European Journal of Radiology 165 (2023) 110924

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Deep learning for assessing image quality in bi-parametric prostate MRI: A feasibility study

Deniz Alis [a,*], Mustafa Said Kartal [b], Mustafa Ege Seker [c], Batuhan Guroz [a], Yeliz Basar [d], Aydan Arslan [e], Sabri Sirolu [f], Serpil Kurtcan [d], Nurper Denizoglu [d], Umit Tuzun [g], Duzgun Yildirim [h], Ilkay Oksuz [i], Ercan Karaarslan [b]

[a] Acibadem Mehmet Ali Aydinlar University, School of Medicine, Department of Radiology, Istanbul, 34457, Turkey
[b] Cumhuriyet University, School of Medicine, Sivas, 581407, Turkey
[c] Acibadem Mehmet Ali Aydinlar University, School of Medicine, Istanbul, 34752, Turkey
[d] Acibadem Healthcare Group, Department of Radiology, Istanbul, 34457, Turkey
[e] Umraniye Training and Research Hospital, Department of Radiology, Istanbul, 34764, Turkey
[f] Istanbul Sisli Hamidiye Etfal Training and Research Hospital, Department of Radiology, Istanbul, 34396, Turkey
[g] Neolife, Radiology Center, Istanbul, 34340, Turkey
[h] Acibadem Mehmet Ali Aydinlar University, School of Vocational Sciences, Department of Radiology, Istanbul, 34457, Turkey
[i] Istanbul Technical University, Department of Computer Engineering, Istanbul, 34467, Turkey

ARTICLE INFO

Keywords:
Deep Learning
Prostate
MRI
Image quality

ABSTRACT

Background: Although systems such as Prostate Imaging Quality (PI-QUAL) have been proposed for quality assessment, visual evaluations by human readers remain somewhat inconsistent, particularly among less-experienced readers.
Objectives: To assess the feasibility of deep learning (DL) for the automated assessment of image quality in bi-parametric MRI scans and compare its performance to that of less-experienced readers.
Methods: We used bi-parametric prostate MRI scans from the PI-CAI dataset in this study. A 3-point Likert scale, consisting of poor, moderate, and excellent, was utilized for assessing image quality. Three expert readers established the ground-truth labels for the development (500) and testing sets (100). We trained a 3D DL model on the development set using probabilistic prostate masks and an ordinal loss function. Four less-experienced readers scored the testing set for performance comparison.
Results: The kappa scores between the DL model and the expert consensus for T2W images and ADC maps were 0.42 and 0.61, representing moderate and good levels of agreement. The kappa scores between the less-experienced readers and the expert consensus for T2W images and ADC maps ranged from 0.39 to 0.56 (fair to moderate) and from 0.39 to 0.62 (fair to good).
Conclusions: Deep learning (DL) can offer performance comparable to that of less-experienced readers when assessing image quality in bi-parametric prostate MRI, making it a viable option for an automated quality assessment tool. We suggest that DL models trained on more representative datasets, annotated by a larger group of experts, could yield reliable image quality assessment and potentially substitute or assist visual evaluations by human readers.

1. Introduction

Prostate MRI has emerged as the standard of care for biopsy-naive men with suspected prostate cancer (PCa), given recent evidence of its benefits in reducing unnecessary biopsies and increasing the detection rate of clinically significant prostate cancer (csPCa) [1]. The Prostate

Abbreviations: ADC, Apparent diffusion coefficient; csPCa, Clinically Significant Prostate cancer; DL, Deep Learning; mpMRI, Multi-parametric MRI; PI-CAI, Prostate Imaging: Cancer AI; PI-QUAL, Prostate Imaging Quality.

* Corresponding author at: Acibadem Mehmet Ali Aydinlar University, School of Medicine, Department of Radiology, Istanbul, Turkey.
E-mail addresses: drdenizalis@gmail.com (D. Alis), serpil.kurtcan@acibadem.com (S. Kurtcan), nurper.denizoglu@acibadem.com (N. Denizoglu), umit.tuzun@neolife.com.tr (U. Tuzun), duzgun.yildirim@acibadem.com (D. Yildirim), ercan.karaarslan@acibadem.edu.tr (E. Karaarslan).

https://doi.org/10.1016/j.ejrad.2023.110924
Received 2 April 2023; Received in revised form 15 May 2023; Accepted 9 June 2023
Available online 11 June 2023

Imaging-Reporting and Data System (PI-RADS) and its subsequent versions outline best practices for acquiring and interpreting prostate MRI scans [2]. PI-RADS stipulates minimum technical requirements for prostate MRI acquisition, as scan quality is crucial for the accurate detection of csPCa. However, compliance with PI-RADS does not always ensure a high-quality MRI scan [3–5].

Recently, the Prostate Imaging Quality (PI-QUAL) score has been introduced as an initial step towards standardizing image quality in prostate MRI [6]. The PI-QUAL score combines an objective assessment (e.g., field-of-view, slice thickness) with a visual assessment based on human readers' perceptions [6]. Studies by the same group have demonstrated high reproducibility of the PI-QUAL score among experienced radiologists [7].

Subsequent research by a separate group revealed that PI-QUAL had a moderate inter-observer agreement, emphasizing the need for automated methods [8]. In fact, the authors of the PI-QUAL suggested that deep learning (DL) techniques could automatically assess image quality, potentially replacing the visual quality assessment [9].

Despite the impressive performance of DL in assessing prostate cancer [10–12], only a few studies have employed DL in the context of quality assessment in prostate MRI [13,14], which could serve as an alternative or complement to visual assessment, promoting the generalizability and widespread adoption of image quality assessment methods in prostate MRI.

In this study, we developed a DL model with ordinal loss to assess the quality of multi-center bi-parametric MRI scans, using expert radiologists' consensus as the ground truth and compared the model's performance to that of less-experienced readers. We hypothesize that our DL model can perform similarly to less-experienced readers in estimating image quality when using expert consensus as the ground truth.

2. Methods

2.1. Study sample

This study used the publicly available Prostate Imaging: Cancer AI (PI-CAI) training data. The PI-CAI public training dataset comprises 1500 bi-parametric prostate MRI scans from 1476 men collected at four tertiary academic centers in the Netherlands and Norway between March 2015 and January 2018. All men in the PI-CAI dataset had undergone whole-mount pathology or biopsy following the MRI scan, or they were MRI-negative (i.e., PI-RADS score of 1 or 2) with a minimum follow-up of 36 months without any clinical, laboratory, or imaging evidence of csPCa. Further details can be found in [15].

DL requires extensive data for optimal performance, and the efficacy of a DL model often increases with more comprehensive training data [16]. However, annotating medical images demands significant time from medical professionals and considerable computational power. Therefore, to efficiently utilize our limited computational and human resources, we randomly sampled 700 out of the 1500 scans. The scans were divided into three groups: a development set (500 scans), a testing set (100 scans), and a radiologist-training set (100 scans), ensuring that the same patient scans were not included in both development and testing sets.

2.2. Bi-Parametric MRI

All bi-parametric MRI scans were performed at 1.5 T Siemens (Aera, Avanto, Siemens Healthcare, Erlangen/Germany), 1.5 T Philips (Achieva, Intera, Philips Healthcare, Eindhoven/The Netherlands), 3 T Siemens (Skyra, TrioTim, Prisma, Siemens Healthcare, Erlangen/Germany), or 3 T Philips (Ingenia, Philips Healthcare, Eindhoven/The Netherlands) units with surface coils following PI-RADS V2. The bi-parametric MRI protocol consisted of tri-planar T2-weighted images, apparent diffusion coefficient (ADC) maps and high b-value DWI images ($b \geq 1400$ s/mm2) calculated from acquired DWI images. The organizers of the challenge did not share dynamic-contrast-enhanced images. Further details regarding the MRI protocols of the study sample can be found in [15].

2.3. Ground truthing by expert readers

Three expert readers (E.K., D.Y., U.T.) established the ground-truth labels in the present work on a dedicated workstation equipped with a dedicated browser-based platform (https://matrix.md.ai) with a 6-megapixel diagnostic color monitor (Radiforce RX 660, EIZO). All reviewed images were in Digital Imaging and Communications in Medicine (DICOM) format. The term expert reader was defined following the consensus statement of the European Society of Urogenital Radiology (ESUR) [17]. All expert readers had over ≥10 years of prostate imaging experience and were reading ≥300 scans/year.

The evaluation by the experts was mainly inspired by the visual assessment criteria proposed in the PI-QUAL [6]. Tri-planar T2W images and ADC maps were evaluated separately. The experts ranked the quality of T2-weighted images based on the following anatomical structures: prostate capsule, seminal vesicles, ejaculatory duct, neurovascular bundles, and sphincter muscle. Additionally, the experts assessed the presence or absence of artifacts. The visual assessment of ADC maps encompassed the presence of adequate ADC maps and the presence or absence of artifacts (e.g., movement or rectal air).

The ground truthing procedure primarily consisted of two sequential steps: (i) The joint reading session, in which expert radiologists familiarized themselves with each other's quality perceptions, shared their experiences, and eventually established a 3-point Likert Scale; (ii) The independent reading session, during which radiologists scored the images individually.

In contrast to the PI-QUAL score, where the visual image assessment was binary (i.e., diagnostic or non-diagnostic), we implemented a 3-point Likert Scale formulated as follows: poor image quality represents non-diagnostic images; moderate image quality represents images with a certain degree of artifacts or lack of clear delineation of some anatomical structures, but still diagnostic; excellent quality images represent images where all of the anatomical structures can be clearly visible without any artifacts.

(i) Joint reading session

In the joint reading sessions, expert radiologists evaluated the 100 scans from the radiologist-training set allocated for their training and ground truthing for the quality assessment using the 3-point Likert Scale. We acknowledge that image reading is somewhat subjective, and radiologists' perceptions can vary significantly due to many factors [18]. Consequently, establishing completely objective ground-truth labels is virtually impossible. Therefore, during the joint session, we did not force expert radiologists to reach a consensus score for each case. The primary aim of this session was to familiarize expert radiologists with each other's experiences and image-reading behaviors. Furthermore, this session allowed radiologists to share their knowledge and image-reading approaches.

(ii) Independent reading session

Following the joint session, expert radiologists independently ranked the T2-weighted and ADC images in the study sample's development and testing sets. The ground-truth labels of the study sample were determined using majority voting. For instance, if two experts ranked a sequence as poor quality while the other ranked it as moderate or excellent quality, the final ground-truth score was considered poor quality. In cases where no image quality category received two votes (i.e., three experts scored poor, moderate, and excellent quality, respectively), the ground-truth label was designated as moderate quality.

2.4. DL models

All DL experiments were carried out using a DL library, TensorFlow (Tensorflow 2.1 Google LLC, Mountain View, CA), on a custom-built workstation equipped with a 12 GB graphical processing unit. The present work used 3D InceptionResNetV2 [19]. Two different DL models were trained for tri-planar T2W images and ADC maps.

The development set was used to train and validate the DL model using an 80%/20% train/validation split. The images were cropped to 192 mm × 192 mm × 81 mm. In this work, we used probabilistic prostate masks to focus the DL model on the prostate and its surroundings, similar to the readers. However, we avoided binary prostate masks to prevent completely depriving the model of global image features. The success of probabilistic prostate masks was demonstrated in a prior study in the context of clinically csPCa detection [20].

We utilized standard data augmentation techniques, including rotation, flipping, and random contrast adjustment with gamma correction. We did not use a small rotation angle, low-resolution, or stretch operation, as they significantly affect the image quality.

The batch size was set to 4. The scholastic gradient descent algorithm was utilized with a learning rate of 0.00005 to minimize the ordinal loss function. We selected the ordinal loss function over the default categorical cross-entropy loss as it considers the ordinal characteristics of the data. For instance, the punishment is the same for incorrect prediction of a poor-quality image as an excellent-quality image or moderate-quality image for the categorical cross-entropy loss, while the ordinal loss punishes the former more compared with the latter. The model underwent training for 30 epochs, with a training time of 6.5 h. Fig. 1 shows the representative image of the DL model used in this study.

2.5. Less-experienced readers

Four less-experienced readers (Y.B, A.A. S.S, N.D) from different centers gave the image-quality scores on the testing set using the same Likert Scale. All radiologists were reading ≤300 mpMRI scans a year with the following experience levels: Y.B. (less-experienced reader 1) 20 years of radiology and 2 years of mpMRI experience; A.A. (less-experienced reader 2) 8 years of radiology and 4 years of mpMRI experience; S. S. (less-experienced reader 3) 6 years of radiology and 2 years of mpMRI experience; and N.D. (less-experienced reader 4) 12 years of radiology and 6 years of mpMRI experience.

Similar to the ground truthing steps, the less-experienced readers first became acquainted with the scoring system and checked the ground-truth labels of the 100 scans from the radiologist-training set during an online consensus meeting. One of the expert readers also attended the online consensus meeting. After the joint session, the less-experienced radiologists independently ranked the T2W images and ADC maps in the testing set.

3. Statistical analyses

The statistical analyses were performed using the SciPy library of Python Version 3. The continuous variables are presented using median and interquartile ranges, and the categorical and ordinal variables are presented with frequencies and percentages. The inter-reader agreement between the DL model and the readers was assessed using weighted Cohen's kappa [21]. The kappa scores were interpreted as follows: a kappa score of <20, a poor agreement; 21–40, a fair agreement; 41–60, a moderate agreement; 61–80, a good agreement; and 81–100, an excellent agreement.

4. Results

A total of 100 bi-parametric MRI scans were included in the testing sample. For T2W images, the majority voting consensus of the expert readers categorized the scans as poor, moderate, and excellent quality in 29/100, 37/100, and 34/100 of the scans, respectively. For ADC maps, the majority voting consensus of the expert readers categorized the scans as poor, moderate, and excellent quality in 29/100, 45/100, and 26/100 of the scans, respectively. The kappa scores among the experts were 0.66 (95% CI, 0.54–0.78) and 0.68 (95% CI, 0.56–0.80) for T2W images and ADC images, representing good levels of agreement.

For T2W images, the DL model categorized the scans as poor, moderate, and excellent quality in 43/100, 17/100, and 40/100 of the scans, respectively. For ADC images, the DL model categorized the scans as poor, moderate, and excellent quality in 19/100, 59/100, and 22/100 of the scans, respectively. The kappa scores between the DL model and the expert-consensus scores for T2W images and ADC maps were 0.42 (95% CI, 0.30–0.54) and 0.61 (95% CI, 0.51–0.73), representing moderate and good levels of agreement.

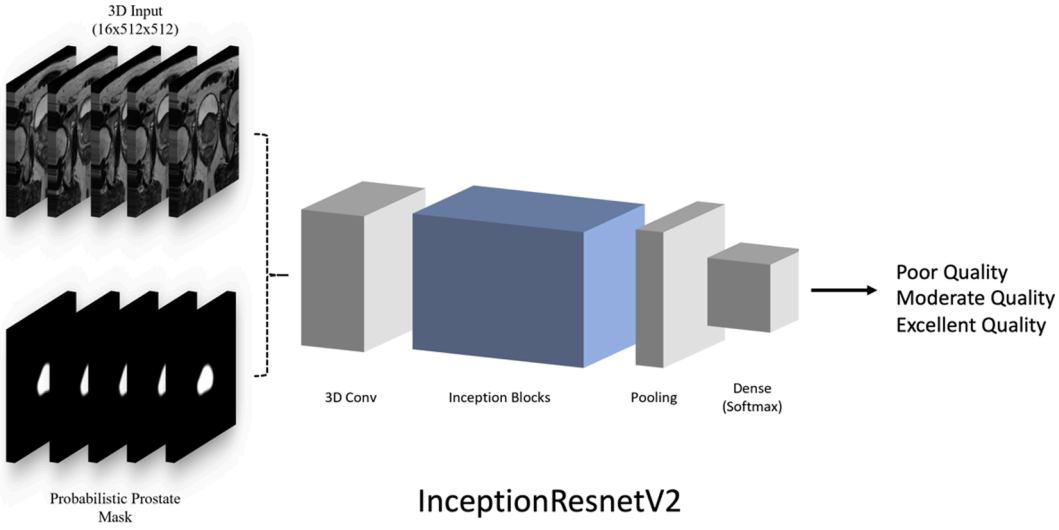

Fig. 1. The deep learning model used in this study. InceptionResNetV2 is a deep neural network architecture that employs a combination of 3D convolutions, inception modules, and pooling layers to reduce the spatial dimensions of the input while capturing richer and denser features for the given task. Our input to the model consisted of tri-planar T2W images or axial ADC maps along with probabilistic prostate masks. We used quality scores provided by a consensus of experts to define our ground-truth labels, which were categorized as poor, moderate, or excellent. To take into account the ordinality of the scores, we utilized the ordinal loss function. Note that only an axial T2W image is shown in this particular case.

For T2W images, in comparison to the ground-truth labels, the less-experienced readers had kappa scores of 0.38 (95% CI, 0.26–0.50), 0.56 (95% CI, 0.44–0.68), 0.34 (95% CI, 0.22–0.46), and 0.44 (95% CI, 0.32–0.56), indicating fair to moderate levels of agreement with the experts. For ADC maps, in comparison to the ground-truth labels, the less-experienced readers had kappa scores of 0.39 (95% CI, 0.27–0.51), 0.47 (95% CI, 0.35–0.59), 0.27 (95% CI, 0.15–0.39), and 0.62 (95% CI, 0.50–0.74), indicating fair to good levels of agreement with the experts.

Figs. 2, 3, and 4 show representative cases for the quality assessment. The quality scores of the DL model and less-experienced readers against the ground truth can be found in Supplementary Document 1. Further representative cases can be found in Supplementary Document 2. Various experiments exploring the relationship between image quality and the performance of DL are presented in Supplementary Document 3.

5. Discussions

In this study, we investigated the feasibility of a DL model trained with an ordinal loss function and probabilistic prostate masks for estimating the image quality of bi-parametric MRI scans. We used the consensus of expert readers as the ground truth and compared the model's performance with that of less-experienced readers using a 3-point Likert scale. Notably, the DL model demonstrated moderate and good levels of agreement with expert readers for T2W images and ADC maps in assessing image quality. In comparison, less-experienced readers achieved fair-to-moderate levels of agreement with the expert consensus for T2W images and fair-to-good levels of agreement for ADC maps. Our results suggest that our DL model is on par with or superior to the performance of the less-experienced readers in evaluating the image quality of bi-parametric prostate MRI scans.

In this study, we aimed to minimize the exposure of the DL model to human readers' biases by using the consensus of expert readers, following several prior studies on medical image quality assessment [22]. However, we acknowledge that it is virtually impossible to eliminate subjectivity entirely. Human-related biases can be further mitigated by involving a diverse group of readers from around the world, as such an approach would likely dilute the element of subjectivity and enhance robustness. Moreover, larger and more inclusive datasets from across the world, comprising low-volume non-specialized and high-volume specialized centers, scanners from different MRI manufacturers with varying magnet strengths, diverse scan protocols, and a range of ethnicities, can improve the generalizability of the DL models in assessing image quality. This would ultimately lead to more accurate and consistent image quality assessments, benefiting both patients and medical professionals alike.

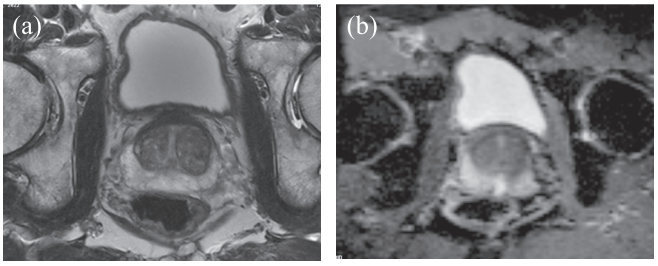

Fig. 3. A representative bi-parametric MRI scan of a man from the testing set. Two expert readers assessed the image quality as moderate, and one as excellent for T2W images (a). For ADC maps (b), two expert readers rated the image quality as excellent, while one assessed it as moderate. The prostate capsule and neurovascular bundle are mostly visible, although there is a certain degree of blurring. Additionally, there is a slight distortion artifact at the rectum-prostate interface in the ADC maps. The deep learning model accurately categorized the T2W images as moderate quality but incorrectly assigned the ADC maps as moderate quality. Among the less-experienced readers, two scored the T2W images as moderate quality, and the remaining two as poor quality. For the ADC maps, three less-experienced readers evaluated them as moderate quality, and the remaining one assigned excellent quality.

Another solution to mitigate bias might involve using synthetically created images by manipulating original images using various methods (e.g., adding Rician or Gaussian noise, blurring, motion). However, the factors affecting the quality of MRI scans are complex, and unique hardware or software-related MRI artifacts (e.g., magnetic susceptibility, zipper artifacts, etc.) can also play a role [23,24]. Consequently, it is not straightforward to artificially create MRI scans with different quality levels.

Our research group has been diligently working on training deep learning (DL) models using datasets that comprise both original and

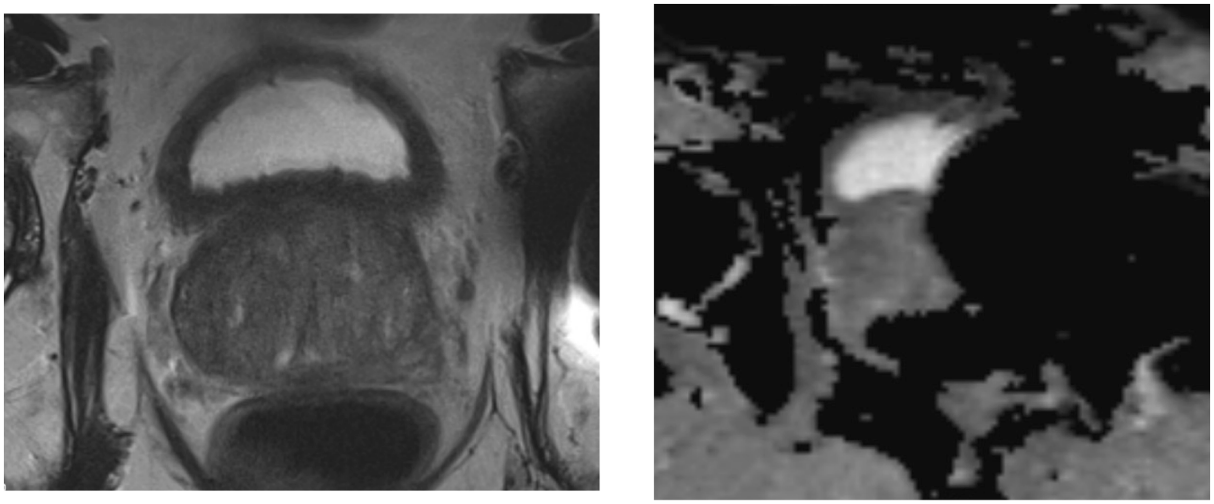

Fig. 2. A representative bi-parametric MRI scan of a man with a left hip prosthesis from the testing set. All expert readers assessed the image quality as poor for both T2W images (a) and ADC maps (b). Consequently, the ground-truth consensus expert scores were designated as poor for T2W images (a) and ADC maps (b). Both images, particularly the ADC maps, are affected by metallic artifacts from the left hip prosthesis. The prostate capsule and neurovascular bundles are not clearly distinguishable on the axial T2W image (a). Additionally, there is significant noise present in the T2W image. In agreement with the expert readers, all less-experienced readers and the deep learning model evaluated the T2W images and ADC maps as poor quality.

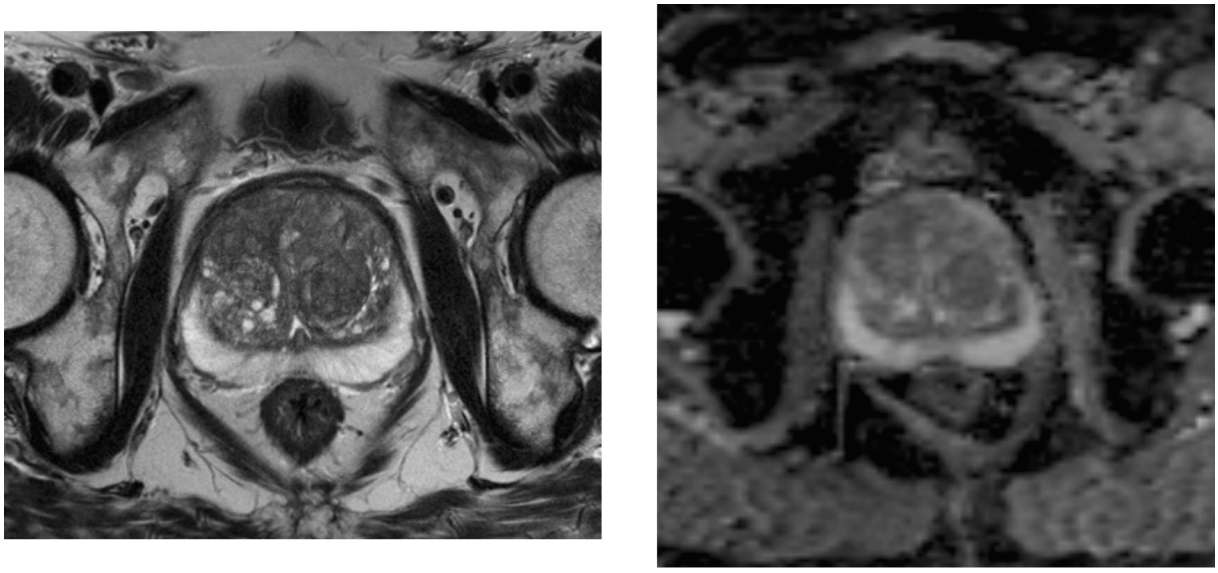

Fig. 4. A representative bi-parametric MRI scan of a man from the testing set. All expert readers assessed the image quality as excellent for both T2W images (a) and ADC maps (b). The prostate capsule and neurovascular bundle are clearly visible, without any discernible artifacts. The deep learning model assigned excellent quality to both T2W images and ADC maps. All less-experienced readers assigned excellent quality to T2W images. For the ADC maps, three less-experienced readers scored the images as excellent quality, while the remaining reader assigned moderate image quality.

manipulated images. We intend to publish our findings shortly. However, it is crucial to recognize that the quality of synthetically manipulated images may not always correspond with human readers' perception of image quality. In practical terms, the latter is often more critical, as radiologists are ultimately responsible for making the final diagnosis. Although it is possible to synthesize more realistic poor-quality images by manipulating k-space, the availability of k-space data remains limited [25].

In the current study, we did not evaluate the influence of image quality on the diagnostic performance of DL models in identifying csPCa. The reason for excluding performance estimation was the inability to objectively determine whether poor-quality images resulted in inferior csPCa identification by the DL model using a retrospectively collected public dataset. For instance, a large and conspicuous PI-RADS Score 5 csPCa might be readily detected by a DL model, even with low-quality images, while a small PI-RADS Score 3 csPCa could be missed. Consequently, we recommend that future research examining the impact of image quality on DL models for csPCa identification should focus on cases where poor and good-quality images exhibit similar lesion characteristics (e.g., scans with comparable PI-RADS scores or lesion sizes).

In a similar vein, our study employed a task-agnostic DL model for quality assessment. This model used human-derived quality criteria as ground truth labels and was optimized to differentiate high-quality images from low-quality images potentially containing artifacts, noise, low resolution, and distortion. In contrast, Saeed et al. utilized a task-specific DL model, which learns to distinguish between poor and good-quality images based on the performance of a sequential DL model in a downstream task, such as (csPCa detection or segmentation [13,14]. Notably, the authors reported that a DL model might perform poorly on some radiologist-assigned high-quality images while performing well on others labeled as low-quality images [13,14].

Given the rapid pace of clinical translation of DL models for prostate cancer diagnostics, we suggest that addressing the issue of image quality, taking into account the performance of downstream DL models—such as clinically significant PCa detection and prostate gland segmentation—is a worthwhile endeavor.

There are several limitations to our study that warrant acknowledgment. First, since the PI-CAI challenge did not provide metadata, we could not confirm the extent to which the scans adhered to the PI-RADS standards. Nevertheless, Sackett et al. investigated a selected sample of bi-parametric MRI scans from 62 institutions and discovered that compliance with PI-RADS did not necessarily result in a higher likelihood of obtaining a high-quality prostate MRI [5]. Moreover, assessing these technical parameters is relatively straightforward when DICOM metadata is available, while the more error-prone and time-consuming aspect lies in visual image quality assessment. This is where DL can play a crucial role.

Secondly, in this study, a somewhat arbitrary 3-point Likert scale was utilized to assess image quality. Although we considered the T2W and ADC map criteria proposed in the PI-QUAL score, we did not directly replicate the scoring system. Indeed, the Likert-Scale used in this study individually considered each sequence (i.e., T2W and ADC map), while the PI-QUAL Score is a standardized scoring method involving a combination of binary diagnostic assessment of each sequence. For example, in the PI-QUAL score, mpMRI sequences were categorized as diagnostic or non-diagnostic based on a set of objective and subjective criteria.

Instead of grouping each sequence into two broad categories, we opted for a more detailed classification system, the 3-point Likert Scale, to provide more information to the DL model using an ordinal loss, allowing it to comprehend the concept of quality better. We acknowledge that the broad categorization of scans as diagnostic or non-diagnostic might be a more straightforward task for human readers, but the DL model may encounter challenges in learning to distinguish scans with artifacts or indistinct anatomical structures that still have diagnostic image quality.

It is important for readers to understand that the 3-point Likert scale used in this study is not intended to propose a new quality assessment for either human readers or machine learning applications. Instead, it serves as a tool within this study's design to effectively train a DL model. As such, we believe this limitation does not detract from the academic value of our study. Our primary objective was not to establish a universal or robust quality scoring system, but rather to demonstrate the feasibility of DL in assessing image quality. Additionally, the 3-point Likert scale

can be easily translated into a binary assessment of image quality, distinguishing between diagnostic and non-diagnostic images.

Third, our study did not evaluate the image quality of dynamic-contrast-enhanced (DCE) images, as they were not included in the dataset provided by the challenge organizers. Although the role of DCE has been diminished with PI-RADS version 2.1, DCE remains crucial in mpMRI and is incorporated in the PI-QUAL score [26,27]. Since DCE images contain temporal information, future studies that use deep learning for image quality assessment of DCE images may utilize methods such as recurrent neural networks or more recent advancements like transformers. Indeed, researchers must include DCE images in their deep learning pipeline to cover each sequence of an mpMRI scan (for example, a combination of binary diagnostic assessments and combining them to obtain the PI-QUAL score or directly estimating the PI-QUAL score using all mpMRI images). In this way, a deep learning model could be integrated into the PI-QUAL score and its successors to fully automate image quality assessment.

Fourth, we employed a relatively simple DL model in this feasibility study. We recognize that more advanced next-generation convolutional neural networks or vision transformers may yield superior performance in quality assessment, but they often require higher computational power and larger sample sizes. However, in this study, our objective was not to develop the best DL model possible but to demonstrate the feasibility of DL in quality assessment for prostate MRI.

In conclusion, this small-scale feasibility study demonstrates that DL can be employed as an automated quality assessment tool for bi-parametric prostate MRI. DL models, trained on larger and more representative datasets annotated by a diverse group of experts from around the world, hold the potential to become robust enough to provide reliable image quality assessments. These models could potentially replace or supplement the visual assessment component of proposed quality assessment systems, such as PI-QUAL. Moreover, DL models can effectively streamline quality assessments and reduce inter-reader inconsistencies.

Declarations

Ethical statement

Open access dataset was used. There was no need for ethical review from the institutional review board.

Consent for publication.

Written informed consent was not required because it was related to identification imaging only, and patient anonymity was maintained.

CRediT authorship contribution statement

Deniz Alis: Supervision, Writing – original draft, Writing – review & editing. **Mustafa Said Kartal:** Data curation, Formal analysis. **Mustafa Ege Seker:** Data curation, Formal analysis. **Batuhan Guroz:** Data curation. **Yeliz Basar:** Data curation. **Aydan Arslan:** Data curation. **Sabri Sirolu:** Data curation. **Serpil Kurtcan:** Data curation. **Nurper Denizoglu:** Data curation. **Umit Tuzun:** Data curation. **Duzgun Yildirim:** Data curation. **Ilkay Oksuz:** Conceptualization, Supervision, Writing – review & editing. **Ercan Karaarslan:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2023.110924.

References

[1] H.U. Ahmed, A.-E.-S. Bosaily, L.C. Brown, R. Gabe, R. Kaplan, M.K. Parmar, Y. Collaco-Moraes, K. Ward, R.G. Hindley, A. Freeman, A.P. Kirkham, R. Oldroyd, C. Parker, M. Emberton, Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study, The Lancet. 389 (2017) 815–822, https://doi.org/10.1016/S0140-6736(16)32401-1.

[2] B. Turkbey, A.B. Rosenkrantz, M.A. Haider, A.R. Padhani, G. Villeirs, K.J. Macura, C.M. Tempany, P.L. Choyke, F. Cornud, D.J. Margolis, H.C. Thoeny, S. Verma, J. Barentsz, J.C. Weinreb, P.I. Reporting, D.S. Version, 2.1,, Update of Prostate Imaging Reporting and Data System Version 2, Eur Urol. 76 (2019) (2019) 340–351, https://doi.org/10.1016/j.eururo.2019.02.033.

[3] S.J. Esses, S.S. Taneja, A.B. Rosenkrantz, Imaging Facilities' Adherence to PI-RADS v2 Minimum Technical Standards for the Performance of Prostate MRI, Academic Radiology. 25 (2018) 188–195, https://doi.org/10.1016/j.acra.2017.08.013.

[4] P.R. Burn, S.J. Freeman, A. Andreou, N. Burns-Cox, R. Persad, T. Barrett, A multicentre assessment of prostate MRI quality and compliance with UK and international standards, Clinical Radiology. 74 (894) (2019) e19–894.e25, https://doi.org/10.1016/j.crad.2019.03.026.

[5] J. Sackett, J.H. Shih, S.E. Reese, J.R. Brender, S.A. Harmon, T. Barrett, M. Coskun, M. Madariaga, J. Marko, Y.M. Law, E.B. Turkbey, S. Mehralivand, T. Sanford, N. Lay, P.A. Pinto, B.J. Wood, P.L. Choyke, B. Turkbey, Quality of Prostate MRI: Is the PI-RADS Standard Sufficient? Acad Radiol. 28 (2021) 199–207, https://doi.org/10.1016/j.acra.2020.01.031.

[6] F. Giganti, C. Allen, M. Emberton, C.M. Moore, V. Kasivisvanathan, PRECISION study group, Prostate Imaging Quality (PI-QUAL): A New Quality Control Scoring System for Multiparametric Magnetic Resonance Imaging of the Prostate from the PRECISION trial, Eur Urol Oncol. 3 (2020) 615–619, https://doi.org/10.1016/j.euo.2020.06.007.

[7] F. Giganti, E. Dinneen, V. Kasivisvanathan, A. Haider, A. Freeman, A. Kirkham, S. Punwani, M. Emberton, G. Shaw, C.M. Moore, C. Allen, Inter-reader agreement of the PI-QUAL score for prostate MRI quality in the NeuroSAFE PROOF trial, Eur Radiol. 32 (2022) 879–889, https://doi.org/10.1007/s00330-021-08169-1.

[8] E. Karanasios, I. Caglic, J.P. Zawaideh, T. Barrett, Prostate MRI quality: clinical impact of the PI-QUAL score in prostate cancer diagnostic work-up, BJR. 95 (2022) 20211372, https://doi.org/10.1259/bjr.20211372.

[9] F. Giganti, S. Lindner, J.W. Piper, V. Kasivisvanathan, M. Emberton, C.M. Moore, C. Allen, Multiparametric prostate MRI quality assessment using a semi-automated PI-QUAL software program, Eur Radiol Exp. 5 (2021) 48, https://doi.org/10.1186/s41747-021-00245-x.

[10] N. Aldoj, S. Lukas, M. Dewey, T. Penzkofer, Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network, Eur Radiol. 30 (2020) 1243–1253, https://doi.org/10.1007/s00330-019-06417-z.

[11] R. Alkadi, F. Taher, A. El-Baz, N. Werghi, A Deep Learning-Based Approach for the Detection and Localization of Prostate Cancer in T2 Magnetic Resonance Images, J Digit Imaging. 32 (2019) 793–807, https://doi.org/10.1007/s10278-018-0160-1.

[12] S. Mehralivand, D. Yang, S.A. Harmon, D. Xu, Z. Xu, H. Roth, S. Masoudi, T. H. Sanford, D. Kesani, N.S. Lay, M.J. Merino, B.J. Wood, P.A. Pinto, P.L. Choyke, B. Turkbey, A Cascaded Deep Learning-Based Artificial Intelligence Algorithm for Automated Lesion Detection and Classification on Biparametric Prostate Magnetic Resonance Imaging, Acad Radiol. 29 (2022) 1159–1168, https://doi.org/10.1016/j.acra.2021.08.019.

[13] S.U. Saeed, W. Yan, Y. Fu, F. Giganti, Q. Yang, Z.M.C. Baum, M. Rusu, R.E. Fan, G. A. Sonn, M. Emberton, D.C. Barratt, Y. Hu, Image quality assessment by overlapping task-specific and task-agnostic measures: application to prostate multiparametric MR images for cancer segmentation 10.48550/arXiv (2022) 2202.09798.

[14] S.U. Saeed, Y. Fu, V. Stavrinides, Z.M.C. Baum, Q. Yang, M. Rusu, R.E. Fan, G. A. Sonn, J.A. Noble, D.C. Barratt, Y. Hu, Image quality assessment for machine learning tasks using meta-reinforcement learning, Med Image Anal. 78 (2022), 102427, https://doi.org/10.1016/j.media.2022.102427.

[15] A. Saha, M. Hosseinzadeh, H. Huisman, End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction, Medical Image Analysis. 73 (2021), 102155, https://doi.org/10.1016/j.media.2021.102155.

[16] M. Hosseinzadeh, A. Saha, P. Brand, I. Slootweg, M. de Rooij, H. Huisman, Deep learning-assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge, Eur Radiol. 32 (2022) 2224–2234, https://doi.org/10.1007/s00330-021-08320-y.

[17] M. de Rooij, B. Israël, M. Tummers, H.U. Ahmed, T. Barrett, F. Giganti, B. Hamm, V. Løgager, A. Padhani, V. Panebianco, P. Puech, J. Richenberg, O. Rouvière, G. Salomon, I. Schoots, J. Veltman, G. Villeirs, J. Walz, J.O. Barentsz, ESUR/ESUI consensus statements on multi-parametric MRI for the detection of clinically significant prostate cancer: quality requirements for image acquisition, interpretation and radiologists' training, Eur Radiol. 30 (2020) 5404–5416, https://doi.org/10.1007/s00330-020-06929-z.

[18] S. Chabert, J.S. Castro, L. Muñoz, P. Cox, R. Riveros, J. Vielma, G. Huerta, M. Querales, C. Saavedra, A. Veloz, R. Salas, Image Quality Assessment to Emulate Experts' Perception in Lumbar MRI Using Machine Learning, Applied Sciences. 11 (2021) 6616, https://doi.org/10.3390/app11146616.

[19] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, Proceedings of the AAAI Conference on Artificial Intelligence. 31 (2017). 10.1609/aaai.v31i1.11231.

[20] A. Karagoz, M.E. Seker, M. Yergin, T.A. Kan, M.S. Kartal, E. Karaarslan, D. Alis, I. Oksuz, Prostate Lesion Estimation using Prostate Masks from Biparametric MRI, (2023). 10.48550/arXiv.2301.09673.
[21] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement. 20 (1960) 37–46, https://doi.org/10.1177/001316446002000104.
[22] Q. Gao, S. Li, M. Zhu, D. Li, Z. Bian, Q. Lyu, D. Zeng, J. Ma, Blind CT Image Quality Assessment via Deep Learning Framework, in: 2019 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2019: pp. 1–4. 10.1109/NSS/MIC42101.2019.9059777.
[23] N. Sinha, A.G. Ramakrishnan, Quality assessment in magnetic resonance images, Crit Rev Biomed Eng. 38 (2010) 127–141, https://doi.org/10.1615/critrevbiomedeng.v38.i2.20.
[24] L.S. Chow, R. Paramesran, Review of medical image quality assessment, Biomedical Signal Processing and Control. 27 (2016) 145–154, https://doi.org/10.1016/j.bspc.2016.02.006.
[25] I. Oksuz, B. Ruijsink, E. Puyol-Antón, J.R. Clough, G. Cruz, A. Bustin, C. Prieto, R. Botnar, D. Rueckert, J.A. Schnabel, A.P. King, Automatic CNN-based detection of cardiac MR motion artifacts using k-space data augmentation and curriculum learning, Medical Image Analysis. 55 (2019) 136–147, https://doi.org/10.1016/j.media.2019.04.009.
[26] E.J. Bass, A. Pantovic, M. Connor, R. Gabe, A.R. Padhani, A. Rockall, H. Sokhi, H. Tam, M. Winkler, H.U. Ahmed, A systematic review and meta-analysis of the diagnostic accuracy of biparametric prostate MRI for prostate cancer in men at risk, Prostate Cancer Prostatic Dis. 24 (2021) 596–611, https://doi.org/10.1038/s41391-020-00298-w.
[27] V. Brancato, G. Di Costanzo, L. Basso, L. Tramontano, M. Puglia, A. Ragozzino, C. Cavaliere, Assessment of DCE Utility for PCa Diagnosis Using PI-RADS v2.1: Effects on Diagnostic Accuracy and Reproducibility, Diagnostics (Basel) 10 (2020) 164, https://doi.org/10.3390/diagnostics10030164.