Automatic No-reference Quality Assessment in Chest Radiographs Based on Deep Convolutional Neural Networks

Yu Meng
Zhejiang Provincial People's Hospital https://orcid.org/0000-0001-6009-3200
Yang Gao
Zhejiang Provincial People's Hospital
Jianqiu Jin
Zhejiang Gongshang University
Jingru Ruan
Zhejiang Provincial People's Hospital
Linyang He
Hangzhou Jianpei Technology Ltd
Qiang Shen
Zhejiang Provincial People's Hospital
Xiangyang Gong (cjr.gxy@hotmail.com)
Zhejiang Provincial People's Hospital

Research article

Keywords: Chest radiography, Quality Assessment, Deep Learning, Gray-level, Sharpness

Posted Date: February 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1213399/v1

Abstract

Background: Chest radiography is the most frequently performed examinations in the department of radiology. The image quality plays an essential role in further decision making of dignosis. The current manual assessment is low-efficiency with significant inter-observer variability. The purpose of this study is to develop and evaluate the performance of deep learning-based model for the automatic quality assessment in chest radiographs.

Methods: A set of 1138 posterior-anterior chest radiographs were included in this retrospective study, which were randomly divided into training (n = 826), validation (n = 207), and testing sets (n = 105). The image quality was evaluated on the aspect of gray-level and sharpness. Ten experienced experts independently assessed all radiographs with a score of 1 to 10 based on their subjective perception for the image quality. On the testing set, three of the ten experts classified chest radiographs as acceptable or non-acceptable based on the image quality further. A neural network model CaHDC-RGA was trained to output the quality score of the gray-level and sharpness automaticlly. The intra-class correlation coefficient (ICC), Pearson correlation coefficient (r), and mean absolute difference (MAD) were used for quantitative scoring. The AUC value, sensitivity, and specificity were used for binary classification. Time spent by model and experts were recorded, respectively.

Results: A statistically significant correlation was observed between the model and experts on the aspect of gray-level (ICC=0.92, r = 0.91, MAD=0.45) and sharpness (ICC=0.90, r = 0.89, MAD=0.44). For the classification of image quality as acceptable or non-acceptable, the model achieved an AUC of 0.972 for gray-level, with a sensitivity of 93.90% and a specificity of 95.69%. The AUC for sharpness was 0.970, with a sensitivity of 87.95% and a specificity of 100%. The average time spent by the model was significantly shorter than the human expert's (3.03 seconds VS. 10.96 seconds, P < 0.05).

Conclusions: *The developed deep learning model could rapidly and automatically evaluate the gray-level and sharpness of chest radiographs,* the performance of the model was comparable with the subjective perception of human experts. *The model may be further applied to automated quality audits of full samples.*

Background

Chest radiographs is one of the most frequently performed examinations in the department of radiology. In 2016, over 110 million chest radiographs were taken in the United States, accounting for 40% of all X-rays in clinical practice[1]. Despite the relative simplicity of the photographic technique in chest radiography, the poor quality of the images are frequently appeared in clinical practice. The chest radiographs were account for the largest proportion of rejection, even up to 38% in some institutions[2]. Since the quality of medical image plays an essential role in further decision making, the mechanism for image quality assessment (IQA) has been established to improve the interpretability of the chest radiographs[3].

Currently, IQA is usually first performed by a radiographer immediately after image acquisition[4]. The technicians need to assess the acceptability of the image quality and decide whether the image needs to be retaken. After the images are transferred to PACS, radiologists also need to evaluate whether the image quality affects the diagnosis. This IQA method has no reference information and is highly dependent on the experience and responsibility of the observers[5, 6]. The manual process of IQA inevitably has significant inter-observer variability[7, 8]. The manual assessment is also a relatively time-consuming, low-efficiency, and tedious task in clinical practice. In addition, a limited sample size included in quality control (QC) programs usually does not reflect the medical institution's overall level of image quality.

Developing an automatic computer-based assessment tool for image quality would be essential for improving the efficiency in clinical practice. Traditional machine learning methods require manually designed features, such as colorfulness, dark channel feature, and entropy, which are not completely express the subjective perception of image quality by radiologists[9]. In recent years, the quality and efficiency of medical image analysis have been greatly improved with the application of deep learning (DL). The deep learning-based algorithms have also been applied in IQA of fundus image[10], MRI artifact detection[11], fetal ultrasound[12], and digital pathology image[13]. At present, automatic IQA based on deep learning has been applied to chest radiography[14]. However, these studies are focused on the position of chest radiographs, lacking in the research of the gray-level and sharpness, which are important for clear visualization of pathologies.

The purpose of this study was to develop an automated no-reference IQA model based on deep learning algorithms to evaluate the gray-level and sharpness of chest radiographs, and to evaluate the performance of the model.

Methods

Image acquisition

A clinical database consisting of adult posterior-anterior chest radiographs was used in this study. All images were acquired between July to September 2019 in three medical centers. According to the European Commission guidelines for image quality (Table 1)[15], two radiology residents (Y.M., and Y.G., each with 3 years of experience in radiologic reading) performed an initial assessment of image quality to obtain chest radiographs with various image quality. The collected images were arbitrarily classified as excellent, moderate, and poor. An image was rated as excellent if it met all quality criteria, moderate if less than half of the criteria were not met, and poor if more than half of the criteria were not met. To avoid any bias, image quality was assessed independently of the presence of lesions, so those patients with pulmonary pathology were also included. All images were taken in digital radiograohy (DR) and saved in DICOM format, images were obtained by four manufacturers (Canon, Siemens, GE, Philips), and patient information was anonymized.

Table 1
Image quality in Radiography: PA chest projection

Dimensions of IQA	Image quality criteria
Gray-level	Reproduction of the vascular pattern in the whole lung particularly the peripheral vessels.
	Visualization of the retrocardiac lung and the mediastinum.
	Visualization of the spine through the heart shadow.
Sharpness	Visually sharp reproduction of the trachea and proximal bronchi.
	Visually sharp reproduction of the borders of the heart and aorta.
	Visually sharp reproduction of the diaphragm and costophrenic angles.

There were 2349 chest radiographs reviewed with 823 excellent images, 1090 moderate images, and 436 poor images. Then, 1200 chest radiographs were randomly extracted with a ratio of excellent: moderate: poor = 1:1:1 (400 cases). After reviewing images together for the second time, 62 cases with repeated image acquisition were excluded. Finally, a total of 1138 chest radiographs were included. 105 out of 1138 chest radiographs were randomly selected as a testing set, the remaining 1033 images were assigned into a training set (80%; n=826) and a validation set (20%; n=207). The training set was used to optimize the model parameters, the validation set was used to tune the model hyperparameters, and the testing set was used to evaluate the performance of the model. The flowchart for the image acquisition and data split is illustrated in Fig. 1.

Subjective assessment of experts

All chest radiographs (n = 1138) were firstly imported into a dedicated web-based platform for the IQA (http://score.healthviewcn.com:9990/#/login). The observers are blinded to the clinical information about the patient, they can operate the image by zooming and panning during the evaluation, and the assessment time of each image will be recorded by the platform.

Ten experienced radiologists participated in the IQA (five diagnostic radiologists and five radiological technologists, with an average of 15-25 years of experience). Without discussing the European Commission guidelines for image quality, all experts independently assessed all radiographs with a score of 1 to 10 based on their subjective perception of the image quality, where 10 represented the best quality and 1 meant the worst quality. The mean opinion score (MOS) of ten experts was taken for each image.

Furthermore, on the testing set (n = 105), three of ten experts (with over 10 years of experience in QC of chest radiographs) classified the images as acceptable or non-acceptable based on whether the quality of gray-level or sharpness affected the diagnosis. The reference standard of binary classification was based on the majority opinion between 3 observers.

Dataset preparation

Python (version 3.6) and Simple-ITK (http://www.simpleitk.org) were used to conduct the process of image preprocessing. The training set was cropped to contain only diagnostic regions of the chest, including the apex of the lung, the rib-diaphragm angle, and the chest wall on both sides. All images were finally preprocessed with a resolution of 768 × 512 pixels.

Development of CNN model

CNN model architecture

A novel neural network model based on CaHDC[16] was designed by introducing ROI (region of interest) guided attention, and we named it CaHDC-RGA. Fig. 2 showed the overall structure of networks. The DR images and MOS of experts were as initial inputs to the networks. The ResNet[17] is used for extracting features as input to the following Hierarchical Convolution Net (HCNet)[18] and Unet[19]. The Hierarchical Convolution Net produces multi-level feature maps, and an attention map is obtained from the UNet expectedly. Then, the attention map is applied on the four levels of feature maps from the HCNet, and the results are as the input of the Side Pooling Nets (SiPNets). Finally, the results of the SiPNets are concatenated and fed into the Regression Net (RegNet) as a regressor to obtain the final quality score.

The residual network consists of five convolutional layers as showed in Fig. 3. The size of convolution kernel (with ReLU activation function) of each layer is 3*3 and the number is 32. The size of the output is same as the input due to the step size of each convolution operation is 1. The role of ResNet is as a feature extractor serving the subsequent UNet and HCNet.

As shown in Fig. 4, the Hierarchical Convolutional Network is composed of a series of convolutional layers including 4 levels to extract hierarchical features O1-O4, which are the inputs of SiPNet1-SiPNet4, respectively. The parameters of these convolution layers from left to right are Conv(3,3,64,2), Conv(3,3,128,2), Conv(3,3,256,2) and Conv(3,3,512,2), respectively.

As shown in Fig. 5, its input is the output from the ResNet, and its output is expected to be an attention map with the channel size of 1. The UNet, which only include convolution and upconv (also called deconvolution) layers, is a fully convolutional neural network. It is the same as ResNet and HCNet. As a result, the trainable parameters of these neural networks are independent of the size of the input image. The attention map from the UNet would be paid to the hierarchical features O1-O4 from the HCNet. Due to the different size of these features, the attention map should be resampling properly to meet the size of the hierarchical features before applying attention mechanism.

As shown in Fig. 6, after the attention is applied, the data are fed into the SiPNets. A SiPNet consists of serval convolutional layers and a max pooling layer. Its first layer with 1*1 kernel and 1*1 stride is to reduce the channel of the input data to 64. After that, a series of repeated convolutional layers with 3*3 kernel and 2*2 stride are adopted for down sampling. The number R of the repeated layers is dependent on the size of input such that the size of output is between 8*8 and 15*15. For example, if the size of

input feature map is 512*512, R should be 6, and in this case the size of output is 8*8. Then, a feature vector with size of 1*1*64 is outputted by a max pooling layer.

RegNet is the abbreviation for Regression Net. As shown in Fig. 7, the RegNet accepts four 64-dimension feature vectors from SiPNet1-4 as input, and output Q1~ Q4,and Q. Each of FC1-4 is a fully connected layer which accepts a 64-dimesion feature vector to output a score. FC5, a fully connected layer, accepts a 256-dimension feature vector concatenated from the outputs of the SiPNets and generates a 128-dimesion feature vector. Then, FC6 regresses it to a quality score Q .

Loss function

Automated quality assessment systems need to be more sensitive to low quality image quality. Therefore, we should pay more attention to the low-score images, and the following loss function is adopted,

$$\mathrm{L}_{\alpha}(\mathrm{s}) = (\mathrm{s} - \bar{\mathrm{s}})^{2-\alpha} / \mathrm{s}^{-\alpha} \quad (1)$$

where α is a super parameter, s is the predicted quality score, and $\bar{\mathrm{s}}$ is the ground truth. It should be noted that the loss function should be adjusted to suit the training dataset. To reduce overfitting, the error between predicted score by each single level and MOS is adopted as auxiliary loss to train our proposed model for hierarchical degradation measurement. Therefore, the overall loss function can be expressed as

$$\mathrm{L} = \sum_{\mathrm{i}=1}^{4} (\mathrm{b}_{\mathrm{i}} \mathrm{L}_{\alpha}\left(\mathrm{Q}_{\mathrm{i}}\right) + \mathrm{L}_{\alpha}(\mathrm{Q})) \quad (2)$$

Training process of networks

In the training process, a mini-batch of random sampled image patches are fed into CaHDC and Adam optimization algorithm is adopted for training. The parameters are set as $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and the learning rate is set as

$$\alpha = \begin{cases} \alpha_0 * \mathrm{d}^{0.01*\mathrm{t}} \alpha > \alpha_{\mathrm{m}} \\ \alpha_{\mathrm{m}} \alpha \leq \alpha_{\mathrm{m}} \end{cases} \quad (3)$$

where $\alpha_0$is the initial learning rate, d is the decay factor, α_{m} is the minimum learning rate. In our typical prosesses, $\alpha_0 = 10^{-3}$, $\alpha_{\mathrm{m}} = 3*10^{-5}$, and $\mathrm{d} = 0.8$. in the loss function Eq. 2, we set $\alpha = 1 \mathrm{and} \mathrm{b}_{\mathrm{i}} = 0.1$, for $\mathrm{i} = 1, \ldots, 4$.

Statistical analysis

On the testing set, the performance of the DL model was appraised through the following three tasks: (a) quantitative scoring of image quality, (b) binary classification of the image quality, and (c) time spent for image quality assessment.

For the first task, the MOS of ten experts on the gray-level and sharpness was used as the reference standard. The performance of the DL model was evaluated based on the pair t test, intra-class correlation coefficient (ICC), Pearson correlation coefficient (r), the mean absolute difference (MAD), and 95% limit of agreement (LoA) between the predictions and reference standard. ICC, which reflects abstract agreement, with values greater than or equal to 0.75 were considered adequate for reliability[20]. MAD is defined as $\frac{1}{n}\sum_{k=1}^{n}\left|A_k - B_k\right|$, where A_K and B_K are the prediction and reference standard for the kth image respectively, and n is the number of images. The mean difference and 95% LoA were determined with the Bland-Altman plot[21].

For the second task, the reference standard of binary classification was based on the majority opinion between 3 observers. Receiver operating characteristic curve (ROC) analysis was performed, and the value of area under curve (AUC) was used as the performance measure[22]. Moreover, we calculated the sensitivity and specificity at the optimal threshold.

Data information was collected and saved as a spreadsheet by using Excel 2019 (Microsoft). All statistical analyses were performed using MedCalc software (Version 15.2.2), and $P < 0.05$ was considered to indicate a statistically significant difference.

Results

Study participants

The study population characteristics for datasets were showed in Table 2. Among the 1138 chest radiographs, the mean ages in the training, validation, and testing sets were 51.90 years ± 17.03 (range 18 to 87 years), 51.45 years ± 18.29 (range 18 to 91 years), and 50.93 years ± 17.76 (range 19 to 89 years), respectively. Between three data sets, ANOVA analysis showed no statistical differences in age composition ($F=0.17$, $P =0.84$), and chi-square tests showed no statistical differences in gender composition ($\chi^2 = 0.57$, $P = 0.75$).

Table 2
Demographic data of the datasets

	Training set	Validation set	Testing set	Groups
No. of images	826	207	105	
age	51.90 ± 17.03	51.45 ± 18.29	50.93 ± 17.76	P = 0.84[a]
Sex				
Male	441 (53.3%)	109 (52.7%)	52 (49.5%)	P = 0.75[b]
Female	385 (46.7%)	98 (47.3%)	53 (49.5%)	

Note: Except where indicated, data are means±standard deviations, with percentages in parentheses.
[a] ANOVA analysis, F=0.18; [b] pearson Chi-Square, c^2 = 0.57.

Subjective assessment of experts

Before testing, the images of the training and validation sets (n=1033) were evaluated by ten radiologists. Fig. 8a showed the MOS of ten experts on the gray-level and sharpness of the chest radiographs, respectively. The mean score for image gray-level was 7.75 with a standard deviation of 1.24, while the mean score for sharpness was 8.01 with a standard deviation of 1.21. Fig. 8b showed some representative examples of chest radiographs with corresponding image quality scores.

Performance of quantitative scoring by model

On the testing set, the MOS of ten experts were used as reference standard. The paired-samples t test showed no statistical difference between DL model and reference standard in terms of gray-level (t=-1.11, P = 0.26) and sharpness (t=3.12, P = 0.66) (Table 3). Compared with the reference standard, the DL model assessed accurately of the gray-level (ICC=0.92, r=0.91, MAD=0.45) and sharpness (ICC=0.90, r=0.89, MAD=0.44). Fig. 9(a/b) showed the correlation of quantitative scores between the model and reference standard. The 95% LoAs and the mean differences determined from the Bland-Altman plots were shown in Fig. 9(c/d).

Table 3
Quantitative analysis of the variability between model and expert scores

	Radiologists	DL model	t-test	ICC	r	MAD
Gray-level	7.84 ± 1.27	7.90 ± 0.87	*P = 0.26[a]*	0.92	0.91	0.45
Sharpness	8.08 ± 1.88	7.93 ± 0.92	*P = 0.66[b]*	0.90	0.89	0.44

Note: Paired-samples t test, [a]t = -1.11, [b]t = 3.12, p <0.05 indicates the statistical correlation between Prediction and reference standard of quality scores. **ICC** (Intra-class Correlation Coefficient), **MAD** (Mean Absolute Difference).

Performance of binary classification by model

In order to classify the image quality as acceptable or non-acceptable, the majority opinion of the three experts were used as the reference standard. The ROC analysis showed that the model performed similarly with experts in both gray-level and sharpness, with AUCs of 0.973 and 0.970, respectively. When the cutoff score of gray-level was 7.0, the model achieved a sensitivity of 93.90% and a specificity of 95.69%. When the cutoff score of sharpness was 7.2, the model achieved a sensitivity of 87.95% and a specificity of 100%. The performance of binary classification for the model was shown in Fig. 10.

Time of image quality assessment

The time spent of image quality assessment by human experts and DL model were showed in Table 4. Among the 1138 chest radiographs, the mean time of experts on the training, validation, and testing sets were 11.32 seconds, 11.08 seconds, and 10.96 seconds, respectively. The time of DL model on the training, validation, and testing sets were 2.99 seconds, 3.01 seconds, and 3.03 seconds, respectively. The paired-samples t test showed statistically significant differences in time between the models and human experts ($P < 0.05$).

Table 4
Time spent by human observers and DL model in IQA

Data sets	Radiologists	Radiographers	Mean	DL Model	t-test
Training	12.0(7.5)	11.5(8.0)	11.32	2.99	*P < 0.05*
Validation	11.5(8.0)	11.0(7.5)	11.08	3.01	*P < 0.05*
Testing	11.0(6.0)	11.0(9.0)	10.96	3.03	*P < 0.05*

Note: Time was measured in seconds, except where indicated, data were expressed as medians, with interquartile

Discussion

In this study, a novel neural network model, CaHDC-RGA, was designed based on deep learning algorithms by introducing ROI-guided attention. The developed DL model was used to assess the image quality of the gray-level and sharpness for chest radiographs automatically. The results showed that the performance of the model is comparable with the subjective perception of radiologist experts and the assessment time of the model is significantly shorter than that of radiologists.

Image quality assessment (IQA) has been an important topic for many years in radiology, which can identify and address the causes of poor image quality[23, 24]. In 1996, the European Union published radiology image quality criteria to unify the practices in Europe. However, many studies have found that image criteria are not objective and intero-bserver variation remains a significant problem in IQA, and the Cohen's weighted kappa values between observers ranged from 0.32 to 0.46[25–27]. In routine clinical practice of IQA, the observer often relies only on individual experience and subjective perception of the

image, with no standard image reference information, such as quality criteria. This approach is more operational and efficient for clinical practice[28]. In this study, in order to reduce the potential impact caused by interobserver variation, 10 experienced experts participated in the evaluation, including 5 radiologists and 5 radiographers. The mean opinion score (MOS) was uesd as the final label for image quality. In addition, unlike most previous studies that used a 5-point scale of preference (qualitative assessment), we used a scoring system ranging from 1 to 10 point because it allows for a quantitative assessment for image quality[29–31]. The results showed that the expert's subjective perception of image quality could be learned and quantified by deep learning algorithms, which can accurately output quantitative scores for images.

Previous studies have attempted to assess image quality with physical metrics such as DQE or the MTF[32]. Although the metrics were objective enough, they only measured the system performance rather than the radiologists' perception of clinical image quality. Recently, some CNN-based IQA has been proposed in the natural images, but they did not take full advantage of the perceptual properties of the human visual system[33–38]. In 2020, Wu et al.[16] proposed an end-to-end cascaded framework, called CaHDC, which can jointly optimize the feature extraction procedures, hierarchical degradation concatenation, and quality prediction. In this study, we designed the CaHDC-GRA by introducing ROI (region of interest) guided attention to focus on low-quality images. Our networks can evaluate hierarchical quality degradation because the number of network parameters was significantly decreased, the optimization of the network was also sped up while alleviating overfitting. The features were integrated with convolutional operations by downsampling them to the same scale, the number of features in this network is reduced, and the spatial information is preserved.

The chest radiographs may need to be retaken for the poor image quality, so it is necessary to determine whether the image quality is acceptable after image acquisition[2]. However, such a decision may be difficult for radiologists and radiographers with little experience. The automated assessment tool would serve as an advisor to improve the efficiency of the clinical work. Several CNN-based models have been proposed to evaluate the position of chest radiographs. Berg et al.[39] used the CNN to detect lung borders, spines, medial clavicular margins, ribs, and diaphragms with a typical 2.5-2.8 mm error. Nousiainen et al.[14] trained ResNet50 and DenseNet121 networks for assessing lung inclusion, rotation, and inspiration. To our knowledge, there is no automated tool for evaluating the degree of visualization in chest radiographs. In this study, a deep learning model has been trained to assess the gray-level and sharpness of chest radiographs. The results showed that the DL model could accurately simulate the subjective perception of radiologists. When the model was set to the optimal cutoff, it showed high sensitivity and specificity for classifying image quality as acceptable or non-acceptable. In future work, we will aim to identify the specific causes of non-acceptable to perform quality analysis better.

The increasing clinical demands on radiographers and radiologists make it imperative to operate as efficiently as possible to provide the highest quality of patient care[1]. A robust automated image quality assessment algorithm would be helpful within the clinic. For example, computer-based algorithms can analyze image data and provide technicians instant quality assurance (QA) feedback[40]. Automatic QA

in teleradiology would ensure assistance in reaching a diagnosis and deciding the best clinical management of the patient[41]. Big data analysis of image quality can be automatically constructed into a QA database, which can be used for trend analysis, education and training, and radiation dose reduction[4]. Our study demonstrated that deep learning algorithms could be used to assess the image quality of chest radiographs rapidly and automatically. The algorithms may be used for fully-sample audit in QC programs and be applied to other examination parts or modalities in the future.

Our study had several limitations. First, although the datasets used were selected from three medical institutions and the sample size was larger than previous studies, it is still not considerable for the deep learning algorithm. Except for training DL models with real clinical images, in further studies, we can artificially add noise or use phantoms to improve the performance of the models. Second, since posterior-anterior chest radiographs in adults are the most common clinical examination, our study did not include lateral, bedside, and pediatric chest radiographs. Future studies of images in more scenarios are necessary. Finally, similar to traditional image assessment methods, our system does not provide information beyond chest radiograph quality parameters, such as the presence or absence of disease in the images. In a follow-up study, we will further incorporate more image quality chest radiographs to improve the robustness and generalizability of the model. We will also apply the developed DL model to clinical examinations, analyze the effectiveness of the developed DL model for timely quality feedback to radiographers, and test the performance of the model in a large-scale image quality audit.

Conclusions

In conclusion, the developed deep learning model could rapidly and automatically evaluate the gray-level and sharpness of chest radiographs, the performance of the model was comparable with the subjective perception of human experts. The model may be further applied to automated quality audits of full samples in clinical practice.

Abbreviations

DL: Deep Learning; CNN: Convolutional Neural Networks; DR: Digital Radiography; IQA: Image Quality Assessment; NR: No-Reference; QC: Quality Control; MOS: Mean Opinion Score; ICC: Intra-class Correlation Coefficient; MAD: Mean Absolute Difference; ROC: Receiver Operating Characteristic Curve; AUC: Area Under the ROC Curve; LoA: Limit of Agreement.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of Zhejiang Provincial People's Hospital, China (No. 2021QT342). A certificate of approval has been provided. The requirement of informed consent was exempted due to the retrospective nature of the study.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and analysed during the current study are not publicly available due to patient privacy concerns but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no conflicts of interest.

Funding

This work was supported by the Key Research and Development Projects of Zhejiang Province (2020C01058). The funding agreements ensured the authors independence in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

YM and YG contributed equally to this work. YM designed the study and drafted the manuscript. YG was responsible for preprocessing the data and checking the results. JQJ and LYH provided software and supported on data analysis and interpretation. JRR and QS collected the data. XYG was responsible for manuscript review and supervision. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank all the involved professional image quality assessment practitioners (radiologists and technicians) for dedicating their time and skill to the comple tion of this study. In addition, we thank Dr. Yuting Yan for her contribution to the manuscript review.

Author details

[1]Department of Radiology, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, 158 Shangtang Road, Gongshu District, Hangzhou 310014, China. [2]School of Computer Science and Information Engineering, Zhejiang Gongshang Univeristy, Hangzhou 310018, China. [3]Hangzhou Jianpei Technology Company Ltd, Hangzhou, 311200, China. [4]Institute of Artificial Intelligence and Remote Imaging, Hangzhou Medical College, Hangzhou 310014, China.

References

1. Mettler FA Jr, Mahesh M, Bhargavan-Chatfield M, Chambers CE, Elee JG, Frush DP, Miller DL, Royal HD, Milano MT, Spelic DC, et al: **Patient Exposure from Radiologic and Nuclear Medicine Procedures**

in the United States: **Procedure Volume and Effective Dose for the Period** 2006-**2016**. *Radiology* 2020, **295**(2):418-427.

2. Little KJ, Reiser I, Liu L, Kinsey T, Sánchez AA, Haas K, Mallory F, Froman C, Lu ZF. Unified Database for Rejected Image Analysis Across Multiple Vendors in Radiography. J Am Coll Radiol. 2017;14(2):208–16.
3. Verschakelen J, Bellon E, Deprez T. Digital chest radiography: quality assurance. J Thorac Imaging. 2003;18(3):169–77.
4. Reiner BI. Automating quality assurance for digital radiography. J Am Coll Radiol. 2009;6(7):486–90.
5. Precht H, Hansson J, Outzen C, Hogg P, Tingberg A. Radiographers' perspectives' on Visual Grading Analysis as a scientific method to evaluate image quality. Radiography (Lond). 2019;25(Suppl 1):14-s18.
6. Lin W, Kuo C-CJ. Perceptual visual quality metrics: A survey. J Vis Comun Image Represent. 2011;22(4):297–312.
7. 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(2):217–26.
8. Lee J, Nishikawa RM, Reiser I, Zuley ML, Boone JM. Lack of agreement between radiologists: implications for image-based model observers. J Med Imaging (Bellingham). 2017;4(2):025502.
9. Varga D. **No-Reference Image Quality Assessment Based on the Fusion of Statistical and Perceptual Features**. *J Imaging* 2020, 6(8).
10. Shen Y, Sheng B, Fang R, Li H, Dai L, Stolte S, Qin J, Jia W, Shen D. Domain-invariant interpretable fundus image quality assessment. Med Image Anal. 2020;61:101654.
11. Oksuz I. Brain MRI artefact detection and correction using convolutional neural networks. Comput Methods Programs Biomed. 2021;199:105909.
12. Wu L, Cheng JZ, Li S, Lei B, Wang T, Ni D. FUIQA: Fetal Ultrasound Image Quality Assessment With Deep Convolutional Networks. IEEE Trans Cybern. 2017;47(5):1336–49.
13. Chen Y, Zee J, Smith A, Jayapandian C, Hodgin J, Howell D, Palmer M, Thomas D, Cassol C, Farris AB. 3rd et al: Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies. J Pathol 2021, 253(3):268–78.
14. Nousiainen K, Mäkelä T, Piilonen A, Peltonen JI. Automating chest radiograph imaging quality control. Phys Med. 2021;83:138–45.
15. European, Commission, **European guidelines on quality criteria for diagnostic radiographic images**. *EUR 16260 ISBN 92-827-7284-5, Brussels* 1996.
16. Wu J, Ma J, Liang F, Dong W, Shi G, Lin W. End-to-End Blind Image Quality Prediction With Cascaded Deep Neural Network. IEEE Trans Image Process. 2020;29:7414–26.
17. He F, Liu T, Tao D. Why ResNet Works? Residuals Generalize. IEEE Trans Neural Netw Learn Syst. 2020;31(12):5349–62.

18. Ma C, Huang J, Yang X, Yang M: **Hierarchical Convolutional Features for Visual Tracking**. In: 2015 *IEEE International Conference on Computer Vision (ICCV): 7-13 Dec. 2015 2015*; 2015: 3074-3082.
19. Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, Böhm A, Deubner J, Jäckel Z, Seiwald K, et al. U-Net: deep learning for cell counting, detection, and morphometry. Nat Methods. 2019;16(1):67–70.
20. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155–63.
21. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307–10.
22. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010;5(9):1315–6.
23. Abujudeh H, Kaewlai R, Shaqdan K, Bruno MA. Key Principles in Quality and Safety in Radiology. AJR Am J Roentgenol. 2017;208(3):W101-w109.
24. Spijker S, Andronikou S, Kosack C, Wootton R, Bonnet M, Lemmens N. Quality assessment of X-rays interpreted via teleradiology for Médecins Sans Frontières. J Telemed Telecare. 2014;20(2):82–8.
25. Lanhede B, Båth M, Kheddache S, Sund P, Björneld L, Widell M, Almén A, Besjakov J, Mattsson S, Tingberg A, et al. The influence of different technique factors on image quality of chest radiographs as evaluated by modified CEC image quality criteria. Br J Radiol. 2002;75(889):38–49.
26. Grewal RK, Young N, Colins L, Karunnaratne N, Sabharwal N. Digital chest radiography image quality assessment with dose reduction. Australas Phys Eng Sci Med. 2012;35(1):71–80.
27. Niemann T, Reisinger C, Rau P, Schwarz J, Ruis-Lopez L, Bongartz G. Image quality in conventional chest radiography. Evaluation using the postprocessing tool Diamond View. Eur J Radiol. 2010;73(3):555–9.
28. Tesselaar E, Dahlström N, Sandborg M. CLINICAL AUDIT OF IMAGE QUALITY IN RADIOLOGY USING VISUAL GRADING CHARACTERISTICS ANALYSIS. Radiat Prot Dosimetry. 2016;169(1-4):340–6.
29. Fink C, Hallscheidt PJ, Noeldge G, Kampschulte A, Radeleff B, Hosch WP, Kauffmann GW, Hansmann J. Clinical comparative study with a large-area amorphous silicon flat-panel detector: image quality and visibility of anatomic structures on chest radiography. AJR Am J Roentgenol. 2002;178(2):481–6.
30. Uffmann M, Neitzel U, Prokop M, Kabalan N, Weber M, Herold CJ, Schaefer-Prokop C. Flat-panel-detector chest radiography: effect of tube voltage on image quality. Radiology. 2005;235(2):642–50.
31. Chae KJ, Goo JM, Ahn SY, Yoo JY, Yoon SH. Application of Deconvolution Algorithm of Point Spread Function in Improving Image Quality: An Observer Preference Study on Chest Radiography. Korean J Radiol. 2018;19(1):147–52.
32. Båth M, Sund P, Månsson LG. Evaluation of the imaging properties of two generations of a CCD-based system for digital chest radiography. Med Phys. 2002;29(10):2286–97.

33. Gao F, Yu J, Zhu S, Huang Q, Tian Q. Blind image quality prediction by exploiting multi-level deep representations. Pattern Recogn. 2018;81:432–42.
34. Wu J, Zeng J, Liu Y, Shi G, Lin W: **Hierarchical Feature Degradation Based Blind Image Quality Assessment**. In: 2017 *IEEE International Conference on Computer Vision Workshops (ICCVW): 22-29 Oct. 2017 2017*; 2017: 510-517.
35. Sun C, Li H, Li W: **No-reference image quality assessment based on global and local content perception**. In: *2016 Visual Communications and Image Processing (VCIP): 27-30 Nov.* 2016 *2016*; 2016: 1-4.
36. Kang L, Ye P, Li Y, Doermann D: **Convolutional Neural Networks for No-Reference Image Quality Assessment**. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition: 23-28 June 2014 2014*; 2014: 1733-1740.
37. Liu X, Weijer J, Bagdanov AD. **RankIQA: Learning from Rankings for No-reference Image Quality Assessment**. *IEEE Computer Society* 2017.
38. Ma K, Liu W, Zhang K, Duanmu Z, Wang Z, Zuo W. End-to-End Blind Image Quality Assessment Using Deep Neural Networks. IEEE Trans Image Process. 2018;27(3):1202–13.
39. Berg JV, Krnke S, Gooen A, Bystrov D, Young S: **Robust chest x-ray quality assessment using convolutional neural networks and atlas regularization**. In: *Image Processing: 2020*; 2020.
40. Kashyap S, Moradi M, Karargyris A, Wu J, Morris M, Saboury B, Siegel E, Syeda-Mahmood T: **Artificial Intelligence for Point of Care Radiograph Quality Assessment**; 2019.
41. Hanna TN, Steenburg SD, Rosenkrantz AB, Pyatt RS Jr, Duszak R Jr, Friedberg EB. Emerging Challenges and Opportunities in the Evolution of Teleradiology. AJR Am J Roentgenol. 2020;215(6):1411–6.

Figures

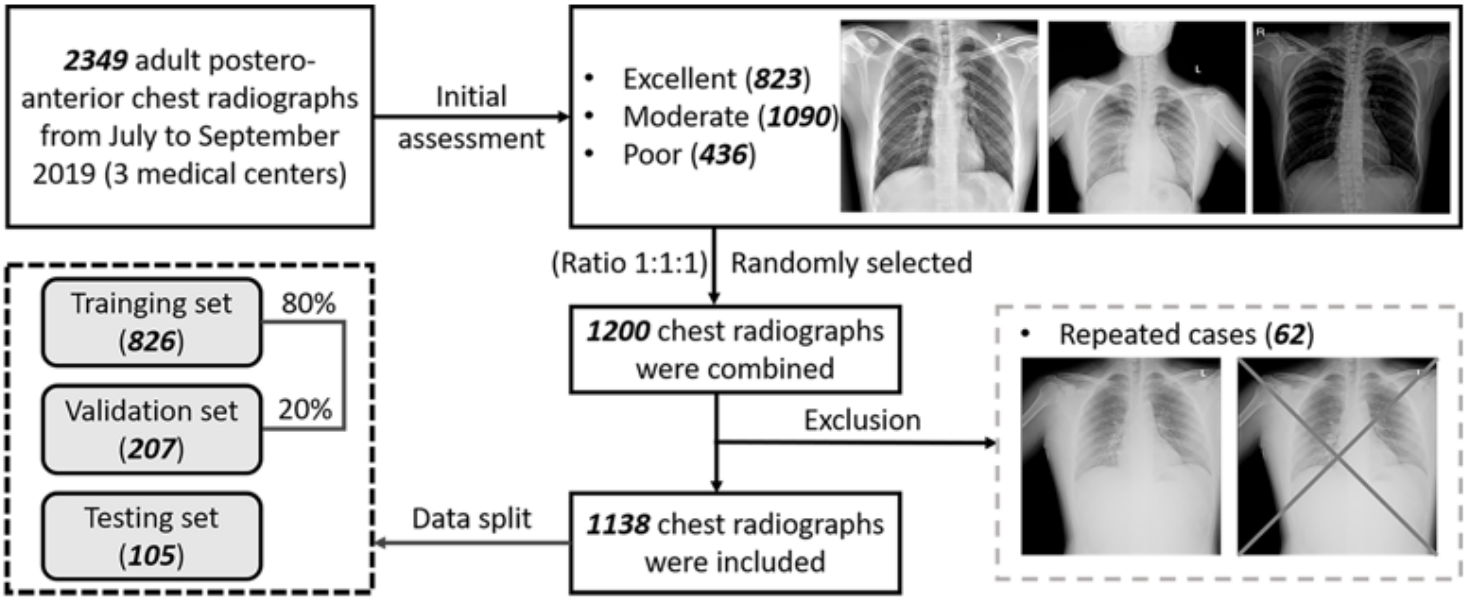

Figure 1

The flowchart of the image acquisition and data split.

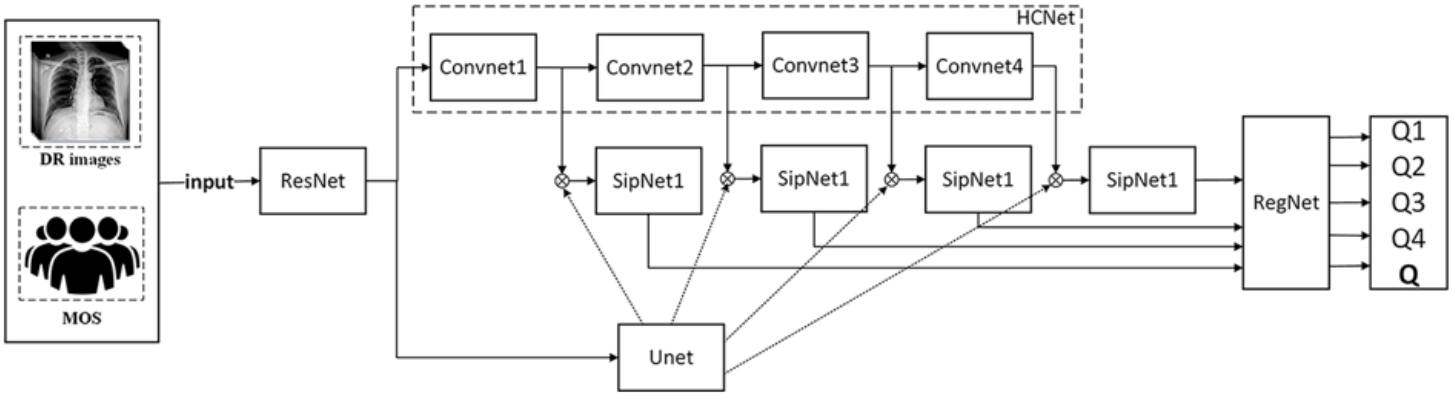

Figure 2

Deep Learning Network Architecture for IQA of Chest Radiographs. DR (Digital radiographic), MOS (Mean Opinion Scores), ResNet (Residual Network), HCNet (Hierarchical Convolutional Net), SipNet (Side Pooling Nets), RegNet (Regression Net)

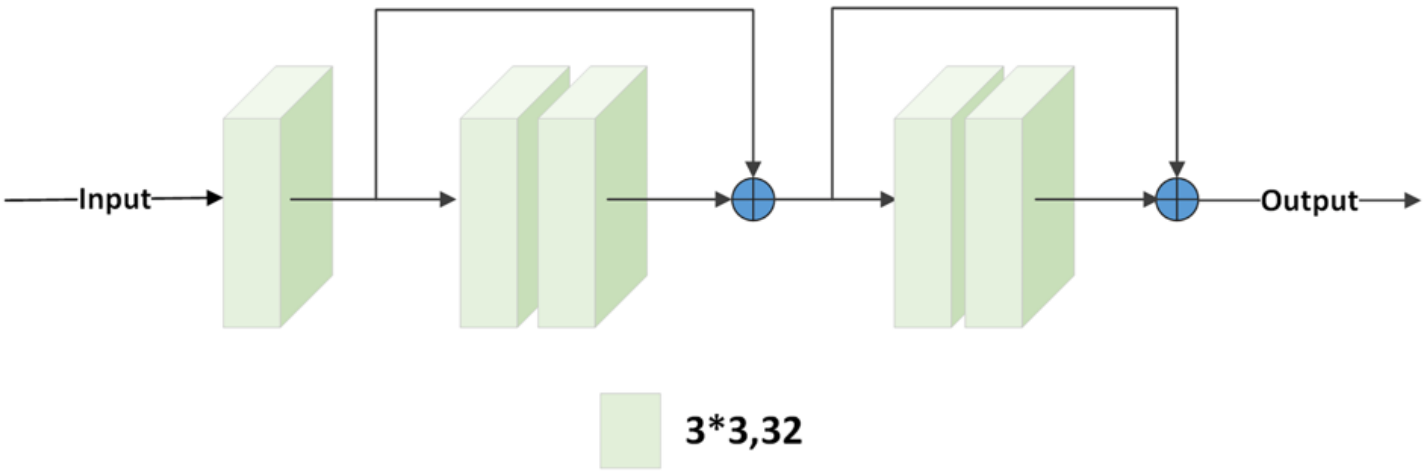

Figure 3

The stucture of ResNet

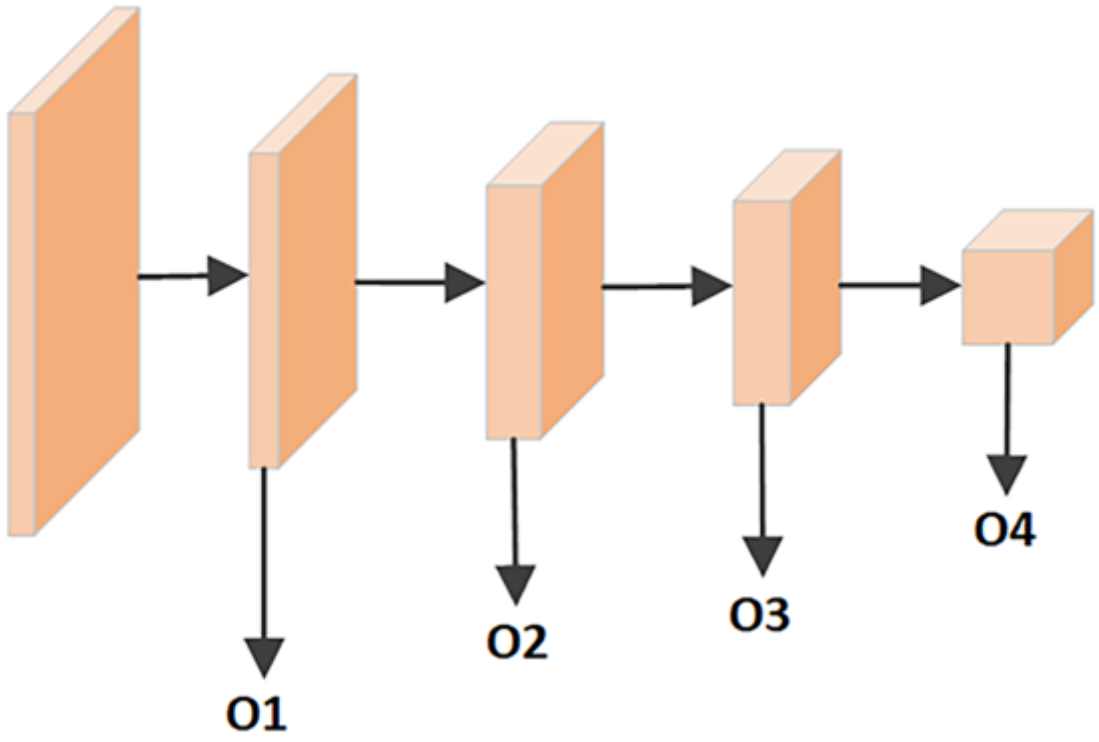

Figure 4

The stucture of HCNet

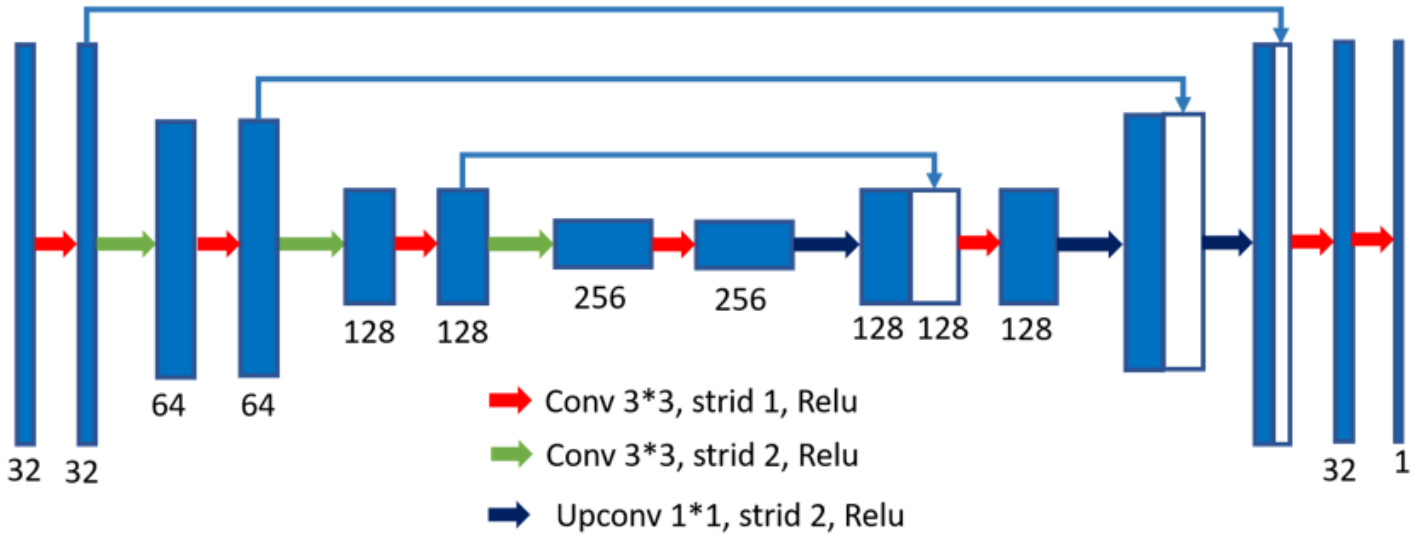

Figure 5

The stucture of Unet

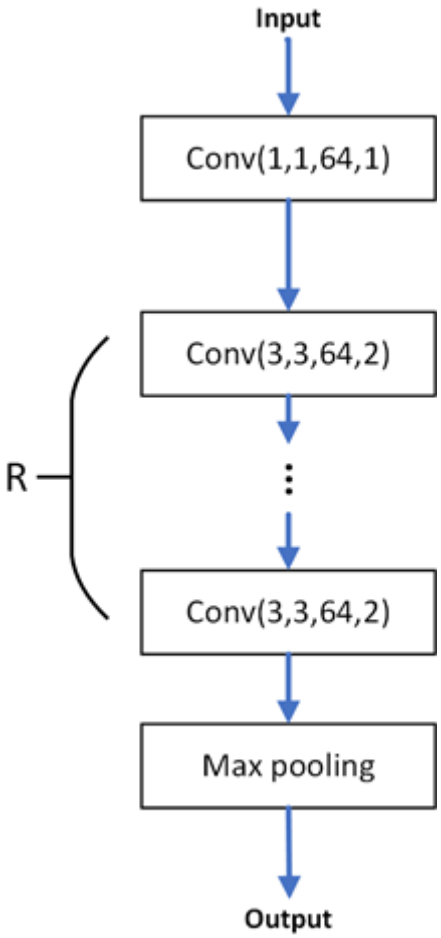

Figure 6

The stucture of SiPNet

Figure 7

The structure of RegNet

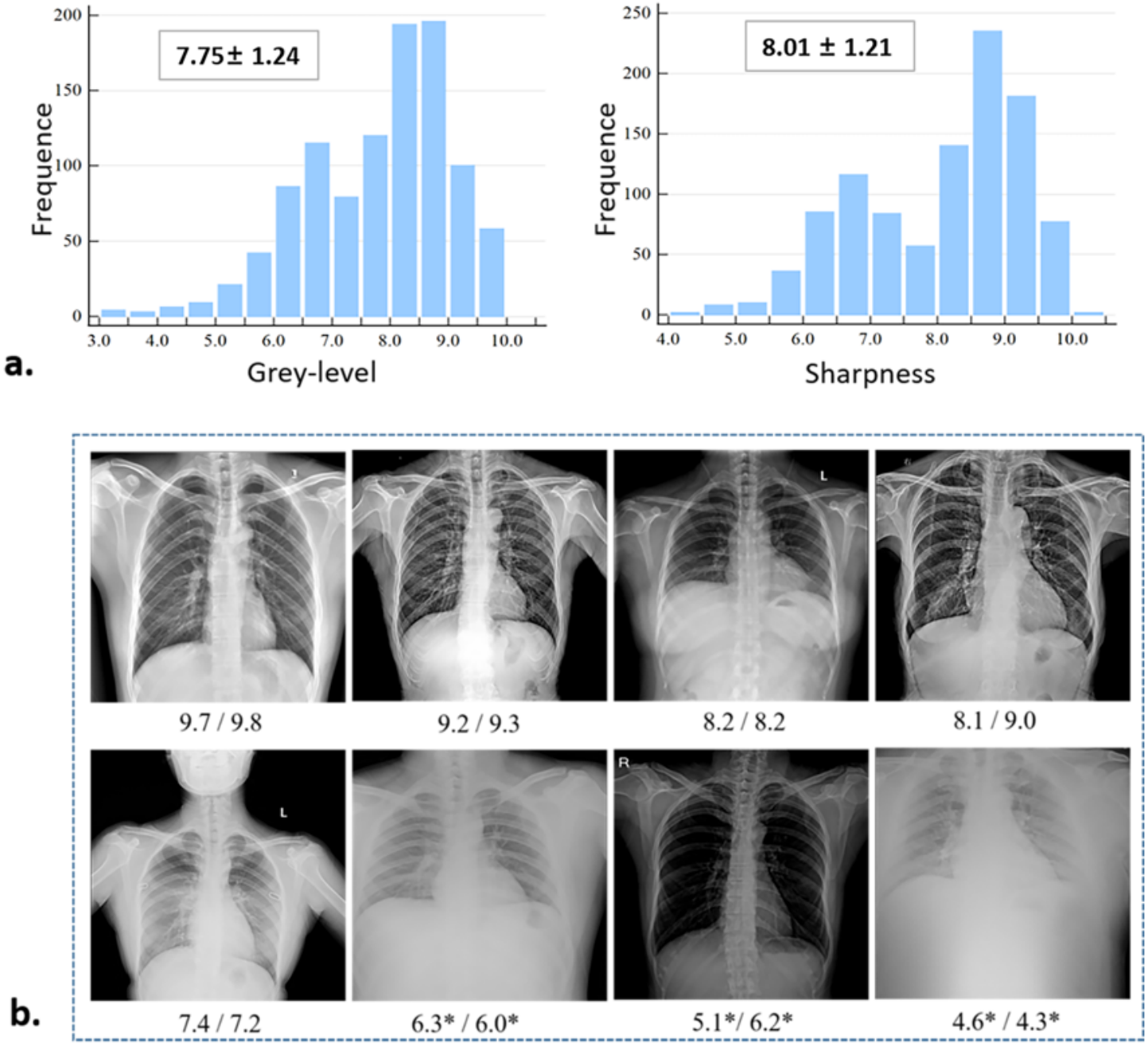

Figure 8

The subjective assessment of chest radiographs by experts. **a** the distribution of the gray-level and sharpness scores, respectively. **b** the examples of the images with scores, the left and right sides of the slash indicated the score of gray-level and sharpness, respectively. * denoted that the experts considered image quality non-acceptable as it affects diagnosis.

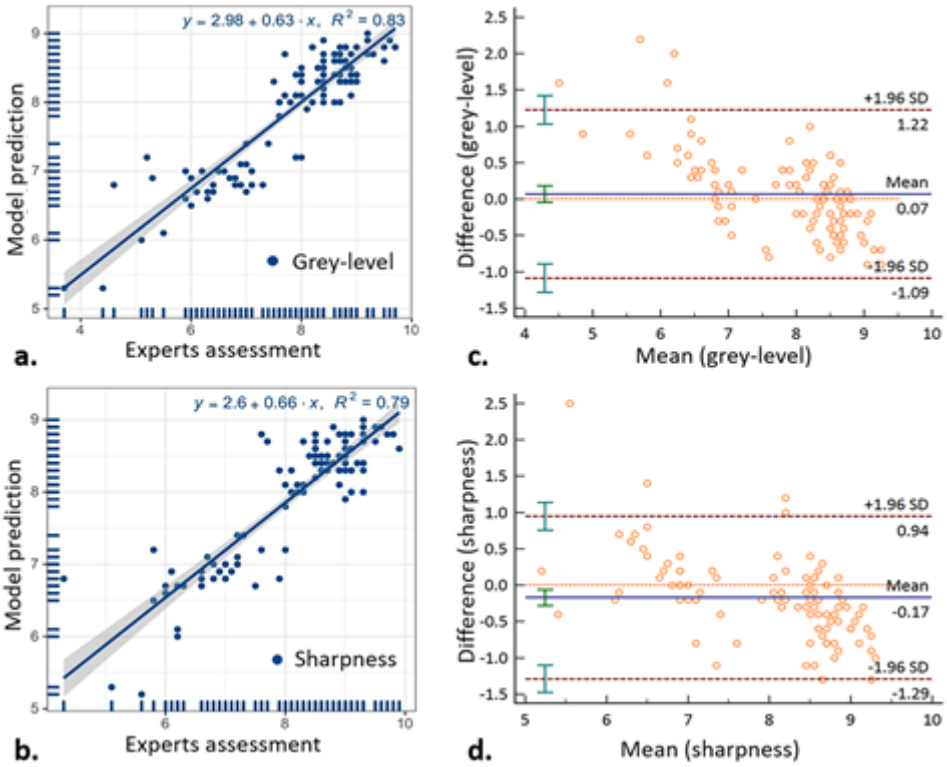

Figure 9

Correlation of quantitative scores between DL model and reference standard. (**a**/**b**) Line regression of quantitative scores between model prediction and experts assessment. (**c**/**d**) Bland-Altman plots for differences in quantitative scores between the model and the reference standard.

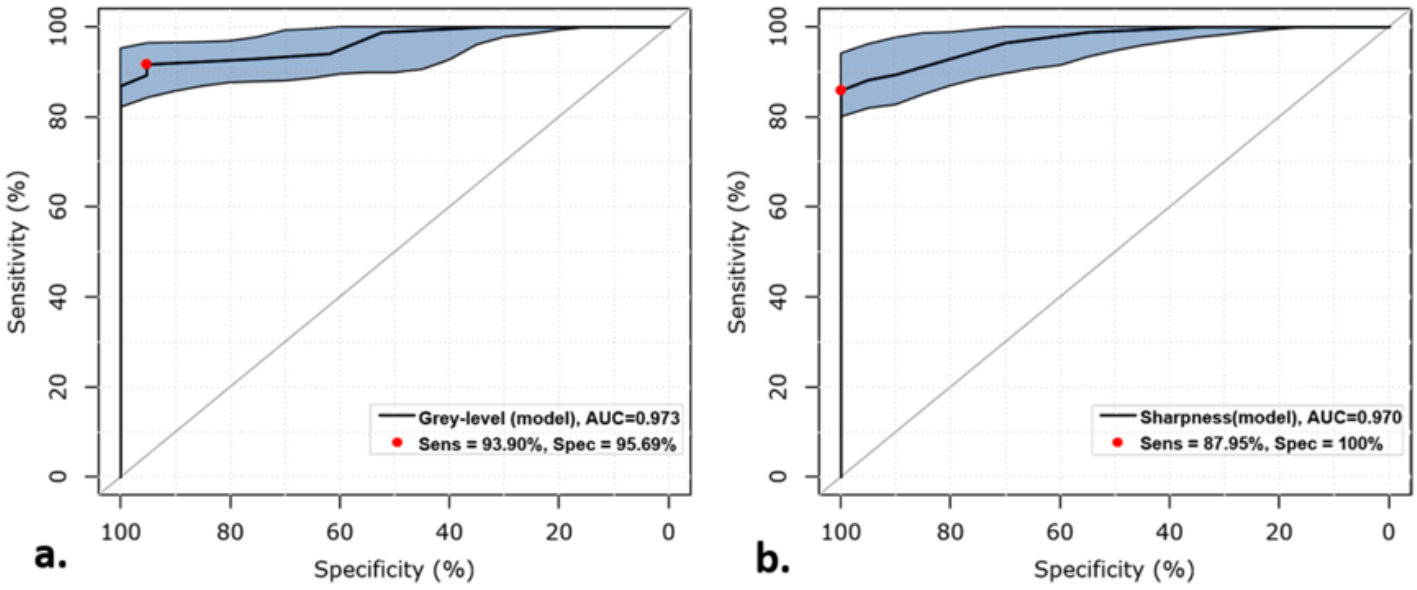

Figure 10

Receiver operating characteristic curves of gray-level (**a**) and sharpness (**b**). The area under the receiver operating characteristic curve was 0.966 (**a**) and 0.969 (**b**). **Sens** (Sensitive), **Spec** (Specific).