Objective quality assessment of medical images and videos: review and challenges

Rafael Rodrigues, Lucie Lévêque, Jesús Gutiérrez, Houda Jebbari, Meriem Outtas, Lu Zhang, Aladine Chetouani, Shaymaa Al-Juboori, Maria G. Martini, Antonio M G Pinheiro

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20292-x

Objective quality assessment of medical images and videos: review and challenges

Rafael Rodrigues[1] · Lucie Lévêque[2] · Jesús Gutiérrez[3] · Houda Jebbari[4] · Meriem Outtas[5] · Lu Zhang[5] · Aladine Chetouani[6,7] · Shaymaa Al-Juboori[8] · Maria G. Martini[8] · Antonio M. G. Pinheiro[1]

Received: 13 December 2022 / Revised: 24 July 2024 / Accepted: 19 September 2024

Abstract
Quality assessment is a key element for the evaluation of hardware and software involved in image and video acquisition, processing, and visualization. In the medical field, user-based quality assessment is still considered more reliable than objective methods, which allow the implementation of automated and more efficient solutions. Regardless of increasing research on this topic in the last decade, defining quality standards for medical content remains a non-trivial task, as the focus should be on the diagnostic value assessed by expert viewers rather than the perceived quality from naïve viewers, and objective quality metrics should aim at estimating the first rather than the latter. In this paper, we present a survey of methodologies used for the objective quality assessment of medical images and videos, dividing them into visual quality-based and task-based approaches. Visual quality-based methods compute a quality index directly from visual attributes, while task-based methods, being increasingly explored, measure the impact of quality impairments on the performance of a specific task. A discussion on the limitations of state-of-the-art research on this topic is also provided, along with future challenges to be addressed.

Keywords Quality assessment · Objective metrics · Medical imaging · Task-based quality

✉ Rafael Rodrigues
rafael.rodrigues@ubi.pt

1 Instituto de Telecomunicações and Universidade da Beira Interior, Covilhã, Portugal

2 Nantes Laboratory of Digital Sciences (LS2N), Nantes University, Nantes, France

3 Universidad Politécnica de Madrid, Madrid, Spain

4 INSERM, LTSI - UMR 1099, University of Rennes, Rennes, France

5 National Institute of Applied Sciences (INSA), Rennes, France

6 PRISME Laboratory, University of Orléans, Orléans, France

7 Laboratoire L2TI - Galilée, Univesity Sorbonne Paris Nord, Villetaneuse, France

8 Kingston University, London, UK

Published online: 17 October 2024

1 Introduction

Medical images and videos provide clinical information from the human body with reduced invasiveness, but also structural and functional outcomes that could not be obtained by other means. Nowadays, medical imaging plays an inexorable part in diagnosis, treatment planning, and patient monitoring. Over the last decades, medical imaging techniques have been constantly developed, updated, and extensively used in numerous medical specialties. The World Health Organization (WHO) estimated around 3.6 billion diagnostic procedures performed worldwide, per year, between 1997 and 2007 [1]. Radiology is the leader in the production of medical imaging content [2], with a vast number of imaging modalities. Data from the Organization for Economic Cooperation and Development (OECD), which considers Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET) scans performed each year, further reveals a clear upward trend in the last decade, with over 253 million exams reported by OECD countries [3]. Moreover, and besides other relevant imaging modalities, medical images and videos are now also transmitted in real time for telemedicine applications. Therefore, large amounts of content from different acquisition methods are continuously created in the medical practice.

Several impairments can affect the quality of the end visual signal, impacting the quality perceived by the viewers (e.g., clinicians, radiologists, etc.), as well as their performance on clinical tasks. This spectrum of impairments highly depends on a wide variety of acquisition and reconstruction-related factors, which are often specific to each imaging modality. Furthermore, external and patient-related factors may induce artifacts in the acquired content, as well. Images and videos may also be subject to different processing, encoding/compression, transmission, and visualization methodologies. For example, in telemedicine applications, video content and real-time interactions are key features, which lead to a high demand of hardware resources and network bandwidth for data storage and transmission [4, 5]. Thus, measuring image and video quality in health applications is both a necessity, towards improving methodologies throughout the clinical workflow, and a wide open challenge, given the diversity of content, applications, and impairments. This paper aims at providing researchers with a broad sense of current boundaries and future opportunities in this field. The authors of this review have been involved in the field of QA, with relevant publications on medical image and video QA. Prior knowledge of research in this domain served as the basis for the considered selected papers.

While a recent review of subjective QA of medical images and videos was published by Lévêque et al. [6], this paper focuses on objective QA methods for medical images and videos. Objective QA relies on the use of image processing and analysis algorithms towards automated quality estimation methods [7]. Although objective methods may be less reliable than subjective assessment, they are more cost-effective, less time-consuming, and reduce observer variability and bias [8]. In another review paper published in 2016, Chow et al. discussed both subjective and objective methods [9], with a primary focus on MRI, CT, and ultrasound imaging. More recently, Raj et al. [10] presented a survey of objective QA methods for fundus images. In this paper, we focus mainly on more recent research, and consider several other imaging modalities than those previous reviews.

Another contribution of this paper is the proposal of a new categorization for medical image and video QA methods, based on their design principle, into visual quality-based and task-based methods. Visual quality-based methods mainly use a traditional approach to quality assessment, which focuses on visual and/or structural attributes of the image. On the other hand, task-based quality assessment targets a specific application of the visual content

in the medical domain [11]. It should be emphasized that medical images may have lower perceptual quality due to acquisition factors but keep all the needed information from a clinical point of view. Some research efforts that fit into this category had already referred to this approach as task-based, e.g., [12–15]. This is a core contribution of this paper and, to the best of our knowledge, it is the first review to formalize this discrimination in two distinct QA categories. Regarding this topic, an overview of model observers used in task-based QA of medical images was published in 2014 by Zhang et al. [16]. Similarly to the reviews mentioned in the last paragraph, this paper focuses on more recent research not included there.

The remainder of the paper is organized as follows: Section 2 presents a brief overview of image and video quality assessment and discusses the main differences between medical imaging and other types of visual content; Sections 3 and 4 provide a review of visual quality-based and task-based methods, respectively; Section 5 provides a discussion on some merits and drawbacks of the reported methods, as well as future research directions; Finally, Section 6 concludes the paper and summarizes the main insights of this contribution.

2 Overview of image and video quality assessment

The quality perceived by the users of image and video content is one of several factors influencing their Quality of Experience (QoE), defined by EU Cost Action 1003 Qualinet [17] as "the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations concerning the utility and enjoyment of the application or service in the light of the user's personality and current state".

Typically, image and video quality is evaluated from a perceptual perspective, either subjectively or objectively, with each approach bearing its own motivations, advantages, and drawbacks. Subjective QA relies on human observers to analyze images or videos and rate their quality following specific methodologies (e.g., [18, 19]). The output of subjective tests is then statistically analyzed, usually by computing the Mean Opinion Score (MOS) or Differential MOS (DMOS) [7]. Depending on the type of content and application, user expertise may be a requirement for an accurate and meaningful evaluation. In most healthcare applications, the quality perception of end-users is likely to be strongly influenced by the clinical utility of the content, rather than strictly aesthetic criteria. Therefore, subjective QA methods may have some limitations in this context, concerning the availability of expert observers, in addition to the inherent intra- and inter-subject scoring variability [8].

The most common design principle in objective image and video QA applications is the computation of a quality index from certain visual and/or structural attributes of the content. On the contrary, task-based QA methods assess the quality in terms of a specific goal of the content by measuring its influence on the performance of certain tasks (e.g., diagnosis or localization of an anatomical structure [13]). This kind of approach has been increasingly explored in the development of quality metrics that are specific for medical images and video. Consequently, we propose to divide the discussed QA methodologies into two major categories, i.e., visual quality-based, and task-based methods, as described in Fig. 1.

The performance of objective methods is usually evaluated through a statistical comparison with subjective results [18, 20]. More precisely, the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-order Correlation Coefficient (SROCC) are used to measure the linear and rank correlations, respectively. A less commonly used rank correlation coefficient is the Kendall Rank Correlation Coefficient (KRCC). These coefficients can take absolute values from 0 to 1, with 1 indicating a perfect correlation. Negative coefficients indicate an inverse correlation between the variables.

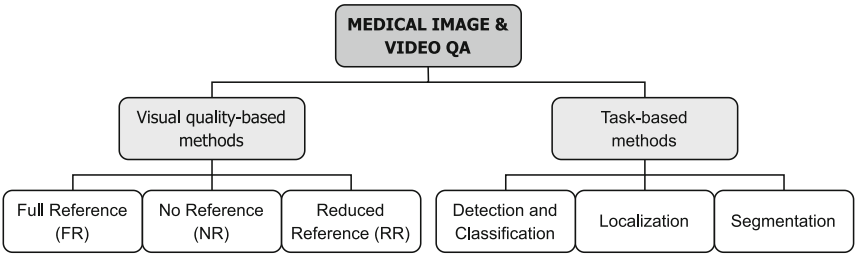

Fig. 1 Categorization of quality assessment methods for medical image and video. As proposed, these are divided into visual quality-based and task-based methods

Metric performance can also be evaluated from a classification standpoint, e.g., if a metric aims at classifying an image as being acceptable or non-acceptable. Accuracy is commonly defined as the ratio of correctly classified instances according to a subjectively defined ground truth over the total number of instances and is typically presented in percentage. Another common performance indicator is the area under the Receiver Operating Characteristic (ROC) curve. For a given classification model, the ROC curve plots the true positive rate (TPR) as a function of the false positive rate (FPR) for a number of candidate decision thresholds. Both the TPR and FPR may vary from 0 to 1, as does the resulting area under the ROC curve (AUC). An AUC of 0.5 is equivalent to random guessing, whereas a value of 1 indicates a perfect performance in classification.

3 Visual quality-based methods

Visual quality-based methods are commonly categorized as Full Reference (FR), No Reference (NR), or Reduced Reference (RR), depending on the availability of an undistorted reference (preferably the original image or video). FR metrics predict the quality by directly comparing the reference and its distorted versions, contrary to NR metrics, which assess the quality only using the distorted image. Finally, RR metrics compare features representative of the distorted and reference images.

The following subsections present an overview of visual quality-based objective QA methods for medical images and video, according to the aforementioned categorization.

3.1 Full-reference approaches

Full-reference QA of medical images and videos had focused almost exclusively on adapting metrics designed for non-medical content to obtain quality predictions for compressed and/or artificially distorted medical content. Table 1 summarizes the works presented in this section.

3.1.1 Magnetic resonance imaging

FR quality assessment studies using MRI includes the works of Chow et al. [21] and Mason et al. [8]. Both works tested a large set of metrics, e.g. [33–38] on datasets with different

Table 1 Overview of visual quality-based approaches to the quality assessment of medical images and video (FR metrics)

Reference	Imaging modality	Quality impairments / Processing	Validation	Objective metrics
[21]	MRI	Rician and Gaussian noise, and Gaussian blur (5 levels); DCT, JPEG and JPEG2000 compression (5 ratios)	MOS: 4 medical observers (SDSCE)	SNR, PSNR, SSIM, MSSIM, FSIM, ICF, NQM, WSNR, VIF, VIFP, UQI, IW-PSNR, IW-SSIM
[8]	MRI	White and Rician noise, Gaussian blur, motion, undersampling and wavelet compression	MOS: 5 radiologists	RMSE, PSNR, SSIM, MS-SSIM, IW-SSIM, GMSD, FSIM, HDRVDP, NQM, VIF
[22]	CT	Compression with DCT and Wavelet using different compression ratios	MOS: Common persons, clinic doctors, and radiologists	PSNR, SSIM, distortion measure from [23]
[24]	CT	JPEG and JPEG2000 compression (5 ratios)	MOS: 6 radiologists (Double-stimulus DCR)	MSE, local MSE, SNR, SSIM, VSNR, VIF
[25]	Ultrasound videos	H.264 Compression with different quantization parameters and packet-loss rates	MOS: 2 medical experts	PSNR, SSIM, VSNR, VIF, VIFP, IFC, NQM, WSNR
[26]	Ultrasound videos	HEVC compression (8 quantization levels)	MOS: 4 medical experts and 16 naïve observers (DSCQS)	PSNR, SSIM, UQI, VQM, NQM, VIF, VSNR
[27]	Ultrasound videos	HEVC compression (8 quantization levels)	DMOS: 4 medical experts (DSCQS)	PSNR, SSIM, UQI, VQM, NQM, VIF, VSNR, CUQI (proposed metric)

Table 1 continued

Reference	Imaging modality	Quality impairments / Processing	Validation	Objective metrics
[28]	Laparoscopic surgery videos	H.264 compression (4 bit rates)	MOS: 9 laparoscopic surgeons and 16 naïve observers (SSCQE)	VQM, HDR-VDP-2, PSNR
[5]	Endoscopic surgery videos	H.264 compression (11 ratios)	MOS: 14 medical observers (DSCQS)	SSIM, UQI, PSNR, WSNR, VSNR, HDR-VDP, IFC, MSE, MS-SSIM, PSNR-HVS, PSNR-HVS-M, VIF, VIFP, NIQE* and BRISQUE*
[29]	Endoscopic videos	HEVC compression (8 quantization levels)	MOS: 6 medical observers and 19 naïve observers (DSCQS)	MSE, PSNR, SSIM, MS-SSIM, VSNR, IFC, VIF, VIFP, UQI, NQM, WSNR
[30]	CT, MRI	Varying bitrates (from 0.1 to 2.0) using SPIHT compression	MOS: 6 medical observers	PSNR, SSIM
[31]	CT, MRI, Ultrasound	Varying bitrates (from 0.02 to 2.0) using SPIHT compression	MOS: 5 medical observers	PSNR, SSIM, proposed SSIM variation
[32]	X-ray, MRI	Gaussian noise and Gaussian blur (5 levels); JPEG and JPEG2000 compression (5 ratios)	MOS: 3 radiologists (Double-stimulus DCR)	16 combinations of SSIM components (multiscale, gradient-based and structural component)

In some works, NR metrics were also tested, which are marked with *

compression settings (Discrete Cosine Transform (DCT) [39], JPEG [40], JPEG2000 [41], wavelets) and simulated artifacts (e.g. Rician noise, Gaussian blur or motion artifacts). The best average correlations with the subjective scores were achieved with Noise Quality Measure (NQM), with PLCC and SROCC around 0.94 [21], and Visual Information Fidelity (VIF) [8].

3.1.2 Computed tomography

In an early work from 2003, Zhou et al. [22] estimated the quality of CT images with both DCT and Wavelet compression, using a back-propagation neural network with Peak Signal-to-noise Ratio (PSNR) and Structural Similarity index (SSIM), and the Block-wise Distortion Measure [23] as inputs. The reported agreement rates with the obtained subjective scores ranged between 85.71% and 96.88%. Kowalik et al. [24] also used CT images with compression - JPEG and JPEG2000, and studied the ROC curves of SSIM, Signal-to-noise Ratio (SNR), VIF, Mean Squared Error (MSE), and Visual Signal-to-noise Ratio (VSNR) [42] in classifying images as having non-noticeable or non-acceptable distortions, after subjective annotation by experts. The largest AUC was obtained with SSIM (0.99 for brain images and 0.96 for body images), closely followed by VIF.

3.1.3 Ultrasonography

In [25], the authors evaluated the performance of a H.264 encoding framework for atherosclerotic plaque ultrasound videos, which resulted in enhanced performance in noisy environments. WSNR obtained the best correlation with the subjective scores (PLCC = 0.69 and SROCC = 0.72). The authors also proposed minimum settings for several parameters, including the frame rate, bit rate, and PSNR in the Regions of Interest (ROI).

The authors of [26] and [27] assessed the quality of ultrasound video excerpts compressed with High Efficiency Video Coding (HEVC) [43], using a set of 7 common FR metrics, as well as the proposed content-specific Cardiac Ultrasound Video Quality Index (CUQI), in [27]. CUQI estimates the diagnostic quality of cardiac ultrasound video using motion and edge information. In terms of correlation with DMOS, all metrics yielded good PLCC and SROCC coefficients in both studies. In [27], CUQI outperformed the other tested metrics, with PLCC and SROCC after DMOS nonlinear regression of 0.94 and 0.93, respectively. Nonetheless, all the tested metrics achieved correlations above 0.9, and SSIM achieved the best result in terms of PLCC without nonlinear regression of DMOS.

3.1.4 Endoscopic / laparoscopic videos

In [28] and [5], the authors performed objective QA of laparoscopic and surgical videos, respectively, compressed with H.264. SSIM, High Dynamic Range Visible Difference Predictor v2 (HDR-VDP-2), and Video Quality Metric (VQM) were used in both studies, but a larger set of metrics was used in [5], including, for example, PSNR-HVS and PSNR-HVS-M [44], MSE, Multi-scale SSIM (MS-SSIM) [45], and two NR metrics - Natural Image Quality Evaluator (NIQE) [46] and Blind/Referenceless Image Spatial Quality Elevator (BRISQUE) [47]. In the latter study, MSE, SSIM, MS-SSIM, and BRISQUE achieved good correlations with MOS (PLCC $>$ 0.9).

Usman et al. [29] were the first to investigate the quality of wireless capsule endoscopy videos. Videos were compressed with HEVC, and 10 metrics were computed to estimate video quality. VIF, Pixel Visual Information Fidelity (VIFP), and Information Fidelity Criterion

(IFC) showed good correlation with the obtained MOS (between 0.88 and 0.92), with VIFP outperforming the other two, both in statistical significance and in computation time.

3.1.5 Multiple imaging modalities

Other interesting works were published in [30] and [31], where the authors studied the performance of PSNR and SSIM on CT, MRI, and ultrasound images with Set Partitioning in Hierarchical Trees (SPIHT) compression [48]. Both PSNR and SSIM quality estimations correlated well with subjective scores. In [31], Kumar et al. proposed an improved version of SSIM, which outperformed PSNR with correlation coefficients of 0.99 (MRI), 0.98 (CT), and 0.98 (ultrasound).

Several variations of SSIM were also tested by Renieblas et al. [32], to estimate the quality of planar X-ray and MRI, with simulated Gaussian blur and noise, and JPEG and JPEG2000 compression. The best overall correlation with subjective scores was obtained with 4-MS-G-SSIM (PLCC $\geq$ 0.75 overall, and $\geq$ 0.86 for MR images), which considers four-component region-based weighting (4-), multiscale (MS-), and gradient-based (G-) computation. The authors concluded that MS- approaches generally improve the performance of single-scale counterparts, and that $r*$ (structural component only) showed a slight advantage over the complete SSIM index.

3.2 Reduced-reference approaches

A few works, described in Table 2, addressed the use of RR metrics to measure the quality of medical images and video. One of the most popular is the RR version of VQM, which, as aforementioned, was used in [28], showing a reasonable performance (PLCC=0.97 and SROCC=0.94) in comparison to other FR metrics. Other approaches are based on the comparison of certain attributes of the images or videos through the computation of a similarity score. For example, Lee and Wang used the similarity between the intensity histograms of reference and distorted fundus images to provide an estimation of their quality [49]. For the same image modality, the authors in [50] proposed a similarity metric based on the comparison of the distribution of edge magnitudes and the local intensity distribution for distorted and reference images. In both cases, the performance of their approaches was not reported with objective measures but through a comparison with SNR and through qualitative results, respectively. The results show that the proposed approaches are useful to discriminate between good and bad images. In addition, the proposed approaches can be extended to other types of images.

Finally, another set of RR metrics is based on the use of watermarking, i.e., adding a payload to the content, and then using similarity metrics to compare the reference and the distorted image or video. An example of this approach was proposed by Planitz and Maeder [51], using SSIM to measure the degradations and watermarking capacity. This study demonstrated that more robust watermarking techniques can be used in less visually sensitive areas while lighter techniques should be applied in more sensitive areas. Nasr and Martini [7] used PSNR and Mean SSIM (MSSIM) to measure the quality of medical ultrasound videos with simulated Gaussian noise, and JPEG2000 and HEVC compression. A predefined reduced-size logo that shares the same features of the original frame of interest (i.e., same organ and layout) was embedded in an unused part of the original image for the purpose of quality assessment. Considering that the logo does not depend on the specific content of the sequence, this method can be considered, to some extent, as NR. The authors concluded that the proposed technique does not require the original frame whilst achieving a high PLCC with the subjective results, with reported average correlations above 0.97 with both full reference PSNR and MSSIM.

Table 2 Overview of visual quality-based approaches to the quality assessment of medical images and video (RR metrics)

Reference	Imaging modality	Quality impairments / Processing	Validation	Objective metrics
[49]	Retinal fundus photography	Reduction of brightness, contrast, and SNR	Comparison with SNR and agreement with human perception	Similarity metric based on intensity histograms (RR)
[50]	Retinal fundus photography	Image acquisition impairments (out of focus, poor illumination, etc.)	Qualitative assessment: 1 ophthalmologist	Similarity metric based on the distribution of edge magnitudes and the local intensity distribution (RR)
[51]	CT, MRI	Compression with DCT and Least Significant Bit (LSB) techniques and different compression ratios	Objective measures of classification and percentage of damages pixels	Structural comparison metric (RR)
[7]	Ultrasound videos	Gaussian noise, JPEG2000 compression (9 ratios), HEVC compression (8 QP)	DMOS: 4 medical experts, 16 naive observers (DSCQS)	PSNR, SSIM, MSSIM

3.3 No-reference approaches

Tables 3 and 4 summarize the articles further presented in this section.

3.3.1 Magnetic resonance imaging

Liebgott et al. [52] proposed a new method to assess the quality of 2D MRI using active learning. A descriptor including features based on contrast, resolution, texture, and intensity information was reduced using Principal Component Analysis (PCA) [63] and classified using a Support Vector Machine (SVM) model [64]. Subjective scores on a 5-level scale were used to train and validate the model. Results showed that active learning reduces the need for training data by around 50%, compared to a previous method by the same authors.

Chow and Rajagopal proposed a modified version of BRISQUE [53], trained with mean-subtracted contrast-normalized MRI and the corresponding DMOS. This proposed metric was evaluated on two separate datasets with both T1-weighted (T1*w*) and T2*w* brain MRI, one with unknown artifacts and another with the same range of distortions of [21]. The proposed modified BRISQUE, the original counterpart, and a JPEG-based model were tested on the first dataset, and the correlation with MOS was obtained with the proposed metric (PLCC of 0.96 and SROCC of 0.93). The proposed BRISQUE also outperformed its original version in terms of correlation with two FR metrics - NQM and Feature Similarity Index (FSIM) -, for the majority of the simulated distortions in the second dataset.

The authors in [54] applied the Bayes theorem to calculate the posterior probability of entropy, given three quality attributes, i.e., contrast, sharpness, and standard deviation. A global quality index for 2D MRI was obtained by averaging those probability values, which were first separately computed for low- and high-entropy feature images. The evaluation dataset included ten T1*w* MRI of the brain acquired with bias fields and twenty-one images acquired without perceived distortions (T1*w*, T2*w*, Proton Density, and Fluid Attenuated Inversion Recovery (FLAIR)). Twenty different levels of Rician noise and motion blur were induced in the second subset. Five radiologists performed subjective QA, which showed that the predicted quality decreased consistently across the twenty distortion levels of noise and blurring. Correlation with MOS was assessed separately for each modality and distortion level using SROCC. The reported coefficients were globally above 0.6, with a tendency to be higher for lower distortion levels.

In [55], the authors presented ENMIQA, an entropy-based metric for the objective QA of MRI, which expresses local intensity differences after non-maximum suppression at various threshold levels. A dataset of T2*w* MRI was used, but no information was provided in the paper on the type of artifacts or noise present in the images. The performance of the proposed metric in quality prediction was compared against a large set of NR, and ENMIQA outperformed them all in terms of correlation with MOS. However, overall correlation coefficients were quite low, especially for rank correlations (PLCC = 0.65, SROCC = 0.35). PLCC was also reported for each anatomical structure separately, with wrist and knee images yielding coefficients near 1 and 0.9, respectively.

In a recent work, Chabert et al. [56] proposed using a set of features reduced using PCA, which included SNR, Contrast-to-noise Ratio (CNR), Foreground Background Energy Ratio (FBER), sharpness in fat, uniformity, the Wang index [65] and Shannon entropy [66], among others, to predict the quality perception of neuroradiologists for lumbar T1*w* and T2*w* MRI, with simulated Gaussian noise and blurring, and contrast manipulation. A few classifiers were tested, with SVM showing a superior overall performance (accuracy above 73%). Despite

Table 3 Overview of visual quality-based approaches to the quality assessment of medical images and video (NR metrics - I)

Reference	Imaging modality	Quality impairments / Processing	Validation	Objective metrics
[52]	MRI	Different imaging sequences, contrast weights and subsampling strategies	5-level qualitative scale: 5 observers	Contrast, resolution, texture and intensity-based features with SVM classifier
[53]	MRI	Rician and Gaussian noise, and Gaussian blur (5 levels); DCT, JPEG and JPEG2000 compression (5 ratios)	MOS: 10 radiologists; FR metrics: NQM and FSIM	BRISQUE (modified and original), JPEG-based model
[54]	MRI	Bias fields, motion blur, Rician noise (20 levels)	MOS: 4 radiologists and 1 experienced MRI reader	Entropy posterior probability
[55]	MRI	Not specified	MOS: 31 radiologists	ENMIQA (proposed entropy-based metric), BPRI, dipIQ, IL-NIQE, MEON, Q index, S index, QENI, SNR-ROI, SSEQ and SISBLIM
[56]	MRI	Gaussian noise, Gaussian blurring, contrast manipulation	MOS: 3 neuroradiologists	Whole image and ROI-based features (e.g. SNR, CNR, uniformity, Wang index, Shannon entropy) with different classifiers (LDA, QDA, SVM, MLP, Logistic Regression)

Table 3 continued

Reference	Imaging modality	Quality impairments / Processing	Validation	Objective metrics
[57]	MRI	Not specified (motion artifacts and inhomogeneous fat supression mentioned in some cases)	Binary quality label: 2 radiologists	AlexNet
[58]	MRI	Motion artifacts	Qualitative binary score: 2 expert observers	6-layer CNN
[59]	MRI	Motion artifacts	Qualitative binary score and 3-level scales: 2 radiologists	4-layer CNN and ResNet-10
[60]	Retinal fundus photography	Gaussian blurring(2 levels) and Gaussian noise (20 levels)	FR metrics: PSNR and SSIM; MOS (no. of observers not given)	Local gradient-based metrics (normal and vessel-guided), CPBD and anisotropy
[61]	Retinal fundus photography	Uneven illumination, blurring, low contrast and color distortion	Composite binary scoring: 3 ophthalmologists	HVS-based feature matrix (multiple channel sensation, JNB, contrast sensitivity function)
[62]	Retinal fundus photography	Not specified	MOS (no. of observers not given)	Texture features from RGB and CIELab channels (Butterworth, Gabor and Wavelet filters, Gaussian Markov random fields, GLCM)

Table 4 Overview of visual quality-based approaches to the quality assessment of medical images and video (NR metrics - II)

Reference	Imaging modality	Quality impairments / Processing	Validation	Objective metrics
[72]	Retinal fundus photography	Not specified	MOS: 3 ophthalmology experts (cross-validation set) + 6 ophthalmology experts (test set)	InceptionV3
[73]	Retinal fundus photography	Not specified	Qualitative 3-level scale from feature-based MOS: 15 ophthalmologists	Multivariate regression-based CNN, with feature extraction using InceptionV3, DenseNet-121, ResNet-101, and Xception
[76]	Retinal fundus photography	Artifacts (e.g. shadows, lens stains), image clarity and field definition	Binary quality label and factor-specific qualitative scores: 3 ophthalmologists	CNN model with global and local feature extraction and ADDA, DeepIQA, DRIQC, MEON, MFIQA
[79]	Retinal fundus photography	Not specified	10-level scale: 2 ophthalmologists	Pre-trained CNN model with Inception-V3 backbone
[80]	AS-OCT	Not specified	Qualitative 3-level scale: 2 medical experts	DDDT-CWT based LBP features with a naïve Bayes classifier
[81]	Laparoscopic surgery videos	Defocus and motion blur, Gaussian noise, uneven illumination, smoke	Labels inferred from simulated distortions	ResNet (18, 34, 50 layers), RankIQA

Table 4 continued

Reference	Imaging modality	Quality impairments / Processing	Validation	Objective metrics
[82]	Endoscopy video	Blur, bubbles, contrast, specularity, saturation, misc. artifacts	MOS: 3 radiologists	YOLOv3 with DeepLabv3+ spatial pyramid pooling
[83]	Ultrasound (Echocardiogram)	Not specified	MES (6-level scale): 1 cardiologist	PSO optimized CNN model: 3 convolutional layers
[84]	CT	1) simulated low-dose CT (Poisson noise); 2) low-dose radiation	1) SSIM of simulated low-dose images; 1) 5-level qualitative scale (5 radiologists)	DistilIQA (Transformer architecture + knowledge distillation framework)
[85]	Fused images (MRI, CT, US, SPECT, PET)	8 image fusion algorithms: NNM, LP-SR, CSCS, GFF, NSCT-PCNN-SF, ISML, CSR, and DTM-PCNN	MOS: 20 radiologists	Pooling of phase congruency and standard deviation (proposed), Gradient-based and Structure-based metrics, Edge information, RSFE, MI
[86]	Fused images (MRI, CT, US, SPECT, PET)	8 image fusion algorithms [85]	MOS: 20 radiologists	PCNN in NSCT (proposed), baseline metrics of [85], Phase congruency, Entropy, OSSI, Tsallis entropy-based metric, BMPRI, MEON
[87]	MRI, Ultrasound	Noise artifacts, median filtering, JPEG 2000 (3 ratios)	Direct comparison of metrics; JAFROC: 5 radiologists (MRI)	NIQE, BiQES, NIQE-K (proposed modified NIQE)

relying on user interaction for ROI-based feature extraction, the results seem promising for online monitoring of the image quality at the moment of acquisition.

The appearance of deep learning models has been establishing a new paradigm in NR medical image quality assesment. Esses et al. [57] trained a Convolutional Neural Networks (CNN) model based on the AlexNet architecture [67] to screen the diagnostic quality of T2w liver MRI. The obtained results indicate a 79% and 73% concordance between CNN predictions and experts 1 and 2, respectively. Moreover, the CNN showed good negative predictive values in the identification of non-diagnostic image quality (94% and 86% for each of the experts).

Sujit et al. implemented a branched custom CNN model to assess the image quality of 3D T1w brain MRI [58]. Each branch received a different imaging plane (i.e., coronal, sagittal, and axial) as input. The global quality score was obtained by averaging the scores of each plane. Training and validation data were collected from the Autism Brain Imaging Data Exchange (ABIDE) multicentric dataset [68], which is publicly available with binary subjective quality scores (i.e., "acceptable" or "unacceptable"). The proposed ensemble model obtained an AUC of 0.90 on a 20% test split. The model was also tested on a separate test set from the CombiRx dataset [69], yielding a lower AUC of 0.71. The authors argued that the lower results may be due to differences in cohorts, as CombiRx data was not included in training.

Ma et al. [59] investigated the use of CNN to estimate the diagnostic quality of abdominal MRI. A custom 4-layer CNN and a ResNet-10 model were trained to classify images on both binary (non-diagnostic vs. diagnostic) and 3-level (non-diagnostic (0), diagnostic (1), or excellent (2)) quality scales. The proposed 4-layer CNN outperformed ResNet-10, with an accuracy of 84% in binary classification (AUC = 0.72) and 65% in 3-label classification (AUC_0 = 0.77, AUC_1 = 0.69, AUC_2 = 0.83). However, the authors argued that the high disagreement between human observers, as well as the probability of label unreliability, could influence the results. Moreover, the activation maps suggest that low-level features have higher discriminative power, whereas deeper features do not have a great impact on these classification tasks, i.e., deeper models may not necessarily improve the results in similar applications.

3.3.2 Retinal fundus images / ophthalmology images

Köhler et al. [60] extended the work of [70], where a global quality index Q was obtained from patch-wise quality indices $q(P)$. These were based on the singular value decompositions of G, a patch-wise gradient matrix. In [60], the authors implemented a spatially weighted version (Q_v), assigning larger weights to patches around blood vessels according to a vesselness measure. The proposed metric was tested as a PSNR and SSIM, on images from the DRIVE database with simulated Gaussian blur and noise. Q_v outperformed Q, with an overall SROCC of 0.89 (PSNR) and 0.91 (SSIM). In a second experiment, the authors performed binary quality prediction (acceptable vs. non-acceptable), considering annotations from experts. Q_v outperformed other NR metrics - Q, Cumulative Probability of Blur Detection (CPBD), and an anisotropy measure - with an AUC of 0.89.

The authors in [61] proposed an Human Visual System (HVS)-based feature extraction algorithm, which relies on multi-channel sensation, Just Noticeable Blur (JNB), and the Contrast Sensitivity Function, to detect illumination and color distortions, blur, and low contrast in retinal fundus images. Extracted features were classified with an SVM and a decision tree to predict image quality, either based on each of the three aforementioned properties or globally. A joint dataset contained the 3 partial binary annotations, given by

ophthalmologists. Average AUC of 0.97, 0.96, and 0.68 for illumination/color, JNB, and color, respectively, was obtained with the decision tree, and 0.93, 0.93, and 0.88 with the SVM. In terms of overall quality, the SVM showed the best performance with sensitivity/specificity of 0.92/0.87.

In [62], the authors used texture features from RGB and CIELab channels, including Wavelet and Gabor filters, and features from the Gray-level Co-occurrence Matrix (GLCM) [71] to compute the quality of retinal fundus images. Several feature selection methods and classifiers were tested for the binary classification of the image quality (good or poor quality). However, no information on the quality impairments nor the subjective evaluation setup is provided in the paper. The best performance in quality prediction was obtained with GLCM features from CIELab channels, filtered using correlation-based feature selection, and classified with an SVM classifier (accuracy = 99.09%).

More recently, Coyner et al. [72] studied the use of CNN models to measure the quality of retinal fundus images, considering their usefulness for a confident detection of retinopathy of prematurity. The model was based on the InceptionV3 architecture and was trained to classify an independent set of thirty images as having acceptable or non-acceptable quality. The testing set was previously ranked by six experts, and the CNN probability output showed good correlation with the consensus ranking from the experts (SROCC = 0.90). SROCC ranged from 0.86 to 0.93 when considering individual rankings.

Focusing on eye fundus images for the diagnosis of diabetic retinopathy, Raj et al. [73] presented a new multivariate regression-based CNN model to predict the image diagnostic quality. The proposed model incorporates four different backbones - InceptionV3, DenseNet-121, ResNet-101, and Xception - trained against subjective scores for six quality parameters, i.e., visibility of the optic disc, macula, and blood vessels, color, contrast, and blur. The top of the model provides a global diagnostic quality score based on the computed features. Results showed a strong correlation with the subjective scores for overall diagnostic quality, with values of 0.94, 0.95, 0.85, and 0.40 obtained for SROCC, PLCC, KRCC, and Root Mean Square Error (RMSE) respectively. The classification accuracy was 95.66% over the FIQuA dataset, presented in the same paper, and 98.96% and 88.43%, respectively, over the two publicly available datasets, DRIMDB [74] and EyeQ [75].

In [76], the authors proposed a CNN-based model with three modules for the quality assessment of retinal images. The model extracts global and local features (optic disc and fovea) using a VGG-16 backbone, which provides an overall quality score and three partial scores for different factors - artifacts, clarity, and field definition. Visual feedback from class activation mapping is also provided to ophthalmologists. ROI detection is performed by a separate module, using a ResNet-50 backbone for center detection and a VGG-16 local encoder for iterative ROI refinement. A third module implements an unsupervised Adversarial Discriminative Domain Adaptation (ADDA) method [77], using a Generative Adversarial Network architecture [78] to address domain shifts between training and test data. Several variations of the proposed model were tested to classify images as adequate or inadequate for the diagnosis of diabetic retinopathy, and the best result was obtained with the full described model, with an AUC of 0.95.

Abramovich et al. developed FundusQ-Net [79], a deep learning model using an Inception-V3 backbone, for automated quality grading of retinal fundus images. The model was trained on an extensive dataset of 89,947 images, of which 1245 were labeled by the two ophthalmologists, while the remaining 88,702 were used for pre-training and semi-supervised learning. Subjective quality scores were provided on a scale from 1 to 10, with 0.5 resolution. Final validation was performed on two test datasets with a reported mean absolute error of 0.61 against the subjective scores on an internal dataset and 99% accuracy on the DRIMDB database.

Niwas et al. [80] developed a metric to assess the quality of Anterior Seg-ment Optical Coherence Tomography (AS-OCT) images using Local Binary Patterns (LBP) in the complex Wavelet domain. The image is first decomposed into 32 wavelet decomposition levels using the Double Density Dual Tree-complex Wavelet transform, and the LBP histogram is obtained for each sub-band. The Minimum Redundancy Maximum Relevance (mRMR) method is then applied to select the most relevant features to classify the image in terms of quality, with three levels considered. Several classifiers were tested, including SVM, random forest, decision tree, and an AdaBoost classifier, but the best results were achieved using a Naïve Bayes classifier with mRMR feature selection. These were compared to the ground truth given by experts, yielding an overall weighted accuracy of 82.9%.

3.3.3 Endoscopic / laparoscopic videos

Based on a new publicly available database of 2D laparoscopic videos (LVQ) [88], a residual network-based method was proposed in [81] to predict the quality score and detect the type of distortion. Particularly, to tackle the problem of limited data, a ranking-based pre-training approach has been proposed. The proposed model with a ResNet-18 backbone obtained a better performance in terms of SROCC with the established quality ranking (0.69) when compared to another deep learning-based method, i.e., RankIQA [89] (0.57). It should be noted that the quality ranking was inferred directly from the severity levels of simulated distortions (i.e., defocus and motion blur, Gaussian noise, uneven illumination, and smoke). Simultaneously, the authors attempted image classification into 20 different labels, considering every distortion-level pair. Different ResNet depths were tested, with ResNet-50 achieving the best accuracy (87.3%).

In a very recent paper, Ali et al. [82] proposed an integrated deep learning approach for both QA and frame restoration in video endoscopy. In the context of this paper, the QA method relied on detecting six different types of impairments, i.e., bubbles, blur, contrast, specularity, saturation, and miscellaneous artifacts, in near real-time, using a YOLOv3 architecture. The dataset included a set of frames from both normal bright field and narrow-band imaging videos of different patients. After a bounding box detection stage, the framework also performed finer segmentation of the artifacts, with the best results being obtained with DeepLabv3+ spatial pyramid pooling.

3.3.4 Ultrasonography

Abdi et al. [83] trained a total of 430 CNN models to attempt the automated QA of transthoracic echocardiograms, using Particle Swarm Optimization (PSO) for hyperparameter optimization. Subjective scoring was based on the visibility of several anatomical structures (Manual Echo Score (MES)). The network resulting from PSO featured three convolutional layers, fed to two fully connected layers, and was trained independently three times. The final model performance was measured by the mean absolute error between the predicted score and the MES, with average reported values of 0.71 ± 0.58. The authors further analyzed the obtained CNN feature maps, which suggested the visibility of the septum and lateral walls to be important factors for a higher predicted image quality.

3.3.5 Computed tomography

Baldeon-Calisto et al. proposed DistilIQA, which combines a Transformer architecture with multi-headed self-attention and a knowledge distillation framework for the QA of CT

images [84]. The authors tested the method in predicting quality scores for two distinct datasets. The first included low-dose chest CT images simulated by adding Poisson noise to data from 11 non-contrast scans. DistilIQA obtained a high correlation with the SSIM scores computed between full-dose and the simulated low-dose images (PLCC = 0.99 / SROCC = 0.98), outperforming several other deep learning methods. In a second experiment, the authors validated the proposed model by predicting the quality of abdominal CT images of the LDCTIQAC2023 challenge [90], which were annotated by five radiologists according to a 5-level qualitative scale based on the visibility of anatomical structures and suitability for diagnosis. On the challenge test set, DistilIQA obtained a PLCC of 0.95 and SROCC of 0.84 with the subjective scores.

3.3.6 Fused images

Tang et al. published two papers on the quality assessment of Multimodal Medical Image Fusion (MMIF). In both papers, the authors tested eight image fusion algorithms, namely, Nuclear Norm Minimization (NNM), Laplacian Pyramid and Sparse Representation (LP-SR), Cross-scale Coefficient Selection (CSCS), Guided Filtering (GFF) Pulse-coupled Neural Network with Modified Spatial Frequency based on Non-subsampled Contourlet Transform (NSCT-PCNN-SF), Improved Sum-Modified-Laplacian (ISML), Convolutional Sparse Representation (CSR), and Discrete Tchebichef Moments and Pulse-coupled Neural Network (DTM-PCNN), with different pairs of imaging modalities, namely, CT and MRI, T1*w* and T2*w* MRI, B-mode Ultrasound and Single-photon Emission Computed Tomography (SPECT), MRI and PET, and MRI and SPECT. In [85], the quality index was obtained by pooling the phase congruency, which measures the correlation between salient features of the source and fused images, and the standard deviation, which provides a measure of sharpness and clarity. In [86], the authors further proposed using a PCNN with NSCT sub-images (both high- and low-frequency components). An overall quality index was then given by pooling the obtained components. The performance of the proposed metrics was compared to several state-of-the-art metrics, including gradient, structure, edge, and entropy-based metrics, considering their correlation to the MOS of twenty radiologists. In [85], the proposed metric achieved average SROCC near 0.8 and KRCC near 0.7, whilst in the more recent paper, the proposed metric yielded the following performance measures: PLCC=0.79, SROCC=0.73, KRCC=0.61, and RMSE=0.27. Although both metrics largely outperformed the baseline metrics, suggesting their relevance in MMIF QA, the obtained correlations could be further improved.

3.3.7 Multiple imaging modalities

In [87], the authors proposed to modify NIQE [46], a NR metric originally developed for natural images, and used for the QA of medical images in some studies. The authors improved its perceptual evaluation using a frequency-domain analysis inspired by the Blind Image Quality Evaluator based on Scales (BIQES) metric [91]. The ratio kurtosis/standard deviation of the log amplitude of the Fourier spectra was taken as a weighting factor. In a first experiment, ultrasound images were corrupted with simulated noises, i.e., Sattar's noise and speckle noise, and then filtered using a median filter. NIQE-K and BIQES achieved comparable performances in terms of quality ranking. Nonetheless, NIQE-K ranked a noisy image (Sattar's noise) better than the original ultrasound image, despite their close quality indices. In the second experiment, three radiologists were asked to detect and localize

multiple sclerosis lesions in MRI. Their performance was quantified by the Jackknife Alternative Free-response Receiver Operating Characteristic (JAFROC). Result analysis showed the consistency of NIQE-K for ultrasound image quality and that, for 40% of the images, the behavior of NIQE-K is the same as that of the radiologists.

4 Task-based methods

Task-based objective QA methods, also often referred to as *numerical observers* (NO) or *model observers* (MO), are designed to approach the performance of human observers (i.e., medical experts) in a given task, as opposed to visual quality-based approaches, which typically aim at predicting the MOS of human observers [11].

The underlying paradigm is to quantify the quality of medical images and videos by their effectiveness with respect to their intended purpose [92]. In this section, we review task-based methods proposed for different tasks, as well as the corresponding Figures of Merit (FOM) for performance evaluation. Table 5 is a summary of these works (note that we focus on works published after 2013, since an overview was already published then [16]).

4.1 Detection and classification tasks

The most common task is the detection task, in which a decision is made in favor of one of two hypotheses, i.e., signal present (H1) or absent (H2). Low-dose reconstruction methods using iterative algorithms in tomography introduced new challenges related to the extraction of image information. In any diagnosis, a good compromise must be found between the dose level and the resulting quality of the medical image in order to reduce the exposure dose while allowing the detection of low contrast textures. In [93], the authors used a Channelized Hotelling Observer (CHO) [16, 94] and developed an internal noise model to compare detectability indices (d') in low-dose CT images. The d' index, chosen as FOM, was calculated from the distribution of signal and noise decision variables (i.e., sum of channel outputs). The experiment was performed with five observers, using CT phantom images, and showed that Iterative Model Reconstruction (IMR, Philips) enables at least a 67% dose reduction comparatively to Filtered Back-projection (FBP).

Racine et al. [95] conducted research to objectively evaluate the low contrast detectability in CT with different radiation doses ($CTDI_{vol}$ of 5, 10, 15, and 20 mGy). Images of a QRM 401 phantom containing 5- and 8-mm diameter spheres, with a contrast level of 10 and 20 HU, were acquired and reconstructed using three algorithms, i.e., FBP, Adaptive Statistical Iterative Reconstruction (ASIR), and Model-based Iterative Reconstruction (MBIR). A CHO model with Dense Difference-of-Gaussian (D-DOG) channels was used to evaluate the image quality for every combination of the aforementioned parameters. The performance of the CHO model was compared with that of six medical students, who provided detectability levels for the test images in four-alternative forced choice tests. A high correlation was found between the results of the human observers and the CHO model, independently of the dose levels or the signals considered, suggesting it might be used to predict expected detectability levels and ensure the diagnostic quality of low-dose CT acquisitions. PLCC was 0.98 for MBIR and 0.93 for FBP. MBIR gave the highest overall detectability index, particularly for low $CTDI_{vol}$.

Greffier et al. [14] used the *imQuest* software to assess the quality of low-dose CT scans, comparing the performance of four manufacturers (i.e., GE, Philips, Siemens, and Canon),

Table 5 Overview of task-based approaches to the quality assessment of medical images and video

Reference	Imaging modality	Quality issue	Model	Task	Figure of merit
[93]	CT phantom	Reconstruction: IMR, FPB/ low dose radiation	CHO	Detection	Detectability (d')
[95]	CT phantom	Reconstruction: FBP, ASIR, MBIR/ low dose radiation	CHO (D-DOG)	Detection	Detectability (d')
[100]	CT phantom	Detection of lesions in CT with different X-ray exposures	SR-MO, CNN-MO, CHO (Gabor)	Detection	ROC / AUC
[14]	CT phantom	Reconstruction: Asir, Asir-V, IMR, iDose SAFIRE, ADMIRE, AIDR3D, First, FPB, H/SIR and MBIR / low dose radiation	NPWMF (imQuest tool)	Detection	Detectability (d')
[101]	Computer-simulated images	Approximation of the IO and the HO by supervised learning methods to evaluate image quality for detection tasks	CNN-IO, SLNN-IO	Detection	ROC / AUC
[102]	Mammography	Study the minimum detectable contrast in mammography scans	CNN-MO	Detection	Contrast threshold
[15]	Digital Breast Tomosynthesis	Reconstruction: model-based CNN regularized reconstruction (MDR)	Proposed CNN-MC observer: VGG-Net, ResNet, and ConvNeXt backbones	Detection	Detectability (d')

Table 5 continued

Reference	Imaging modality	Quality issue	Model	Task	Figure of merit
[97]	Simulated Planar Scintigraphy images	Comparison of CNN-based denoising methods	IO, CNN-MO, HO, RHO, NPWMF	Detection	ROC / AUC
[103]	Ultrasound	Quality assessment of acquisition of fetal ultrasound scans	Custom CNN-based model	Localization	Multiple structure visibility
[104]	Computer-simulated images	Approximation of the IO and the HO by supervised learning methods to evaluate image quality for localization tasks	Analytical IO, Scanning HO, MCMC-IO, CNN-IO	Localization	LROC
[105]	Computer-simulated images	Approximation of a human observer in defect localization tasks	U-Net	Localization	Accuracy
[106]	Retinal fundus photography	Quality assessment of acquisition of retinal images	Measures from segmented areas with SVM or decision tree classifier	Segmentation	ROC / AUC
[107]	MRI	Compare the performance of the segmentation model with several no-reference quality metrics	Texture-based features with AdaBoost classifier	Segmentation	Dice overlap coefficient
[108]	Retinal fundus photography	Image quality in macular region	Custom CNN, U-Net	Segmentation and localization	Macula visibility

with different reconstruction algorithms (i.e., FBP, MBIR and Hybrid/Statistical Iterative reconstruction (H/SIR)), and five dose levels (i.e., 0.5, 1.5, 3.0, 7.0, and 12.0 mGy). A Non-Prewhitening Matched Filter (NPWMF) model observer with an eye filter [96] was used to calculate the detectability index, considering two detection tasks: a large mass in the liver and a small calcification. The d' index was obtained from the Noise Power Spectrum (NPS)s and the Task Transfer Function (TTF), which were measured on a ACR QA phantom with acrylic inserts. The reported results showed that the use of an optimization algorithm (either H/SIR or MBIR) improves d' for low-dose acquisition, when compared to FBP. When directly comparing H/SIR and MBIR, the second led to an increase in d' (measured at 3 mGy), as well as potential dose reductions for Siemens, GE, and Philips systems, with features sizes. Potential dose reduction using Philips (IMR) reached 62% and 78% for the small and large features, respectively.

Supervised learning-based approaches for implementing model observers have become possible substitutes for numerical observers, particularly through the use of CNN. Li et al. [97] evaluated different CNN-based denoising methods on simulated planar scintigraphy images. The images were simulated for Signal Known Exactly / Background Known Statistically (SKE/BKS) lesion detection with a Gaussian signal and a lumpy object as random background. Mixed Poisson and Gaussian noise was added to the simulated images, both signal absent and signal present, to produce the final dataset. Three denoising encoder-decoder networks were tested, i.e., a linear CNN (without ReLU activation layers), a non-linear CNN with MSE loss, and a ResNet-based model with a perceptual loss function. Their performance was then assessed using several observers for the detection of the lesion signals, i.e., a Bayesian Ideal Observer IO [98, 99], a CNN-based observer, a Hotelling Observer (HO) [16, 98], a CHO, a Regularized HO (RHO), and a NPWMF. The authors performed ROC analysis and further computed a detection efficiency measure, given by $e \equiv AUC_{denoised}/AUC_{noisy}$, observing that, while an increase in network depth improved SSIM and RMSE measures, the performance of the detection task generally dropped, thus suggesting that denoising methods might cause a loss of statistical information in the image. For the CNN-based observer, HO and RHO, the performance decreased with the increase in network depth. It is also argued that using non-task-based loss functions to optimize CNN-based denoising models might play an important part in that loss of information.

The authors in [100] tested two antropomorphic model observers to detect liver lesions, one based on *softmax* regression (SR-MO) and a CNN-MO. A phantom with contrast targets of different diameters was scanned on a CT scanner at different X-ray exposures, and the images were reconstructed using both FBP, which was used for the test dataset, and iterative reconstruction, used in training and validation. One radiologist provided confidence levels considering the detection task for a total of 7488 images. Model performance evaluation relied on computing the JAFROC for the reader study, a CHO with Gabor channels, and the two proposed models, which were trained using two strategies, i.e., separated models for each lesion size or a common model for all diameters. The PLCC, the χ^2 goodness-of-fit, and the mean absolute percentage difference (MAPD) were then used to compare the models' AUC values. The authors concluded that the CNN-MO can accurately approximate human performance. This model outperformed the CHO (PLCC = 0.95, MAPD = 2.2%) using the first training strategy (PLCC = 0.98, MAPD = 1.2%), and the SR-MO with both strategies. With the second strategy, the CNN-MO achieved a PLCC of 0.92 and MAPD of 3.0%.

Alnowami et al. [102] aimed at studying the minimum detectable contrast in mammography screenings using a deep learning-based MO. They first trained a data-driven CNN model to classify image patches from clinical routine screenings as normal tissue or containing a lesion. In this SKS detection task, the model achieved a sensitivity of 0.90 and a specificity of

0.92 on the test set. In a second experiment, its performance was compared with two groups of human observers (experts and non-experts), using a four-alternative forced choice setup with simulated images to assess the minimum detectable contrast. Contrast threshold values obtained with the proposed MO approximated the human performance on spherical targets, and even outperformed it for 4mm targets. For lesion targets, human performance was better, but the CNN-MO still achieved a comparable performance (12% difference).

Zhou et al. [101] used deep learning methods to approximate an IO and a HO for binary signal detection tasks, i.e., a CNN and a Single Layer Neural Network (SLNN), respectively. Computer-simulated images were generated using continuous-to-discrete mapping with a Gaussian kernel and considering four binary signal detection tasks: a Signal Known Exactly / Background Known Exactly (SKE/BKE) task, where the IO and HO are analytically determined, two SKE/BKS tasks, with lumpy background and clustered lumpy background object models, and a Signal Known Statistically / Background Known Statistically SKS/BKS task with a lumpy background. Overall, the CNN-IO and SLNN-HO closely approximated the results of the IO and HO, in terms of AUC. In the case of the SKE/BKE task, the obtained AUC with IO and CNN-IO was 0.89, whereas with HO and SLNN-HO it was 0.83. For the SKE/BKS task with lumpy background, the IO was computed using Markovchain Monte Carlo (MCMC) techniques [109], and achieved an AUC of 0.91. CNN-IO closely approximated this performance, with AUC=0.91. The traditional HO and SLNN-HO yielded an AUC of 0.81. A similar outcome was observed in the SKE/BKS task with clustered lumpy background for HO and SLNN-HO (AUC=0.85). CNN-IO achieved an AUC of 0.89, but its performance could not be compared to the IO, with the authors arguing that MCMC application to clustered lumpy background object models had not been reported to date. Finally, for the SKS/BKS task, the performance of both the traditional HO and SLNN-HO is close to a random estimate (AUC $\approx$ 0.5), as linear observers are generally unable to detect signals with random locations. As for MCMC-IO and CNN-IO, the performance was again very close (AUC=0.86).

Recently, Gao et al. [15] proposed a reconstruction model for Digital Breast Tomosynthesis (DBT) images which incorporated two CNN-based QA observers: CNN-NE, which estimates the root-mean-square (RMS) noise in image patches, and CNN-MC, which evaluates the detectability of clustered microcalcifications (MC) in human DBT images. In the context of this paper, we focus on the latter. The reported results indicate that using the proposed CNN-MC observer can effectively substitute human observers in ranking imaging systems and may ultimately lead to lower dose acquisition with enhanced sensitivity and specificity for MC detections.

4.2 Localization tasks

An interesting work was published in [103], where the authors proposed a localization-based approach to assess the quality of acquisition of fetal ultrasound scans. First, the authors implemented a localization CNN (L-CNN) using pre-learned AlexNet low-level cues, which identifies a ROI containing the fetal abdomen. A sliding window strategy was used to feed local inputs to the L-CNN and obtain a ROI probability map throughout the entire scan. Image quality was then assessed by considering the quality of depiction of both the stomach bubble (SB) and umbilical vein (UV) within the ROI. These two structures may be labeled as satisfactory, not good, or absent. While the first label translates to S_{SB} or S_{UV}=1, the other two led to 0. From the combinations of these two binary scores, a 4-class output was obtained by either human observers or a classification CNN (C-CNN), with knowledge transferred

from the L-CNN to its encoding layers. The final quality score of the Fetal Ultrasound Image Quality Assessment (FUIQA) scheme is given by the sum of S_{SB}, S_{UV} and S_{ROI}, which is also a binary score, determined by the ROI to field of view ratio. The authors provided an extensive discussion on the proposed methods, for example, comparing the AUC for SB and UV detection with different input layers based on local phase analysis. The reported quality outputs of the described FUIQA model were highly coherent with those from 3 radiologists, with agreement values of 0.91, 0.89, and 0.88.

Zhou et al. [104] extended their proposed supervised learning method [101] to approximate an IO in joint detection and localization tasks. The LROC curves produced by the proposed method were compared with those of traditional observers when computationally feasible. The signal can be localized in nine different locations, and the considered signal detection-localization tasks are BKE, BKS with a lumpy background model, and BKS with a clustered lumpy background model. Overall, the CNN-IO was able to come close to the analytical IO for the BKE task, the MCMC-IO for the BKS task with a lumpy background, and outperformed the scanning HO for the BKS task with CLB since MCMCs have not been reported to date for this type of background. Note that approximating an IO using supervised learning requires a large amount of training data, which can be a challenge when only a limited amount of experimental data is available.

Lorente et al. [105] proposed a MO based on U-Net [110] for defect localization on simulated images with three levels of correlated noisy backgrounds. Two network configurations (i.e., kernel sizes 3×3 and 5×5) and two loss functions (i.e., MSE and binary cross-entropy) were tested, and their accuracy was compared with that of a human observer. Accuracy was described as the ratio between correctly located defects, i.e., within a 5 pixel distance of the actual defect, and the total number of images. The authors concluded that the models trained with a binary cross-entropy loss function provided results closer to the human observer. For example, for the third level of correlated noisy background, the accuracy of the human observer was 0.8, and the 3×3 and 5×5 MO trained with MSE loss yielded an accuracy of 0.92 and 0.91, respectively. The MO trained with binary cross-entropy loss got accuracy values of 0.89 and 0.85, respectively.

4.3 Segmentation tasks

In [106], the authors studied a segmentation-based QA framework for retinal images. Unsupervised vessel segmentation was performed on a dataset of 800 images from the UK Biobank [111], using QUARTZ (Quantitative Analysis of Retinal Vessel Topology and Size) [112]. The algorithm relies on a multi-scale line detector, based on the average gray-level intensities around each target pixel, complemented with high-intensity pixel thresholding, masking of the fovea, and suppression of small objects. Image quality was then measured by extracting three features from the binary segmentation image, selected to mimic manual QA, i.e., area, fragmentation, and complexity, which were finally classified using both a SVM classifier with a radial basis function kernel and an ensemble decision tree classifier. The dataset had been labeled by two observers as adequate or inadequate for epidemiological studies. The SVM achieved the best performance in classification, with an AUC of 0.98. Although method validation is not directly related to the human segmentation performance, the method uses the segmentation task outcome to predict image quality.

The authors in [107] studied the correlation between texture-based muscle segmentation in Dixon MRI and a set of NR quality metrics, i.e., Variance, Laplacian, Gradient, Autocorrelation, Frequency Threshold metric (FTM), Marziliano Blurring metric (MarzBM), HP metric,

Kurtosis-based metric, and Riemannian Tensor-based metric (RTBM), to assess the feasibility of using texture segmentation methods as a content-specific quality measure for MRI. The authors implemented a pixel-wise binary segmentation method, using AdaBoost [113], with a local texture descriptor consisting of the histogram of oriented gradients (HOG) [114], 3-level Haar Wavelet coefficients, and statistical measures from the original grayscale image and the Laplacian of Gaussian (LoG) filter [115]. The Dice overlap coefficient with manual segmentations was chosen as FOM. Overall, the segmentation output showed reasonable correlation with the variance metric (PLCC = 0.72, SROCC = 0.74) and RTBM (PLCC = 0.71, SROCC = 0.73). Considering only cases with poor segmentation (Dice < 0.7), all metrics performed better, even though it should be noted that this analysis relied on a smaller number of points. RTBM (PLCC = 0.93, SROCC = 0.86) and FTM (PLCC = 0.84, SROCC = 0.96) yielded the best correlations.

More recently, Alais et al. [108] proposed a CNN that segments the macular region, which consists of 3×3 convolutional layers only, has very few parameters, and makes decisions considering a threshold t to obtain a binary image. If the obtained area is greater than a value A, then the algorithm considers that the macula is visible, and the image quality is considered sufficient for the macula location. The authors extracted 6098 eye fundus images from the *e-ophtha* database [116], with more than half of the images containing a visible macula. The chosen parameters, t and A, that offer a trade-off between sensitivity and specificity, gave an accuracy of 96.4% on the test set. The proposed CNN wrongly predicted that the macula was visible in four images among 304 images (i.e., 1.3% of false positives), whereas U-Net had nine false positives. Regarding the fovea localization, the authors obtained an average error of 0.95 pixels for their network versus 1.22 pixels for U-Net. The same tests, conducted on the ARIA database [117], revealed a prediction of the fovea with 1.4 pixels of mean error (0.1mm) and 6 pixels of maximum error. Finally, the network was tested on pathological images, with one unsuccessful detection due to the presence of macular hemorrhages or exudates.

5 Discussion and future work

Medical images and video refer to a wide variety of different acquisition methods, clinical applications, and quality issues, as is clear from the papers reviewed in this contribution. Thus, it is almost impossible to establish meaningful comparisons between the objective methodologies used, i.e., their merits and drawbacks. Nonetheless, some global considerations may be drawn, particularly on future research directions.

Figure 2 summarizes the discussed QA metrics for medical image and video. Regarding FR and RR visual quality-based metrics, all reviewed papers reported the use of metrics that were originally designed for natural content, or variations of those metrics. PSNR and SSIM were the most common across these studies, and both were used in a vast majority of them [5, 7, 8, 21, 22, 24–27, 29–31]. Renieblas et al. [32] also reported the use of SSIM, along with several variations of that metric. Other metrics, such as VIF and NQM were commonly tested as well. The only FR metric specifically designed for medical imaging was proposed by [27], who described a QA metric for cardiac ultrasound videos based on cardiac motion and structural information. Considering the reviewed NR visual quality-based approaches, most methods are tailored for specific imaging modalities and/or artifacts. From our analysis, it is also clear that deep learning methods are becoming a staple in NR medical image and video

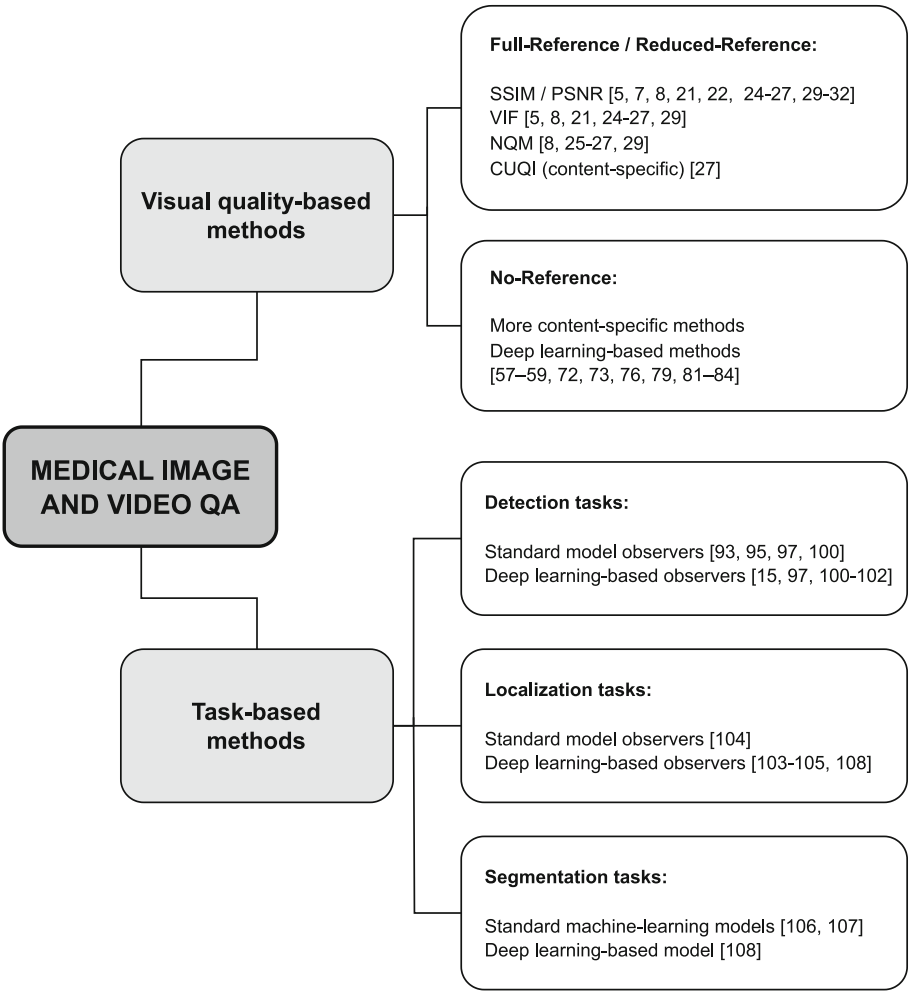

Fig. 2 Summary of the discussed QA metrics for medical image and video

QA, as most recent papers used CNN instead of handcrafted features, with some interesting results [57–59, 72, 73, 76, 79, 81–84].

As stated in [6], there seems to be a lack of publicly available databases with subjective quality annotations in the medical imaging domain. Some notable examples are described in [68, 88, 118, 119]. In order to develop new reliable objective quality metrics for the medical imaging field, more training data is needed. With a view to addressing this issue, human-in-the-loop machine learning techniques could be considered [120, 121]. In this sense, although automated and semi-automated techniques have been proposed for segmentation purposes [106–108], there is still a lack of annotated databases and studies to support the development of reliable methods, for example, incorporating models of how clinicians perform diagnosis from images and videos [122]. Artificial intelligence promises a strong breakthrough in medical imaging objective QA.

Another common challenge in objective QA studies for medical images and videos has to do with artifact simulation. While some of the reported visual quality-based studies used databases containing images or video with real artifacts (e.g., [50, 52, 57–59, 61, 76, 82]), others used simulated distortions and/or artifacts over real data assumed to be clean (e.g., [7, 8, 21, 32, 54, 56, 60, 81, 87]). Collecting data with real artifacts may not always be possible or end up being impractical. On the other hand, simulated artifacts are often generic and limited in range, which might hinder the application of the developed QA methods to real clinical data [123]. Some efforts towards the simulation of content-specific and realistic artifacts, to be applied in healthcare QA research, have been reported (e.g., [123–126]), and future work will likely approach this question more often, leveraging on the continuous development of deep-learning methods, such as GAN [127].

A similar question can be raised regarding task-based QA research. Traditional model observers are based on the total or partial knowledge of statistical characteristics of the images (i.e., signal and background). Hence, many studies (e.g., [14, 93, 95, 101, 104, 105, 128]) are based on simulated or phantom images since real medical images suffer from a lack of statistical information. Currently, there is no evidence that QA studies conducted on simulated images ensure sufficient confidence to draw relevant conclusions on real images. Deep learning methods could allow to go through this weakness. Indeed, any task that could be carried out with deep learning provides an assessment of the quality. In fact, the more efficiently the task is done, the better the quality. There is still room for improvement in the field, the whole purpose being to define the most relevant tasks, i.e., those that can be reliably delegated to a model. In task-based QA, the key problem lies in modeling the task performed by human observers. So far, to our knowledge, existing models are limited in task range. For example, no model observer has been proposed for characterization tasks, which focus on analyzing certain properties of abnormalities (e.g., contour or texture) for differential diagnosis and generally involve a linguistic response (e.g., benign vs. malignant), given its high complexity [16]. Other tasks could also be further explored in future work, such as estimation tasks (or joint detection / classification / estimation tasks), which aim at determining a scalar value or range of values from a given object to be used in diagnosis (e.g., tumor diameter or radiotracer uptake [13]).

Although 3D visualization of medical images and videos is emerging, research on relevant quality assessment aspects is still behind. Stereoscopic medical imaging and, more recently, light field medical imaging, open new opportunities, for instance in surgery training, also at a distance [129]. Compression [130] and transmission [129] of 3D stereoscopic medical images and videos, as well as of light field medical data, require suitable metrics for the assessment of their performance. Recently, studies on quality assessment for light field medical images have started (e.g, [131]). However, objective metrics in this domain are still missing, or their development is still ongoing. Future research might focus on assessing whether existing metrics, developed for generic 3D images and videos (e.g., [132–135]), and light field data (e.g., [136–139]), are suitable to assess the quality of medical data represented in these formats. While efforts are ongoing in this direction, the availability of wider datasets of medical data in stereoscopic 3D and light field 3D formats would definitely be useful towards this effort.

Several of the reported studies considered coding distortions [5, 7, 8, 21, 22, 24–32, 51, 53, 87]. There are a number of applications, mostly based on telemedicine applications, where lossy compression of medical images or videos might be acceptable. However, in most cases, medical imaging applications cannot rely on images with lossy compression, as no one can be sure of the influence that those losses can have in a diagnosis. Multiple times, radiologists use almost imperceptible textures to define their diagnosis, and, in such cases,

lossy compression can have a major impact. Hence, several studies on medical image quality are quite questionable as they use medical modalities where no radiologist would accept any kind of lossy coding.

6 Conclusion

This paper presented a review of the literature on objective quality assessment of medical images and video, considering various imaging modalities and application purposes. It covers a wide range of approaches to the quality assessment of medical visual content, including the use of preexisting metrics for natural images and the development of content-specific metrics either based on handcrafted features or using deep learning-based models. This contribution aimed to be as exhaustive as possible, including research efforts considered to be the most relevant in their application field. However, it should be noted that this does not reduce the merit of any work not included.

Drawn conclusions include that deep-learning methods are gaining prominence in the objective QA of medical visual content, with a rapid increase in use over the last few years. In traditional visual quality-based QA, FR metrics such as PSNR, SSIM and VIF, which are not specific for medical content, are among the most widely tested. As for NR metrics, most works proposed content-specific methods. Regarding task-based QA, existing models are still limited in their task range.

As a key contribution for future research, this paper formalizes a new categorization of QA methods for medical images and video into visual quality-based and task-based methods. Moreover, some challenges were identified, such as the lack of publicly available databases with subjective annotations and the lack of research data with content-specific and realistic artifacts. Emerging 3D visualization modalities will likely require suitable QA methods, which are still lacking.

Acknowledgements R. Rodrigues and A. Pinheiro acknowledge FCT - Fundação para a Ciência e Tecnologia, I.P. for funding this research under the doctoral grant SFRH/BD/130858/2017 and the project UIDB/50008/2020 of Instituto de Telecomunicações, with DOI identifier https://doi.org/10.54499/UIDB/50008/2020, respectively. They would also like to acknowledge the project CENTRO-01-0145-FEDER-000019 - C4 Cloud Computing Competence Centre.

Funding Open access funding provided by FCT|FCCN (b-on).

Data Availability This manuscript has no associated data.

Declarations

Conflict of Interest The authors have no conflicts of interest to declare.

References

1. World Health Organisation (2016) Communicating radiation risks in paediatric imaging: information to support health care discussions about benefit and risk. World Health Organisation
2. Krupinski E (2010) Current perspectives in medical image perception. Atten Percept Psychophys 72(5):1205–1217
3. OECD (2021) Health care utilisation. https://doi.org/10.1787/data-00542-en. https://www.oecd-ilibrary.org/content/data/data-00542-en
4. Lévêque L, Zhang W, Cavaro-Ménard C, Le Callet P, Liu H (2017) Study of video quality assessment for telesurgery. IEEE Access 5:9990–9999
5. Chaabouni A, Gaudeau Y, Lambert J, Moureaux J-M, Gallet P (2016) H.264 medical video compression for telemedicine: a performance analysis. Innovation and Research in BioMedical engineering 37(1):40–48
6. Lévêque L, Outtas M, Liu H, Zhang L (2021) Comparative study of the methodologies used for subjective medical image quality assessment. Phys Med Biol 66(15)
7. Nasr KM, Martini MG (2017) A visual quality evaluation method for telemedicine applications. Signal Process Image Commun 57:211–218
8. Mason A, Rioux J, Clarke SE, 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
9. Chow LS, Paramesran R (2016) Review of medical image quality assessment. Biomed Signal Process Control 27:145–154
10. Raj A, Tiwari AK, Martini MG (2019) Fundus image quality assessment: survey, challenges, and future scope. IET Image Process 13(8):1211–1224
11. Zhang L, Cavaro-Ménard C, Le Callet P, Tanguy J-Y (2012) A perceptually relevant channelized joint observer (PCJO) for the detection-localization of parametric signals. IEEE Trans Med Imaging 31(10):1875–1888
12. He X, Song X, Frey EC (2008) Application of three-class ROC analysis to task-based image quality assessment of simultaneous dual-isotope myocardial perfusion SPECT (MPS). IEEE Trans Med Imaging 27(11):1556–1567
13. Barrett HH, Myers KJ, Hoeschen C, Kupinski MA, Little MP (2015) Task-based measures of image quality and their relation to radiation dose and patient risk. Phys Med Biol 60(2):1
14. Greffier J, Frandon J, Larbi A, Beregi J, Pereira F (2020) CT iterative reconstruction algorithms: a task-based image quality assessment. Eur Radiol 30(1):487–500
15. Gao M, Fessler JA, Chan H-P (2023) Model-based deep CNN-regularized reconstruction for digital breast tomosynthesis with a task-based CNN image assessment approach. Phys Med Biol 68(24):245024
16. Zhang L, Cavaro-Ménard C, Le Callet P (2014) An overview of model observers. IRBM 35(4):214–224
17. Le Callet P, Moller S, Perkis A (2013) Qualinet white paper on definitions of Quality of Experience (QoE). In: Output from the fifth qualinet meeting p 8
18. ITU-R (2019) Methodology for the subjective assessment of the quality of television pictures. Recommendation BT.500-14
19. ITU-T (2016) Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment. Recommendation p 913
20. ITU-T (2012) Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. Recommendation p 1401
21. Chow LS, Rajagopal H, Paramesran R (2016) Alzheimer's Disease Neuroimaging Initiative et al Correlation between subjective and objective assessment of magnetic resonance (MR) images. Magn Reson Imaging 34(6):820–831
22. Zhou Y, Chen D, Li C-f, Li X-o, Feng H-q (2003) A practice of medical image quality evaluation. In: International conference on neural networks and signal processing, vol 1, pp 204–207
23. Fränti P (1998) Blockwise distortion measure for statistical and structural errors in digital images. Signal Process Image Commun 13(2):89–98
24. Kowalik-Urbaniak IA, Castelli J, Hemmati N, Koff D, Smolarski-Koff N, Vrscay ER, Wang J, Wang Z (2015) Modelling of subjective radiological assessments with objective image quality measures of brain and body CT images. In: International conference image analysis and recognition, Springer, pp 3–13
25. Panayides A, Pattichis MS, Pattichis CS, Loizou CP, Pantziaris M, Pitsillides A (2011) Atherosclerotic plaque ultrasound video encoding, wireless transmission, and quality assessment using H.264. IEEE Transactions on Information Technology in Biomedicine 15(3):387–397
26. Razaak M, Martini MG, Savino K (2014) A study on quality assessment for medical ultrasound video compressed via HEVC. IEEE J Biomed Health Inform 18(5):1552–1559

27. Razaak M, Martini MG (2016) CUQI: cardiac ultrasound video quality index. J Med Imaging 3(1):011011
28. Kumcu AE, Bombeke K, Chen H, Jovanov L, Platisa L, Luong HQ, Van Looy J, Van Nieuwenhove Y, Schelkens P, Philips W (2014) Visual quality assessment of H.264/AVC compressed laparoscopic video. In: Medical imaging 2014: image perception, observer performance, and technology assessment, vol 9037, p 90370. International society for optics and photonics
29. Usman MA, Usman MR, Shin SY (2017) Quality assessment for wireless capsule endoscopy videos compressed via HEVC: from diagnostic quality to visual perception. Comput Biol Med 91:112–134
30. Kumar B, Singh SP, Mohan A, Singh HV (2009) MOS prediction of SPIHT medical images using objective quality parameters. In: International conference on signal processing systems, pp 219–223
31. Kumar B, Kumar SB, Kumar C (2013) Development of improved SSIM quality index for compressed medical images. In: IEEE International conference on image information processing, pp 251–255
32. Renieblas GP, Nogués AT, González AM, León NG, Del Castillo EG (2017) Structural similarity index family for image quality assessment in radiological images. J Med Imaging 4(3):035501
33. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
34. Zhang L, Zhang L, Mou X, Zhang D (2011) FSIM: a feature similarity index for image quality assessment. IEEE Trans Image Process 20(8):2378–2386
35. Sheikh HR, Bovik AC, De, (2005) Veciana G: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans Image Process 14(12):2117–2128
36. Damera-Venkata N, Kite TD, Geisler WS, Evans BL Bovik AC (2000) Image quality assessment based on a degradation model. IEEE Trans Image Process 9(4):636–650
37. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15(2):430–444
38. Wang Z, Bovik AC (2002) A universal image quality index. IEEE Signal Process Lett 9(3):81–84
39. Ahmed N, Natarajan T, Rao KR (1974) Discrete Cosine Transform. IEEE Trans Comput C-23(1):90–93
40. Wallace GK (1992) The JPEG still picture compression standard. IEEE Trans Consum Electron 38(1):18–34
41. Skodras A, Christopoulos C, Ebrahimi T (2001) The JPEG 2000 still image compression standard. IEEE Signal Process Mag 18(5):36–58
42. Chandler DM, Hemami SS (2007) VSNR: A wavelet-based visual signal-to-noise ratio for natural images. IEEE Trans Image Process 16(9):2284–2298
43. Sullivan GJ, Ohm J-R, Han W-J (2012) Wiegand T: Overview of the high efficiency video coding (HEVC) standard. IEEE Trans Circuits Syst Video Technol 22(12):1649–1668
44. Egiazarian K, Astola J, Ponomarenko N, Lukin V, Battisti F, Carli M (2006) New full-reference quality metrics based on HVS. In: Proceedings of the second international workshop on video processing and quality metrics, vol 4
45. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The thrity-seventh asilomar conference on signals, systems & computers, 2003, IEEE, vol 2, pp 1398–1402
46. Mittal A, Soundararajan R, Bovik AC (2013) Making a "completely blind" image quality analyzer. IEEE Signal Process Lett 20(3):209–212
47. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708
48. Said A, Pearlman WA (1996) A new, fast, and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans Circ Syst Vid Technol 6(3):243–250
49. Lee SC, Wang Y (1999) Automatic retinal image quality assessment and enhancement. In: Medical imaging 1999: Image Processing, vol 3661, pp 1581–1590. International Society for Optics and Photonics
50. Lalonde M, Gagnon L, Boucher M-C et al (2001) Automatic visual quality assessment in optical fundus images. In: Proceedings of vision interface, Ottawa, vol 32, pp 259–264
51. Planitz BM, Maeder AJ (2005) A study of block-based medical image watermarking using a perceptual similarity metric. In: Digital image computing: techniques and applications
52. Liebgott A, Küstner T, Gatidis S, Schick F, Yang B (2016) Active learning for magnetic resonance image quality assessment. In: 2016 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 922–926
53. Chow LS, Rajagopal H (2017) Modified-BRISQUE as no reference image quality assessment for structural MR images. Magn Reson Imaging 43:74–87
54. Osadebey M, Pedersen M, Arnold D, Wendel-Mitoraj K et al (2017) Bayesian framework inspired no-reference region-of-interest quality measure for brain MRI images. Journal of Medical Imaging 4(2):025504

55. Obuchowicz R, Oszust M, Bielecka M, Bielecki A, Piórkowski A (2020) Magnetic resonance image quality assessment by using non-maximum suppression and entropy analysis. Entropy 22(2):220
56. Chabert S, Castro JS, Muñoz L, Cox P, Riveros R, Vielma J, Huerta G, Querales M, Saavedra C, Veloz A et al (2021) Image quality assessment to emulate experts' perception in lumbar MRI using machine learning. Appl Sci 11(14):6616
57. Esses SJ, Lu X, Zhao T, Shanbhogue K, Dane B, Bruno M, Chandarana H (2018) Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture. J Magn Reson Imaging 47(3):723–728
58. Sujit SJ, Coronado I, Kamali A, Narayana PA, Gabr RE (2019) Automated image quality evaluation of structural brain mri using an ensemble of deep learning networks. J Magn Reson Imaging 50(4):1260–1267
59. Ma JJ, Nakarmi U, Kin CYS, Sandino CM, Cheng JY, Syed AB, Wei P, Pauly JM, Vasanawala SS (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
60. Köhler T, Budai A, Kraus MF, Odstrčilik J, Michelson G, Hornegger J (2013) Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems, IEEE, pp 95–100
61. Wang S, Jin K, Lu H, Cheng C, Ye J, Qian D (2015) Human visual system-based fundus image quality assessment of portable fundus camera photographs. IEEE Trans Med Imaging 35(4):1046–1055
62. Remeseiro B, Mendonça AM, Campilho A (2017) Objective quality assessment of retinal images based on texture features. In: 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, pp 4520–4527
63. Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2(4):433–459
64. Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):1–27
65. Wang Z, Sheikh HR, Bovik AC (2002) No-reference perceptual quality assessment of JPEG compressed images. In: Proc of the international conference on image processing
66. Hadjidemetriou E, Grossberg MD, Nayar SK (2004) Multiresolution histograms and their use for recognition. IEEE Trans Pattern Anal Mach Intell 26(7):831–847
67. Imagenet classification with deep convolutional neural networks (2012) Krizhevsky A, Sutskever I, Hinton G.E. Adv Neural Inf Process Syst 25:1097–1105
68. Di Martino A, Yan C-G, Li Q, Denio E, Castellanos FX, Alaerts K, Anderson JS, Assaf M, Bookheimer SY, Dapretto M et al (2014) The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19(6):659–667
69. Lublin FD, Cofield SS, Cutter GR, Conwit R, Narayana PA, Nelson F, Salter AR, Gustafson T, Wolinsky JS, Investigators C (2013) Randomized study combining interferon and glatiramer acetate in multiple sclerosis. Ann Neurol 73(3):327–340
70. Zhu X, Milanfar P (2010) Automatic parameter selection for denoising algorithms using a no-reference measure of image content. IEEE Trans Image Process 19(12):3116–3132
71. Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Syst Man Cybern 6:610–621
72. Coyner AS, Swan R, Campbell JP, Ostmo S, Brown JM, Kalpathy-Cramer J, Kim SJ, Jonas KE, Chan RP, Chiang MF et al (2019) Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. Ophthalmology Retina 3(5):444–450
73. Raj A, Shah NA, Tiwari AK, Martini MG (2020) Multivariate regression-based convolutional neural network model for fundus image quality assessment. IEEE Access 8:57810–57821
74. Sevik U, Kose C, Berber T, Erdol H (2014) Identification of suitable fundus images using automated quality assessment methods. J Biomed Opt 19(4):046006
75. Fu H, Wang B, Shen J, Cui S, Xu Y, Liu J, Shao L (2019) Evaluation of retinal image quality assessment networks in different color-spaces. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 48–56
76. Shen Y, Sheng B, Fang R, Li H, Dai L, Stolte S, Qin J, Jia W, Shen D (2020) Domain-invariant interpretable fundus image quality assessment. Med Image Anal 61:101654
77. Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial Discriminative Domain Adaptation. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2962–2971
78. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Advances in Neural Information Processing Systems 27

79. Abramovich O, Pizem H, Van Eijgen J, Oren I, Melamed J, Stalmans I, Blumenthal EZ, Behar JA (2023) FundusQ-Net: a regression quality assessment deep learning algorithm for fundus images quality grading. Comput Methods Programs Biomed 239:107522
80. Niwas SI, Jakhetiya V, Lin W, Kwoh CK, Sng CC, Aquino MC, Victor K, Chew PTK (2016) Complex wavelet based quality assessment for AS-OCT images with application to Angle Closure Glaucoma diagnosis. Comput Methods Programs Biomed 130:13–21
81. Khan ZA, Beghdadi A, Kaaniche M, Cheikh FA (2020) Residual networks based distortion classification and ranking for laparoscopic image quality assessment. In: 2020 IEEE International conference on image processing (ICIP), IEEE, pp 176–180
82. Ali S, Zhou F, Bailey A, Braden B, East JE, Lu X, Rittscher J (2021) A deep learning framework for quality assessment and restoration in video endoscopy. Med Image Anal 68:101900
83. Abdi AH, Luong C, Tsang T, Allan G, Nouranian S, Jue J, Hawley D, Fleming S, Gin K, Swift J et al (2017) Automatic quality assessment of echocardiograms using convolutional neural networks: feasibility on the apical four-chamber view. IEEE Trans Med Imaging 36(6):1221–1230
84. Baldeon-Calisto M, Rivera-Velastegui F, Lai-Yuen SK, Riofrío D, Pérez-Pérez N, Benítez D, Flores-Moyano R (2024) DistilIQA: Distilling vision transformers for no-reference perceptual CT image quality assessment. Comput Biol Med p 108670
85. Tang L, Tian C, Qian J, Li L (2018) No reference quality evaluation of medical image fusion. Int J Imaging Syst Technol 28(4):267–273
86. Tang L, Tian C, Li L, Hu B, Yu W, Xu K (2020) Perceptual quality assessment for multimodal medical image fusion. Signal Process Image Commun 85:115852
87. Outtas M, Zhang L, Deforges O, Hammidouche W, Serir A, Cavaro-Menard C (2016) A study on the usability of opinion-unaware no-reference natural image quality metrics in the context of medical images. In: 2016 International symposium on signal, image, video and communications (ISIVC), pp 308–313
88. Khan ZA, Beghdadi A, Cheikh FA, Kaaniche M, Pelanis E, Palomar R, Fretland ÅA, Edwin B, Elle OJ (2020) Towards a video quality assessment based framework for enhancement of laparoscopic videos. In: Medical imaging 2020: image perception, observer performance, and technology assessment, vol 11316, pp 113160. International society for optics and photonics
89. Liu X, Van De Weijer J, Bagdanov AD (2017) RankIQA: Learning from rankings for no-reference image quality assessment. In: Proceedings of the IEEE international conference on computer vision, pp 1040–1049
90. Lee W, Wagner F, Maier A, Wang A, Baek J, Hsieh SS, Choi J-H (2023). Low-dose Computed Tomography Perceptual Image Quality Assessment Grand Challenge Dataset (MICCAI)
91. Saha A, Wu QMJ (2015) Utilizing image scales towards totally training free blind image quality assessment. IEEE Trans Image Process 24(6):1879–1892
92. Zhang L, Cavaro-Ménard C, Le Callet P (2012) Key issues and specificities for the objective medical image quality assessment. In: Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), pp 1–6
93. Eck BL, Fahmi R, Brown KM, Zabic S, Raihani N, Miao J, Wilson DL (2015) Computational and human observer image quality evaluation of low dose, knowledge-based CT iterative reconstruction. Medical Physics 42(10):6098–6111
94. Brankov JG (2013) Evaluation of the channelized hotelling observer with an internal-noise model in a train-test paradigm for cardiac spect defect detection. Phys Med Biol 58(20):7159
95. Racine D, Ba AH, Ott JG, Bochud FO, Verdun FR (2016) Objective assessment of low contrast detectability in computed tomography with Channelized Hotelling Observer. Physica Medica 32(1):76–83
96. Richard S, Siewerdsen JH (2008) Comparison of model and human observer performance for detection and discrimination tasks using dual-energy x-ray images. Med Phys 35(11):5043–5053
97. Li K, Zhou W, Li H, Anastasio MA (2021) Assessing the impact of deep neural network-based image denoising on binary signal detection tasks. IEEE Trans Med Imaging 40(9):2295–2305
98. Barrett HH, Yao J, Rolland JP, Myers KJ (1993) Model observers for assessment of image quality. Proc Natl Acad Sci 90(21):9758–9765
99. Cormack LK (2005) Computational models of early human vision. In: BOVIK A (ed) Handbook of Image and Video Processing (2nd edn), Second edition edn. Communications, Networking and Multimedia, Academic Press, Burlington, p 325
100. Kopp FK, Catalano M, Pfeiffer D, Fingerle AA, Rummeny EJ, Noël PB (2018) CNN as model observer in a liver lesion detection task for X-ray computed tomography: a phantom study. Med Phys 45(10):4439–4447
101. Zhou W, Li H, Anastasio MA (2019) Approximating the Ideal Observer and Hotelling Observer for binary signal detection tasks by use of supervised learning methods. IEEE Trans Med Imaging 38(10):2456–2468

102. Alnowami M, Mills G, Awis M, Elangovanr P, Patel M, Halling-Brown M, Young K, Dance DR, Wells K (2018) A deep learning model observer for use in alterative forced choice virtual clinical trials. In: Medical imaging 2018: image perception, observer performance, and technology assessment, vol 10577, p 105770. International society for optics and photonics
103. Wu L, Cheng J-Z, Li S, Lei B, Wang T, Ni D (2017) FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans Cybern 47(5):1336–1349
104. Zhou W, Li H, Anastasio MA (2020) Approximating the Ideal Observer for joint signal detection and localization tasks by use of supervised learning methods. IEEE Trans Med Imaging 39(12):3992–4000
105. Lorente I, Abbey CK, Brankov JG (2020) Deep learning based model observer by U-Net. In: Medical imaging 2020: image perception, observer performance, and technology assessment, vol 11316, pp 113160. International society for optics and photonics
106. Welikala R, Fraz M, Foster P, Whincup P, Rudnicka AR, Owen CG, Strachan D, Barman SA et al (2016) Automated retinal image quality assessment on the UK Biobank dataset for epidemiological studies. Comput Biol Med 71:67–76
107. Rodrigues R, Pinheiro AM (2019) A quality of recognition case study: texture-based segmentation and MRI quality assessment. In: 2019 27th European signal processing conference
108. Alais R, Dokládal P, Erginay A, Figliuzzi B, Decencière E (2020) Fast macula detection and application to retinal image quality assessment. Biomed Signal Process Control 55:101567
109. Kupinski MA, Hoppin JW, Clarkson E, Barrett HH (2003) Ideal-observer computation in medical imaging with use of Markov-chain Monte Carlo techniques. JOSA A 20(3):430–438
110. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 234–241
111. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M et al (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12(3):1001779
112. Fraz MM, Welikala R, Rudnicka AR, Owen CG, Strachan D, Barman SA (2015) QUARTZ: quantitative Analysis of Retinal Vessel Topology and size–an automated system for quantification of retinal vessels morphology. Expert Syst Appl 42(20):7221–7234
113. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: International conference on machine learning, vol 96, pp 148–156
114. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Computer vision and pattern recognition, 2005. CVPR 2005. IEEE Computer Society Conference On, IEEE, vol 1, pp 886–893
115. Agaian S, Almuntashri A (2009) Noise-resilient edge detection algorithm for brain MRI images. In: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, IEEE, pp 3689–3692
116. Decenciere E, Cazuguel G, Zhang X, Thibault G, Klein J-C, Meyer F, Marcotegui B, Quellec G, Lamard M, Danno R et al (2013) TeleOphta: machine learning and image processing methods for teleophthalmology. IRBM 34(2):196–203
117. Farnell DJ, Hatfield F, Knox P, Reakes M, Spencer S, Parry D, Harding SP (2008) Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. J Franklin Inst 345(7):748–765
118. Suad J, Jbara W (2013) Subjective quality assessment of new medical image database. Int J Comput Eng Technol 4:155–164
119. Outtas M, Zhang L, Deforges O, Serir A, Hamidouche W (2018) Subjective and objective evaluations of feature selected multi output filter for speckle reduction on ultrasound images. Phys Med Biol 63(18)
120. Willemink M, Koszek W, Hardell C, Wu J, Fleischmann D, Harvey H, Folio L, Summers R, Rubin D, Lungren M (2020) Preparing medical imaging data for machine learning. Radiology 295(1)
121. Li Y, Ercisli S (2023) Explainable human-in-the-loop healthcare image information quality assessment and selection. CAAI Trans Intell Technol
122. Alexander RG, Waite S, Macknik SL, Martinez-Conde S (2020) What do radiologists look for? Advances and limitations of perceptual learning in radiologic search. J Vis 20(10):17
123. Oh G, Lee JE, Ye JC (2021) Unpaired MR motion artifact deep learning using outlier-rejecting bootstrap aggregation. IEEE Trans Med Imaging 40(11):3125–3139
124. Yang J, Faraji M, Basu A (2019) Robust segmentation of arterial walls in intravascular ultrasound images using Dual Path U-Net. Ultrasonics 96:24–33
125. Oktaviana A, Pawiro S, Siswatining T, Soejoko D (2019) Preliminary study of ring artifact detection in SPECT imaging using Jaszczak phantom. In: Journal of physics: conference series, IOP Publishing, vol 1248, pp 012030

126. Hu R, Yang R, Liu Y, Li X (2021) Simulation and mitigation of the wrap-around artifact in the MRI image. Front Comput Neurosci 15:89
127. Makhlouf A, Maayah M, Abughanam N, Catal C (2023) The use of generative adversarial networks in medical image augmentation. Neural Comput Appl 35(34):24055–24068
128. Kalayah MM, Marin T, Brankov JG (2013) Generalization evaluation of machine learning numerical observers for image quality assessment. IEEE Trans Nucl Sci 60(3):1609–1618
129. Martini MG, Hewage CT, Nasralla MM, Smith R, Jourdan I, Rockall T (2013) 3D robotic tele-surgery and training over next generation wireless networks. In: 2013 35th Annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, pp 6244–6247
130. Nagoor OH, Whittle J, Deng J, Mora B, Jones MW (2020) Lossless compression For volumetric medical images using deep neural network with local sampling. In: 2020 IEEE international conference on image processing (ICIP), IEEE, pp 2815–2819
131. Kara PA, Kovacs PT, Vagharshakyan S, Martini MG, Imre S, Barsi A, Lackner K, Balogh T (2017) Perceptual quality of reconstructed medical images on projection-based light field displays. eHealth 360°. Springer, Cham, pp 476–483
132. Han Y, Yuan Z, Muntean G-M (2016) An innovative no-reference metric for real-time 3D stereoscopic video quality assessment. IEEE Trans Broadcast 62(3):654–663
133. Hewage CT, Martini MG (2013) Quality of experience for 3D video streaming. IEEE Commun Mag 51(5):L101-107
134. Hewage CT, Martini MG (2011) Reduced-reference quality assessment for 3D video compression and transmission. IEEE Trans Consum Electron 57(3):1185–1193
135. Battisti F, Bosc E, Carli M, Le Callet P, Perugia S (2015) Objective image quality assessment of 3D synthesized views. Signal Process Image Commun 30:78–88
136. Ak A, Le Callet P (2019) Investigating epipolar plane image representations for objective quality evaluation of light field images. In: European workshop on visual information processing, pp 135–139
137. Tamboli RR, Kara PA, Cserkaszky A, Barsi A, Martini MG, Appina B, Channappayya SS, Jana S (2018) 3D objective quality assessment of light field video frames. In: 3DTV-conference: the true vision-capture, transmission and display of 3D video
138. Tamboli RR, Cserkaszky A, Kara PA, Barsi A, Martini MG (2018) Objective quality evaluation of an angularly-continuous light-field format. In: International conference on 3D immersion
139. Viola I, Řeřábek M, Bruylants T, Schelkens P, Pereira F, Ebrahimi T (2016) Objective and subjective evaluation of light field image compression algorithms. In: Picture coding symposium