Original Article

A deep learning-based automatic image quality assessment method for respiratory phase on computed tomography chest images

Jialin Su[1#], Meifang Li[2,3#], Yongping Lin[1], Liu Xiong[1], Caixing Yuan[2], Zhimin Zhou[2], Kunlong Yan[2]

[1]School of Optoelectronic and Communication Engineering, Xiamen University of Technology, Xiamen, China; [2]Department of Medical Imaging, Affiliated Hospital of Putian University, Putian, China; [3]School of Clinical Medicine, Fujian Medical University, Fuzhou, China

Contributions: (I) Conception and design: Y Lin; (II) Administrative support: Y Lin, M Li; (III) Provision of study materials or patients: M Li, C Yuan, Z Zhou, K Yan; (IV) Collection and assembly of data: J Su, L Xiong; (V) Data analysis and interpretation: J Su, L Xiong; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Yongping Lin, PhD. School of Optoelectronic and Communication Engineering, Xiamen University of Technology, 600 Ligong Road, Jimei District, Xiamen 361024, China. Email: yplin@t.xmut.edu.cn.

Background: Computed tomography (CT) chest scans have become commonly used in clinical diagnosis. Image quality assessment (IQA) for CT images plays an important role in CT examination. It is worth noting that IQA is still a manual and subjective process, and even experienced radiologists make mistakes due to human limitations (fatigue, perceptual biases, and cognitive biases). There are also kinds of biases because of poor consensus among radiologists. Excellent IQA methods can reliably give an objective evaluation result and also reduce the workload of radiologists. This study proposes a deep learning (DL)-based automatic IQA method, to assess whether the image quality of respiratory phase on CT chest images are optimal or not, so that the CT chest images can be used in the patient's physical condition assessment.

Methods: This retrospective study analysed 212 patients' chest CT images, with 188 patients allocated to a training set (150 patients), validation set (18 patients), and a test set (20 patients). The remaining 24 patients were used for the observer study. Data augmentation methods were applied to address the problem of insufficient data. The DL-based IQA method combines image selection, tracheal carina segmentation, and bronchial beam detection. To automatically select the CT image containing the tracheal carina, an image selection model was employed. Afterward, the area-based approach and score-based approach were proposed and used to further optimize the tracheal carina segmentation and bronchial beam detection results, respectively. Finally, the score about the image quality of the patient's respiratory phase images given by the DL-based automatic IQA method was compared with the mean opinion score (MOS) given in the observer study, in which four blinded experienced radiologists took part.

Results: The DL-based automatic IQA method achieved good performance in assessing the image quality of the respiratory phase images. For the CT sequence of the same patient, the DL-based IQA method had an accuracy of 92% in the assessment score, while the radiologists had an accuracy of 88%. The Kappa value of the assessment score between the DL-based IQA method and radiologists was 0.75, with a sensitivity of 85%, specificity of 91%, positive predictive value (PPV) of 92%, negative predictive value (NPV) of 93%, and accuracy of 88%.

Conclusions: This study develops and validates a DL-based automatic IQA method for the respiratory phase on CT chest images. The performance of this method surpassed that of the experienced radiologists on the independent test set used in this study. In clinical practice, it is possible to reduce the workload of radiologists and minimize errors caused by human limitations.

Keywords: Automatic image quality assessment (automatic IQA); deep learning (DL); chest computed tomography respiratory phase images (chest CT respiratory phase images)

Submitted Sep 08, 2023. Accepted for publication Jan 04, 2024. Published online Mar 07, 2024.
doi: 10.21037/qims-23-1273
View this article at: https://dx.doi.org/10.21037/qims-23-1273

Introduction

Computed tomography (CT) is one of the most widely used medical imaging techniques, and CT chest scans have become commonly used in clinical diagnosis (1). Considering that CT scans are relatively high-dose procedures, it can be damaging to patients' health if an excessive number of CT scans are performed [due to potential carcinogenesis from medical imaging (2)]. Therefore, medical imaging should adhere to the principle of 'AS Low As Reasonably Achievable' (ALARA). The ALARA principle involves limiting the radiation dose received by patients during medical imaging examination while ensuring optimal image quality (3). Image quality assessment (IQA) plays an important role in balancing between image quality and radiation dose in medical imaging. CT IQA is of great significance for optimizing CT software, hardware design and controlling patient scanning dose. As pointed out by Cai *et al.* (4), image quality is influenced by various factors, including CT hardware, scanning protocol, patient motion, reconstruction algorithms, and image post-processing algorithms. Throughout the entire medical process, image quality can be evaluated from the physical, algorithmic, diagnostic, and retrieval layers. Parameters set in each layer differ, leading to varying impact on the image quality.

This study focuses on IQA at the diagnostic level, which is task specific IQA aiming at controlling patient scanning dose by accurately and efficiently assessing CT image quality. Inadequate respiratory during the scan of patients will lead to artifacts in CT images, which will reduce the quality and diagnosability of the images, and may even cause problems such as misdiagnosis or incomplete scanning (5). Therefore, patients may need to be rescanned, leading to increase in radiation dose, wasted resources and a growing workload on radiologists (6). IQA is crucial for controlling the rate of patient rescans (7). By accurately evaluating the image quality of patients, IQA aims to minimize the risk of misclassifying patients with acceptable image quality as unacceptable, thereby reducing the likelihood of unnecessary rescans for patients.

The quality of CT images is one of the factors that affect the accuracy of clinical diagnosis decisions after the examination (8,9). In most cases, whether a CT image quality is good or bad, it is influenced by many factors such as inspiration (10), the field of view (11), and position (12), etc. A major factor influencing the image quality is due to insufficient inspiration during a CT chest examination (5). There are two aspects that can be used to judge whether the degree of inspiration is sufficient or not, i.e., the tracheal carina and bronchial beam. The shape of tracheal carina and the clarity of the bronchial beam are both visualizations of the patient's inspiration, i.e., tracheal morphology and the clarity of bronchial beam changes as the patient breathes. Tracheal carina morphology is classified as convex, flat, or bowed inward toward the tracheal lumen, based on the configuration of the posterior membrane, and a normal tracheal carina is ovoid (13). A clearer bronchial beam is visualized when the inspiration of the patient is sufficient (14). Insufficient inspiration run counter to the requirement of taking a CT image which is guided by radiographers may lead to time wasted in rescanning and an increase in the dose of radiation to the patient (15). Thus, it is important to determine whether CT images meet clinical diagnostic requirements (i.e., image quality standards), and it is worth noting that is still a manual process, such as collecting data through Likert scale questionnaire for subjective image quality analysis, and has kinds of biases because of poor consensus among observers (16).

With the development of deep learning (DL), DL as a Computer Aided Diagnosis (CAD) tool has been more and more extensively applied in a wide range of fields in recent years (17-19). In the field of medical image processing, that is mainly applied on processing radiographic images such as magnetic resonance (MR) images, CT images, and ultrasonic images. People use DL algorithms because the DL works similarly to the human brain, especially in the applications of medical image processing such as classification (20), segmentation (21), and detection (22). DL algorithms have shown high performance in evaluating not only the field of view (FOV) (23), but also rotation, inspiration, and patient position (24). Although Poggenborg *et al.* (24) had an assessment of inspiration, the study mainly showed a relatively large improvement of 28% in images with optimal collimation, and only had a relative improvement of 4% in images with optimal inspiration. There was no obvious improvement in selecting the CT images with optimal inspiration compared to manual assessment by radiologists. Nousiainen *et al.* (25) used two convolutional neural networks (CNNs) to estimate lung inclusion, rotation, and inspiration but there were still limitations by the ambiguous score (i.e., the model is most likely to concur with the observer whose annotations were used to train the model, which raises some bias to model performance).

To our knowledge, there is currently no research that combines tracheal carina and bronchial beams for the automatic assessment of respiratory phase image quality in CT chest images and achieving high performance. Therefore, in this retrospective study, a DL-based automatic IQA method is proposed based on the tracheal carina and bronchial beams. This method was used to evaluate whether the respiratory phase image quality of CT chest images were optimal or not so that the CT chest images can be used for diagnosing the patient, thereby determining the presence of any diseases. The sufficient degree of inspiration was used to qualitatively measure the image quality, and multiple evaluation metrics were employed to validate the performance of the method.

The major contributions of this study are summarized as follows:

(I) An image quality evaluation method based on DL is proposed to evaluate the image quality of respiratory phase CT images. Compared to traditional IQA methods, our proposed method can automatically select CT images with region of interest (ROI) from a series of CT images of patients and evaluate their image quality, reducing the workload of radiologists and the errors caused by human limitations such as fatigue, perceptual biases, and cognitive biases.

(II) An area-based method is proposed to improve the segmentation performance of the segmentation model. Unlike traditional segmentation models, which directly use the segmentation results as the final classification results, considering that the morphological similarity of ROI (tracheal carina) may cause classification errors in segmentation results, our proposed method can help reduce such errors and improve the performance of the segmentation model.

(III) A score-based method is proposed to adapt object detection models for morphology-based organ IQA tasks (i.e., the clarity assessment of bronchial beams). Traditional detection models often directly utilize the detection results for a single image as the final classification outcome. However, when it comes to a morphology-based assessment of image quality for organs, it often involves reviewing multiple slices since the organ may span across these slices. Therefore, our proposed score-based method aims to enable traditional detection models to review multiple slices when performing morphology-based organ IQA tasks.

In conclusion, the area-based approach and the score-based approach proposed in this paper improve the accuracy and reliability of automatic IQA of respiratory CT images, which can potentially aid radiologists in clinical diagnosis and decision-making.

Methods

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board (IRB) of Fujian Putian Hospital in China (No. 202132) and individual consent for this retrospective analysis was waived.

In many target detection tasks, You Only Look Once (YOLO) series (26-29) plays an important role in one-stage detectors. U-Net (30) is a semantic segmentation model based on fully convolutional networks (FCN), which is widely used in medical image segmentation. In this study, the image quality of the High-Resolution CT respiratory phase images was assessed from the perspective of patient's sufficient degree of inspiration, which was mainly considered from two aspects: the morphology of tracheal carina and the clarity of the bronchial beam. The flowchart in *Figure 1* shows the entire research process. The whole process from the input of the CT image sequence to the output of classification is automated. In terms of morphological classification of the tracheal carina, the state-of-the-art (SOTA) model YOLOv8 (29) as an image selection model was used to select the image with a just separated trachea from the original CT sequence. Then, the image with a spacing between slices of about 10 mm relative to the CT image selected by YOLOv8 was fed into the

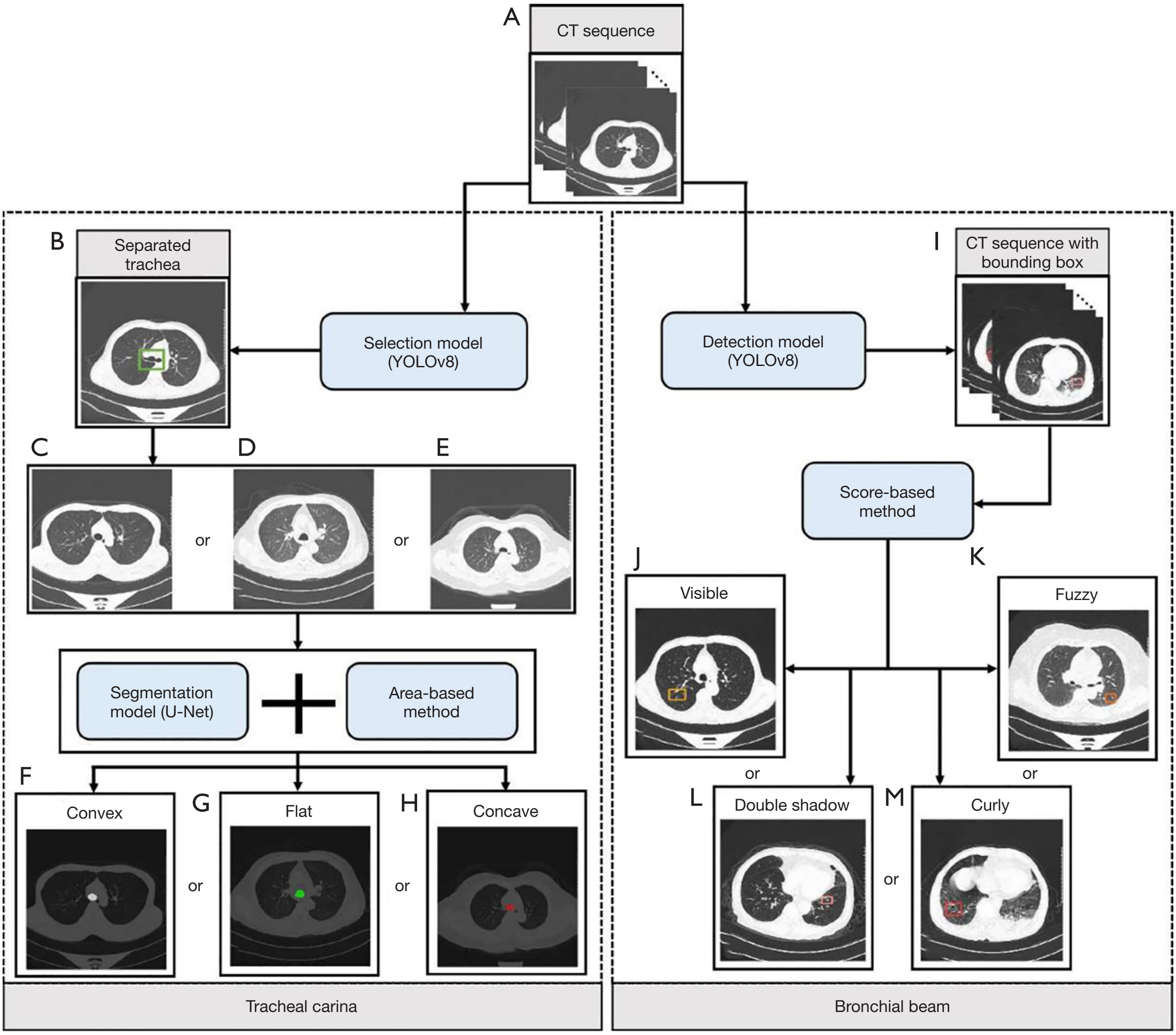

Figure 1 A flowchart that shows the entire automatic evaluation process. (A) is the original CT image sequence. (B) is a CT image selected by YOLOv8, where the green bounding box contains the separated trachea. (C-E) are the tracheal carina CT image which has a spacing between slices of about 10mm relative to the YOLOv8 selected CT image get from (B). (F-H) are the CT image that combines U-Net with the area-based method to segment from (C-E), the white, green, and red region represent the convex, flat, and concave tracheal carina respectively. (I) is the CT sequence with bounding box detected by YOLOv8. (J-M) are the CT image predicted by combining YOLOv8 with the score-based method, where the yellow, orange, pink, and red bounding box contains the visible, fuzzy, double shadow, and curly bronchial beam respectively. CT, computed tomography.

segmentation model (U-Net) to segment the tracheal carina region. It is worth noting that due to the morphological fuzziness of tracheal carina, the area-based method was used to further classify the segmentation result, and finally, the classification result of tracheal carina morphology was obtained. To judge the clarity of the bronchial beam better, YOLOv8 was used again to detect the bronchial beam from the original CT sequence, due to the excellent detection ability in the ablation experiment. The bronchial beam was bounded with detection boxes representing different clarity

Table 1 Annotation and rating standard

Type	Classification	Score
Separated trachea	–	–
Tracheal carina	Convex	10
	Flat	5
	Concave	0
Bronchial beam	Visible	15
	Fuzzy	10
	Double shadow	5
	Curly	0
Inspiration sufficient degree	Sufficient	20, 25
	Insufficient	0, 5, 10, 15

Table 2 Patient's information (n=188) for the data sets used in the three models training

	n (%)	Gender, n (%)		Age (year)[a]
		Male	Female	
Separated trachea	n=80	46 (57.50)	34 (45.50)	50±19.8
Tracheal carina	n=53			47.2±27.6
Convex	18 (33.96)	11 (61.11)	7 (38.89)	
Flat	18 (33.96)	11 (61.11)	7 (38.89)	
Concave	17 (32.08)	9 (52.94)	8 (47.06)	
Bronchial beam	n=55			56.1±17.4
Visible	10 (18.18)	7 (70.00)	3 (30.00)	
Fuzzy	16 (29.09)	7 (43.75)	9 (56.25)	
Double shadow	14 (25.45)	7 (50.00)	7 (50.00)	
Curly	15 (27.27)	6 (40.00)	9 (60.00)	

[a], the data are presented in the table as mean ± standard deviation.

classifications. Then, the final clarity detection result was obtained through the score-based method. The above three DL-based models were used to teach machines to assess the image quality of respiratory phase CT images like an experienced radiologist.

Imaging protocol

All patients were scanned on a SIEMENS SOMATOM DEFINITION DUAL SOURCE CT scanner (manufacturer: SIEMENS) with a tube voltage of 120 kV and a tube current of 50–800 mA. The nominal slice thickness was 1 mm.

Table 3 Dataset split

Type	Training set	Validation set	Testing set
Separated trachea	64	8	8
Tracheal carina	43	5	5
Bronchial beam	43	5	7

Data preparation and processing

Inspired by the anatomy of the trachea, carina and bronchi in the literature (31), a DL-based automatic IQA method for respiratory phase on CT chest images was proposed. After discussions of the annotation and rating standard for evaluating the sufficient degree of inspiration with a focus group that consisted of radiology residents, radiologists, and technicians, the annotation and rating standard are summarized in *Table 1*.

In this study, the CT chest images of 212 patients were used as the original dataset. Among them, 188 patients were employed for training the three models as follows: image selection model, semantic segmentation model, and detection model. The remaining 24 patients were used for the observer study, which aimed to validate the reliability of our IQA method. The ROIs in the CT chest images of all 212 patients were annotated manually as the ground truth by three radiologists (with five, eight, and ten years of experience in radiology, respectively) according to the criteria in *Table 1*.

The three models were trained and validated using 168 patients and were tested using 20 patients. Thus, a total of 188 patients were used to determine the degree of inspiration. As shown in *Table 2*, we used 80 patients (mean age, 50±19.8 years; 46 males, 34 females) as the dataset for the image selection model, 53 patients (mean age, 47.2±27.6 years; 31 males, 22 females) for the detection model, and 55 patients (mean age, 56.1±17.4 years; 27 males, 28 females) for the segmentation model. To train all three models, determine the hyper-parameters of all three models, and evaluate the performance of all three models, we used the dataset split shown in *Table 3*.

A patient's CT image with concave tracheal carina, double shadow, or curly bronchial beam was judged to be an insufficient inspiration. Thus, the sufficient degree of inspiration was determined to be sufficient (score: 20, 25) or insufficient (score: 0, 5, 10, 15) according to the rating

Table 4 Patient's information for the independent test set used in an observer study

Parameters	Observer study (n=24)		P value
	Sufficient inspiration	Insufficient inspiration	
Population	12 (50%)	12 (50%)	
Age (years)[a]	53.6±17.8	68.6±13.4	0.029
Sex			1
Male	7 (58.3%)	5 (41.7%)	
Female	5 (41.7%)	7 (58.3%)	
Tracheal carina type			1
Convex	12 (100%)	0 (0%)	
Flat	0 (0%)	9 (75%)	
Concave	0 (0%)	3 (25%)	
Bronchial beam type			1
Visible	6 (50%)	0 (0%)	
Fuzzy	6 (50%)	5 (41.7%)	
Double shadow	0 (0%)	4 (33.3%)	
Curly	0 (0%)	3 (25.0%)	

[a], the data are presented in the table as mean ± standard deviation.

standard as shown in *Table 1*. An independent test set of 24 patients was used in the observer study, which consisted of 12 patients (mean age, 53.6±17.8 years; 7 males, 5 females) of whom the inspiration was sufficient, and 12 patients (mean age, 68.6±13.4 years; 5 males, 7 females) of whom the inspiration was insufficient. The details of the independent test set are shown in *Table 4*, and it should be noted that patients with sufficient inspiration led to possessing relatively high score classification, and those with insufficient inspiration led to possessing relatively low score classification in the judgment of tracheal carina morphology and bronchial beam clarity.

Due to limited computational resources, the images read from the digital imaging and communications in medicine (DICOM) files were resized according to the 'Window Center' tag and the 'Window Width' tag using window-level algorithm (32) into 512×512 pixels. Afterward, data augmentation was performed in the preprocessing stage to improve the model robustness, which by multi-combination means as followed: (I) horizontal flipping, (II) random scaling, (III) random cropping and padding, (IV) perspective transformation, (V) gaussian noise, and (VI) gaussian blur. All methods are visualized in *Figure 2*.

After data augmentation of the raw image, an additional 880 images, 294 images, and 720 images were generated for the image selection model, segmentation model and detection model, respectively.

Models training

In this study, YOLOv8 was used as the image selection model and detection model, while U-Net acted as the segmentation model. The transfer learning strategy was used on all the three models. The Common Objects in Context (COCO) (33) dataset is an extensive dataset for object recognition, segmentation, and labeling, and the weight pre-trained on it was used to train YOLOv8. The ImageNet (34) is an image database that uses the hierarchical structure of WordNet (35), and the weight pre-trained on it was used to train U-Net. All images were standardized before being fed into the model, by subtracting them with the mean and then dividing them by the standard deviation of the image.

Different strategies were employed to train different models. For training YOLOv8 and to improve the model robustness, a mosaic data augmentation was used with fifty percent probability, followed by using a mix up data augmentation on images with mosaic data augmentation being used, which had a fifty percent probability. The stochastic gradient descent (SGD) algorithm with an initial learning rate of 0.01 was used to minimize the binary-cross-entropy loss, the complete intersection over union (CIoU) (36) loss, and the distribution focal loss (DFL), it is worth noting that for all bounding boxes that were considered true positive must have a CIoU greater than or equal to 0.70. For training the U-Net, the adaptive moment estimation (ADAM) algorithm with an initial learning rate of 0.0001 was used to minimize the cross-entropy loss and dice loss.

The segmentation model was trained using Tensorflow (version: 2.9.2, Google) while the selection and detection models were trained using Pytorch (version: 1.12.1, Facebook), and all of them were programmed in Python (version 3.10.6). The experiments were conducted on a workstation equipped with two NVIDIA RTX 2080Ti GPUs.

Determination of the sufficient degree of inspiration

An area-based approach was used to perform tracheal

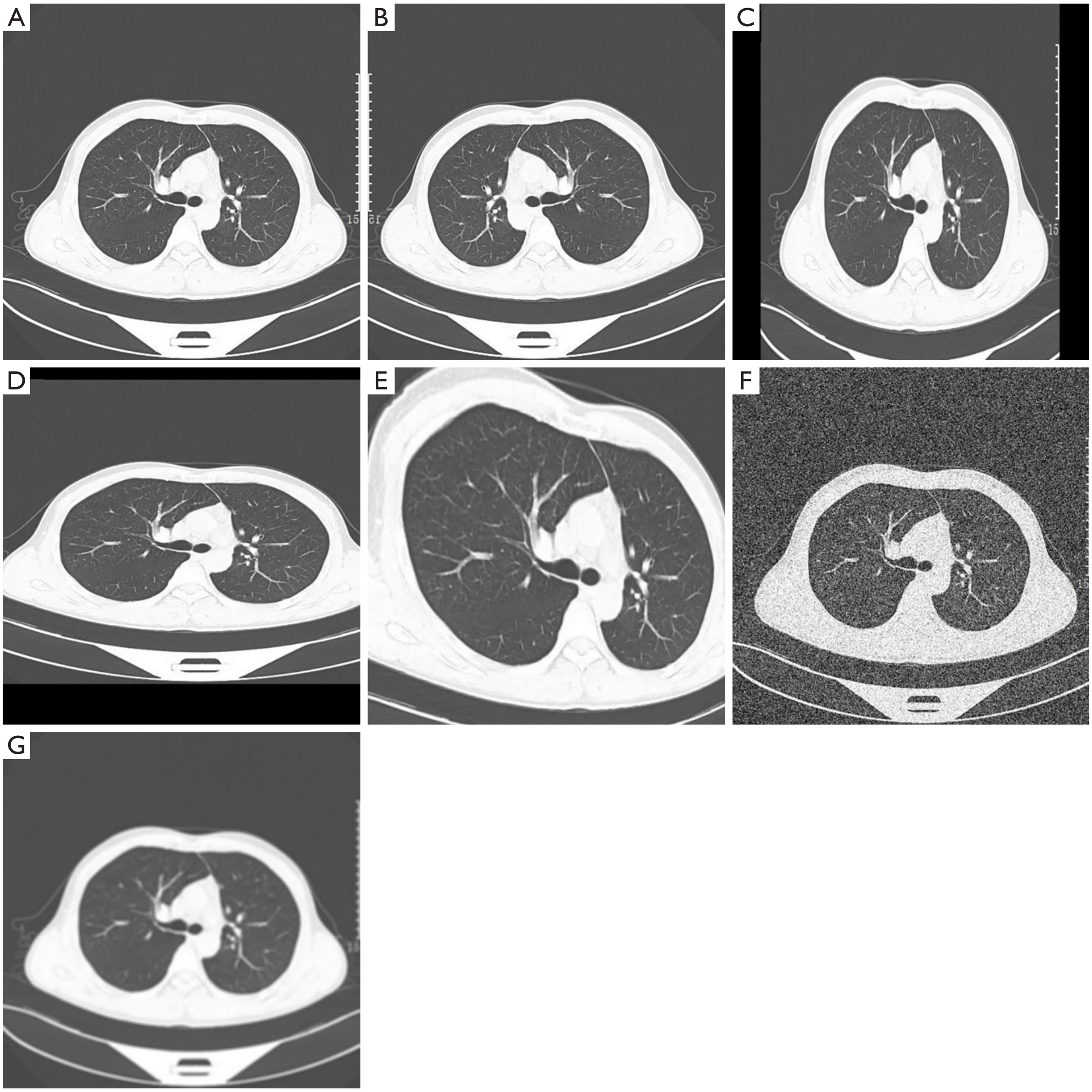

Figure 2 The original CT image and its results after processing with different data augmentation methods. (A) is the original CT image, while (B), (C), (D), (E), (F), and (G) represent the image of (A) following horizontal flipping, random scaling, random cropping and padding, perspective transformation, gaussian noise and gaussian blur processing, respectively. CT, computed tomography.

carina classifications on the segmented images predicted by the segmentation model in the test set (the segmentation results are shown in *Figure 3*). Due to the possibility that there were morphological similarity between different classifications of tracheal carina, the segmentation model may predict multiple colors which represented different classifications for tracheal carina of which its morphology was difficult to distinguish, that kind of prediction result is shown in *Figure 3H*. Therefore, the classification of that tracheal carina can be determined by calculating the proportion of the area of each color area to the whole segmentation region area, which is calculated as:

$$\mathrm{R}_i\left(Cls\right)=\frac{A_i\left(Cls\right)}{\sum_{i=1}^{n}A_i\left(Cls\right)} \quad [1]$$

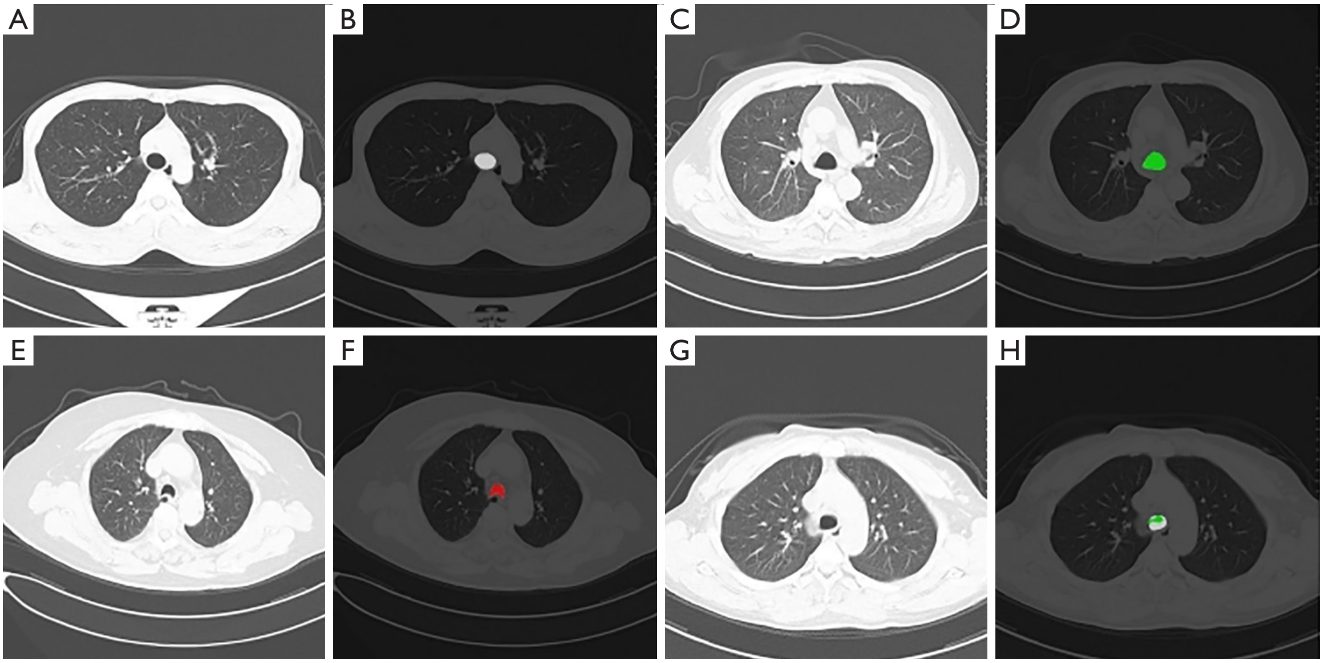

Figure 3 The segmentation results of tracheal carina which are segmented by the segmentation model. (A), (C), (E), and (G) are four CT images with tracheal carina selected from the continuous sequences of four patients. (B), (D), (F), and (H) are the predictive tracheal carina images by the segmentation model. White, green, and red represent convex, flat, and concave respectively. (H) is the predictive tracheal carina image, which contains two colors in the ROI and needs the area-based approach used to further classify. CT, computed tomography; ROI, region of interest.

where $A_i(Cls)$ indicates the area of the classification region. Concretely, the tracheal carina has three morphological classifications, so n is set to 3 in Eq. [1]. i indicates the classification currently being calculated, and the classification corresponding to i is represented as i = *{1: convex, 2: flat, 3: concave}*. In detail, the number of pixels is used as an index to calculate the size of region area. After the $R_i(Cls)$ of each classification is calculated, the morphology of tracheal carina will be classified as convex, flat, or concave when the largest value of $R_i(Cls)$ is $R_1(Cls)$, $R_2(Cls)$, or $R_3(Cls)$, respectively.

A score-based approach combined with the detection model was used to perform bronchial beam classifications on each patient in the test set. Each patient had multiple CT images with bronchial beam, and experienced radiologists usually determine the patient's bronchial beam classification by reviewing multiple images. Therefore, the prediction results of the detection model for the bronchial beam classification of a single CT image were not enough to be used as the patient's bronchial beam classification. A comprehensive consideration of the results predicted by the detection model for all bronchial beam classifications on CT images was needed to determine the patient's bronchial beam classification. The number of prediction boxes for categories in all CT images are calculated as follows:

$$\mathrm{N}(k)=\sum_{j=1}^{n}V_j(k) \qquad [2]$$

where V_j represents the number of predicted bounding boxes on the jth CT image, k = *{1: Visible, 2: Fuzzy, 3: Double shadow, 4: Curly}* since the clarity of the bronchial beam can be divided into four classifications. After calculating N, the classification corresponding to *max* (N (k)) will be determined as the final result for the patient.

The sufficient degree of inspiration of a patient can be determined through the comprehensive consideration of the above two approaches. The general application of the above two approaches are shown in *Figure 1*. It is worth noting that only images with a bounding box that reached CIoU greater than or equal to 0.92 were considered as

Figure 1B.

Evaluation and statistical analysis

To evaluate the performance of our method, patients in the independent test set were used for a blind test. First, the CT image sequence of the patient was input into the trained image selection model to obtain the CT image of the separated trachea and input into the trained detection model to obtain the bounding box containing the bronchial beam in all CT images. Subsequently, the CT image with tracheal carina was found based on the CT image of the separated trachea and then fed into the segmentation model to obtain the region of the tracheal carina. Finally, according to the area-based approach and score-based approach, the region was classified as convex, flat, or concave, and the bronchial beam of the patient was classified as visible, fuzzy, double shadow or curly.

To analyze the results of our method quantitatively, an observer study was conducted by testing on the independent test set, the details of this independent test set are shown in *Table 4*. Four blinded experienced radiologists participated in this observer study, they were not the same group of individuals as the three radiologists who manually annotated the ROIs in the CT chest images as ground truth, and all of them rated the score of the patient's sufficient degree of inspiration respectively on the grounds of the rating standard shown in *Table 1*, and to acquire a more reliable result, the mean opinion score (MOS) of the four radiologists was used as the final rating result. Subsequently, to obtain a binary classification result according to the radiologist's rating result, the sufficient degree of inspiration with ratings 0, 5, 10, and 15 were categorized into the inadequate group, and the other sufficient degree of inspiration were categorized into the adequate group.

The recall, precision, and F1 Score were used as evaluation metrics for three models. Subsequently, the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were compared between our DL-based IQA method, radiologists, and ground truth. Finally, the Kappa value was to reflect the relevance of our DL-based IQA method, ground truth, and the MOS of the four radiologists. Kappa values within the ranges (0.60, 0.80] and (0.81, 1.00] were considered substantial agreement and almost perfect agreement, respectively. All statistical analyzes were performed on excel (version 11.1., KINGSOFT) and P values were obtained by *t*-test, and $P<0.05$ was considered a significant difference.

Results

The segmentation results and the detection results are shown in *Figure 3* and *Figure 4* respectively, the edges of the tracheal carina are clear and differentiate into three classifications as shown in *Figure 3B*,*3D*,*3F*, and all four classifications of the bronchial beam are detected clearly, and distinguished by different colored bounding boxes. *Table 5* shows the recall, precision, and F1 Score of the U-Net and YOLOv8. As shown in *Table 5*, the tracheal carina is well segmented by U-Net, with a recall of 90%, a precision of 92%, and an F1 Score of 91%. The bronchial beam was detected by YOLOv8, and the same evaluation metrics were 92%, 90%, and 91%, respectively.

Three confusion matrices were used to compare the similarities and differences in the evaluation results between this study's DL-based IQA method, radiologists, and ground truth. *Table 6* shows a series of evaluation metrics obtained by confusion matrices shown in *Figure 5*. The Kappa value indicated the high consistency of evaluation results, with 0.83 between our DL-based IQA method and ground truth (P=0.93), 0.75 between radiologists and ground truth (P=0.67), and 0.75 between our DL-based IQA method and radiologists (P=0.75). Furthermore, the assessment accuracy of our method was 0.92, which was higher than that of the radiologists with 0.88, and the probability of making the same judgment as the radiologists was 0.88 on the test set. The P value reflected the difference in scores among all patients in the independent test set. For any pair of our DL-based IQA method, radiologists, and ground truth, a significance difference in the assessment scores between the two was considered when P value was less than 0.05.

Discussion

In this study, a DL-based IQA method has been proposed, where large public datasets which consisted of natural images such as COCO and ImageNet were used as pre-trained data, which can improve the performance of the model to a certain extent.

Three DL-based models were used to evaluate the sufficient degree of inspiration of the patient after a CT chest examination, as shown in *Figure 1*. The image selection model was used to quickly select the visible image of the tracheal carina from a sequence of CT images, then the tracheal carina was segmented on the selected CT image

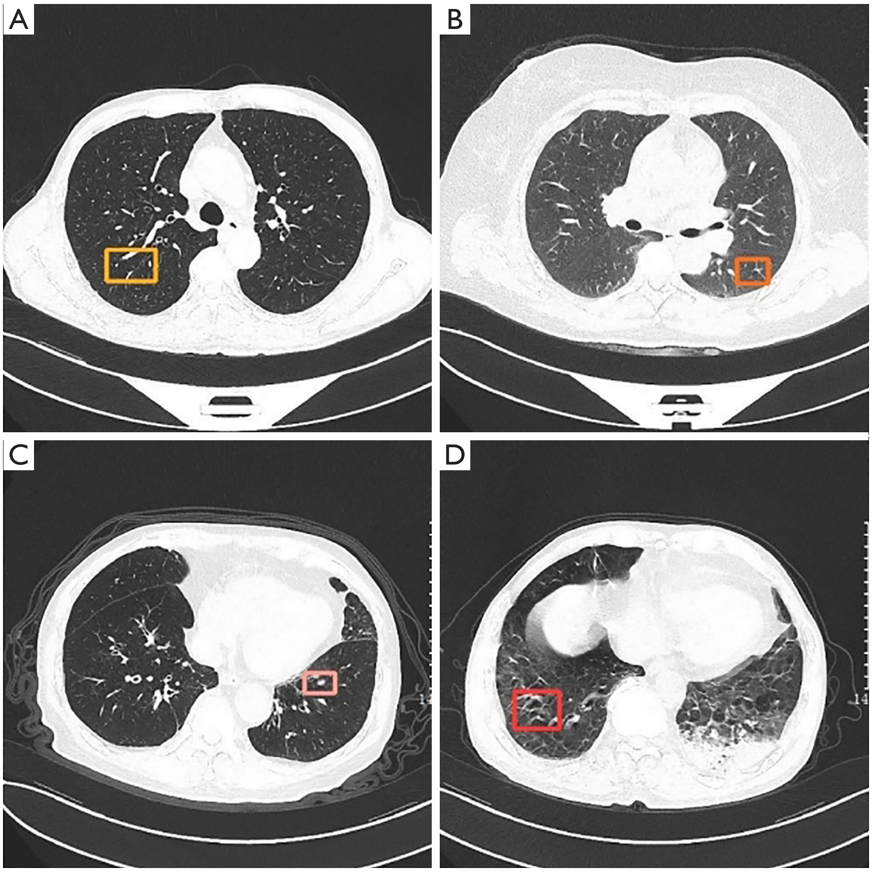

Figure 4 The detection results of bronchial beam which are detected by the detection model. (A), (B), (C), and (D) is the CT image where the yellow, orange, pink, and red bounding box contains the visible, fuzzy, double shadow, and curly bronchial beam respectively. CT, computed tomography.

Table 5 The recall, precision, and F1 Score of segmentation model and detection model

Model	Recall[a]	Precision[a]	F1 Score[a]
U-Net	0.90±0.09	0.92±0.07	0.91±0.05
YOLOv8	0.92±0.01	0.90±0.09	0.91±0.05

[a], the data are presented in the table as mean ± standard deviation.

by using a segmentation model, and all the bronchial beams were detected on the sequence of CT images by using a detection model. In terms of model training, due to the lack of existing data, a series of data augmentation methods were used as shown in *Figure 2*. After data augmentation, additional data were obtained for model training, which can improve the performance and robustness of the model. For our method (DL-based IQA) to evaluate the quality of respiratory phase CT images like a radiologist, the patient's sufficient degree of inspiration under clinical judgment was employed as the golden standard, whereby two approaches (i.e., area-based approach and score-based approach) were used to process the results of the segmentation model and the detection model and a rating score was provided. The proposed DL-based IQA method is end-to-end and the whole procedure is automated, which can reduce the radiologist's workload, and the comparison in *Table 6* shows that it can save time without losing the accuracy of image quality assessment.

The recall, precision, and F1 Score in *Table 5* shows that this study achieved excellent segmentation performance in the task of tracheal carina segmentation. The primary reason behind this is that the morphology of trachea carina is relatively simple, which enables the model to better learn the feature region parameters. However, due to the similarities between different morphologies of the

Table 6 The sensitivity, specificity, PPV, NPV, accuracy, Kappa value, and P value of three confusion matrices, respectively

Confusion matrix	Sensitivity	Specificity	PPV	NPV	Accuracy	Kappa value	P value
DL-based IQA method *vs.* ground truth	0.92	0.92	0.92	0.92	0.92	0.83	0.93
Radiologists *vs.* ground truth	0.92	0.83	0.85	0.91	0.88	0.75	0.67
DL-based IQA method *vs.* radiologists	0.85	0.91	0.92	0.83	0.88	0.75	0.75

PPV, positive predictive value; NPV, negative predictive value; DL, deep learning; IQA, image quality assessment.

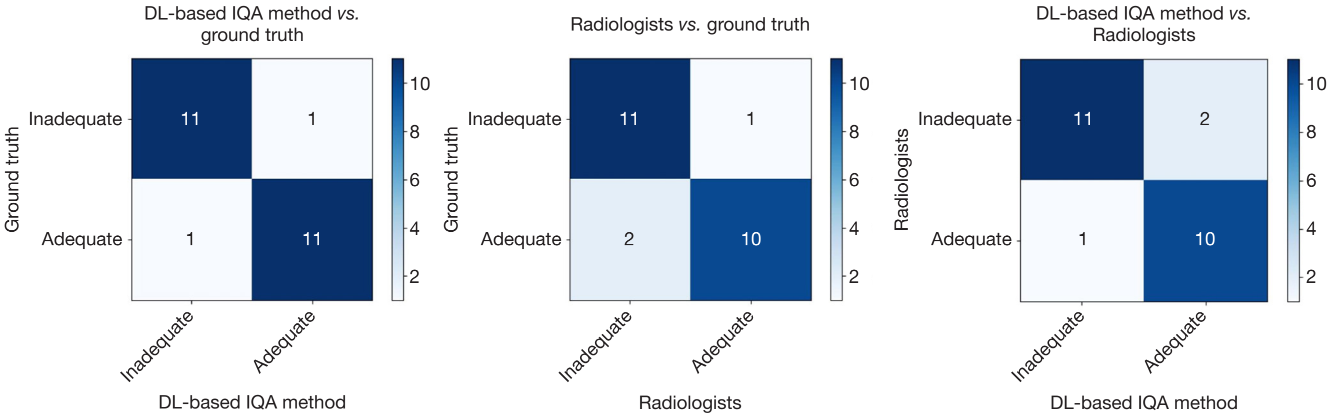

Figure 5 Three confusion matrices that compare the similarities and differences in evaluation results between our DL-based IQA method, radiologists, and ground truth. DL, deep learning; IQA, image quality assessment.

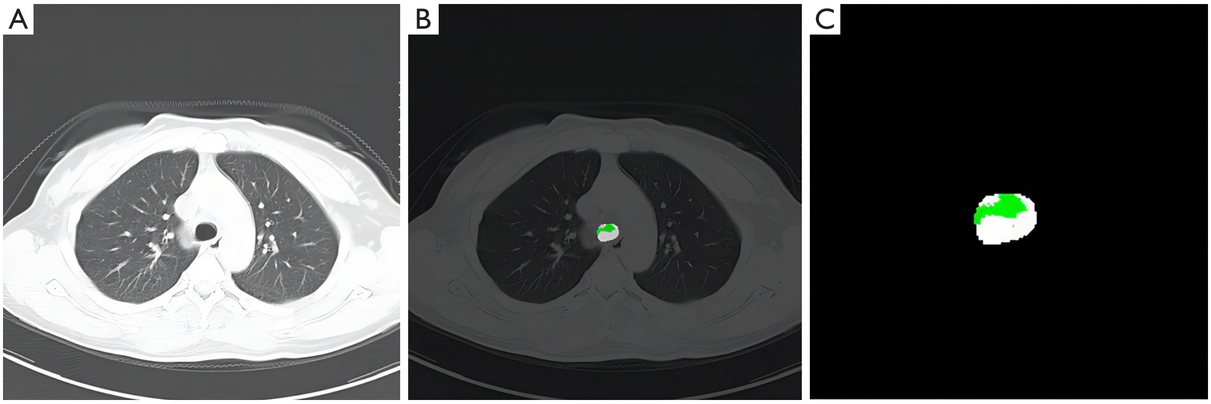

Figure 6 An uncertain predictive result that contains two colors in the ROI and needs the area-based approach used to further classification. (A) is the CT image with tracheal carina from the continuous sequence of a patient, while (B) is a visualization consisting of the predicted result of the segmentation model combined with (A). (C) is an enlarged view of the predictive result. It is worth noting that the green region represents flat classification, while the white region represents convex classification. ROI, region of interest; CT, computed tomography.

trachea carina, it could lead to some errors. In particular, as *Figure 6A* shows, due to the ambiguous morphology of the tracheal carina, the model's prediction results were uncertain, as seen in *Figure 6B,6C*. To address this issue, an area-based approach was proposed, and it turned out that this successfully avoided the ambiguity of tracheal carina's morphology from affecting our final evaluation results.

The performance of models reflects the accuracy of our IQA method to a certain extent, the recall, precision, and F1 Score were calculated to evaluate the performance of two

Table 7 The recall, precision, and F1 Score of different detection model

Model	Recall[a]	Precision[a]	F1 Score[a]
Faster R-CNN	0.44±0.15	0.37±0.04	0.40±0.09
CenterNet	0.36±0.25	0.60±0.43	0.44±0.29
RetinaNet	0.88±0.03	0.86±0.07	0.87±0.03
YOLOv7	0.62±0.09	0.63±0.10	0.62±0.10
YOLOv8	0.92±0.01	0.90±0.09	0.91±0.05

[a], the data are presented in the table as mean ± standard deviation. CNN, convolutional neural network.

main models (i.e., segmentation model and detection model) which are shown in *Table 5*. The two approaches (i.e., area-based approach and score-based approach) we proposed were used to finalize the evaluation results of the model. Image quality assessment is still a subjective problem and there is no unique standard for evaluating whether medical images can be used for clinical diagnosis until now (37). Therefore, an observer study on independent test set was conducted to validate the reliability of our IQA method, and due to human limitations such as fatigue, perceptual biases, and cognitive biases, even experienced radiologists make mistakes in evaluation (38), thus the MOS was used to acquire more reliable reference standard radiologists. A multi-angle analysis is conducive to proving a conclusion and discovering a problem. To this end, with the rating scores given by the radiologists, three confusion matrices as shown in *Figure 5* were constructed and a series of evaluation metrics (i.e., sensitivity, specificity, PPV, NPV, accuracy, Kappa value, P value) were calculated as shown in *Table 6*. In comparison to the ground truth, our DL-based IQA method achieved a Kappa value of 0.83 and a P value of 0.93, while the radiologists achieved a Kappa value of 0.75 and a P value of 0.67, respectively. Since a P value less than 0.05 was considered as significance difference, our DL-based IQA method and the radiologists did not show significant difference from the ground truth. However, the larger P value of our DL-based IQA method (0.26 difference) compared to radiologists suggests that the agreement of our DL-based IQA method might be superior to that of radiologists. The Kappa value supported our hypothesis, as it fell within the ranges of (0.60, 0.80] and (0.81, 1.00] and therefore were considered substantial agreement and almost perfect agreement, respectively. Therefore, our DL-based IQA method exhibited almost perfect agreement with the ground truth (Kappa value of 0.83), while radiologists demonstrated only substantial agreement with the ground truth (Kappa value of 0.75). It is worth noting that our DL-based IQA method exhibited superior evaluation capabilities compared to radiologists in our independent test set, and this performance is notably commendable. However, it is important to acknowledge that the independent test set utilized in this study consisted of data from only 24 patients, which can be considered relatively small. Therefore, the evaluation capabilities of our DL-based IQA method have only been validated on a limited dataset. To obtain more convincing verification, it would be beneficial to test our method on a larger dataset.

Although Poggenborg *et al.* (24) had an assessment of inspiration, the study mainly showed a relatively large improvement of 28% in images with optimal collimation, and only had a relatively improvement of 4% in images with optimal inspiration. There was no obvious improvement in selecting the CT images with optimal inspiration compared to manual assessment by radiologists. Azour *et al.* (13) proposed a quantitative and easily reproducible metric to quantitatively measure the change in lung volume between inspiratory and expiratory phase CT images, which reflected the sufficient degree of inspiration. The metric proposed by Azour *et al.* (13) divided tracheal carina morphologies into three categories. Based on the three classifications of tracheal carina morphologies, the tracheal carina morphologies under different sufficient agreement of inspiration were defined as convex, flat, and concave, which further verified the reliability of the method used in this study to assess tracheal carina morphologies. It is important to note that there is no literature to automate this classification method, therefore, our study is also groundbreaking. Deep learning-based methods were used to achieve automation, and an ablation experiment was conducted to compare and analyze the evaluation effect of multiple DL models, to obtain the DL model with the best effect. As shown in *Table 7*, the same configured dataset was used to train and test five detection models, and the faster R-CNN, CenterNet, RetinaNet, and YOLOv7 were compared with the YOLOv8 which we finally used. Among the many anchor-based detectors, R-CNN series (39-41) is featured in two-stage detectors, and the faster R-CNN (41) is the newest model structure with the best performance in R-CNN series, but it performed poorly in the ablation experiment. As mentioned above, YOLO series plays an important role in one-stage detectors, but the performance of YOLOv7 in the ablation experiment was dissatisfactory, while YOLOv8 showed the best performance than others.

The RetinaNet (42) and CenterNet (43) are also one-stage detectors, the performance of RetinaNet was second only to YOLOv8, but CenterNet had the worst performance with high standard deviation in the ablation experiment, which we would rather not see. Our ablation experiment indicated that the YOLOv8 had a good detection ability in detecting the bronchial beam in our study.

The automatic DL-based IQA method also has some limitations. First, although our IQA method had an excellent performance in this study, in the era of rapid development of artificial intelligence, our IQA method still needs further development to improve deep learning models for even better performance, and only six models had been assessed in this study, there may be better models that we have not tried. Secondly, the patients that formed part of the study, their medical conditions/diseases relating to their respiratory system were not considered. Normally if a patient has severe respiratory diseases, a patient's breathing ability to some extent is affected negatively, which makes their inspiratory adequacy worse than healthy people, and it is not necessarily that they did not follow the breathing instructions for the CT chest examination. Thirdly, the number of patients used in this study to train the three models was relatively small and the positive and negative samples were not well balanced, although some data augmentation techniques were used to make up for this, the performance of the three models did not surpass that of training with the original data. Furthermore, the number of patients used to validate the performance of our DL-based IQA method was small, although our IQA method performed well in it, it is no doubt that performing well on a larger test data set will make our IQA method more trustworthy. Lastly, the labels (i.e., ROIs manually annotated as the ground truth) were based on subjective human assessment, although an annotated group consisting of three radiologists with different experiences was formed for the purpose of obtaining a more robust label, there may be different opinions for anyone but these three. For future research studies more evenly distributed data should be collected, and the patient's disease relating to the respiratory system should be considered. The annotation opinions of more experts should also be taken into account, to increase the robustness of the IQA method, and to adopt diverse strategies to enhance the performance of our IQA method.

Conclusions

In conclusion, our proposed DL-based IQA method in this study demonstrates an assessment accuracy of 0.92, a Kappa value of 0.83, and a P value of 0.93 on the independent test set. These results indicate that it exhibits excellent assessment capabilities, offering significant clinical relevance for evaluating the quality of respiratory phase images in CT chest. Moreover, it has the potential to reduce the workload of radiologists.

Acknowledgments

Funding: This work was supported by the Natural Science Foundation of Fujian Province (No. 2021J011216), the Natural Science Foundation of Putian City (No. 2021S3F002), and Scientific Research Project of Putian University (No. 2023062).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-1273/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board (IRB) of Fujian Putian Hospital in China (No. 202132) and individual consent for this retrospective analysis was waived.

References

1. Garvey CJ, Hanlon R. Computed tomography in clinical practice. BMJ 2002;324:1077-80.
2. Mayo-Smith WW, Hara AK, Mahesh M, Sahani DV,

Pavlicek W. How I do it: managing radiation dose in CT. Radiology 2014;273:657-72.

3. Baert AL. Encyclopedia of Imaging. Springer Science & Business Media, 2008.
4. Cai J, Chen X, Huang W, and Mou X. Image Quality Assessment on CT Reconstruction Images: Task-specific vs. General Quality Assessment. Fully3D, 2017.
5. Chand RB, Thapa N, Paudel S, Pokharel GB, Joshi BR, Pant DK. Evaluation of image quality in chest radiographs. Journal of Institute of Medicine Nepal 2013;35:50-2.
6. Mettler FA Jr, Wiest PW, Locken JA, Kelsey CA. CT scanning: patterns of use and dose. J Radiol Prot 2000;20:353-9.
7. Reiner BI. Hidden costs of poor image quality: a radiologist's perspective. Journal of the American College of Radiology 2014;11:974-8.
8. Alpert HR, Hillman BJ. Quality and variability in diagnostic radiology. J Am Coll Radiol 2004;1:127-32.
9. Reiner BI, Siegel EL, Siddiqui KM, Musk AE. Quality assurance: the missing link. Radiology 2006;238:13-5.
10. Guckenberger M, Weininger M, Wilbert J, Richter A, Baier K, Krieger T, Polat B, Flentje M. Influence of retrospective sorting on image quality in respiratory correlated computed tomography. Radiother Oncol 2007;85:223-31.
11. Miyata T, Yanagawa M, Hata A, Honda O, Yoshida Y, Kikuchi N, Tsubamoto M, Tsukagoshi S, Uranishi A, Tomiyama N. Influence of field of view size on image quality: ultra-high-resolution CT vs. conventional high-resolution CT. Eur Radiol 2020;30:3324-33.
12. Andersen ER, Jorde J, Taoussi N, Yaqoob SH, Konst B, Seierstad T. Reject analysis in direct digital radiography. Acta Radiol 2012;53:174-8.
13. Azour L, Mendelson DS, Rogers L, Salvatore MM. Diaphragmatic excursion: Quantitative measure to assess adequacy of expiratory phase CT chest images. Eur J Radiol 2021;136:109527.
14. Little BP. Approach to chest computed tomography. Clin Chest Med 2015;36:127-45, vii.
15. Moghadam N, Rehani MM, Nassiri MA. Assessment of patients' cumulative doses in one year and collective dose to population through CT examinations. Eur J Radiol 2021;142:109871.
16. Whaley JS, Pressman BD, Wilson JR, Bravo L, Sehnert WJ, Foos DH. Investigation of the variability in the assessment of digital chest X-ray image quality. J Digit Imaging 2013;26:217-26.
17. Deng L, Yu D. Deep learning: methods and applications. Foundations and trends® in signal processing, 2014;7:197-387.
18. Messay T, Hardie RC, Rogers SK. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Med Image Anal 2010;14:390-406.
19. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500-10.
20. Gao XW, Hui R, Tian Z. Classification of CT brain images based on deep learning networks. Comput Methods Programs Biomed 2017;138:49-56.
21. Yuan Z, Puyol-Antón E, Jogeesvaran H, Smith N, Inusa B, King AP. Deep learning-based quality-controlled spleen assessment from ultrasound images. Biomed Signal Process Control 2022;76:103724.
22. Riquelme D, Akhloufi MA. Deep learning for lung cancer nodules detection and classification in CT scans. AI 2020;1:28-67.
23. Meng Y, Ruan J, Yang B, Gao Y, Jin J, Dong F, Ji H, He L, Cheng G, Gong X. Automated quality assessment of chest radiographs based on deep learning and linear regression cascade algorithms. Eur Radiol 2022;32:7680-90.
24. Poggenborg J, Yaroshenko A, Wieberneit N, Harder T, Gossmann A. Impact of AI-based real time image quality feedback for chest radiographs in the clinical routine. medRxiv 2021. doi: 10.1101/2021.06.10.21258326.
25. Nousiainen K, Mäkelä T, Piilonen A, Peltonen JI. Automating chest radiograph imaging quality control. Phys Med 2021;83:138-45.
26. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition 2016:779-88.
27. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition 2017:7263-71.
28. Dang F, Chen D, Lu Y, Li Z. YOLOWeeds: a novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems. Comput Electron Agric 2023;205:107655.
29. Glenn J. Ultralytics YOLOv8. Available online: https://github.com/ultralytics/ultralytics, (2023)
30. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical segmentation//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015:234-41.

31. Minnich DJ, Mathisen DJ. Anatomy of the trachea, carina, and bronchi. Thorac Surg Clin 2007;17:571-85.
32. Kamauu A, DuVall S, Liimatta A, Avrin D, Wiggins R. 3rd: Automatic window-level calculation for DICOM to JPEG conversion for vendor-neutral teaching file web application, Radiological Society of North America (RSNA). McCormick Place, Chicago, IL, 2005.
33. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014:740-55.
34. Li FF, Deng J, Li K. ImageNet: Constructing a large-scale image database. J Vis 2009;9:1037.
35. Fellbaum C. WordNet: An electronic lexical database. MIT press, 1998.
36. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: Faster and better learning for 2020, bounding box regression. Proceedings of the AAAI Conference on Artificial Intelligence 2020;34:12993-3000.
37. Chow LS, Paramesran R. Review of medical image quality assessment. Biomedical Signal Processing and Control 2016;27:145-54.
38. Fitzgerald R. Error in radiology. Clin Radiol 2001;56:938-46.
39. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition 2014:580-7.
40. Girshick R. Fast R-CNN. Proceedings of the IEEE international conference on computer vision 2015:1440-8.
41. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell 2017;39:1137-49.
42. Zhou X, Wang D, and Krähenbühl P. Objects as points. arXiv preprint arXiv: 1904.07850, 2019.
43. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision 2017:2980-8.

Cite this article as: Su J, Li M, Lin Y, Xiong L, Yuan C, Zhou Z, Yan K. A deep learning-based automatic image quality assessment method for respiratory phase on computed tomography chest images. Quant Imaging Med Surg 2024;14(3):2240-2254. doi: 10.21037/qims-23-1273