

Evaluation of No-Reference Image Quality Measurements for Medical Imaging

Dilek Özer

Institute of Biomedical Engineering
Boğaziçi University
Istanbul, TÜRKİYE

Prof. Dr. Albert Güveniş

Institute of Biomedical Engineering
Boğaziçi University
Istanbul, TÜRKİYE

Abstract— Medical imaging technologies are essential for making clinical diagnosis and treatments, but issues like blurriness, noise, and loss of contrast can impair diagnostic precision. This research study assesses the effectiveness of three commonly used No-Reference Image Quality Assessment(NR-IQA) models – PIQE, BRISQUE, and NIQE-on X-ray computed Tomography(CT) images. A collection of 50 images, consisting of both ideal and artificially distorted CT scans, was evaluated subjectively by three doctors and objectively using NR-IQA models. The results reveal that PIQE demonstrates strong agreement (82%) with human assessment, while BRISQUE and NIQE show a poor alignment (30% and 28%, respectively). These findings suggest that PIQE is more appropriate for applications in medical imaging, recommending its incorporation into automated quality control systems to enhance diagnostic results.

Keywords— medical imaging, image distortion, image quality assessment(IQA), subjective IQA, no-reference IQA, clinical decision support.

I. INTRODUCTION

The advent of medical imaging technologies has transformed diagnostics, allowing for precise and prompt medical interventions[1]. Nevertheless, the quality of these images is essential for facilitating accurate clinical decision-making[2]. No-Reference Image Quality Assessment(NR-IQA) models are crucial in this area, as they gauge image quality without the need for reference images[3].

Despite their common application, NR-IQA models encounter challenges within the medical field due to the specific imaging demands. Medical images frequently exhibit localized artifacts that differ from those typically seen in natural images, where conventional NR-IQA models have mainly been tested[4]. Prior research highlights the challenges of applying IQA models that were initially developed for natural scenes to clinical datasets, emphasizing the need for specialized solutions that consider the structural intricacies of medical data[5],[8]. This reinforces the significance of thoroughly examining how effectively NR-IQA models align with expert evaluations in clinical settings.

Among the frequently used NR-IQA models, BRISQUE[1] and NIQE[2] are often employed; however, they have shown varying effectiveness in the context of medical imaging[8]. On the other hand, the Perception-based Image Quality Evaluator(PIQE) has been developed as a block-wise, unsupervised method that focuses on localized distortions and does not depend on Natural Scene Statistics(NSS)[3],[9]. These features render PIQE especially pertinent for clinical imaging, where issues such as blur, loss of contrast, and noise usually appear in specific areas with clear diagnostic significance[5].

A. Quality of Medical Images

Medical imaging, particularly Computed Tomography(CT), plays a vital role in clinical diagnostics by offering a non-invasive view of anatomical structures[5]. The reliability of CT images in diagnosis is heavily dependent on their quality, as issues such as blur, noise, and contrast loss can hinder physicians from interpreting images correctly. Consequently, Image Quality Assessment(IQA) is essential to guarantee that only diagnostic-quality images are used in clinical settings[6].

B. No-Reference Image Quality Assessment(IQA)

Conventional Image Quality Assessment(IQA) techniques rely on reference images[7], which may frequently be unavailable in numerous clinical scenarios. As a result, No-Reference Image Quality Assessment(NR-IQA) methods have gained traction for automated quality control. Among the most commonly used NR-IQA methods are BRISQUE[1], which utilizes Natural Scene Statistics(NSS) along with opinion-aware training, and NIQE [2], an opinion-unaware approach that assesses statistical deviations from naturalness. However, both methods were mainly created for natural images, which restricts their applicability in medical settings[8].

C. PIQE and Considerations Relevant to Specific Domains

Recently, the Perception-based Image Quality Evaluator (PIQE) has been introduced as a block-wise, unsupervised approach that focuses on localized artifacts[3]. In contrast to BRISQUE and NIQE, PIQE operates independently of NSS assumptions, which may make it a better fit for medical images where distortions are often localized and have clinical relevance[9]. To explore this theory, a dataset of CT scans was created using images from NIH DeepLesion repository[4], featuring both pristine and synthetically distorted images.

D. Motivation and Contribution

Prior research indicates the challenges of utilizing natural-image IQA models for medical datasets[10],[11]. To tackle this issue, we assess PIQE, BRISQUE, and NIQE on CT images, using subjective ratings from physicians as the benchmark. The contributions include:

- Creation of a CT dataset incorporating clinically relevant synthetic distortions.
- Comparative analysis of NR-IQA models(BRISQUE, NIQE, and PIQE) against the ratings provided by physicians.
- Determining the model that aligns most closely with clinical requirements.

Table 1: Essential Features of NR-IQA Models.

Model	Evaluation Method	Key Features	Assumptions
PIQE	Block-wise evaluation	Localized artifact detection	None
BRISQUE	Statistical measures	Uses NSS for quality assessment	NSS-dependent
NIQE	Statistical measures	Works without training on datasets	NSS-dependent

Table 1 summarizes PIQE has become a strong contender because of its block-wise assessment approach, which does not depend on NSS, thus making it more suitable for medical images that frequently exhibit localized distortions[3].

II. MATERIALS AND METHODS

The assessment of NR-IQA models centers on their effectiveness and precision in medical imaging, specifically CT scans. This research presents a systematic approach to evaluate the PIQE, BRISQUE, and NIQE models, highlighting their clinical significance. The study has four phases, Dataset creation, Subjective and Objective assessment, Comparative Analysis. As shown in Table 2, the procedure initiated with dataset creation consisting of 20 pristine and 30 distorted images, which was then followed by subjective evaluations in a clinical environment and objective assessments using BRISQUE, NIQE, and PIQE. The outcomes were subsequently analyzed and compared to delivering an extensive evaluation of the quality of CT images.

A. Dataset

A collection of 20 unmodified CT scans was obtained from the NIH DeepLesion repository[4]. To mimic actual degradations, synthetic distortions such as Gaussian blur, noise, and contrast were introduced[12], leading to a final count of 50 images. These images constituted the evaluation dataset(FYPDataset). Table 2 provides an overview of the evaluation dataset.

Table 2. Composition of the Evaluation Dataset.

Image Type	Source	Number of Images
Pristine	NIH DeepLesion repository	20
Artificially Distorted	Synthetic distortions applied	30

Figure 1 displays processed images sourced from the identical original CT scan (a). The samples feature Gaussian blur (b, $\sigma=2.25$), grayscale contrast alteration (c, range [0.2–0.4]), and salt-and-pepper noise (d, density=0.02). These examples illustrate the various distortions implemented in the FYPDataset.

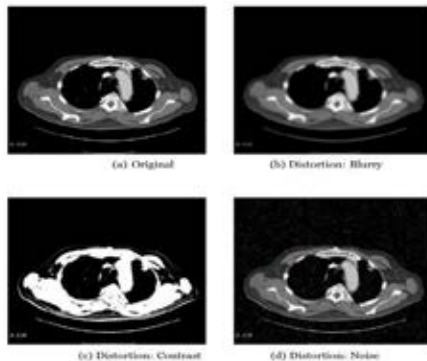


Fig. 1. A set of processed images as a sample.

B. Subjective Assessment

For reliable evaluation, subjective assessments were conducted by three physicians, which included two clinicians and one radiologist, and assessed the dataset with the help of a 5-point Likert scale. The subjective evaluation utilized the Single Stimulus Method, where each image was assessed on its own without any comparison to a reference image[21]. Each image was rated according to four structured diagnostic quality

inquiries, concentrating on aspects such as clarity, diagnostic utility, and noticeable artifacts[13]. The consensus and average ratings were considered the subjective gold-standard baseline.

Table 3. Subjective IQA Observers.

Observer Index	Position	Experience(years)	Time Cost(min)	Data Consumption
1	Clinician	13+	31	Nominative
2	Clinician	3	40	Testing
3	Radiologist	13+	23	Nominative

Table 3 shows information about the observers, including observer index, position, experience, time cost and data usage. A preliminary assessment questionnaire was created to collect the viewpoints of doctors, with the last question encouraging involvement in Subjective IQA. From the responses, three doctors were chosen as observers (Table 3) although the selection was constrained by their type and level of expertise.

C. Objective NR-IQA Models

All three NR-IQA models(BRISQUE, NIQE, and PIQE) produce numerical ratings from 0 to 100, with scores close to 0 indicating optimal image quality and scores close to 100 representing the lowest quality. Three models for No-Reference Image Quality Assessment(NR-IQA) were developed using Python: BRISQUE[1], NIQE[2], and PIQE[3]. Each model was utilized on the dataset to generate objective scores, which were subsequently compared to the ratings given by physicians to evaluate their results. In particular, BRISQUE[1] operates as an opinion-aware model that has been trained using natural scene statistics (NSS) features, NIQE[2] functions as an opinion-unaware model that relies on statistical naturalness, while PIQE[3] serves as a block-based unsupervised evaluator that focuses on localized distortions and does not rely on NSS[9], rendering it more appropriate for applications in medical imaging.

D. Agreement Calculation

Agreement between subjective and objective evaluations was quantitatively assessed as the ratio of instances where subjective and objective rankings aligned. For each image group g_i , denote $R_s(g_i)$ as the subjective ranking and $R_o(g_i)$ as the objective ranking. The indicator function can be expressed as equation (1):

$$I(g_i) = \{1, \text{ if } R_s(g_i) = R_o(g_i); 0, \text{ otherwise}\} \quad (1)$$

The overall agreement is calculated as equation (2):

$$A = (1/N) * \sum I(g_i) \times 100\% \quad (2)$$

N denotes the total count of assessed images. This metric was chosen to offer an intuitive indication of how often model-generated rankings align with expert assessments, serving as a complement to the analysis. According to this calculation, PIQE achieved 41 out of 50 matches(82%), while BRISQUE and NIQE had significantly lower rates of 30% and 28%, respectively.

III. RESULTS

The correlation in Table 4 indicates that PIQE aligns most closely with physician evaluations, where lower objective scores are typically associated with higher subjective ratings(Pearson $r = -0.16$, Spearman $\rho = -0.18$). BRISQUE shows a more moderate yet inconsistent correlation, which is evidenced by its minimal correlation coefficients(Pearson $r = 0.03$, Spearman $\rho = 0.07$). In terms of performance, NIQE ranks the lowest, characterized by poorly clustered scatter points and weak positive correlation(Pearson $r = 0.11$, Spearman $\rho = 0.09$), which goes against the anticipated pattern. In summary, PIQE stands out as the most dependable NR-IQA model when compared to the other two models.

Table 4. Correlation of NR-IQA Models with Subjective Scores.

	Model	Pearson r	Spearman p
1	PIQE	-0.163655602	-0.178549706
2	BRISQUE	0.029057302	0.070579188
3	NIQE	0.106123323	0.089692466

To confirm previous results and eliminate potential mistakes or coincidences, further tests were carried out with a new set of 20 medical images, including 8 originals, 7 blurred images, 4 with contrast issues, and 1 containing noise. The subjective Image Quality Assessment (IQA) relied on ratings from the secondary observer (index 002) in Table 3, while the Objective IQA utilized the same three models, placing particular focus on PIQE. The analysis comparing Subjective and Objective findings produced results matching with the initial study, confirming that the previous experiments did not contain major errors or discrepancies, thus enhancing the reliability of the results across primary, clinical, and computational assessments.

The degree of consensus between NR-IQA models and expert results was evaluated. Among these models, PIQE showed the highest result with subjective evaluations at 82%, whereas BRISQUE and NIQE scored 30% and 28%, respectively. This indicates that PIQE is a more reliable option for evaluating CT images in clinical settings.

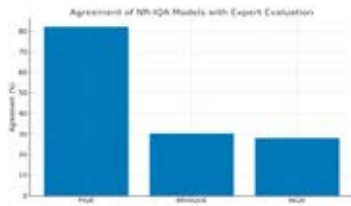


Fig. 2. Proportions of agreement between PIQE, BRISQUE, and NIQE and evaluations from experts.

Figure 2 illustrates the agreement rates between three NR-IQA models and expert evaluations. PIQE exhibits the highest performance at roughly 82%, whereas BRISQUE and NIQE show much lower agreement rates of about 30% and 28%, respectively. This indicates that piqe aligns more closely with physicians' evaluations of CT image quality than the other two models.

Table 5. Examples of Match Case.

	Scores	
	Subjective IQA	Objective IQA
Original	53	5
Blurry	52	4
Contrast	46	3

Table 6. Examples of Mismatch Case.

	Scores	
	Subjective IQA	Objective IQA
Original	46	5
Blurry	52	4
Contrast	53	3

Table 5 and Table 6 are directly linked to Equation(1) and (2). Instances labeled as "Match" receive a value of 1 in the indicator function, while "Mismatch" instances are given a value of 0. These tables illustrate cases where subjective and objective evaluations either align or differ, providing clarity on the methodology used to determine agreement percentages. A match indicates that subjective and objective rankings align for a particular group of images, whereas a mismatch shows a lack of agreement. Collectively, these examples offer tangible insight into how the reported agreement rates-82% for PIQE, 30%for BRISQUE, and 28% for NIQE- were calculated.

Table 7. Comparison of IQA Models.

Criterion	PIQE	NIQE	BRISQUE
Advantage	Fairly effective; appropriate for medical imaging [3],[9].	Excels in processing natural images[2].	Commonly utilized[1].
Disadvantage	Restrictions in different areas[3],[9].	Inappropriate for medical images[2],[8].	Insufficiently effective for medical images[1],[5].
Performance Evaluation	High perf [3],[14].	Low perf [2],[15].	Medium perf[1],[15].
Performance with X-ray CT Scans	Superior perf[3],[5].	Inconsistent[2][5]	Inconsistent[1],[5].
Applicability to Medical Imaging	Fairly efficient[3],[5].	Limited, similar to BRISQUE[2],[5].	Restricted, difficulties with texture/contrast[1],[5].
Dependence on NSS	Independent[3][9]	High[2],[8].	High[1],[8].
Sensitivity to Local Distortions	High [3],[9],[16].	Low[2],[15].	Low[1],[5].
Correlation with Subjective Evaluations	Positive (~82%)[3],[5],[14]	Poor (~28%)[2],[5],[14]	Poor (~30%)[1],[5],[14]
Requirement for Training Data	Minimal[3]	Uses training data[2]	Significant[1]
Performance in This Study	Good [5]	Poor [5]	Poor [5]

Table 7 summarizes the comparison of different IQA models. The findings indicate that the BRISQUE model exhibited a low relation with expert reviews, attaining only around %30. Likewise, the NIQE model also fell short, achieving a result of nearly 28%. On the other hand, the PIQE model exhibited significantly better performance in this study, reaching 82% when compared to expert evaluations. This enhanced result with human assessments implies that PIQE may be more adept at accurately measuring image quality concerning perceptual distortions.

IV. DISCUSSION

A. Findings

Despite their widespread use, NR-IQA models are widely used, they face difficulties in the medical domain because of the unique imaging requirements. Medical images often contain localized artifacts that are distinct from those found in natural images, where traditional NR-IQA models have primarily been evaluated[4]. Earlier research underscores the obstacles in adapting IQA models designed for natural images to clinical datasets, stressing the necessity for customized solutions. Thus, understanding how effectively these models correspond with clinical evaluations is crucial for enhancing their applicability in medical imaging.

The results show that PIQE significantly performs both BRISQUE and NIQE in evaluating the quality of CT images[14]. Unlike other approaches, PIQE does not rely on Natural Scene Statistics(NSS), which often fail to capture the complex structural characteristics found in medical imaging[9]. This represents a significant advantage for PIQE, as it can detect distortions that directly impact the diagnostic quality of CT images. The poor performance of BRISQUE and NIQE highlights the statistical differences between NSS and the structural features of medical images, indicating that these models- originally designed for natural images- are ineffective at recognizing distortions critical for clinical use[15]. In contrast, PIQE's block-wise, unsupervised method effectively identifies blur and localized noise[16], making it more capable of detecting degradations that influence physicians' evaluations of diagnostic value. Instances of match and mismatch

evaluations can be found in Table 5 and Table 6, which further clarifies how agreement is calculated.

B. Comparison with Previous Research

The performance of BRISQUE and NIQE underscores the statistical disparities between NSS and the structural characteristics of medical images, which aligns with earlier research reporting comparable shortcomings of NSS-based techniques. Conversely, the effectiveness of PIQE reinforces prior observations that highlight the benefits of localized, block-wise, and unsupervised approaches over models constructed on natural image statistics[9],[15],[16].

Previous research[3],[5] has also indicated that PIQE outperforms other NR-IQA models in analyzing X-ray CT scans; however, those findings were primarily derived from broad evaluations or review analysis. In contrast, this study builds on that evidence by presenting a direct comparison of PIQE with BRISQUE and NIQE using a specially curated CT dataset, while also validating the outcomes against assessments by physicians. This approach provides stronger clinical relevance and empirical validation, illustrating that PIQE not only corresponds more closely with expert evaluations but also offers improved practicality for clinical use. This signifies important progress toward dependable automated quality control in CT imaging.

C. Potential Clinical Applications

The proven capability of PIQE to identify distortions that impact diagnostic quality has real-world implications for medical imaging processes. Given its greater alignment with physicians' subjective evaluations, PIQE could be incorporated into automated quality assurance systems. This integration would facilitate the prompt identification of low-quality CT images prior to their application in clinical diagnosis, consequently decreasing the chances of diagnostic errors and improving reliability in healthcare decision-making.

D. Limitations of the Current Research

The findings of this study should be interpreted with certain limitations. Firstly, the dataset comprised only 50 CT images, which could limit its broader applicability; larger datasets are necessary to account for more extensive variability[5]. In addition, the application of synthetic distortions may not fully capture the intricacies of actual clinical artifacts, underscoring the importance of validation with authentic medical data[14]. This research was also confined to CT images and assessments from only three clinicians, which may lead to potential bias[5],[11]. Lastly, the research study did not investigate real-time processing, which is vital for clinical settings and should be considered in future research. [16].

E. Future Work

Future research should focus on improving NR-IQA models and exploring hybrid approaches that incorporate deep learning techniques to advance the field of medical image quality assessment[11]. The effect of the extent of distortions may also be investigated. Such advancements could help bridge the gap between algorithmic evaluations and the more nuanced assessments made by human experts, facilitating the integration of technology into healthcare systems. Broadening research to encompass larger datasets and various imaging modalities will enhance the validation of PIQE's versatility across different clinical scenarios, while also promoting the advancement of deep learning-driven IQA models specifically designed for medical imaging[17]-[20].

V. CONCLUSION

The evaluation of NR-IQA models indicates that PIQE significantly outperforms both BRISQUE and NIQE in assessing the quality of CT images, which is crucial for clinical purposes. Integrating PIQE into imaging workflows could

enhance automated quality control, support the early detection of poor scans, and boost diagnostic reliability.

REFERENCES

- [1] A. Mittal, A. Moorthy, and A. Bovik, "BRISQUE: Blind/Referenceless Image Spatial Quality Evaluator," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 3449–3461, 2012.
- [2] A. Mittal, R. Soundararajan, and A. Bovik, "NIQE: A Natural Scene Statistic-based Quality Evaluator," *IEEE Signal Processing Letters*, vol. 20, no. 11, pp. 154–158, 2013.
- [3] S. Venkatanath, R. Praneeth, B. Chandrasekhar, et al., "PIQE: Perception Based Image Quality Evaluator," in *Proc. IEEE National Conf. of Communications (NCC)*, 2015, pp. 1–6.
- [4] K. Yan et al., "DeepLesion: Automated Mining of Large-Scale Lesion Annotations and Universal Lesion Detection with Deep Learning," *Journal of Medical Imaging*, vol. 5, no. 3, 036501, 2018.
- [5] R. J. Ferrari et al., "Medical image quality assessment: A review," *Journal of Digital Imaging*, vol. 30, pp. 449–459, 2017.
- [6] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. IEEE Int. Conf. on Pattern Recognition(ICPR)*, 2010, pp. 2366–2369.
- [7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [8] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.
- [9] P. Le Callet, S. Möller, and A. Perkis, "Qualinet white paper on subjective methods for image quality assessment," *COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services*, 2012.
- [10] M. Narwaria and W. Lin, "Objective image quality assessment based on support vector regression," *IEEE Transactions on Neural Networks*, vol. 21, no. 3, pp. 515–519, 2010.
- [11] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [12] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Proc. IEEE Int. Conf. on Image Processing(ICIP)*, 2017.
- [13] X. Liu, J. van de Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *Proc. IEEE Int. Conf. on Computer Vision(ICCV)*, 2017, pp. 1040–1049.
- [14] Y. Zhang, J. Gao, and F. Zhu, "Saliency-inspired blind image quality assessment using a naturalness model," *Neurocomputing*, vol. 172, pp. 392–402, 2016.
- [15] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.
- [16] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Computer Vision and Pattern Recognition(CVPR)*, 2017.
- [17] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.
- [18] C. Li, A. C. Bovik, and X. Wu, "Blind image quality assessment using a general regression neural network," *IEEE Transactions on Neural Networks*, vol. 17, no. 5, pp. 1317–1329, 2006.
- [19] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 517–526, 2012.
- [20] S. Winkler and R. Subramanian, "Overview of eye tracking datasets," in *Proc. Int. Workshop on Quality of Multimedia Experience (QoMEX)*, 2013.
- [21] International Telecommunication Union, "Methodology for the Subjective Assessment of the Quality of Television Pictures," ITU-R BT.500-13, 2012.