Evaluating a Deep Learning-Based CT Image Quality Model Using Radiologist Assessment in Clinical Abdominal CT

Azmul Siddique[1], Gary Y. Ge, PhD[1], Trae Brooks, MD[1], Zain Hassan, MD[1], Wilfred R Furtado, MD[1], Bryson Jones, MD[1], Elias Nassar[1], James Lee, MD[1], Halemane Ganesh, MD[1], Abdullah-Al-Zubaer Imran, PhD[1] and Jie Zhang, PhD[1]
[1]University of Kentucky

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Purpose: Deep learning-based image quality assessment (DL-IQA) models are commonly trained using radiologist ratings and may provide a more objective approach to CT image quality evaluation. However, clinical deployment may be limited by institutional variation in scanners, protocols, and reader preferences. This study evaluated how well a DL-IQA model aligns with radiologists' subjective assessment of image quality for routine abdominal CT at our institution.

Methods: A retrospective dataset of 100 adult abdomen-pelvis CT examinations (50 contrast-enhanced, 50 non-contrast) acquired on two Siemens scanners (Edge, Force) was reconstructed using FBP and ADMIRE with five strengths (IR1-IR5). Slice-level IQA was computed using a reference-free DL-IQA model that predicts perceptual quality scores (0-4) using natural-image pretraining and CT fine-tuning within a hybrid CNN-transformer architecture. Seven radiologists (four residents, one fellow, two attendings with >15 years' experience) independently rated image quality using the model's grading rubric for a BMI-balanced subset of 40 cases. Per-slice water-equivalent-diameter (WED), mAs, and DL-IQA scores were recorded. Model alignment was assessed using Spearman correlation versus individual and mean reader scores. Inter-reader agreement was quantified using Spearman's ρ, Kendall's W, and Fleiss' κ, with subgroup analysis by training level.

Results: Inter-reader agreement was low (Fleiss' κ≈0.14; Kendall's W≈0.17), highlighting substantial subjectivity in clinical image quality assessment. DL-IQA showed mild correlation with the mean radiologist score (ρ=0.31), comparable to inter-reader correlations. Stronger alignment was observed for two early-career readers (1-3 years; ρ=0.54 and 0.49). DL-IQA exhibited weak dependence on patient size despite expected mAs scaling with WED (R^2=0.41 non-contrast; 0.13 contrast). No significant IQA differences were observed between contrast and non-contrast scans (p>0.05).

Conclusion: Radiologist assessment remains the clinical reference for CT image quality yet is inherently variable. The evaluated DL-IQA model demonstrated mild agreement with consensus ratings and stronger alignment with early-career readers, supporting further validation across institutions and protocols.

CONTACT

[Azmul Siddique]
[University of Kentucky]
[Lexington, KY]
[ahsi223@uky.edu]

INTRODUCTION

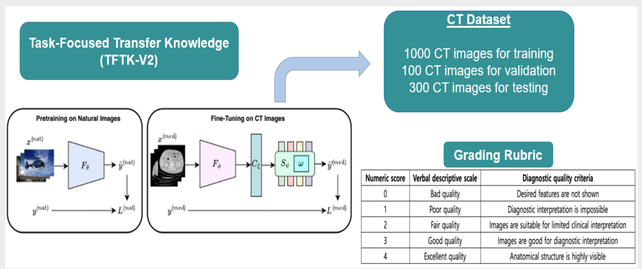
Computed tomography (CT) image quality assessment (IQA) is essential for balancing diagnostic confidence with radiation dose optimization. Traditional IQA methods include objective metrics such as image noise, signal-to-noise ratio (SNR), and contrast-to-noise ratio (CNR), full-reference image similarity metrics, and task-specific approaches such as lesion detectability, anatomical visibility, and quantitative accuracy. While valuable, these methods often require reference images, have limited generalizability across clinical applications, or fail to fully capture radiologists' subjective perception of image quality. Consequently, radiologist assessment remains the clinical reference standard despite being time-consuming and subject to interobserver variability.

Recent advances in deep learning have enabled automated, reference-free CT IQA methods capable of predicting perceptual image quality directly from clinical images. A recently developed task-focused knowledge transfer (TFKT) model leverages transfer learning from natural image quality datasets followed by CT-specific fine-tuning to generate perceptual IQA scores without requiring high-dose reference images (Figure 1). Although the model has demonstrated excellent benchmark performance, its agreement with radiologist assessments in routine clinical CT remains to be established.

This study evaluated the agreement between deep learning-derived CT IQA scores and radiologist image quality assessments for routine abdominal and pelvic CT examinations reconstructed with filtered back projection (FBP) and multiple iterative reconstruction (IR) strengths. Agreement across readers with varying levels of experience was also assessed to investigate the clinical applicability of automated IQA for objective image quality evaluation and CT protocol optimization.

METHODS AND MATERIALS

A retrospective dataset of 100 adult abdomen-pelvis CT examinations (50 contrast-enhanced and 50 non-contrast) acquired on Siemens SOMATOM Edge and Force scanners was evaluated. Each examination was reconstructed using FBP and five IR strengths (ADMIRE IR1-IR5), generating six reconstruction series per examination. Slice-level image quality assessment (IQA) scores were computed using a previously developed reference-free deep learning (DL-IQA) model based on task-focused knowledge transfer (TFKT), which predicts perceptual image quality scores (0-4) through natural-image pretraining followed by CT-specific fine-tuning.



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

Figure 1: Overview of the task-focused knowledge transfer (TFKT V2) deep learning image quality assessment (DL-IQA) framework and the 5-point image quality grading rubric used for radiologist evaluation and model prediction.

METHODS AND MATERIALS CONT.

Seven radiologists (four residents, one fellow, and two attending radiologists with >15 years of experience) independently evaluated a BMI-balanced subset of 40 examinations using the same 5-point image quality grading rubric (Figure 1) employed during model development. For each CT slice, water-equivalent diameter (WED), tube current-time product (mAs), and DL-IQA scores were recorded to investigate the effects of patient size and acquisition technique on model performance.

Model agreement with individual radiologists and the mean radiologist score was evaluated using Spearman correlation. Inter-reader reliability was assessed using Spearman's ρ, Kendall's coefficient of concordance (W), and Fleiss' κ. Additional analyses examined model performance across reader experience levels, contrast-enhanced versus non-contrast examinations, scanner platforms, and slice-wise IQA trends to explore potential clinical applications for automated CT image quality assessment.

RESULTS

Substantial inter-reader variability was observed among the seven radiologists, with only slight agreement by Fleiss' κ (≈0.14) and low concordance by Kendall's W (≈0.17), highlighting the subjective nature of clinical CT image quality assessment. Across all 40 evaluated examinations, the DL-IQA model consistently produced image quality scores that fell within the range of radiologist ratings and closely tracked the mean radiologist score (Figure 2).

Pairwise Spearman correlation analysis demonstrated mild agreement between the DL-IQA model and the mean radiologist assessment (ρ=0.31), comparable to the level of agreement observed among individual readers (Figure 3). The strongest model-reader correlations were observed for two early-career radiologists (1-3 years of experience; ρ=0.54 and 0.49), suggesting closer alignment between the model and less experienced readers.

Analysis of acquisition-related factors demonstrated weak dependence of DL-IQA scores on patient size despite the expected increase in tube current with water-equivalent diameter (R^2=0.41 for non-contrast and 0.13 for contrast examinations). Additionally, no significant differences in DL-IQA scores were observed between contrast-enhanced and non-contrast CT examinations (p>0.05), indicating that model predictions were not biased by contrast enhancement.

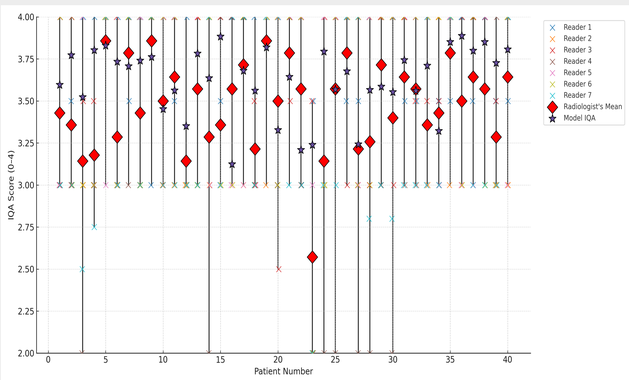


Figure 2: Distribution of model and radiologist IQA scores across 40 evaluated patients. Despite substantial inter-reader variability (Fleiss's κ≈0.14; Kendall's W≈0.17), the DL-IQA model scores consistently fall within the range of radiologist assessments and closely track the mean radiologist score.

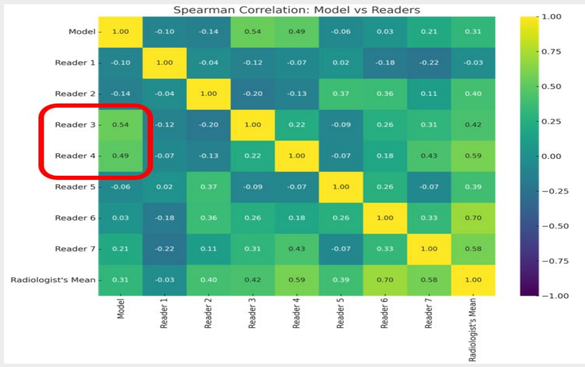


Figure 3: Spearman correlation heat map comparing the DL-IQA model, individual radiologists, and the mean radiologist score. The model demonstrated mild agreement with the mean radiologist score (ρ = 0.31) and stronger agreement with two early-career readers as shown by the red box (ρ = 0.54 and 0.49).

DISCUSSION

This study evaluated the clinical applicability of a reference-free deep learning CT image quality assessment (DL-IQA) model by comparing its predictions with radiologist evaluations of routine abdominal CT examinations. The low inter-reader agreement highlights the inherent subjectivity of perceptual image quality assessment and underscores the need for objective, reproducible tools to complement expert review.

Despite this variability, the DL-IQA model demonstrated mild agreement with the consensus radiologist assessment, comparable to the agreement observed among individual readers. Stronger agreement was observed with two early-career radiologists, which may indicate that the current model has not yet fully captured the image quality preferences of experienced abdominal radiologists. Further fine-tuning using larger, body region-specific clinical datasets may improve agreement across readers with varying levels of experience.

The weak dependence of DL-IQA scores on patient size and the absence of significant differences between contrast-enhanced and non-contrast examinations further support the robustness of the model across routine clinical imaging conditions. These findings suggest potential applications in objective image quality monitoring, quality assurance, and protocol optimization.

This study was limited to abdominal CT examinations from a single institution and a relatively small cohort of radiologists. Future work will focus on validation across additional anatomic regions and scanner platforms, as well as further refinement of the model using larger clinical datasets.

CONCLUSIONS

Radiologist assessment remains the clinical reference standard for CT image quality but is inherently subjective and variable. The evaluated reference-free DL-IQA model demonstrated mild agreement with consensus radiologist ratings and stronger alignment with early-career readers, supporting its potential as an objective adjunct for CT image quality assessment and protocol optimization.

REFERENCES

1. Rifa, K. R., Ahamed, M. A., Zhang, J., & Imran, A. (2025). TFKT V2: Task-focused knowledge transfer from natural images for computed tomography perceptual image quality assessment. *Journal of Medical Imaging*, 12(5), 051805. https://doi.org/10.1117/1.JMI.12.5.051805

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com