 0.5 L_{reg}(y_s, y_t). \tag{A.3}$$

A.2. Consideration of uncertainties

In this paper, we did not consider the uncertainty for individual samples. However, certain samples exhibit disagreement among radiologists regarding quality scores. Therefore, it is crucial to account for the uncertainty of individual samples when calculating correlation scores. Specifically, samples with high uncertainties should be weighted less compared to those with low or zero uncertainty.

To address this, we first calculated the uncertainty for each sample by determining the standard deviation of the image quality scores given by each radiologist. Upon analyzing all test images, the minimum standard deviation was zero, indicating unanimous agreement among radiologists, while the maximum value was one, as shown in Fig. 3. Subsequently, we defined the importance weight by subtracting the uncertainty value from one. Consequently, samples with zero uncertainty receive an importance weight of one, and those with the highest uncertainty value of one receive a weight of zero in the correlation score calculations. Using these uncertainties and their associated weights, we computed weighted versions of the PLCC, SROCC (Bailey et al., 2018), and KROCC (Shieh, 1998). Table A.9 displays the resulting scores for each algorithm.

References

AAPM, 2017. Low dose CT grand challenge. URL: https://www.aapm.org/GrandChallenge/LowDoseCT/.

Abbey, C.K., Barrett, H.H., 2001. Human-and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. J. Opt. Soc. Amer. A 18 (3), 473–488.

Abbey, C.K., Bochud, F., 2000. Modeling visual detection tasks in correlated image noise with linear model observers. In: Handbook of Medical Imaging. Vol. 1, SPIE press, Bellingham, Washington, pp. 629–654.

Arndt, S., Turvey, C., Andreasen, N.C., 1999. Correlating and predicting psychiatric symptom ratings: Spearmans r versus Kendalls tau correlation. J. Psychiatr. Res. 33 (2), 97–104.

Bailey, P., Emad, A., Zhang, T., Xie, Q., Sikali, E., 2018. Weighted and Unweighted Correlation Methods for Large-Scale Educational Assessment: wCorr Formulas. AIR–NAEP Working Paper No. 2018-01, In: NCES Data R Project Series# 02, American Institutes for Research.

Barrett, H.H., Yao, J., Rolland, J.P., Myers, K.J., 1993. Model observers for assessment of image quality. Proc. Natl. Acad. Sci. 90 (21), 9758–9765.

Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Boone, J.M., Hendee, W.R., McNitt-Gray, M.F., Seltzer, S.E., 2012. Radiation exposure from CT scans: how to close our knowledge gaps, monitor and safeguard exposure—proceedings and recommendations of the Radiation Dose Summit, sponsored by NIBIB, February 24–25, 2011. Radiology 265 (2), 544–554.

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., Samek, W., 2018. Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans. Image Process. 27 (1), 206–219. http://dx.doi.org/10.1109/TIP.2017.2760518.

Burgess, A., 1994. Statistically defined backgrounds: performance of a modified nonprewhitening observer model. J. Opt. Soc. Amer. A 11 (4), 1237–1242.

Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., Wang, G., 2017. Low-dose CT denoising with convolutional neural network. In: 2017 IEEE 14th International Symposium on Biomedical Imaging. ISBI 2017, pp. 143–146. http://dx.doi.org/10.1109/ISBI.2017.7950488.

Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P., 2018. Development and validation of deep learning algorithms for detection of critical findings in head CT scans. arXiv preprint arXiv:1803.05854.

Chow, L.S., Paramesran, R., 2016. Review of medical image quality assessment. Biomed. Signal Process. Control 27, 145–154.

Chow, L.S., Rajagopal, H., Paramesran, R., Initiative, A.D.N., et al., 2016. Correlation between subjective and objective assessment of magnetic resonance (MR) images. Magn. Reson. Imaging 34 (6), 820–831.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.-A., 2018. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. arXiv preprint arXiv:1804.10916.

Epstein, E.S., 1969. A scoring system for probability forecasts of ranked categories. J. Appl. Meteorol. (1962-1982) 8 (6), 985–987.

Fahrig, R., Dixon, R., Payne, T., Morin, R.L., Ganguly, A., Strobel, N., 2006. Dose and image quality for a cone-beam C-arm CT system. Med. Phys. 33 (12), 4541–4550.

Galdran, A., 2023. Performance metrics for probabilistic ordinal classifiers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 357–366.

Galdran, A., Verjans, J.W., Carneiro, G., González Ballester, M.A., 2023. Multi-head multi-loss model calibration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 108–117.

Gholizadeh-Ansari, M., Alirezaie, J., Babyn, P., 2020. Deep learning for low-dose CT denoising using perceptual loss and edge detection layer. J. Digit. Imaging 33 (2), 504–515.

Gong, H., Yu, L., Leng, S., Dilger, S.K., Ren, L., Zhou, W., Fletcher, J.G., McCollough, C.H., 2019. A deep learning-and partial least square regression-based model observer for a low-contrast lesion detection task in CT. Med. Phys. 46 (5), 2052–2063.

Gu, J., Cai, H., Dong, C., Ren, J.S., Timofte, R., Gong, Y., Lao, S., Shi, S., Wang, J., Yang, S., et al., 2022. NTIRE 2022 challenge on perceptual image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 951–967.

Han, M., Baek, J., 2020. A convolutional neural network-based anthropomorphic model observer for signal-known-statistically and background-known-statistically detection tasks. Phys. Med. Biol. 65 (22), 225025.

Haralick, R.M., Shanmugam, K., Dinstein, I.H., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. (6), 610–621.

Hosu, V., Lin, H., Sziranyi, T., Saupe, D., 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Trans. Image Process. 29, 4041–4056.

Huh, J., Khan, S., Choi, S., Shin, D., Lee, J.E., Lee, E.S., Ye, J.C., 2023. Tunable image quality control of 3-D ultrasound using switchable CycleGAN. Med. Image Anal. 83, 102651.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pmlr, pp. 448–456.

Jinjin, G., Haoming, C., Haoyu, C., Xiaoxing, Y., Ren, J.S., Chao, D., 2020. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer, pp. 633–651.

Kak, A.C., Slaney, M., 2001. Principles of Computerized Tomographic Imaging. SIAM.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25. Springer, pp. 597–609.

Kim, W., Lee, J., Choi, J.-H., 2023. An unsupervised two-step training framework for low-dose computed tomography denoising. Med. Phys..

Kim, B., Shim, H., Baek, J., 2021. Weakly-supervised progressive denoising with unpaired CT images. Med. Image Anal. 71, 102065.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N., 2020. Big transfer (bit): General visual representation learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer, pp. 491–507.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (2), 155–163.

Kowalski, C.J., 1972. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. J. R. Stat. Soc. Ser. C. Appl. Stat. 21 (1), 1–12.

Lee, W., Cho, E., Kim, W., Choi, H., Beck, K.S., Yoon, H.J., Baek, J., Choi, J.-H., 2022. No-reference perceptual CT image quality assessment based on a self-supervised learning framework. Mach. Learn.: Sci. Technol. 3 (4), 045033.

Lee, W., Wagner, F., Maier, A., Wang, A., Baek, J., Hsieh, S.S., Choi, J.-H., 2023. Low-dose computed tomography perceptual image quality assessment grand challenge 2023. http://dx.doi.org/10.5281/zenodo.7861744.

Li, M., Hsu, W., Xie, X., Cong, J., Gao, W., 2020. SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network. IEEE Trans. Med. Imaging 39 (7), 2289–2301.

Liang, T., Jin, Y., Li, Y., Wang, T., 2020. Edcnn: Edge enhancement-based densely connected network with compound loss for low-dose ct denoising. In: 2020 15th IEEE International Conference on Signal Processing. ICSP, IEEE, pp. 193–198.

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.

Löfstedt, T., Brynolfsson, P., Asklund, T., Nyholm, T., Garpebring, A., 2019. Gray-level invariant Haralick texture features. PLoS One 14 (2), e0212110.

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

Maier, A., Hofmann, H.G., Berger, M., Fischer, P., Schwemmer, C., Wu, H., Müller, K., Hornegger, J., Choi, J.-H., Riess, C., et al., 2013. CONRAD—A software framework for cone-beam imaging in radiology. Med. Phys. 40 (11), 111914.

Mason, A., Rioux, J., Clarke, S.E., Costa, A., Schmidt, M., Keough, V., Huynh, T., Beyea, S., 2019. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images. IEEE Trans. Med. Imaging 39 (4), 1064–1072.

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

Mittal, A., Soundararajan, R., Bovik, A.C., 2012b. Making a "completely blind" image quality analyzer. IEEE Signal Process. Lett. 20 (3), 209–212.

Moen, T.R., Chen, B., Holmes III, D.R., Duan, X., Yu, Z., Yu, L., Leng, S., Fletcher, J.G., McCollough, C.H., 2021. Low-dose CT image and projection dataset. Med. Phys. 48 (2), 902–911.

Nefzger, M.D., Drasgow, J., 1957. The needless assumption of normality in Pearson's r. Am. Psychol. 12 (10), 623.

Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D., 2023. Causality-inspired single-source domain generalization for medical image segmentation. IEEE Trans. Med. Imaging 42 (4), 1095–1106. http://dx.doi.org/10.1109/TMI.2022.3224067.

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

Pooch, E.H., Ballester, P., Barros, R.C., 2020. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In: Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2. Springer, pp. 74–83.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Scherl, H., Keck, B., Kowarschik, M., Hornegger, J., 2007. Fast GPU-based CT reconstruction using the common unified device architecture (CUDA). In: 2007 IEEE Nuclear Science Symposium Conference Record. IEEE, pp. 4464–4466.

Schober, P., Boer, C., Schwarte, L.A., 2018. Correlation coefficients: appropriate use and interpretation. Anesth. Analg. 126 (5), 1763–1768.

Schuppert, C., Rospleszcz, S., Hirsch, J.G., Hoinkiss, D.C., Köhn, A., von Krüchten, R., Russe, M.F., Keil, T., Krist, L., Schmidt, B., et al., 2023. Automated image quality assessment for selecting among multiple magnetic resonance image acquisitions in the German National Cohort study. Sci. Rep. 13 (1), 22745.

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

Shieh, G.S., 1998. A weighted Kendall's tau statistic. Statist. Probab. Lett. 39 (1), 17–24.

Sodickson, A., Baeyens, P.F., Andriole, K.P., Prevedello, L.M., Nawfel, R.D., Hanson, R., Khorasani, R., 2009. Recurrent CT, cumulative radiation exposure, and associated radiation-induced cancer risks from CT of adults. Radiology 251 (1), 175–184.

Solomon, J., Samei, E., 2016. Correlation between human detection accuracy and observer model-based image quality metrics in computed tomography. J. Med. Imaging 3 (3), 035506.

Tan, M., Le, Q., 2021. Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning. PMLR, pp. 10096–10106.

Vallat, R., 2018. Pingouin: statistics in Python. J. Open Source Softw. 3 (31), 1026.

Verdun, F., Racine, D., Ott, J., Tapiovaara, M., Toroi, P., Bochud, F., Veldkamp, W., Schegerer, A., Bouwman, R., Giron, I.H., et al., 2015. Image quality in CT: From physical measurements to model observers. Phys. Medica 31 (8), 823–843.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612. http://dx.doi.org/10.1109/TIP.2003.819861.

Wu, Y., He, K., 2018. Group normalization. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Wu, Q., Ji, X., Gu, Y., Xiang, J., Quan, G., Li, B., Zhu, J., Coatrieux, G., Coatrieux, J.-L., Chen, Y., 2023. Unsharp structure guided filtering for self-supervised low-dose CT imaging. IEEE Trans. Med. Imaging.

Wu, J., Ma, J., Liang, F., Dong, W., Shi, G., Lin, W., 2020. End-to-end blind image quality prediction with cascaded deep neural network. IEEE Trans. Image Process. 29, 7414–7426. http://dx.doi.org/10.1109/TIP.2020.3002478.

Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y., 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1191–1200.

Yao, J., Barrett, H.H., 1992. Predicting human performance by a channelized Hotelling observer model. In: Mathematical Methods in Medical Imaging. SPIE, pp. 161–168.