OPEN

Fully Automated Region-Specific Human-Perceptive-Equivalent Image Quality Assessment: Application to ^{18}F-FDG PET Scans

Mehdi Amini, MSc, Yazdan Salimi, MSc,* Ghasem Hajianfar, MSc,* Ismini Mainta, MD,* Elsa Hervier, MD,* Amirhossein Sanaat, PhD,* Arman Rahmim, PhD,†‡§ Isaac Shiri, PhD,* and Habib Zaidi, PhD*//¶***

Introduction: We propose a fully automated framework to conduct a region-wise image quality assessment (IQA) on whole-body ^{18}F-FDG PET scans. This framework (1) can be valuable in daily clinical image acquisition procedures to instantly recognize low-quality scans for potential rescanning and/or image reconstruction, and (2) can make a significant impact in dataset collection for the development of artificial intelligence–driven ^{18}F-FDG PET analysis models by rejecting low-quality images and those presenting with artifacts, toward building clean datasets.

Patients and Methods: Two experienced nuclear medicine physicians separately evaluated the quality of 174 ^{18}F-FDG PET images from 87 patients, for each body region, based on a 5-point Likert scale. The body regisons included the following: (1) the head and neck, including the brain, (2) the chest, (3) the chest-abdomen interval (diaphragmatic region), (4) the abdomen, and (5) the pelvis. Intrareader and interreader reproducibility of the quality scores were calculated using 39 randomly selected scans from the dataset. Utilizing a binarized classification, images were dichotomized into low-quality versus high-quality for physician quality scores ≤3 versus >3, respectively. Inputting the ^{18}F-FDG PET/CT scans, our proposed fully automated framework applies 2 deep learning (DL) models on CT images to perform region identification and whole-body contour extraction (excluding extremities), then classifies PET regions as low and high quality. For classification, 2 mainstream artificial intelligence–driven approaches, including machine learning (ML) from radiomic features and DL, were investigated. All models were trained and evaluated on scores attributed by each physician, and the average of the scores reported. DL and radiomics-ML models were evaluated on the same test dataset. The performance evaluation was carried out on the same test dataset for radiomics-ML and DL models using the area under the curve, accuracy, sensitivity, and specificity and compared using the Delong test with P values <0.05 regarded as statistically significant.

Results: In the head and neck, chest, chest-abdomen interval, abdomen, and pelvis regions, the best models achieved area under the curve, accuracy, sensitivity, and specificity of [0.97, 0.95, 0.96, and 0.95], [0.85, 0.82, 0.87, and 0.76], [0.83, 0.76, 0.68, and 0.80], [0.73, 0.72, 0.64, and 0.77], and [0.72, 0.68, 0.70, and 0.67], respectively. In all regions, models revealed highest performance, when developed on the quality scores with higher intrareader reproducibility. Comparison of DL and radiomics-ML models did not show any statistically significant differences, though DL models showed overall improved trends.

Conclusions: We developed a fully automated and human-perceptive equivalent model to conduct region-wise IQA over ^{18}F-FDG PET images. Our analysis emphasizes the necessity of developing separate models for body regions and performing data annotation based on multiple experts' consensus in IQA studies.

Key Words: PET, image quality assessment, image analysis, radiomics, deep learning

(*Clin Nucl Med* 2024;49: 1079–1090)

Quantitative metabolic imaging using ^{18}F-FDG PET is a cornerstone of cancer care and has become an important tool in disease management, particularly for oncology applications involving diagnosis, staging, prognosis, and monitoring of treatment response.[1,2] A number of factors in the pipeline of the PET imaging process can influence the quality of the produced images.[3,4] These factors can emerge from the scanner hardware (eg, the miscalibration of the equipment), the radiopharmaceutical injection process (eg, IV injection, ^{18}F-FDG injected dose), patient characteristics (eg, patients' weight, size, and motion during imaging), acquisition procedure (eg, patient and bed positioning, scan duration per bed position), as well as image reconstruction and attenuation and scatter correction (ASC) algorithms.[3–5] In addition, the appearance of different artifacts that emerge from the PET image or propagated from the concomitant anatomical CT image used for ASC can degrade the image quality and reduce the diagnostic confidence of the interpreters.[6] Due to the quantitative nature of PET images, small variations in the images can bias the SUV, thus leading to misinterpretation by physicians, potentially altering the diagnosis.[7] Hence, prior to clinical use of PET images, high image quality is essential for a precise qualitative interpretation and accurate quantitative analysis.

Received for publication July 5, 2024; revision accepted August 31, 2024.
From the *Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland; †Department of Physics and Astronomy, Faculty of Science, University of British Columbia, Vancouver, British Columbia, Canada; ‡Department of Radiology, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, Canada; §Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, British Columbia, Canada; ||Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; ¶Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark; and **University Research and Innovation Center, Óbuda University, Budapest, Hungary.
Conflicts of interest and sources of funding: none declared. This work was supported by the Swiss National Science Foundation under grant SNSF 320030-231742 and the Private Foundation of Geneva University Hospitals under grant RC-06-01.
Correspondence to: Habib Zaidi, PhD, Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva, Switzerland. E-mail: habib.zaidi@hug.ch.
Data availability: The data used in this work is not available.
Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.
Consent to participate: Informed consent was obtained from all individual participants included in the study.
Consent to publish: All authors approved the final version of the manuscript and consent to give the publisher the permission to publish the work.
Supplemental digital content is available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal's Web site (www.nuclearmed.com).

ISSN: 1536-0229/24/4912–1079
DOI: 10.1097/RLU.0000000000005526

Artificial intelligence (AI) techniques are increasingly being used in various applications of nuclear medicine imaging.[8,9] ^{18}F-FDG PET, serving as the cornerstone of PET imaging, has become one of the priorities in AI-driven research within nuclear medicine.[10] A pile of literature explores the use of various machine learning (ML) and deep learning (DL) algorithms to automate and/or improve different tasks in PET imaging.[11,12] AI methodologies have found their way in all stages of nuclear medicine imaging, from instrumentation,[13] to image reconstruction and corrections,[14] segmentation,[15] and eventually harnessing the information in the images toward improved diagnosis,[16,17] prognosis,[18,19] and treatment outcome prediction.[20,21] The efficacy of these models is directly linked to the quality of the input images.[14,22] As a result, researchers and developers often dedicate significant time and effort to search in large collections of images to identify high-quality and artifact-free scans. In addition, in the case of public datasets, although they have significantly propelled the field forward, they have introduced a disconnect between data collection and its application.[23] As such, maintaining quality assessment over ^{18}F-FDG PET images is crucial to avoid bias in physicians' interpretation and training process of AI-based algorithms.

Conventional methods used in the clinic for quality assessment of PET images include subjective and objective approaches.[24,25] In some imaging centers, the quality of the PET images is approved subjectively by the radiologists, which represents an inefficient allocation of high-cost labor and is extremely time-consuming.[26] Additionally, subjective image quality assessment (IQA) cannot always be performed immediately after image acquisition, potentially necessitating a second scan, which increases the patient's radiation dose and discomfort and decreases scanner throughput.[27] It also lacks reproducibility due to various environmental (eg, change in the lighting and workstation hardware/software) and workforce-related (eg, physician experience and fatigue) factors.[28] Alternatively, image quality can be evaluated objectively by extracting metrics from PET images, such as signal-to-noise ratio or standard deviation of the SUV from a manually defined region of interest.[28] Other metrics include mean liver and blood pool SUVs, intact and normalized by lean body mass, which are expected to be within certain limits suggested in the literature.[29,30] However, these metrics are not comprehensive enough to capture image quality as they do not fully align with human perception and are independent of the image content.[31]

Conversely, AI-based approaches, such as radiomics-ML and DL techniques, systematically capture comprehensive and content-specific patterns in the target images and, by learning from the data, identify features well-aligned with human perception.[32,33] Previous studies have investigated the application of AI algorithms in conducting QA over different medical images, such as T2-weighted abdomen MR images,[26] 3D whole-heart MR images,[27] high-frequency ultrasound images,[34] mammography,[35] and optical coherence tomography angiography.[36] Regarding ^{18}F-FDG PET images, Schwyzer et al[37] investigated the application of a residual convolutional network (Res-Net) on MIP images with simulated decrease of acquisition time. Reynés-Llompart et al[38] investigated ML classifiers fed with different conventional biomarkers (biological and physical) and radiomic features to evaluate the quality of abdominal ^{18}F-FDG images. Qi et al[28] applied convolutional neural network algorithms to a combination of predefined 2D axial slices of ^{18}F-FDG PET images with simulating decreasing scanning times. Zhang et al[39] applied a Dense Convolutional Network (DenseNet) algorithm on whole-body ^{18}F-FDG PET images to classify them into binary low- and high-quality images.

Although the previously proposed models revealed promising results, they lack a comprehensive assessment of whole-body ^{18}F-FDG PET image quality, considering the input of the proposed models. In a whole-body ^{18}F-FDG PET scan, various body regions may exhibit varying qualities due to differences in size, shape, tissue composition, radiopharmaceutical uptake, and motion patterns in the body. Additionally, distinct imaging artifacts may appear in different body regions, prohibiting the utilization of a single model for a whole-body scan.

In this study, we propose a fully automated framework to conduct a region-wise IQA on ^{18}F-FDG PET scans. This includes a comprehensive analysis of 2 commonly used AI approaches in medical imaging, including radiomics-ML and DL frameworks. This fully automated framework can be valuable in (1) daily clinical image acquisition procedures to promptly recognize low-quality scans for potential rescanning and/or adjusting image reconstruction parameters; (2) it can play a beneficial role in dataset collection for the development of AI-driven ^{18}F-FDG PET analysis models, by rejecting low-quality images and those with artifacts, to deliver a clean dataset.

PATIENTS AND METHODS

Dataset Description

Patient Demographics

The dataset used in this study was retrospectively collected from patients undergoing whole-body ^{18}F-FDG PET/CT scans at Geneva University Hospital, Geneva, Switzerland, from July 2019 to September 2021.[14] The collection of this dataset was part of a previously conducted study approved by the institutional ethics committee (CCER ID: 2017-00922).[14] The dataset included 87 patients scanned on 2 scanners, namely, Biograph mCT and Biograph Vision (Siemens Healthineers, Knoxville, TN). Table 1 presents the demographics of the enrolled patients, in addition to information regarding PET/CT image acquisition and reconstruction protocols.

PET images were attenuation and scatter corrected once conventionally using CT images (CT-ASC), and once with a deep learning model (DL-ASC) proposed by Shiri et al.[14] The later model performs quality assurance by detecting and disentangling artifacts while performing attenuation/scatter corrections.[14,40]

Subjective Quality Assessment

The overall quality of 174 ^{18}F-FDG PET scans (87 CT-ASC + 87 DL-ASC) was qualitatively evaluated for each body region, separately by 2 experienced board-certified nuclear medicine physicians: rater 1 (I.M.) with 11 years (higher experienced physician) and rater 2 (E.H.) with 5 years (lower experienced physician) of experience (raters' evaluation was separated, and they were blind to each other's evaluation).[14] Randomly selected from the dataset, 39 of the scans were presented to the physicians twice to be reevaluated for intrareader repeatability measurement purposes (the physicians were blind to this information).[14]

The subjective IQA was performed separately for each body region. Regions included (1) head and neck, including the brain; (2) chest, including the lungs; (3) chest-abdomen interval (diaphragm region); (4) abdomen; and (5) pelvis.[14] The physicians ranked the image quality for each body region based on the Likert scale on their standard workstation, OSIRIX.[41] The scores ranged from 1 to 5, reflecting very poor, poor, average, high, and excellent image quality.

To evaluate the consistency of the quality assessment by physicians, interreader and intrareader reproducibility was calculated using an intraclass correlation coefficient (ICC) test with a 2-way mixed average measure and 95% confidence interval.[42,43] In this study, the ICC scores were classified as indicating excellent agreement for ICC ≥ 0.75, good agreement for 0.60 ≤ ICC < 0.75, fair agreement for 0.40 ≤ ICC < 0.60, and poor agreement for ICC < 0.40.[44]

TABLE 1. The Infographics of the Patients Enrolled in This Study Protocol

Characteristic	Info
Scan date	July 2019–Sep 2021
Sex	M [#48], F [#39]
Age, mean ± SD	62.4 ± 15.5 y
Weight,* mean ± SD	74 ± 16.6 kg
Radiopharmaceutical	^{18}F-FDG
Injected activity,* mean ± SD	265 ± 58 MBq
Injection to PET scan time, mean ± SD	71 ± 19 min
PET time per bed, mean ± SD	2.5 ± 0.5 min
PET reconstruction protocol	OSEM3D + TOF 2i21s [#14], PSF + TOF 2i21s [#20], PSF + TOF 5i5s [#53]
PET pixel spacing,* [minX-maxX], [minY-maxY], [minZ-maxZ]	[1.6–4] mm, [1.6–4] mm, [1.4–3] mm
CT pitch factor	0.8
CT kVp, mean ± SD	112 ± 13 kVp
CT tube current,* mean ± SD	131 ± 46 mAs
CTDI, mean ± SD	5.2 ± 2.9 mGy
CT pixel spacing,* [minX-maxX], [minY-maxY], [minZ-maxZ]	[0.98–0.98] mm, [0.98–0.98] mm, [1–1.5] mm
Scanner	SIMENS Biograph Vision [#53], SIMENS Biograph mCT [#34]

*These information were missing for 33 patients.

Automatic Image Quality Assessment

The workflow of the proposed automatic PET IQA model is illustrated in Figure 1. The proposed model for automatic quality assessment of PET images operates in a binary manner, that is, classifying images as high-quality and low-quality. To this end, images were dichotomized into low-quality versus high-quality for physician quality scores ≤3 versus >3, respectively.

For each body region, models were once developed on the scores given by the higher experienced physician (rater 1 models), once based on scores given by the lower experienced physician (rater 2 models), and once based on the average score (Avg-Rate model). For each of the body regions, the dataset was distributed in 3-folds while preserving the ratio of low- and high-quality images within each fold. To keep a fair comparison, models developed for each region were trained and cross-validated on the same 3-folds.

Pre-Processing

Whole-body contouring, excluding the extremities, was performed to define the volume of interest (VOI). To this end, we used an automatic DL-based model publicly available as "TotalSegmentator.*" [45] TotalSegmentator can extract the whole-body contour without the extremities from the CT images. This model was trained on a wide range of CT images from various scanners and institutions.[45] The generated masks were edited/verified by a nuclear medicine physicist.

For the automatic split of whole-body scans to different regions, we used a publicly available DL-based model† proposed by Schuhegger.[46] Their model performs body part regression for CT images by anatomically annotating various axial slices. Their proposed model was trained and validated on more than 2000 CT images from various scanners, institutions, and imaging protocols to ensure generalizability.[46] Based on their defined annotations, we extracted axial slices from the slice annotated "pelvis-start" to "lumbar vertebrae 5 (L5)" for pelvis region, "L5" to "lung start" for the abdomen, "lung-start" to "liver-end" for the chest-abdomen interval, "liver-end" to "thoracic vertebrae 2 (Th2)" for chest, and "thorasic-2" to "head-end" for the head and neck region including the brain.[46] After applying the model, a nuclear medicine physicist edited/verified the extracted regions.

Radiomics and Machine Learning Framework

All PET images were interpolated to an isotropic voxel size of $3 \times 3 \times 3$ mm^3 to make them rotationally invariant. The gray-level intensities were discretized to a fixed bin size of 0.2 SUV. For each region, features were extracted from the whole-body mask, excluding the arms. Standardized Environment for Radiomics Analysis,‡[47,48] an Image Biomarker Standardized Initiative–compliant package,[49,50] was used for feature extraction. A total of 174 features were extracted, including morphological, intensity-based, and texture features (GLCM, GLRLM, GLDZM, GLSZM, NGTDM, and NGLDM). The list of the calculated features with their description is presented in Supplementary Table S1, http://links.lww.com/CNM/A510.

Features with a standard deviation of zero were removed, and the Pearson test was applied to identify and remove highly correlated features (Pearson coefficient >0.95) to address nonredundancy. The predefined 3-folds (same as the folds used for DL models) were used for training and testing. Features in the training dataset underwent Z-score normalization, and the calculated mean and standard deviation were applied to normalize the testing dataset. For feature selection, we used a wrapper method named Boruta.[51]

The well-known nonparametric K-nearest neighbors[52] algorithm was used to classify images into low-quality and high-quality classes. The k-value in K-nearest neighbors was optimized during hyperparameter optimization with internal 3-fold cross-validation on the training dataset. Area under the receiver operating characteristic curve (AUC), sensitivity (SEN), specificity (SPE), and accuracy (ACC) metrics were calculated for each testing dataset, and the average value of the 3 folds was reported.

*https://github.com/wasserth/TotalSegmentator

†https://github.com/MIC-DKFZ/BodyPartRegression

‡https://github.com/ashrafinia/SERA

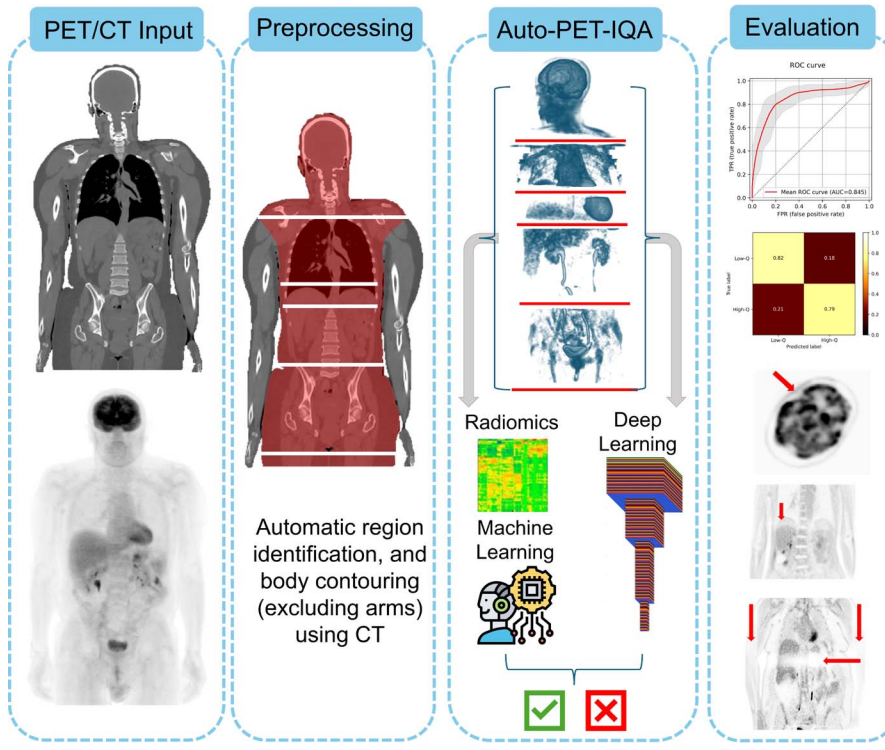

FIGURE 1. The workflow of the proposed automatic and human-perceptive ^{18}F-FDG PET IQA model.

Deep Learning Framework

A Densenet169 DL architecture[53] was utilized to classify the images. The images were resized in right-to-left and anterior-posterior directions and were padded in the craniocaudal direction to reach the target size of 128 × 96 × 64, then normalized by dividing them into 5 (SUV). The predefined 3-folds (same as the folds used in radiomics-ML models) were used for training and testing, whereas 10% of the training dataset was separated as a validation set during the training phases. Training and validation datasets were augmented by flipping and zooming to achieve a balanced number of images in each class. Data augmentation techniques that change the image contrast or add noise to images were not used to mislead the network regarding the image quality. The training was continued for 500 epochs using the cross-entropy loss function. AUC, SEN, SPE, and ACC metrics were calculated during training, and the model with the best average validation metrics was saved for inference. To avoid using an initial random weight leading to a high average metric, the first 20 epochs were neglected during the model selection phase. The training process was performed using PyTorch and MONAI[54] platforms using an RTX4090 GPU and Intel Corei913900KF CPU with an average inference time of less than a second per image.

To compare model performances, we utilized the DeLong test (for each body region: DL vs radiomics-ML, and rater 1 vs rater 2 vs Avg-Rate model). Resulting P values <0.05 were considered as significant.

RESULTS

Subjective Image Quality Assessment

Figure 2 shows the number of images associated with each quality score given by 2 physicians for each of the regions. In general, most of the images achieved scores of 3 and 4 (average and good quality). From a dichotomized point of view, the number of images in low-quality class (scores 1, 2, and 3) versus high-quality class (scores 4, and 5) seemed well balanced in head and neck, chest, and pelvis regions, but is skewed toward low quality in chest abdomen interval and abdomen regions.

The results of the intrareader and interreader reproducibility as evaluated by the ICC test (95% confidence intervals) are presented in Table 2. Rater 1 showed better intrareader reproducibility (higher ICC) in comparison to rater 2. The intrareader reproducibility in head and neck and chest regions was excellent for both raters; in chest/abdomen interval, it was excellent for rater 1 and fair for rater 2; in abdomen, it was good for rater 1 and excellent for rater 2; and in pelvis, it was excellent for both raters. The interreader reproducibility was excellent for the head and neck region, good for chest-abdomen interval and abdomen regions, and fair for chest and pelvis regions.

Automatic Quality Assessment

Model Performances

Table 3 reports the performance of all models for each body region based on AUC, ACC, SEN, and SPE. Figure 3 illustrates the receiver operating characteristic (ROC) curves of the models. The ROC curves are illustrated for each region, for radiomics-ML and DL models, and for rater 1, rater 2, and Avg-Rate models. Supplementary Figures S1 and S2, http://links.lww.com/CNM/A510, show the same ROC curves with their confidence interval generated with bootstrapping. In addition, the confusion matrices of all the models are presented in Supplementary Figures S3 and S4, http://links.lww.com/CNM/A510. In the following, for each body region, the performances of the best model selected by the authors (among rater 1, rater 2, and Avg-Rate models developed with radiomics-ML and DL approaches) are reported in the format of [AUC, ACC, SEN, and SPE].

For the head and neck region, the models showed excellent results. The best model was rater 1 model with DL approach achieving [0.97, 0.95, 0.96, and 0.95]. In the chest region as well, the rater

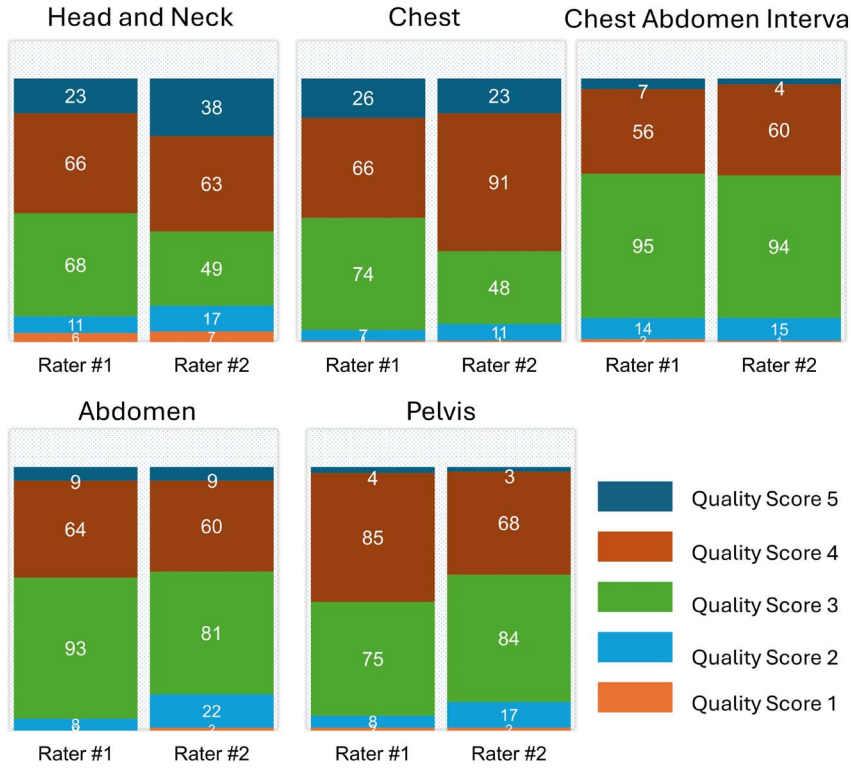

FIGURE 2. The number of images associated with each quality score attributed by 2 physicians for each of the regions.

1 model with DL approach had the best performance with [0.85, 0.82, 0.87, and 0.76]. For the chest-abdomen interval, rater 1 model with DL algorithm again achieved the best results with [0.83, 0.76, 0.68, and 0.80]. In the abdomen region, rater 2 model with DL approach was selected with [0.73, 0.72, 0.64, and 0.77]. Finally, for the pelvis region, the rater 2 model developed with DL approach was selected with [0.72, 0.68, 0.70, and 0.67].

Observer-Based Comparison of Models

As shown in Table 3, there are apparent differences in the performance of the models developed on scores given by rater 1 (higher experienced physician) and rater 2 (lower experienced physician). Figure 4 statistically compares the performance of rater 1, rater 2, and Avg-Rate models, separately for radiomics-ML and DL methods, for each body region. In the head and neck region, regardless of DL or radiomics-ML approach, the performance of the rater 1 model was significantly higher than both rater 2 and the Avg-Rate model. In the chest region, in the DL group, the rater 1 model significantly outperformed rater 2 and had the same results as the Avg-Rate model, whereas in the radiomics-ML group, the differences were nonsignificant. Among models developed for chest abdomen interval, for both radiomics-ML and DL groups, the rater 1 model showed significantly better performance than the other two. In the abdomen region, no model showed significantly different results. Regarding the models developed for the pelvis region, rater 2 and the Avg-Rate model had similar results, whereas both significantly outperformed the rater 1 model.

Comparison of Radiomics-ML and DL Strategies

Table 4 shows the results of the comparison of the performance of radiomics-ML and DL models for each region. The comparison was performed using the Delong test on the AUC of the models. Although, in general, the performance metrics achieved by the DL models were slightly higher than radiomics-ML models, the differences were insignificant for all regions (with a 95% confidence interval).

TABLE 2. Results of the Intrareader and Interreader Variabilities as Evaluated for Each Body Region by the ICC Test (95% Confidence Intervals)

Region	Rater 1 Intrareader ICC	Rater 2 Intrareader ICC	Interreader ICC
Head and neck	0.92 (0.84–0.96)	0.78 (0.58–0.88)	0.79 (0.72–0.84)
Chest	0.80 (0.61–0.89)	0.78 (0.57–0.88)	0.59 (0.44–0.69)
Chest abdomen	0.79 (0.60–0.89)	0.59 (0.21–0.79)	0.65 (0.53–0.74)
Abdomen	0.73 (0.49–0.86)	0.76 (0.54–0.88)	0.60 (0.47–0.71)
Pelvis	0.80 (0.62–0.90)	0.81 (0.63–0.90)	0.58 (0.43–0.70)

TABLE 3. Performances of the DL and Radiomics-ML Models Developed Based on the Scores Given by 2 Physicians (Rater 1 and Rater 2 Models) and the Average Score (Avg-Rate Model) for Each Body Region Based on AUC, ACC, SEN, and SPE.

		Deep Learning				Radiomics			
Region	**Rater**	**AUC**	**ACC**	**SEN**	**SPE**	**AUC**	**ACC**	**SEN**	**SPE**
Head and neck	Rater 1	0.97	0.95	0.96	0.95	0.96	0.91	0.93	0.88
	Rater 2	0.71	0.70	0.71	0.67	0.78	0.70	0.81	0.53
	Avg-Rate	0.84	0.80	0.79	0.82	0.88	0.86	0.90	0.83
Chest	Rater 1	0.85	0.82	0.87	0.76	0.77	0.73	0.79	0.66
	Rater 2	0.70	0.69	0.75	0.57	0.69	0.66	0.76	0.47
	Avg-Rate	0.81	0.76	0.71	0.82	0.78	0.74	0.78	0.70
Chest-abdomen	Rater 1	0.83	0.76	0.68	0.80	0.75	0.73	0.62	0.79
	Rater 2	0.61	0.62	0.63	0.62	0.61	0.60	0.36	0.75
	Avg-Rate	0.60	0.58	0.56	0.58	0.56	0.71	0.29	0.87
Abdomen	Rater 1	0.70	0.67	0.66	0.67	0.70	0.68	0.57	0.75
	Rater 2	0.73	0.72	0.64	0.77	0.71	0.64	0.59	0.68
	Avg-Rate	0.73	0.74	0.44	0.84	0.69	0.73	0.40	0.84
Pelvis	Rater 1	0.63	0.64	0.65	0.64	0.65	0.61	0.61	0.61
	Rater 2	0.72	0.68	0.70	0.67	0.72	0.71	0.56	0.81
	Avg-Rate	0.73	0.75	0.54	0.83	0.69	0.71	0.44	0.82

Qualitative Analysis of the Models

Figure 5 shows the performance of the proposed ^{18}F-FDG PET IQA framework on 5 representative cases. Case A illustrates a patient with a mismatch artifact between the CT and PET scans in the head and neck region due to head motion. This discrepancy can result in erroneous ASC and ultimately inaccurate quantification (red arrows point at regions being overshot in the PET image). Case B shows the chest region of a patient with high levels of noise in the PET scan, making interpretation and clinical diagnosis challenging for the physicians. Case C shows the chest-abdomen interval of a patient with severe respiratory motion artifacts appearing in the diaphragmatic region (lungs and liver), making SUV quantification and assignment of lesions to the correct organ challenging for the physicians. Case D refers to a patient with severe truncation artifact in the abdomen region due to the extremities (arms) being out of the field in the CT scan, leading to inaccurate ASC and corruption of the PET image, preventing the physician from confidence in reporting. Finally, in case E, a patient with severe metallic artifact in the CT image at the pelvis region is shown, which leads to wrong quantification of the regions in the vicinity of the metallic objects due to errors in ASC. The proposed IQA model successfully identified the mentioned cases as low-quality images.

Figure 6 shows 2 other patients with low quality images caused by misadministration of ^{18}F-FDG and external contamination in the scanner's field of view (FOV). Case A shows a patient with ^{18}F-FDG extravasation, causing low and/or disrupted uptake and clearance in different tissues leading to a noisy and low-quality image. Both physicians reported low quality and diagnostic confidence for all body regions. Our model successfully identified the image as being of low quality in the head and neck, abdomen, and pelvis regions. In the chest and chest-abdomen regions, the Avg-Rate model (model based on average of the scores given by physicians) classified the images incorrectly. However, the model trained on the scores given by rater 1 and rater 2 correctly classified chest and chest-abdomen interval regions, respectively. Case B involves an image with contamination (activity outside the patient's body) within the FOV, resulting in disrupted image reconstruction and artifacts in the chest region, thus preventing the physicians from reaching a diagnosis. The proposed IQA model accurately classified the chest region of this scan as being of low quality.

DISCUSSION

In this study, we developed a fully automated framework to conduct a region-wise IQA on ^{18}F-FDG PET images. In a clinical setting, it can serve as a valuable tool to reduce the cumbersome workload of QC on the scans and accelerate the diagnostic and monitoring processes. Additionally, patients would benefit from immediate recognition of the need for a second scan in case the initial one was inconclusive. Another significant application for such a model would be in data management procedures. The day-by-day increasing adoption of AI-driven models in medical imaging necessitates platforms to automate and accelerate the data collection procedure. The presented framework can assist researchers in selecting high-quality and artifact-free ^{18}F-FDG PET images, based on the concerned region of interest. Various body regions in the same PET image can reflect different image quality due to differences in underlying tissue compositions, motion patterns, and susceptibility to artifacts.[40] The proposed ^{18}F-FDG PET IQA framework was trained separately on each body region to fit specific training procedures following extraction of content-specific features from the regions and enable region-wise usage for big-data analysis purposes.

Models developed for the head and neck region showed excellent performance. A common artifact in the PET scan of this region is the mismatch between CT and PET caused by head motion, which is frequent in pediatrics, elderlies, and patients suffering from dementia.[14] Case A in Figure 5 shows a case with severe head motion artifact. This artifact can lead to a corrupted ASC and cause underestimation or overestimation of SUVs, misleading physicians in their interpretation.[14,40,55] For the chest region as well, the proposed models showed desirable performance. High levels of noise in ^{18}F-FDG PET images can occur due to physical (inaccuracy/miscalibration of the hardware or errors in reconstruction propagating noise) or physiological reasons (low injected activity or extravasation or low uptake due to high glucose levels). High levels of noise in PET images

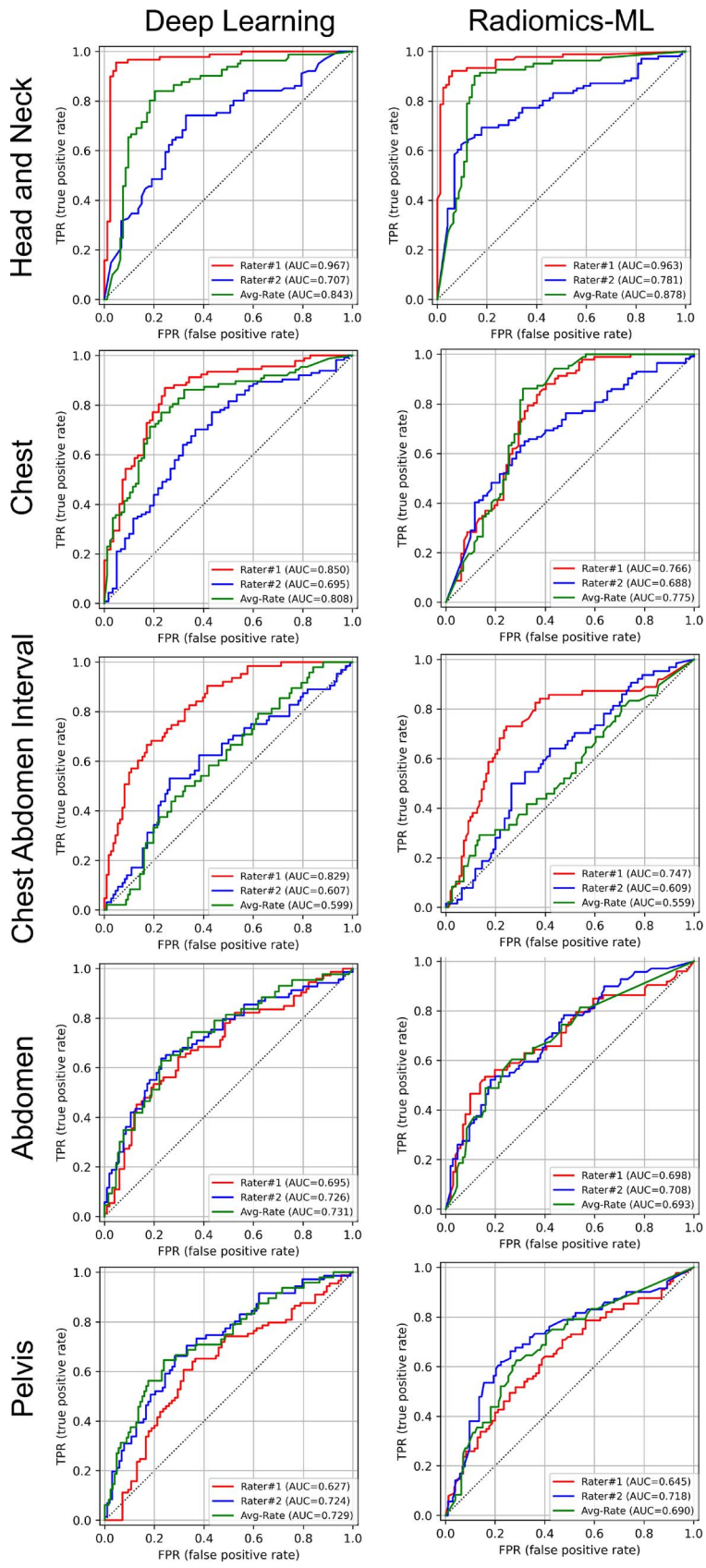

FIGURE 3. The ROC curves of the DL and radiomics-ML models developed for each body region.

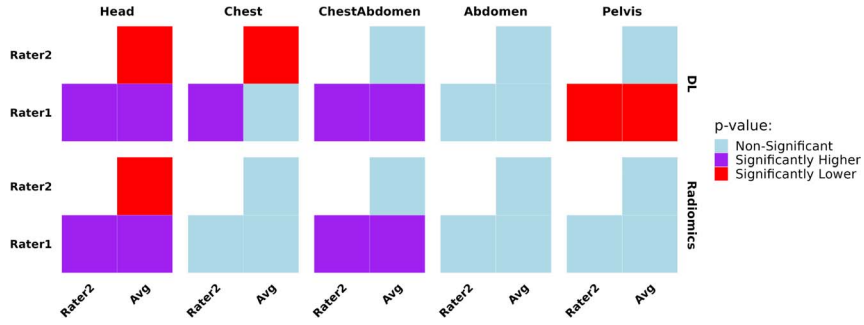

FIGURE 4. Comparison of the performance of models (models annotated on rows are compared with models annotated on the columns) when developed based on scores from rater 1 (experienced physician) versus rater 2 (physician with less experience) and average scores (Avg-Rate models). The comparison was performed with the Delong test on the AUC metric, and *P* values <0.05 were considered significant.

reduce the diagnostic confidence and lower the accuracy of the interpretation.[56] Case B in Figure 5 shows a noisy PET scan in the chest region, with low diagnostic value.

The chest-abdomen interval was purposely separated from the chest and abdomen regions due to the high susceptibility of the diaphragmatic region to respiratory motion artifacts (Fig. 2 shows the relative lower quality of images in this region compared with other regions). Uncontrolled respiratory motion potentially causes misalignment between organs in the diaphragm interval (lungs, liver, and spleen), which can be spanned up to 4.5 cm, and results in a curvilinear cold artifact in the upper lobe of the liver (Fig. 5, Case C).[14,57] This common artifact introduces challenges in attributing lesions to the correct organ and can lead to misinterpretation or misdiagnosis.[14,57] The proposed IQA model for the chest abdomen interval region achieved significantly higher performances in rater 1 models, compared with rater 2, and Avg-Rate models. It is worth mentioning that the intrareader reproducibility in scoring of this region was excellent for rater 1, and only fair for rater 2. The interreader reproducibility in chest-abdomen interval achieved an ICC of 0.65, showing nonnegligible disagreement between quality scores attributed to this region by the raters. This might explain the nonconsistent performance of the models between the raters for this region.

The proposed IQA model showed acceptable performance for the abdomen region. This region can suffer from noise following lower photon counts often aggravated by the larger sizes, plus artifacts such as truncation,[14,58] leading to average lower quality compared with other regions (Fig. 2). The truncation artifact (Fig. 5, Case D) occurs when a portion of the patient's body (mostly arms in obese patients or patients in arms-down position on the bed) is present in the PET FOV but not within the CT FOV.[59] This would result in overshooting around the CT image rim and undershooting in regions outside of the CT FOV.[60] Moreover, for the pelvis region, the proposed model revealed acceptable performance. Metallic artifacts can occur in CT images due to the presence of metal objects in various parts of the body. However, specifically in elderly patients, the pelvis region is more susceptible to metallic artifacts because of the common metallic prosthesis for hip arthroplasties (Fig. 5, Case E). Metallic objects do not directly affect the PET signals but can interfere with PET quantification because of artifacts in CT images used for ASC.[61] This can overshoot the SUVs in adjacent regions up to 20%, potentially leading to misinterpretation as pathology.[62]

In Figure 5, we showed 5 ^{18}F-FDG PET cases in our cohort (along with their CT and fused images), which were tagged as low-quality by both physicians due to the common reasons (noise and artifact) mentioned in previous paragraphs. The proposed ^{18}F-FDG PET models correctly classified all these images from different body regions, which shows their ability in identifying various artifacts and quantification anomalies in ^{18}F-FDG PET images.

Erroneous delivery of ^{18}F-FDG to the patient's body, referred to as ^{18}F-FDG infiltration or extravasation, is one of the factors affecting PET image quality. In a standard ^{18}F-FDG administration, the prescribed dose used for SUV calculations completely enters the systemic circulation. However, extravasation leads to a mismatch between the injected and actual circulating activity, as some of the radiotracer remains near the injection site and reenters the circulation via lymphatic reabsorption at an unpredictable rate. This alters the radiotracer's supply and clearance in tissues, affecting the resulting SUV. Infiltration can degrade image quality mainly due to inadequate tracer delivery and the ongoing release of the radiotracer from the injection site, without sufficient time for clearance.[63] The literature suggests a 15% occurrence for the misadministration of the radiotracer, which by no means is uncommon.[63–65] Our proposed IQA model applies a body contour and excludes the extremities from the input images, and as such, it does not directly see and detect this effect. However, if extravasation leads to reduced radiopharmaceutical uptake in the organs and subsequently results in a low-quality image, it would be classified accordingly by the proposed model. Case A in Figure 6 is an example of the abovementioned scenario. Another cause of quality degradation is external activity contamination in the scanner's FOV. Case B in Figure 6 illustrates an image with external contamination, which disrupted image reconstruction and prevented physicians from achieving diagnosis in the chest region. Our proposed IQA model successfully classified the chest image as low-quality.

TABLE 4. The Delong Test *P* Value for Comparison Between Deep Learning and Radiomics-Machine Learning Models.

	Radiomics-ML vs DL (*P*)		
	Rater 1	**Rater 2**	**Rate_Avg**
Head and neck	0.820	0.161	0.398
Chest	0.077	0.915	0.497
Chest-abdomen interval	0.113	0.982	0.562
Abdomen	0.958	0.747	0.559
Pelvis	0.772	0.913	0.533

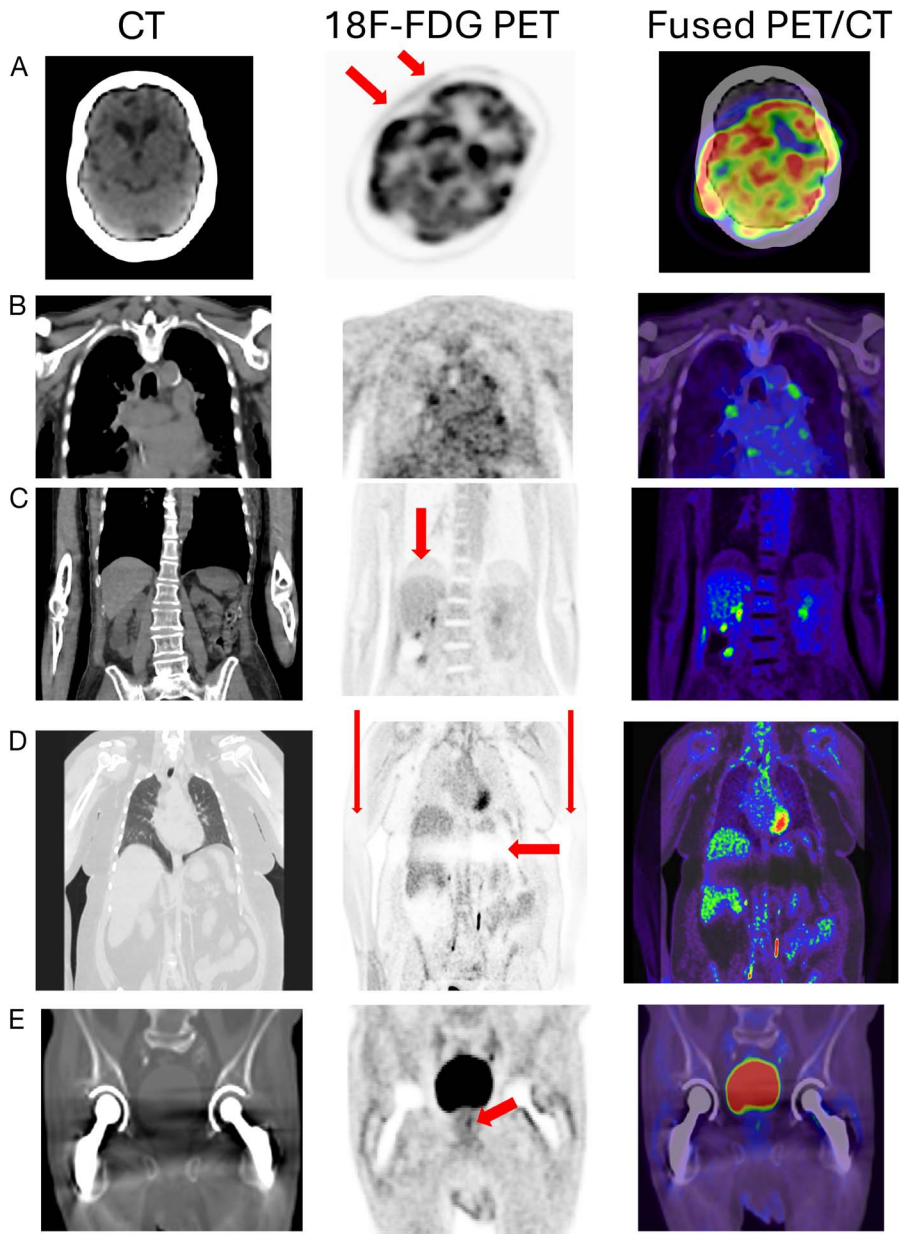

FIGURE 5. Five low-quality ^{18}F-FDG PET images suffering from various artifacts and noise shown along their CT and fused PET/CT images. **A**, Image illustrates a patient with a mismatch artifact between the CT and PET scans in the head and neck region due to head motion resulting in erroneous ASC and ultimately inaccurate quantification (red arrows point at regions being overshot in the PET image). **B**, Image shows the chest region of a patient with high levels of noise in the PET scan, making the diagnosis challenging for the physicians. **C**, Image shows the chest-abdomen interval of a patient with severe respiratory motion artifact in diaphragmatic organs (lungs, liver), resulting in a curvilinear cold artifact in the upper lobe of the liver, indicated by the red arrow. **D**, Image refers to a patient with severe truncation artifact in the abdomen region due to the arms being out of the field in the CT scan (the vertical red arrows depict the arms present in the PET image and cropped from the CT FOV), leading to inaccurate ASC and corruption of the PET image (the horizontal arrow shows the resulting artifact in the PET image), preventing the physician from any diagnosis. **E**, A patient with severe metal artifact in CT at the pelvis region is shown, which leads to error-prone quantification of the regions in the vicinity of the metallic objects due to errors in ASC, consequently uncertainty in lesion classification (the red arrow shows the affected lesion). The proposed ^{18}F-FDG PET IQA model successfully identified the shown cases as low-quality images.

The subjective quality assessments of the images were performed separately by 2 physicians. Models were developed separately based on rater 1 (physician with higher experience), rater 2 (physician with lower experience), and on the average of the scores. The comparison of the models within each region showed statistically significant differences in the performances (Fig. 4). In general, our results showed a good alignment between the performance of the models and the intrareader reproducibility of the scores which

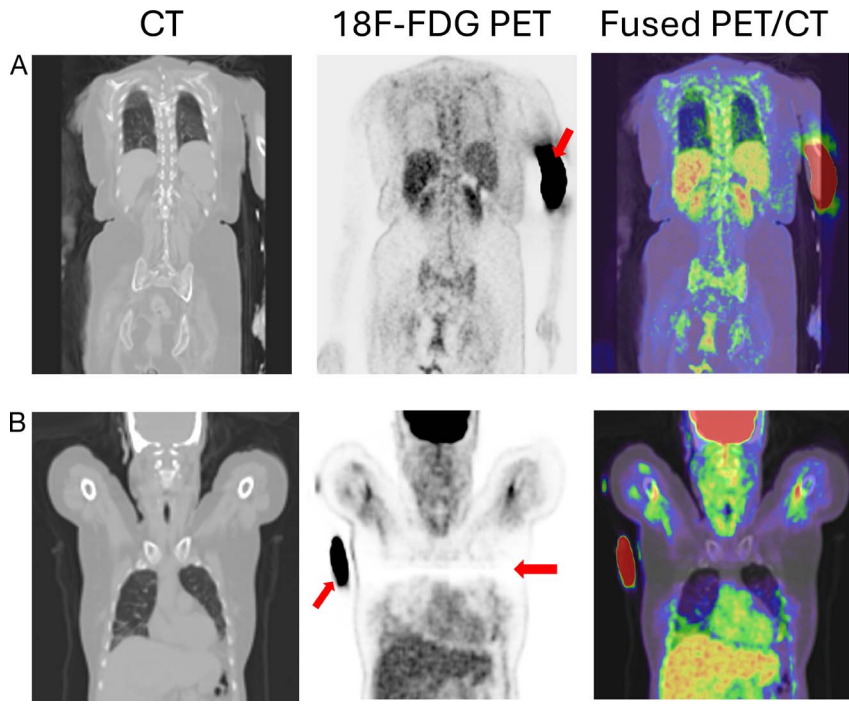

FIGURE 6. Two low-quality ^{18}F-FDG PET images suffering from misadministration of the radiopharmaceutical, and external activity contamination in scanner's FOV. **A**, A patient with ^{18}F-FDG extravasation (red arrow is pointing at the injection site), causing low and/or disrupted uptake and clearance in different tissues leading to a noisy and low-quality image. Both physicians reported low quality and diagnostic confidence for all the body regions. **B**, An image with contamination (activity outside the patient's body, shown by the oblique red arrow) within the FOV, resulting in disrupted image reconstruction and artifacts in the chest region (shown by the horizontal red arrow), which prevented the physicians from making a diagnosis. The proposed IQA model correctly classified both cases as low quality.

the model was developed on. In other words, the performance of the models decreased whenever there was inconsistency in scoring by the physician (low intrareader ICC) (Tables 2 and 3, and Fig. 4). In the head and neck, chest, and chest-abdomen regions, the models trained on the scores given by the physician with higher experience (rater 1 models) revealed superior performance. In these regions, rater 1 also showed better intraobserver agreement in scoring (Table 2). This underscores the importance of using scores from multiple physicians in similar IQA models for better alignment of the model with common human perception and to improve the generalizability.[66] Additionally, conducting educational group sessions with expert readers is strongly recommended to achieve consensus in quality assessment.

In this study, we approached the task with 2 well-explored AI-driven frameworks in the field, namely, radiomics-ML and DL frameworks. For each region, models were developed with both frameworks, and in none they were statistically different (Table 4). Although not statistically significant, the performance of the DL models generally exhibited a slight superiority. Moreover, the DL models are superior in the sense that they are not dependent on a predefined VOI. We used automatic models to delineate the whole-body VOI and exclude the extremities using CT images. Nevertheless, it makes radiomics models to be dependent on CT images for VOI extraction and may potentially propagate any error in the segmentation process to the overall performance of the IQA models.

Automatic and human-perceptive-equivalent quality assessment of ^{18}F-FDG images using AI-based algorithms has been previously studied by various research teams.[28,37–39] However, their models lack comprehensiveness as they extract information from MIP images,[37] or defined axial slices,[28,38] instead of the 3D images. To the best of our knowledge, this is the first study reporting on fully automatic and human-perceptive ^{18}F-FDG PET QA individually for each body region.

Bradshaw et al[66] provided recommendations and guidelines on best practices for the development of trustworthy AI-driven models in nuclear medicine. They classify data annotation quality into 4 levels. Labeling by multiple experts has the second highest quality, just below ground truth labels[66] which is not available for IQA studies. In the absence of ground truth labels, they recommend preparation of the experts (physicians) with clear and comprehensive instructions for annotating and performing the labeling in a consensual manner (through discussions between the physicians).[66] Although in this study, the physicians were well-instructed for quality scoring, the annotation procedure was not performed in consensus, which is one of the limitations of this study. Although the present study focused on supervised models, given the abundance of unlabeled data in clinical PACS systems, exploring unsupervised or semisupervised approaches for IQA would be of great interest as annotating the images is the most time-consuming and main bottleneck of similar studies. Our analysis did not include any PET images acquired with low activity injected protocols. This was by design, as our model was specifically intended to identify scans with low image quality, rather than to assess the effects of low-dose protocols. However, we believe that if a low-dose scan results in a noisy or low-quality image, our model would still be capable of identifying it as such. Another limitation is the limited size of the dataset used. Although images were from 2 PET scanners, a large dataset from multiple imaging centers, acquired on various PET scanners using different acquisition and reconstruction protocols, may significantly increase models' generalizability and reproducibility.[8,66]

CONCLUSIONS

We proposed a fully automated human-perceptive model to conduct region-wise IQA over ^{18}F-FDG PET images. Implementation of the presented framework in clinical settings would alleviate the workload and expedite the interpretation process by delivering high-quality images to physicians. In addition, this can be a powerful tool to select high-quality and artifact-free images toward generating curated ^{18}F-FDG PET datasets for AI-driven research.

REFERENCES

1. Schwenck J, Sonanini D, Cotton JM, et al. Advances in PET imaging of cancer. *Nat Rev Cancer*. 2023;23:474–490.
2. Pinker K, Riedl C, Weber WA. Evaluating tumor response with FDG PET: updates on PERCIST, comparison with EORTC criteria and clues to future developments. *Eur J Nucl Med Mol Imaging*. 2017;44:55–66.
3. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J Nucl Med*. 2009;50:11S–20S.
4. Zaidi H, Karakatsanis N. Towards enhanced PET quantification in clinical oncology. *Br J Radiol*. 2018;91:20170508.
5. Delbeke D, Coleman RE, Guiberteau MJ, et al. Procedure guideline for tumor imaging with ^{18}F-FDG PET/CT 1.0. *J Nucl Med*. 2006;47:885–895.
6. Cook GJ, Wegner EA, Fogelman I. Pitfalls and artifacts in ^{18}FDG PET and PET/CT oncologic imaging. *Semin Nucl Med*. 2004;34:122–133.
7. Thie JA. Understanding the standardized uptake value, its methods, and implications for usage. *J Nucl Med*. 2004;45:1431–1434.
8. Saboury B, Bradshaw T, Boellaard R, et al. Artificial intelligence in nuclear medicine: opportunities, challenges, and responsibilities toward a trustworthy ecosystem. *J Nucl Med*. 2023;64:188–196.
9. Arabi H, AkhavanAllaf A, Sanaat A, et al. The promise of artificial intelligence and deep learning in PET and SPECT imaging. *Phys Med*. 2021;83:122–137.
10. Seifert R, Weber M, Kocakavuk E, et al. Artificial intelligence and machine learning in nuclear medicine: future perspectives. *Semin Nucl Med*. 2021;51:170–177.
11. Sadaghiani MS, Rowe SP, Sheikhbahaei S. Applications of artificial intelligence in oncologic (18)F-FDG PET/CT imaging: a systematic review. *Ann Transl Med*. 2021;9:823.
12. Zaidi H, El Naqa I. Quantitative molecular positron emission tomography imaging using advanced deep learning techniques. *Annu Rev Biomed Eng*. 2021;23:249–276.
13. Sanaat A, Amini M, Arabi H, et al. The quest for multifunctional and dedicated PET instrumentation with irregular geometries. *Ann Nucl Med*. 2024;38:31–70.
14. Shiri I, Salimi Y, Hervier E, et al. Artificial intelligence–driven single-shot PET image artifact detection and disentanglement: toward routine clinical image quality assurance. *Clin Nucl Med*. 2023;48:1035–1046.
15. Shiri I, Amini M, Yousefirizi F, et al. Information fusion for fully automated segmentation of head and neck tumors from PET and CT images. *Med Phys*. 2024;51:319–333.
16. Amini M, Pursamimi M, Hajianfar G, et al. Machine learning–based diagnosis and risk classification of coronary artery disease using myocardial perfusion imaging SPECT: a radiomics study. *Sci Rep*. 2023;13:14920.
17. Hajianfar G, Sabouri M, Salimi Y, et al. Artificial intelligence–based analysis of whole-body bone scintigraphy: the quest for the optimal deep learning algorithm and comparison with human observer performance. *Z Med Phys*. 2024;34:242–257.
18. Salimi Y, Hajianfar G, Mansouri Z, et al. Organomics: a concept reflecting the importance of PET/CT healthy organ radiomics in non-small cell lung cancer prognosis prediction using machine learning. *Clin Nucl Med*. 2024;49:899–908.
19. Khodabakhshi Z, Amini M, Hajianfar G, et al. Dual-centre harmonised multimodal positron emission tomography/computed tomography image radiomic features and machine learning algorithms for non-small cell lung cancer histopathological subtype phenotype decoding. *Clin Oncol*. 2023;35:713–725.
20. Mohebi M, Amini M, Alemzadeh-Ansari MJ, et al. Post-revascularization ejection fraction prediction for patients undergoing percutaneous coronary intervention based on myocardial perfusion SPECT imaging radiomics: a preliminary machine learning study. *J Digit Imaging*. 2023;36:1348–1363.
21. Mansouri Z, Salimi Y, Amini M, et al. Development and validation of survival prognostic models for head and neck cancer patients using machine learning and dosiomics and CT radiomics features: a multicentric study. *Radiat Oncol*. 2024;19:12.
22. Amini M, Salimi Y, Sabouri M, et al. Auto-PET-IQA: a fully automated region-specific PET image quality assessment tool. *IEEE Xplore*. 2023;1–1.
23. Oakden-Rayner L. Exploring large-scale public medical image datasets. *Acad Radiol*. 2020;27:106–112.
24. Amini M, Hajianfar G, Salimi Y, et al. Automatic human perception-driven PET image quality and diagnostic confidence assessment using radiomics and machine learning algorithms. *Soc Nuclear Med*. 2023.
25. Barrett HH, Myers KJ, Hoeschen C, et al. Task-based measures of image quality and their relation to radiation dose and patient risk. *Phys Med Biol*. 2015;60:R1–R75.
26. Ma JJ, Nakarmi U, Kin CYS, et al. Diagnostic image quality assessment and classification in medical imaging: opportunities and challenges. *Proc IEEE Int Symp Biomed Imaging*. 2020;2020:337–340.
27. Piccini D, Demesmaeker R, Heerfordt J. Deep learning to automate reference-free image quality assessment of whole-heart MR images. *Radiol Artif Intell*. 2020;2:e190123.
28. Qi C, Wang S, Yu H, et al. An artificial intelligence–driven image quality assessment system for whole-body [^{18}F] FDG PET/CT. *Eur J Nucl Med Mol Imaging*. 2023;50:1318–1328.
29. Boktor RR, Walker G, Stacey R, et al. Reference range for intrapatient variability in blood-pool and liver SUV for ^{18}F-FDG PET. *J Nucl Med*. 2013;54:677–682.
30. Wahl RL, Jacene H, Kasamon Y, et al. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med*. 2009;50:122S–150S.
31. Chow LS, Paramesran R. Review of medical image quality assessment. *Biomed Signal Process Control*. 2016;27:145–154.
32. Chen X, Wang X, Zhang K, et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med Image Anal*. 2022;79:102444.
33. Amini M, Nazari M, Shiri I, et al. Multi-level PET and CT fusion radiomics-based survival analysis of NSCLC patients. *IEEE*. 2020;1–4.
34. Czajkowska J, Juszczyk J, Piejko L, et al. High-frequency ultrasound dataset for deep learning–based image quality assessment. *Sensors*. 2022;22:1478.
35. Kretz T, Müller KR, Schaeffter T, et al. Mammography image quality assurance using deep learning. *IEEE Trans Biomed Eng*. 2020;67:3317–3326.
36. Lauermann J, Treder M, Alnawaiseh M, et al. Automated OCT angiography image quality assessment using a deep learning algorithm. *Graefes Arch Clin Exp Ophthalmol*. 2019;257:1641–1648.
37. Schwyzer M, Skawran S, Gennari AG, et al. Automated F18-FDG PET/CT image quality assessment using deep neural networks on a latest 6-ring digital detector system. *Sci Rep*. 2023;13:11332.
38. Reynés-Llompart G, Sabaté-Llobera A, Llinares-Tello E, et al. Image quality evaluation in a modern PET system: impact of new reconstructions methods and a radiomics approach. *Sci Rep*. 2019;9:10640.
39. Zhang H, Liu Y, Wang Y, et al. Deep learning model for automatic image quality assessment in PET. *BMC Med Imaging*. 2023;23:1–13.
40. Shiri I, Salimi Y, Maghsudi M, et al. Differential privacy preserved federated transfer learning for multi-institutional (68)Ga-PET image artefact detection and disentanglement. *Eur J Nucl Med Mol Imaging*. 2023;51:40–53.
41. Rosset A, Spadola L, Ratib O. OsiriX: an open-source software for navigating in multidimensional DICOM images. *J Digit Imaging*. 2004;17:205–216.
42. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428.
43. Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*. 2002;58:1020–1027.
44. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6:284–290.
45. Wasserthal J, Breit HC, Meyer MT, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol Artif Intell*. 2023;5:e230024.
46. Schuhegger S. Body part regression for CT images. *arXiv preprint arXiv: 211009148*. 2021.
47. Ashrafinia S. *Quantitative Nuclear Medicine Imaging Using Advanced Image Reconstruction and Radiomics*. Baltimore, MD: The Johns Hopkins University; 2019.

48. Salmanpour MR, Shiri I, Hosseinzadeh M, et al. ViSERA: visualized & standardized environment for radiomics analysis—a shareable, executable, and reproducible workflow generator. *IEEE*. 2023;1–2.
49. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295:328–338.
50. Whybra P, Zwanenburg A, Andrearczyk V, et al. The image biomarker standardization initiative: standardized convolutional filters for reproducible Radiomics and enhanced clinical insights. *Radiology*. 2024;310:e231319.
51. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. 2010;36:1–13.
52. Peterson LE. K-nearest neighbor. *Scholarpedia*. 2009;4:1883.
53. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. *Proc IEEE Conf Comput Vision Pattern Recognit*. 2017;4700–4708.
54. Cardoso MJ, Li W, Brown R, et al. Monai: an open-source framework for deep learning in healthcare. *arXiv preprint arXiv:221102701*. 2022.
55. Yang J, Sohn JH, Behr SC, et al. CT-less direct correction of attenuation and scatter in the image space using deep learning for whole-body FDG PET: potential benefits and pitfalls. *Radiol Artif Intell*. 2020;3:e200137.
56. Mehranian A, Wollenweber SD, Walker MD, et al. Image enhancement of whole-body oncology [^{18}F]-FDG PET scans using deep neural networks to reduce noise. *Eur J Nucl Med Mol Imaging*. 2022;49:539–549.
57. Shiri I, Arabi H, Geramifar P, et al. Deep-JASC: joint attenuation and scatter correction in whole-body ^{18}F-FDG PET using a deep residual network. *Eur J Nucl Med Mol Imaging*. 2020;47:2533–2548.
58. Shiri I, Sanaat A, Salimi Y, et al. PET-QA-NET: towards routine PET image artifact detection and correction using deep convolutional neural networks. *IEEE*. 2021;1–3.
59. Mawlawi O, Erasmus JJ, Pan T, et al. Truncation artifact on PET/CT: impact on measurements of activity concentration and assessment of a correction algorithm. *Am J Roentgenol*. 2006;186:1458–1467.
60. Beyer T, Bockisch A, Kühl H, et al. Whole-body ^{18}F-FDG PET/CT in the presence of truncation artifacts. *J Nucl Med*. 2006;47:91–99.
61. Abdoli M, Dierckx RA, Zaidi H. Metal artifact reduction strategies for improved attenuation correction in hybrid PET/CT imaging. *Med Phys*. 2012; 39:3343–3360.
62. van der Vos CS, Arens AI, Hamill JJ, et al. Metal artifact reduction of CT scans to improve PET/CT. *J Nucl Med*. 2017;58:1867–1872.
63. Kiser JW, Crowley JR, Wyatt DA, et al. Impact of an ^{18}F-FDG PET/CT radiotracer injection infiltration on patient management—a case report. *Front Med*. 2018;5:143.
64. Osman MM, Muzaffar R, Altinyay ME, et al. FDG dose extravasations in PET/CT: frequency and impact on SUV measurements. *Front Oncol*. 2011;1:41.
65. Hall N, Zhang J, Reid R, et al. Impact of FDG extravasation extravasation on SUV measurements in clinical PET/CT. Should we routinely scan the injection site? *J Nucl Med*. 2006;47:115.
66. Bradshaw TJ, Boellaard R, Dutta J, et al. Nuclear medicine and artificial intelligence: best practices for algorithm development. *J Nucl Med*. 2022;63: 500–510.