Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.243102

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Pix2Pix generative-adversarial network in improving the quality of T2-weighted prostate magnetic resonance imaging: a multi-reader study

Yeliz Başar[1]
Mustafa Said Kartal[2]
Mustafa Ege Seker[3]
Deniz Alis[4]
Delal Seker[5]
Müjgan Orman[1]
Sabri Şirolu[6]
Serpil Kurtcan[1]
Aydan Arslan[7]
Nurper Denizoğlu[8]
İlkay Öksüz[9]
Ercan Karaarslan[4]

[1]Acıbadem Healthcare Group, Department of Radiology, İstanbul, Türkiye

[2]Cumhuriyet University Faculty of Medicine, Sivas, Türkiye

[3]University of Wisconsin-Madison, School of Medicine, Department of Radiology, Madison, USA

[4]Acıbadem Mehmet Ali Aydınlar University Faculty of Medicine, Department of Radiology, İstanbul, Türkiye

[5]Dicle University Faculty of Engineering, Department of Electrical-Electronics Engineering, Diyarbakır, Türkiye

[6]University of Health Sciences Türkiye, Şişli Hamidiye Etfal Training and Research Hospital, Clinic of Radiology, İstanbul, Türkiye

[7]Ümraniye Training and Research Hospital, Clinic of Radiology, İstanbul, Türkiye

[8]University of Health Sciences Türkiye, Sultan [2]. Abdulhamid Han Training and Research Hospital, Clinic of Radiology, İstanbul, Türkiye

[9]İstanbul Technical University Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye

Corresponding author: Deniz Alis

E-mail: drdenizalis@gmail.com

Received 28 October 2024; revision requested 23 November 2024; last revision received 22 January 2025; accepted 02 February 2025.

Epub: 05.05.2025

Publication date: 06.11.2025

DOI: 10.4274/dir.2025.243102

PURPOSE

To assess the performance and feasibility of generative deep learning in enhancing the image quality of T2-weighted (T2W) prostate magnetic resonance imaging (MRI).

METHODS

Axial T2W images from the prostate imaging: cancer artificial intelligence dataset (n = 1,476, biologically males; n = 1,500 scans) were used, partitioned into training (n = 1300), validation (n = 100), and testing (n = 100) sets. A Pix2Pix model was trained on original and synthetically degraded images, generated using operations such as motion, Gaussian noise, blur, ghosting, spikes, and bias field inhomogeneities to enhance image quality. The efficacy of the model was evaluated by seven radiologists using the prostate imaging quality criteria to assess original, degraded, and improved images. The evaluation also included tests to determine whether the images were original or synthetically improved. Additionally, the model's performance was tested on the in-house external testing dataset of 33 patients. The statistical significance was assessed using the Wilcoxon signed-rank test.

RESULTS

Results showed that synthetically improved images [median score (interquartile range) 4.71 (1)] were of higher quality than degraded images [3.36 (3), $P = 0.0001$], with no significant difference from original images [5 (1.14), $P > 0.05$]. Observers equally identified original and synthetically improved images as original (52% and 53%), proving the model's ability to retain realistic attributes. External testing on a dataset of 33 patients confirmed a significant improvement ($P = 0.001$) in image quality, from a median score of 4 (2.286)–4.71 (1.715).

CONCLUSION

The Pix2Pix model, trained on synthetically degraded data, effectively improved prostate MRI image quality while maintaining realism and demonstrating both applicability to real data and generalizability across various datasets.

CLINICAL SIGNIFICANCE

This study critically assesses the efficacy of the Pix2Pix generative-adversarial network in enhancing T2W prostate MRI quality, demonstrating its potential to produce high-quality, realistic images indistinguishable from originals, thereby potentially advancing radiology practice by improving diagnostic accuracy and image reliability.

KEYWORDS

Deep learning, generative artificial intelligence, prostate, magnetic resonance imaging, prostate imaging quality

You may cite this article as: Başar Y, Kartal MS, Seker ME, et al. Pix2Pix generative-adversarial network in improving the quality of T2-weighted prostate magnetic resonance imaging: a multi-reader study. *Diagn Interv Radiol.* 2025;31(6):547-565.

The prostate imaging reporting and data system (PI-RADS) and its updates prescribe best practices for the acquisition and interpretation of prostate magnetic resonance imaging (MRI) scans,[1] emphasizing minimum technical requirements to ensure scan quality, which is crucial for the accurate detection of clinically significant prostate cancer. However, adherence to PI-RADS guidelines does not invariably guarantee high-quality MRI scans, as evidenced by various studies.[2-4]

Deep learning (DL)-based reconstruction techniques can speed up image acquisition and improve quality beyond traditional MRI methods.[5-8] Nonetheless, variables such as patient characteristics, equipment quality, and the expertise of the radiology team can still result in suboptimal images.[3,9] Moreover, DL-based reconstruction typically requires newer scanner models and significant initial investments, limiting its accessibility. However, limited research has been conducted on applying DL techniques to enhance prostate MRI quality. Existing studies are often constrained by the use of single-center datasets and proprietary scoring systems, which can affect the reproducibility of their outcomes.[10]

In this study, we employed a generative-adversarial network (GAN) model, Pix2Pix, to enhance the quality of axial T2-weighted (T2W) prostate MRI. We used a large-scale, multi-center, and publicly available dataset, prostate imaging: cancer artificial intelligence (PI-CAI),[11] allowing us to overcome some of the limitations noted in previous studies. Image quality was evaluated by multiple readers from different centers using a scoring system adopted from the newly introduced Prostate imaging quality (PI-QUAL):[12] which provided a standardized assessment method. We also examined the realism of the generated images and tested the model's performance on the in-house external testing dataset to evaluate its generalizability.

Main points

- The Pix2Pix generative-adversarial network (GAN) significantly improved T2-weighted prostate magnetic resonance imaging (MRI) quality while maintaining realism.
- Synthetically improved images scored higher than degraded ones when compared with the original images.
- External testing confirmed significant image quality improvement.
- Radiologists could not distinguish between original and synthetically improved images.
- The study demonstrated GANs' potential for realistic, high-quality prostate MRI enhancement.

Methods

Study sample

The Acıbadem University and Healthcare Institution's Medical Research Ethics Committee approved this retrospective study and waived the requirement for informed consent for the retrospective collection, analysis, and presentation of anonymized medical data (date: 11.02.2021, decision no: 2021-03/12).

This study utilized the publicly available PI-CAI training dataset, which consisted of 1,500 bi-parametric prostate MRI scans obtained from 1,476 biologically male individuals at 4 tertiary academic centers in the Netherlands and Norway between March 2015 and January 2018. The data from these four centers were stratified across the training, validation, and internal testing sets to ensure representation from each center in all data partitions. The examinations were stratified into three distinct groups: a development set (1400 scans, 1300 training, and 100 validation) and a testing set (100 exams). This stratification was done with careful consideration to ensure that scans from the same patient were not included across the development and testing sets. The flowchart of the study is given in Figure 1.

We also included the in-house dataset of 33 bi-parametric MRI examinations from 33 biologically male individuals as an external testing dataset in this study. The overall workflow of the study is shown in Figure 2.

Bi-parametric magnetic resonance imaging examinations

All bi-parametric MRI scans of the PI-CAI dataset were conducted using 1.5T units (n = 82) from Siemens (Aera and Avanto models, Siemens Healthcare, Erlangen, Germany) and Philips (Achieva and Intera models, Philips Healthcare, Eindhoven, the Netherlands), as well as 3T units (n = 1418) from Siemens (Skyra, TrioTim, and Prisma models, Siemens Healthcare, Erlangen, Germany) and Philips (Ingenia model, Philips Healthcare, Eindhoven, the Netherlands). These scans utilized surface coils and adhered to the PI-RADS V2 guidelines. Additional specifications regarding the MRI protocols used for the study sample are detailed in.[11]

The examinations of the in-house testing dataset were performed using 1.5T units from Siemens (Avanto-fit, Siemens Healthcare, Erlangen, Germany). These scans were also performed with surface coils. For this study, only axial T2W images were used for further analysis. Table 1 shows the imaging protocol for the in-house testing dataset.

Synthetic data creation

In this study, a crucial step involved the creation of a robust training dataset by applying clinically relevant MRI artifacts to create realistic low-quality T2W images. For this purpose, TorchIO library[13] was used: a powerful tool specifically designed for data augmentation in medical imaging. A variety of techniques were employed to simulate commonly encountered artifacts, including motion, Gaussian noise, blur, ghosting, spikes, and bias field inhomogeneities (detailed in Supplementary S1).

All images were normalized and resized to uniform dimensions to facilitate consistent neural network training, ensuring each image had intensity values within a specific range for optimal input standardization.

The training set included images manipulated with each artifact individually, as well as in specific combinations, enabling the Pix2Pix model to learn from a wide variety of possible artifact scenarios and improve its ability to generalize across different image quality corruption. Internal testing was

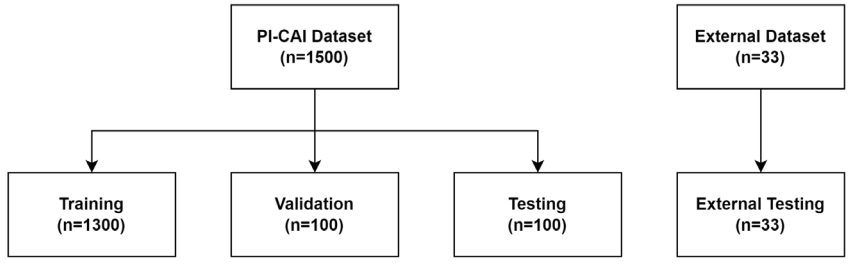


Figure 1. Flowchart of the study. PI-CAI, prostate imaging: cancer artificial intelligence.

created using either a single artifact or a predefined combination of multiple artifacts. This setup allowed for a controlled evaluation of the model's performance in enhancing images with known quality issues. Detailed descriptions of the data pre-processing are given in Supplementary Document S1.

Pix2Pix model

The Pix2pix model, a conditional GAN, was utilized for image-to-image translation using paired images to improve accuracy. It consisted of a generator, employing a U-Net architecture to maintain anatomical features in medical images, and a discriminator, which used a PatchGAN classifier to focus on high-frequency details and realism by evaluating small patches within the images.[14,15]

The training involved an adversarial process where the generator tried to create increasingly realistic images, whereas the discriminator improved at detecting synthetically improved images. The process was governed by a combined loss function: adversarial loss ensured the images were visually indistinguishable from the original ones, and L1 loss maintained structural integrity, reducing blurring and preserving crucial details. This setup enhanced the model's ability to produce clinically useful MRI images while retaining essential diagnostic features.

The Pix2pix model was trained using the Adam optimizer with a learning rate of 0.0002, focusing on 200 epochs where the L1 loss was emphasized initially to enhance accuracy. The training involved an adversarial setup where the generator aimed to produce images close to the original ones by minimizing both L1 and adversarial loss, whereas the discriminator sought to identify whether the image patches were original or synthetic, aiming to maximize the adversarial loss. A detailed description of the model is given in Supplementary Document S1.

The model's performance during training was monitored using the mean absolute error (MAE) calculated between the reconstructed images and the original high-quality images. The model with the lowest MAE on the validation data was selected as the best-performing model for subsequent evaluation and application to the test set.

The best-performing model was then applied to synthetically degraded internal testing data (n = 100) and original images from the in-house external testing dataset (n = 33) to assess performance, as detailed in the subsequent sections.

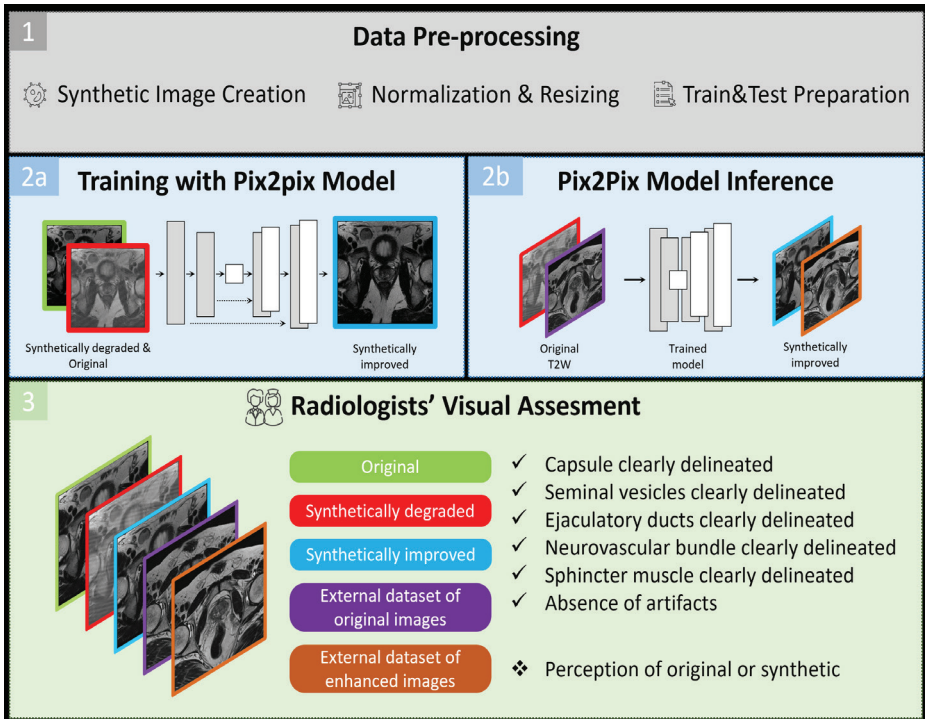


Figure 2. This figure illustrates the three main stages of the study. First, during data pre-processing, synthetically degraded T2-weighted images were created to mimic real-world artifacts, then normalized and resized along with the original images before being split into training and testing sets. Next, in the model training and inference stage, the Pix2Pix model was trained on paired original and synthetically degraded images **(2a)**, with the generator learning to produce improved images from degraded inputs while the discriminator differentiated between original and synthetically improved images. The trained model was then applied to an external in-house testing dataset, generating improved versions of the original images **(2b)**. Finally, radiologists visually assessed a set of images, including original, synthetically degraded, synthetically improved, and improved images from the external in-house testing dataset, evaluating them based on predefined criteria encompassing anatomical delineation, artifact presence, and perceived realism.

Table 1. Prostate T2-WI acquisition parameters of the in-house external testing dataset

Parameters	Values
TR (ms)	5454
TE (ms)	101
FOV (mm)	268 × 400
Matrix size	206 × 512
Slice thickness (mm)	4
Slice gap (mm)	0.6
Flip angle	150°
Temporal resolution (s)	-

T2WI, T2-weighted imaging; TR, repetition time; TE, echo time; FOV, field of view.

Study readers

Seven readers participated in the analysis of the scans for this study. Reader 1, an expert prostate radiologist, interpreted over 300 cases annually for more than 10 years. Readers 2–7 were basic prostate readers, each handling 150–200 cases per year for 2–7 years. The classification of the readers adhered to the consensus statement of the European Society of Urogenital Radiology.[16] Readers one and three were from the same center, whereas the others were based in various other hospitals, ranging from academic to non-academic settings.

Assessment criteria

The evaluation by the readers was adopted from the visual assessment criteria proposed in the PI-QUAL for T2W imaging.[12] Specifically, the readers assessed the clarity with which they could delineate the capsule, seminal vesicles, ejaculatory ducts, neurovascular bundle, and sphincter mus-

cle, awarding one point for each positively identified structure (i.e., if the structure could be seen clearly). Additionally, they awarded one point in the absence of artifacts and zero points if artifacts were present. Thus, the total score for each examination ranged from zero (the worst quality) to six points (the best quality).

Before the reading sessions, several online meetings were conducted to familiarize the readers with the PI-QUAL criteria through examples from published papers[17] and to acquaint them with the reading platform. The primary aim of these sessions was to enhance their understanding of the PI-QUAL.

We used only axial T2W images for the reading sessions, as the model employed in the current study was designed to work with axial T2W images. Although this may be considered a limitation, it was consistent with earlier work, which primarily focused on axial images as they were the primary sequence used in PI-RADS assessments.

Case reading sessions

The readers used a dedicated workstation equipped with a 6-megapixel diagnostic color monitor (Radiforce RX 660, EIZO) and a dedicated browser-based platform (https://matrix.md.ai). All reviewed images were in the Digital Imaging and Communications in Medicine format.

Initially, 7 readers evaluated 300 T2W series in the internal testing set of 100 patients, consisting of 100 original, 100 synthetically degraded, and 100 synthetically improved series. The readers independently assessed the cases in a random order to minimize bias, not knowing which images were original, degraded, or improved. They assigned points to each examination based on the previously described criteria and judged whether the images were original or synthesized.

Subsequently, to further evaluate the model's performance and its ability to enhance image quality on real data, the readers assessed the scans in the in-house external testing dataset of 33 patients, which included 33 original and 33 synthetically improved T2W series.

Statistical analysis

Statistical analyses were performed using the SciPy library in Python version 3. Continuous variables were presented using medians and interquartile ranges, whereas categorical and ordinal variables were presented with frequencies and percentages.

The structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR) were used as quantitative metrics to assess image quality. The SSIM evaluated perceptual similarity by comparing luminance, contrast, and structure between images, with a range from −1 to 1, where 1 indicated perfect similarity. The PSNR measured the ratio between the maximum possible signal value and the distortion introduced, expressed in decibels, where higher values indicated better quality.

For comparing image quality assessments across original, synthetically degraded, and synthetically improved images, pairwise comparisons were conducted using the Friedman test and post-hoc Durbin-Conover test due to the matched nature of the data. For the pairwise comparison of the in-house external testing dataset, the Wilcoxon signed-rank test was used.

To evaluate the performance of radiologists in correctly identifying original versus synthetically improved images, accuracy was calculated. To analyze the differences in radiologists' ability to detect synthetically improved versus original images, McNemar's test was used. A P value less than 0.05 was considered statistically significant.

Results

Image quality assessment

We included 100 examinations in the testing set, each paired with their synthetically degraded and improved versions from the PI-CAI testing set. The PSNR and SSIM values of synthetically improved images [PSNR: 28.79 (32.54), SSIM: 0.92 (0.16)] were statistically significantly higher than those of the degraded image forms [PSNR: 24.87 (15.27), SSIM: 0.78 (0.13)] (PSNR: $P < 0.001$, SSIM: $P < 0.001$).

During the random blinded assessment, the observers gave median scores of 5 (1.14) to the original images, 3.36 (3) to the synthetically degraded images, and 4.71 (1) to the synthetically improved images ($P = 0.0001$). Pair-wise comparisons revealed that original images had a significantly higher median quality score than the synthetically degraded images ($P < 0.0001$). Likewise, synthetically improved images also had a higher image quality than synthetically degraded images ($P < 0.0001$). No statistically significant difference was found between the median image quality of the original and synthetically improved images ($P = 0.37$) (Figure 3a). A detailed breakdown of each reader's median scores for original, synthetically degraded, and synthetically improved images is given in Supplementary Document S2.

Figure 4 shows a representative example of original, synthetically degraded, and synthetically improved images. More representative examples can be found in Supplementary Document 2.

Original vs. synthetic assessment

We evaluated whether the observers could discriminate between original and synthetically improved T2W images using a majority voting scheme from the PI-CAI testing set. In this test, the observers identified 52% of the original and 53% of the synthetically improved images as original, with no statistical difference, indicating that the observers could not reliably discriminate between original and synthetically improved images ($P = 0.62$). A detailed breakdown of each reader's assessments on whether the images are original or synthetic is provided in Supplementary Document S2.

External testing

We evaluated whether the proposed model could also improve original images from the in-house external testing dataset. This set consisted of T2W images of 33 patients from the in-house center, where prostate images were obtained using a 1.5T scanner. The observers gave a median score of 4 (2.286) for the original images in the in-house external testing dataset. The median image quality score for this dataset was statistically lower than that for the original T2W images from the PI-CAI testing set ($P = 0.009$).

The proposed model improved the image quality of the original images from 4 (2.2) to 4.71 (1.7), demonstrating a statistically significant improvement ($P = 0.001$) (Figure 3b). Notably, after the improvement, we found no statistical difference in median image quality between the original images from the PI-CAI dataset [median: 5 (1.14)] and the synthetically improved images from the in-house dataset [median: 4.71 (1.7)] ($P = 0.16$). A detailed breakdown of each reader's median scores for original and synthetically improved images for the in-house external testing dataset is given in Supplementary Document S2.

Figure 5 shows representative examples of original and synthetically improved images of a patient along with observers' ratings from the in-house external testing dataset. More representative examples can be found in Supplementary Document S2.

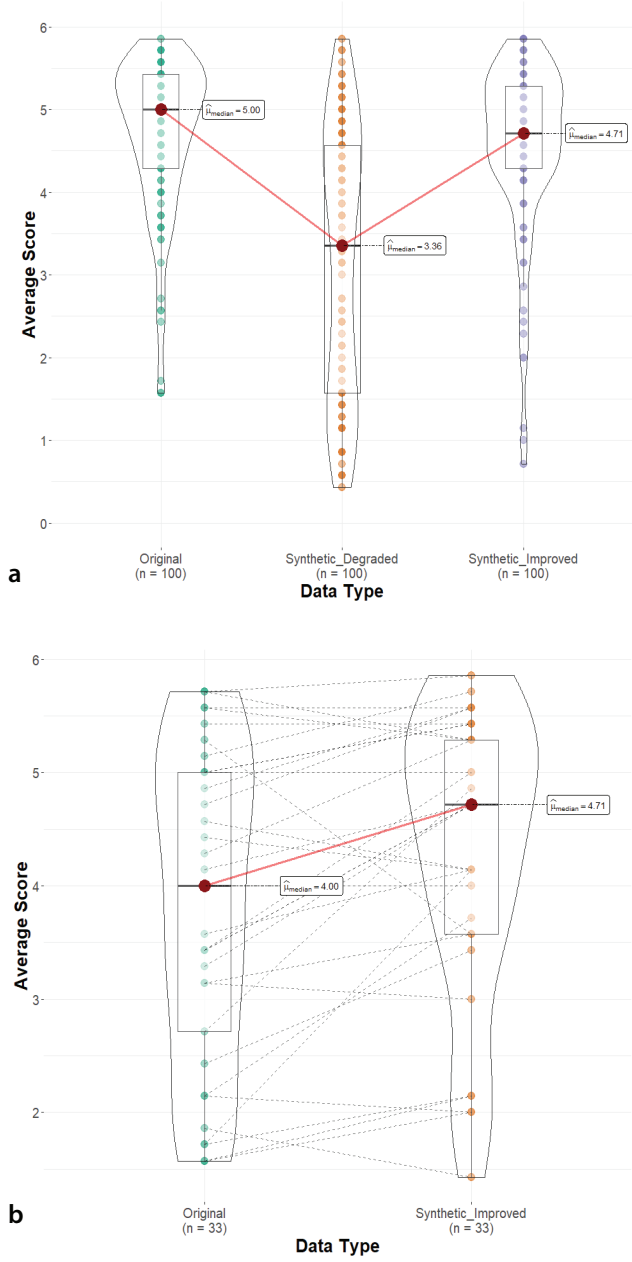


Figure 3. Comparison of the radiologists' scores for original, synthetically degraded, and synthetically improved T2-weighted images **(a)**. Comparison of the radiologists' average score for original and synthetically improved prostate magnetic resonance imaging scans from the in-house external testing dataset **(b)**.

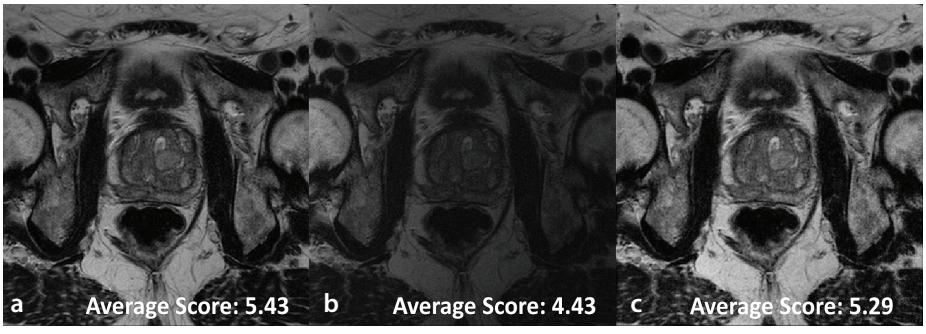


Figure 4. A representative prostate magnetic resonance imaging scan **(a)** original image with the average score of 5.43, **(b)** synthetically degraded image with the average score of 4.43, and **(c)** a synthetically improved image with the average score of 5.29.

Discussion

We found that the Pix2Pix model significantly improved the quality of synthetically degraded images evidenced by quantitative metrics and assessments of multiple readers with different experience levels from different institutions following the criteria adopted from the PI-QUAL. Notably, the synthetically improved images showed no statistical difference in image quality compared with the original images.

We further tested the performance of the proposed model on an external testing dataset, where it substantially increased the image quality. This demonstrates that the model not only works across different datasets but is also effective in improving image quality for original images that have not been synthetically manipulated. This finding is promising as it suggests that DL models can be trained on available datasets without the need for actual poor-quality prostate MRIs. It is important to note that the PI-CAI dataset is derived from centers in the Netherlands and Norway. This geographical restriction could limit the generalizability of our findings to other populations. Future studies should include data from more diverse geographical regions.

Another important finding was that the readers were not unambiguously able to discriminate original from synthetically improved images regardless of their experience levels, showing the proposed model did not only improve image quality but was also able to generate realistic looking images without introducing over-smoothness or plastic appearance.

Our findings diverge from those of the study by Belue et al.[10], where the authors observed no qualitative improvement and the readers mostly opted for original images over synthetically improved images evidenced by expert radiologists. Belue et al.[10] utilized a Cycle-GAN model and tested it using paired original images of both poor and good quality from the same patients. Moreover, they employed bespoke qualitative criteria, which they acknowledged as a significant limitation of their study.[10] We propose that by systematically incorporating a variety of artifacts, our model may better learn the representations of both poor- and good-quality images, thereby effectively transforming poor-quality images into good-quality ones in a realistic manner.

The tendency of DL methods in over-smoothing diagnostic images has also been documented in studies using

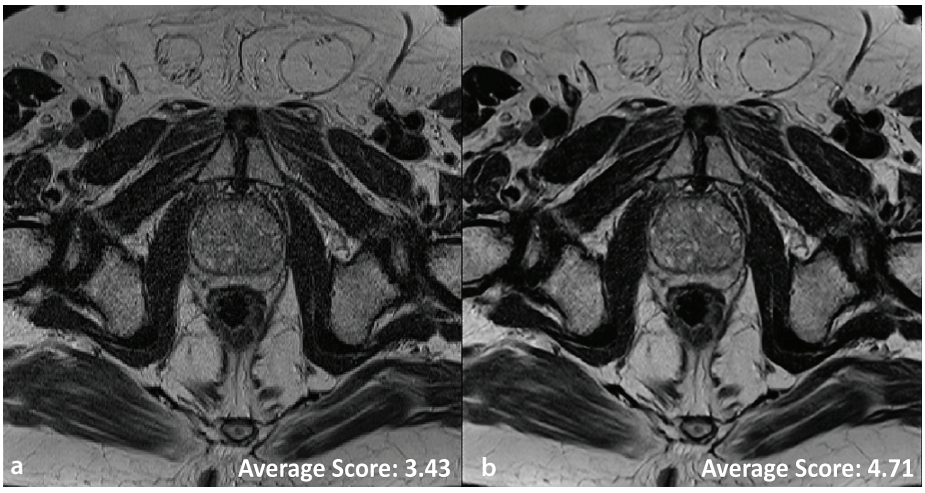


Figure 5. A representative prostate magnetic resonance imaging scan from the in-house external testing dataset, **(a)** an original image with an average radiologists' assessment score of 3.43, and **(b)** a synthetically improved image with an average radiologists' assessment score of 4.71.

DL-based reconstruction methods.[18] This smoothness can cause radiologists to feel uncertain about their interpretations, fearing potential loss of diagnostic information, such as the disruption of lesion appearance or visibility.[18,19] In contrast, our Pix2Pix model, trained on a meticulously prepared dataset, successfully generated realistic images, addressing these concerns by maintaining critical image details essential for accurate diagnosis. The training data included various levels of corruption for each augmentation as well as a combination of these augmentations with the corresponding good quality data. Including a combination of ghosting, spike artifacts, and bias field inhomogeneities with general Gaussian blur and noise in the training regime of the Pix2Pix model increased the robustness of our model against over-smoothing. However, our study did not explicitly evaluate the impact of image enhancement on lesion detection or characterization, which represented an essential area for future investigation.

In reflecting on the methods and results of our study, particularly in terms of experts identifying whether the images were original or synthetic, it is crucial to acknowledge the potential impact of bias. To minimize bias, we did not show the readers both the original and synthetic images simultaneously. Instead, the images were presented in a random order, and the readers were asked to determine their authenticity. A potential limitation is that readers one and three were from the same institution. Although this could introduce bias, the inclusion of readers from other centers helped mitigate this potential issue. Future work could incorporate strategies such as stratified sampling based on institutional affiliation to further address this. Intriguingly, the results suggested that the readers were essentially guessing, indicating no clear distinction between the original and synthetically improved images. However, this design may have inadvertently introduced another form of bias.

Knowing the study's objective—to assess the realism of synthetically generated images—likely predisposed the readers to scrutinize each image more critically. This awareness could have heightened sensitivity to any minor imperfections, predisposing the readers to identify these as indicators of synthetic origin. Admittedly, it is virtually impossible to completely isolate this information from the readers since the core of our evaluation involved discerning the nature of the images, thus directly revealing the study's design.

We openly acknowledge that the design of our study might have influenced the readers' judgments. Recognizing this does not diminish the validity of our findings but rather enhances the transparency and integrity of our analysis. This situation underscores the need for further research to quantify and adjust for such bias, ensuring that the conclusions drawn are robust and applicable in real-world diagnostic settings. This will help in developing methodologies that better emulate the blind assessments typically conducted in clinical practice.

Several other limitations to our study warrant acknowledgment. First, our model was limited to axial T2W images and excluded other crucial sequences. Future studies could explore enhancing image quality across all sequences and integrating them into a single DL pipeline for more effective improvements.[20] Although our study employed PI-QUAL V1, we acknowledge that V2.0 was released during our study period. Future studies should utilize the updated version for assessment.

Second, we used Pix2Pix due to its established use and relatively lower computational demands compared with the state-of-the-art diffusion denoising probabilistic models, which required significantly more resources. Future work will include applying advanced architectures, including transformers and diffusion models for image enhancement.

In conclusion, we demonstrated that a GAN model, Pix2Pix, trained on synthetically degraded axial T2W prostate MRI, can substantially improve image quality as evidenced by quantitative metrics and assessments from multiple readers with varying levels of experience following PI-QUAL criteria, showing no statistical difference in image quality compared with the original images. Additionally, the readers were unable to distinguish between original and synthetic images, indicating that the model did not introduce any unnatural appearance. Furthermore, the same model was able to improve image quality in an external testing dataset of original images, demonstrating its generalizability across datasets and its capability to improve both original and synthetically degraded images.

Acknowledgements

This paper was produced with support from the 1001 Science and Technology Grant Program National Program of TUBITAK (Project No: 122E022). However, the entire responsibility for the publication rests with the author. The financial support received from TUBITAK does not imply that the content of the publication is scientifically endorsed by TUBITAK.

Footnotes

Conflict of Interest

Deniz Alis is the CEO and co-founder of Hevi AI Health Tech. None of Hevi AI's products were used or mentioned in the current work. Furthermore, this paper did not use any commercially available DL software. Other authors have nothing to disclose.

References

1. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging

reporting and data system version 2. *Eur Urol*. 2019;76(3):340-351. [CrossRef]

2. Esses SJ, Taneja SS, Rosenkrantz AB. Imaging facilities' adherence to PI-RADS v2 minimum technical standards for the performance of prostate MRI. *Academic Radiology*. 2018;25(2):188-195. [CrossRef]
3. Burn PR, Freeman SJ, Andreou A, Burns-Cox N, Persad R, Barrett T. A multicentre assessment of prostate MRI quality and compliance with UK and international standards. *Clinical Radiology*. 2019;74(11):894. [CrossRef]
4. Sackett J, Shih JH, Reese SE, et al. Quality of Prostate MRI: Is the PI-RADS Standard Sufficient? *Acad Radiol*. 2021;28(2):199-207. [CrossRef]
5. Ueda T, Ohno Y, Yamamoto K, et al. Deep learning reconstruction of diffusion-weighted MRI improves image quality for prostatic imaging. *Radiology*. 2022;303(2):373-381. [CrossRef]
6. Gassenmaier S, Afat S, Nickel D, Mostapha M, Herrmann J, Othman AE. Deep learning-accelerated T2-weighted imaging of the prostate: reduction of acquisition time and improvement of image quality. *Eur J Radiol*. 2021;137:109600. [CrossRef]
7. Wang X, Ma J, Bhosale P, et al. Novel deep learning-based noise reduction technique for prostate magnetic resonance imaging. *Abdom Radiol*. 2021;46(7):3378-3386. [CrossRef]
8. Lebel RM. Performance characterization of a novel deep learning-based MR image reconstruction pipeline. 2020. [CrossRef]
9. Lin Y, Yilmaz EC, Belue MJ, Turkbey B. Prostate MRI and image quality: it is time to take stock. *Eur J Radiol*. 2023;161:110757. [CrossRef]
10. Belue MJ, Harmon SA, Masoudi S, et al. Quality of T2-weighted MRI re-acquisition versus deep learning GAN image reconstruction: a multi-reader study. *Eur J Radiol*. 2024;170:111259. [CrossRef]
11. Saha A, Twilt JJ, Bosma JS, et al. The PI-CAI challenge: public training and development dataset. Published online May 5, 2022. [CrossRef]
12. Giganti F, Allen C, Emberton M, Moore CM, Kasivisvanathan V, 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*. 2020;3(5):615-619. [CrossRef]
13. Pérez-García F, Sparks R, Ourselin S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Programs Biomed*. 2021;208:106236. [CrossRef]
14. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer Assisted Intervention*; 2015;234-241. [CrossRef]
15. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. [CrossRef]
16. de Rooij M, Israël B, Tummers M, et al. 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*. 2020;30(10):5404-5416. [CrossRef]
17. Giganti F, Dinneen E, Kasivisvanathan V, et al. Inter-reader agreement of the PI-QUAL score for prostate MRI quality in the NeuroSAFE PROOF trial. *Eur Radiol*. 2022;32(2):879-889. [CrossRef]
18. Lee KL, Kessler DA, Dezonie S, et al. Assessment of deep learning-based reconstruction on T2-weighted and diffusion-weighted prostate MRI image quality. *Eur J Radiol*. 2023;166:111017. [CrossRef]
19. Barrett T, Lee KL, de Rooij M, Giganti F. Update on Optimization of Prostate MR Imaging Technique and Image Quality. *Radiologic Clinics of North America*. 2024;62(1):1-15. [CrossRef]
20. Karagoz A, Seker ME, Yergin M, et al. Prostate Lesion estimation using prostate masks from biparametric MRI. *arXiv*. Published online January 11, 2023. [CrossRef]

Supplementary Document S1

1. data pre-processing: creating a robust training set for prostate MRI enhancement

A crucial step in this study involved the creation of a robust training dataset comprised of T2W images realistically mimicking various MRI artifacts. This approach, utilizing the TorchIO library for medical image augmentation, aimed to enhance the robustness and generalizability of the trained model by exposing it to a wide range of image degradations commonly encountered in clinical practice.[1]

1.1. Motion artifacts: simulating the impact of patient movement

Motion artifacts leads to blurring or ghosting that can obscure anatomical details.[2] To make our deep learning model more robust to these artifacts, we simulated realistic motion during the MRI scan. This simulation involves randomly generating a series of small movements, representing the kind of positional shifts a patient might make.

Each simulated movement is mathematically represented as a 3D transformation. These transformations include both rotation (turning) and translation (shifting) components, and their parameters are randomly varied to create a diverse range of plausible motions. To ensure that these simulated movements don't unrealistically displace the prostate from its average position within the image, each movement is adjusted using a "de-meaning" process. This process ensures that the simulated motion primarily degrades image quality through blurring or ghosting, rather than causing a significant shift in the prostate's overall location.

The simulated movements are then applied to the original image data, which is resampled to maintain smooth image features despite the introduced motion. Finally, the motion-corrupted image is synthesized by processing the image data in k-space, the frequency domain representation of the MRI signal. This process incorporates the temporal dynamics of the simulated movements, resulting in a realistic depiction of a motion-affected T2-weighted prostate MRI.

1.2. Gaussian noise: simulating inherent acquisition noise

Inherent noise is an unavoidable part of MRI acquisition, creating random fluctuations in signal intensity that can obscure subtle details within prostate images. To make our synthetically degraded prostate images more realistic, we incorporated random Gaussian noise, simulating this inherent noise.

Instead of using fixed noise levels, we varied the amount of noise added to each image. This mimics the range of noise levels encountered in real-world prostate MRI scans. For each image, the parameters controlling the noise distribution were randomly selected, ensuring our model is exposed to a variety of noise profiles during training.

This random noise is added to each individual voxel within the image. The amount of noise added is determined by scaling a random value with a randomly chosen standard deviation and shifting it by a randomly selected mean. This process creates realistic noise patterns, reflecting the noise characteristics inherent to real-world prostate MRI.

1.3. Gaussian blur: simulating loss of sharpness in prostate MRI

Blurring is often caused by factors like imperfect scanner focus or slight patient movements. This loss of sharpness can make it difficult to see fine anatomical details, potentially affecting diagnosis. To prepare our model for this real-world challenge, we introduced Gaussian blur into our synthetic image degradation process.

We didn't apply the same amount of blurring to each image. Instead, we randomly varied the degree of blurring, mimicking the range of sharpness variations seen in real prostate MRIs. This exposes our model to a wider range of blurring artifacts during training, making it more robust to blurry images in real-world scenarios.

The blurring is implemented by convolving each image with a Gaussian filter. The size of this filter, which controls how much blurring is applied, is randomly chosen and scaled based on the resolution of each image. This ensures the blurring effect is appropriately applied relative to the size of the details in the prostate image.

1.4. Ghosting artifacts: simulating periodic motion effects

Ghosting artifacts are often caused by rhythmic motions like blood flow or bowel movement. These artifacts appear as faint copies or "ghosts" of anatomical structures, shifted along a specific direction in the image. To simulate ghosting artifacts, we manipulate the image data in its frequency domain representation, known as k-space.

In k-space, periodic motions like those causing ghosting affect specific frequency bands. We simulate this by selectively suppressing the strength of certain frequencies in k-space. The amount of suppression controls the intensity of the ghosting effect, while the spacing between the suppressed frequencies determines how often the ghosting pattern repeats. By adjusting these parameters, we can create a wide range of ghosting artifacts with varying appearances.

Once the k-space data has been modified to include the simulated ghosting, we transform the data back to its original spatial representation, resulting in a prostate MRI image containing realistic ghosting artifacts.

1.5. Spike artifacts: simulating radio-frequency interference

Spike artifacts, often called herringbone or corduroy artifacts, can create unwanted stripes in MRI images. These artifacts are caused by radio-frequency interference during the scan, which introduces spikes in the k-space representation of the image.

To simulate these artifacts, we directly add spikes to the k-space data of the prostate MRI. Each spike's location in k-space determines the direction and frequency of the stripe pattern that will appear in the final image.

We randomize both the number and location of these spikes to simulate the unpredictable nature of real-world spike artifacts. The number of spikes controls how severe the artifact is, while their random locations create stripes in various directions and positions within the image. The intensity of each spike is also randomized, resulting in stripes with varying prominence. After adding the spikes to the k-space data, we transform the data back to its normal spatial representation, creating a prostate MRI image with realistic spike artifacts.

1.6. Bias field inhomogeneities: simulating magnetic field imperfections

Bias field inhomogeneity is a common artifact in MRI, caused by imperfections in the scanner's magnetic field.[3] This artifact creates gradual changes in image brightness across the prostate, making some areas appear brighter or darker than others, even if the tissues are the same. This can make it harder to distinguish between different tissues and interpret the image accurately.

We simulated this artifact using a mathematical model based on polynomials. This

model can create smooth, gradual changes in image brightness similar to those seen in real bias field artifacts. The model uses a set of coefficients to control the intensity variations, and by randomly generating these coefficients, we create a variety of bias field patterns.

The complexity of these patterns can be adjusted by changing the order of the polynomial used in the model. A higher-order polynomial allows for more intricate and spatially varying intensity changes. The generated bias field pattern is then applied to the original prostate MRI image, simulating the effect of magnetic field imperfections.

1.7. Combining noise types: enhancing training realism

To create a more challenging and realistic training scenario, each image in the training and validation sets was augmented with every type of synthetic degradation individually (motion, Gaussian noise, blur, ghosting, spike artifacts, bias field inhomogeneity). This ensured the model learned to handle each artifact in isolation.

In addition to individual augmentations, combined noise augmentations were also applied to enhance the model's robustness and generalizability. For 20% of the images in the training and validation sets, we randomly selected a combination of two or more of the aforementioned noise types and applied them together to the same image. This approach aimed to simulate the more complex and diverse noise profiles that are representative of real-world clinical scenarios where multiple artifacts can co-occur. The specific combinations of noise types were randomly selected for each image, ensuring a wide variety of combined degradations within the training data.

For the test set, to evaluate the model's ability to generalize to unseen combinations of artifacts, these noise types were combined randomly with varying ratios. This rigorous testing procedure helped assess the model's performance under conditions that more closely reflect real-world clinical prostate MRI.

2. The Pix2Pix model: image-to-image translation for prostate MRI enhancement

To address the challenge of improving the quality of degraded T2-weighted prostate MRI images, we employed the Pix2Pix model, a conditional Generative Adversarial Network (cGAN) renowned for its efficacy in image-to-image translation tasks.[4] Unlike Cycle-GAN, which relies on unpaired data from two different domains, Pix2Pix leverages paired images for training. This makes it particularly suitable for our study, where we have access to ground truth data in the form of original, high-quality images corresponding to the synthetically degraded images.

The Pix2Pix architecture comprises two key components: a generator and a discriminator, which are simultaneously trained in an adversarial manner. The generator, structured as a U-Net,[5] takes the degraded T2-weighted image as input and endeavors to generate a high-quality image that closely resembles the original, artifact-free image. The U-Net architecture, with its encoder-decoder structure and skip connections, is particularly advantageous in medical image processing. The encoder progressively downsamples the input image, extracting features at multiple scales, while the decoder upsamples the encoded representation, reconstructing the output image. The skip connections, linking corresponding encoder and decoder layers, facilitate the direct flow of low-level information across the network, preserving crucial anatomical details and preventing excessive blurring often associated with traditional encoder-decoder networks.

The discriminator, on the other hand, employs a PatchGAN classifier,[4] which evaluates individual N x N image patches rather than the entire image at once. This patch-based approach encourages the generator to focus on generating realistic high-frequency details crucial for maintaining the clarity and realism of medical images. During training, the generator and discriminator are engaged in a continuous adversarial loop. The generator strives to produce increasingly realistic images to deceive the discriminator, while the discriminator becomes more adept at discerning real images from the synthetic images generated by the generator. The proposed Pix2Pix architecture is given in Figure 1.

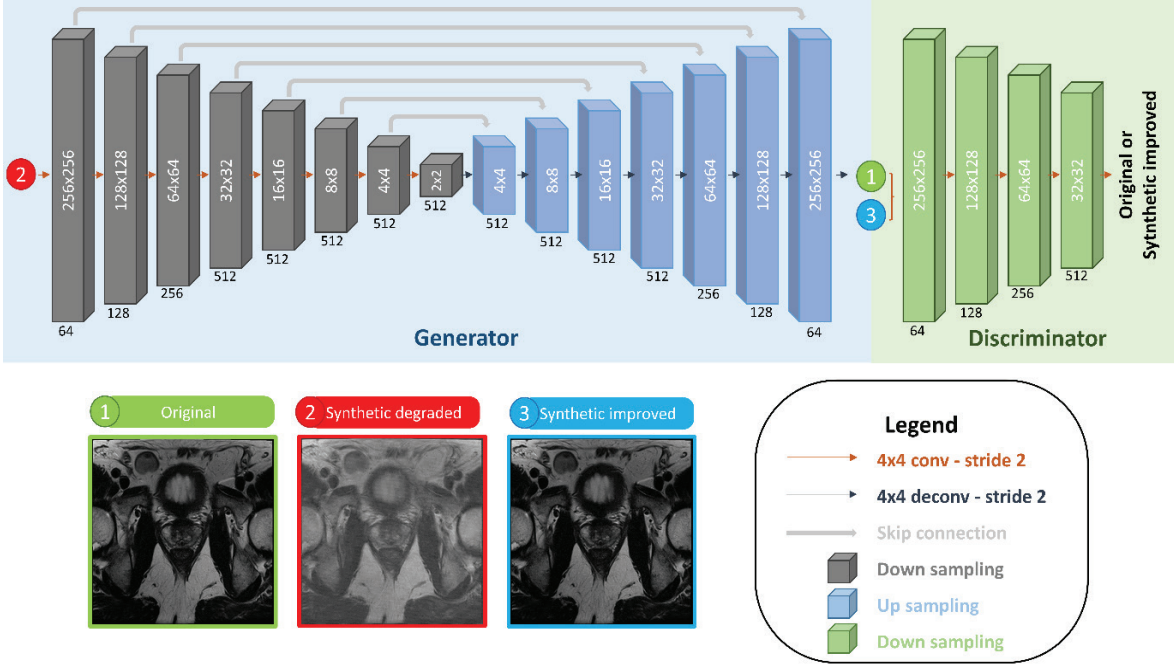


Figure 1. The proposed Pix2Pix architecture in current work.

A U-Net architecture represented by gray and blue blocks generate a synthetically improved image in generator. PatchGAN architecture shown in green blocks evaluate image patches to differentiate between original and synthetic images in discriminator and promotes high-frequency detail generation in the synthetic images. The image examples at the bottom of Figure 1. showcase (1) the original image, (2) the synthetically degraded image, and (3) the synthetically improved image respectively. Moreover, the numbers on the blocks denote the number of feature maps at each layer, and the image dimensions are displayed above each block. Each convolution operations, skip connections, down and up sampling operations are attended to corresponding arrows and blocks with specified colors under legend section in Figure 1.

This adversarial process is guided by a combined loss function encompassing both an adversarial loss and an L1 loss. The adversarial loss, determined by the discriminator's ability to classify image patches as real or fake, ensures that the generated images are visually indistinguishable from the real images in the training dataset. The L1 loss, calculated as the mean absolute error between the generated image and the ground truth target image, promotes structural similarity, preventing excessive blurring and preserving the anatomical integrity crucial for accurate diagnosis.

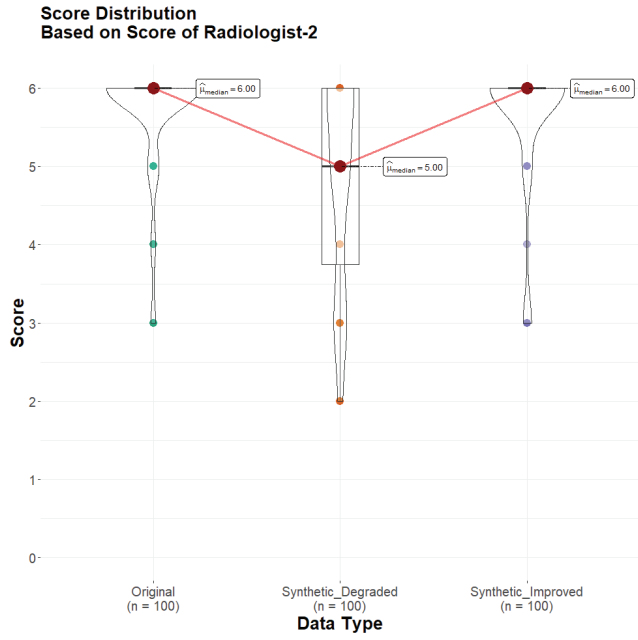
We optimized the Pix2Pix model using the Adam optimizer with a learning rate of 0.0002 for both the generator and the discriminator. The model was trained for 200 epochs, with a heavier emphasis placed on the L1 loss during the early stages of training. This prioritizes the generation of structurally accurate images over purely visually realistic ones, particularly crucial in the context of medical imaging, where diagnostic accuracy is paramount.

References

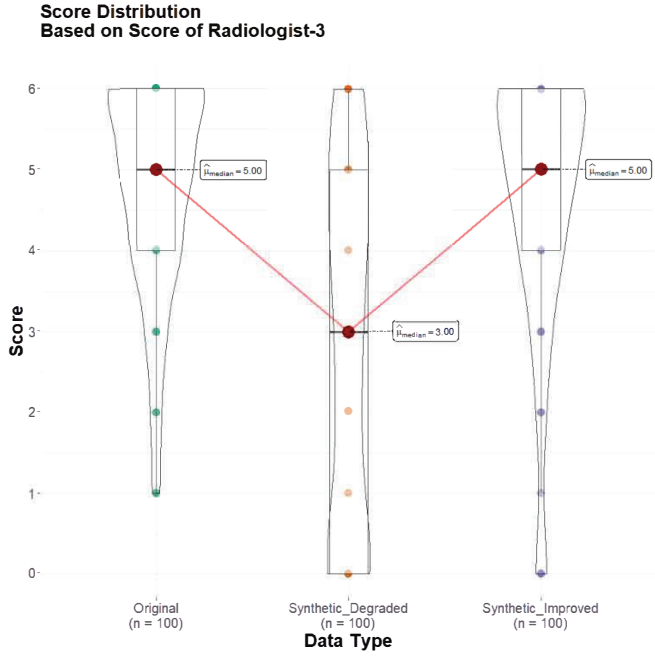
1. Pérez-García F, Sparks R, Ourselin S. TorchIO: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput Methods Programs Biomed. 2021;208:106236. [CrossRef]
2. Shaw R, Sudre C, Ourselin S, Cardoso MJ. MRI k-Space motion artefact augmentation: model robustness and task-specific uncertainty, in: Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning. PMLR. 2019:427-436. Accessed July 1, 2024. [CrossRef]
3. Van Leemput K, Maes F, Vandermeulen D, Suetens P. Automated model-based bias field correction of MR images of the brain. IEEE Trans Med Imaging. 1999;18(10):885-896. [CrossRef]
4. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. 2018. [CrossRef]
5. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015:234-241. [CrossRef]

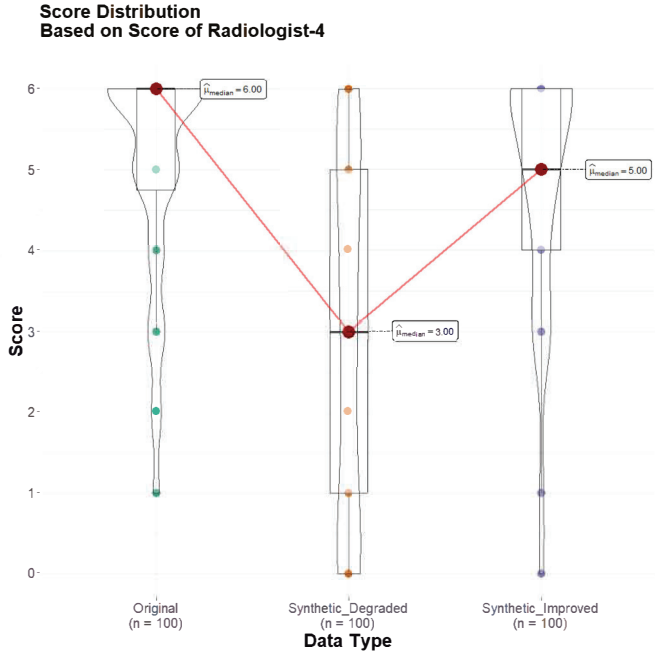
Supplementary Document S2

Comparison of radiologists' score for original, synthetic degraded, and synthetic improved prostate MRI images.

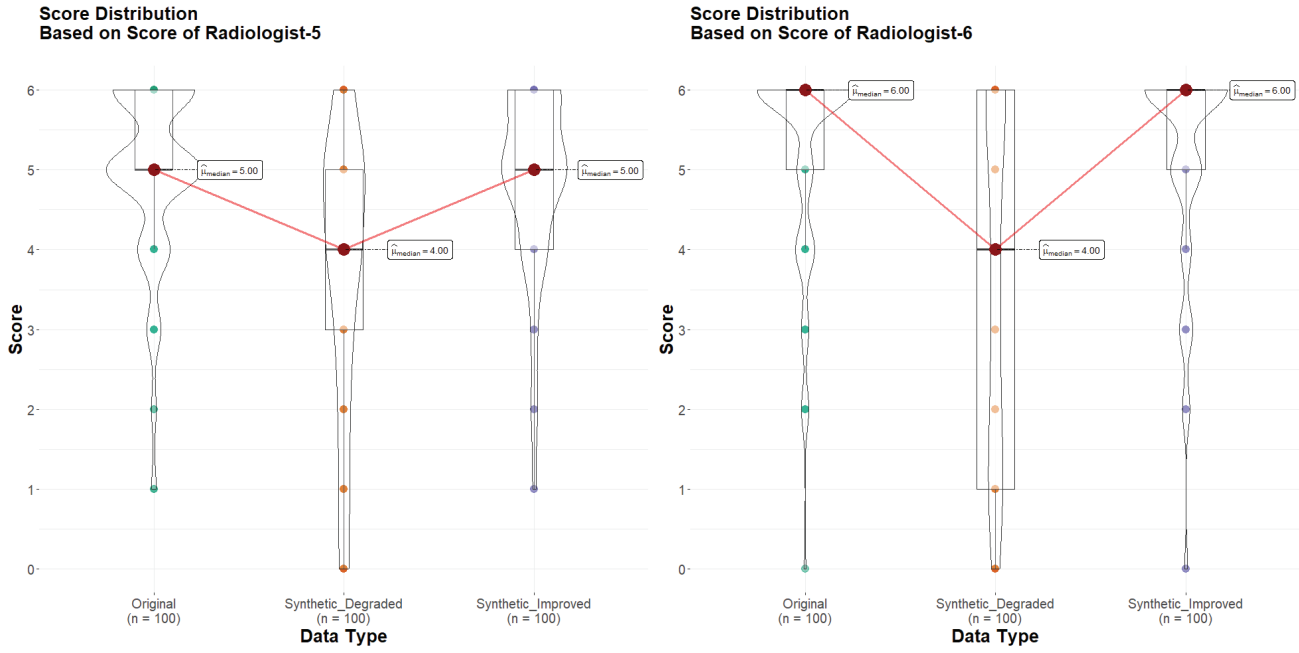


Comparison of radiologists' score for original, synthetic degraded, and synthetic improved prostate MRI images.

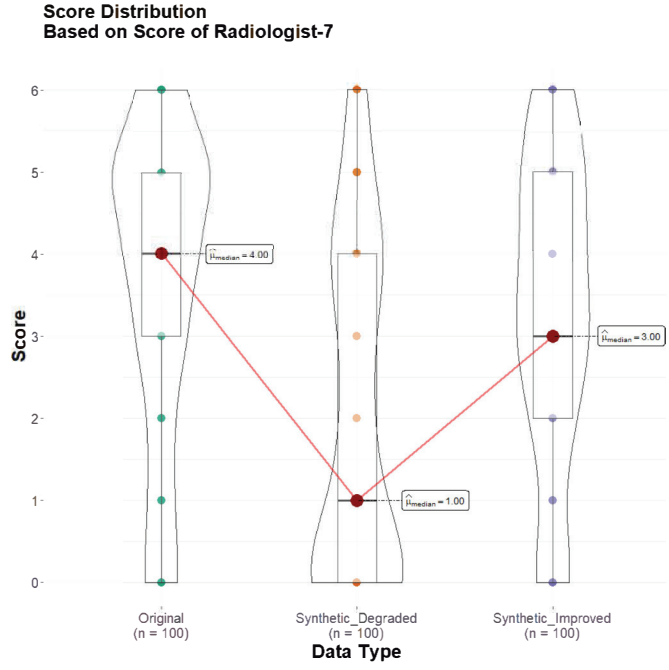




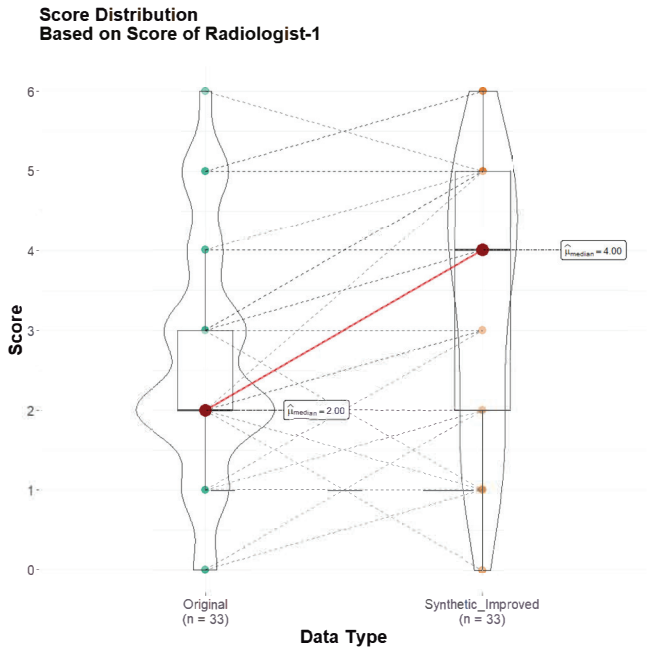
Comparison of radiologists' score for original, synthetic degraded, and synthetic improved prostate MRI images.

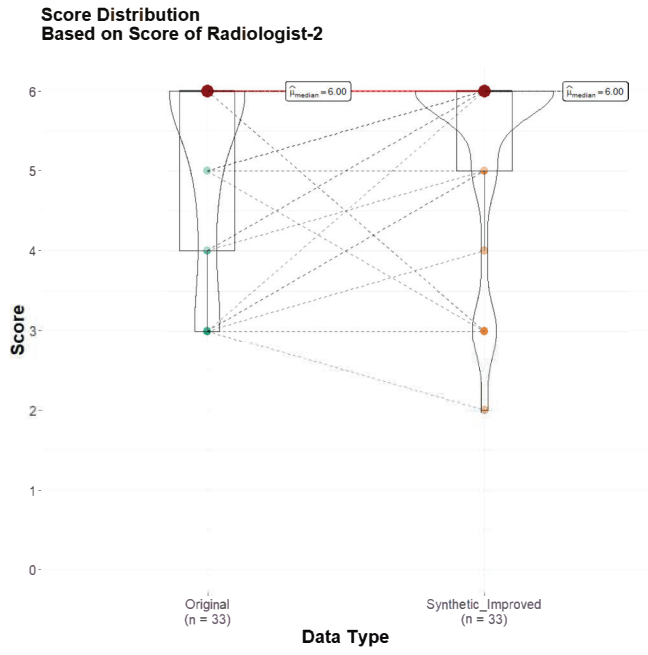


Comparison of radiologists' score for original, synthetic degraded, and synthetic improved prostate MRI images.

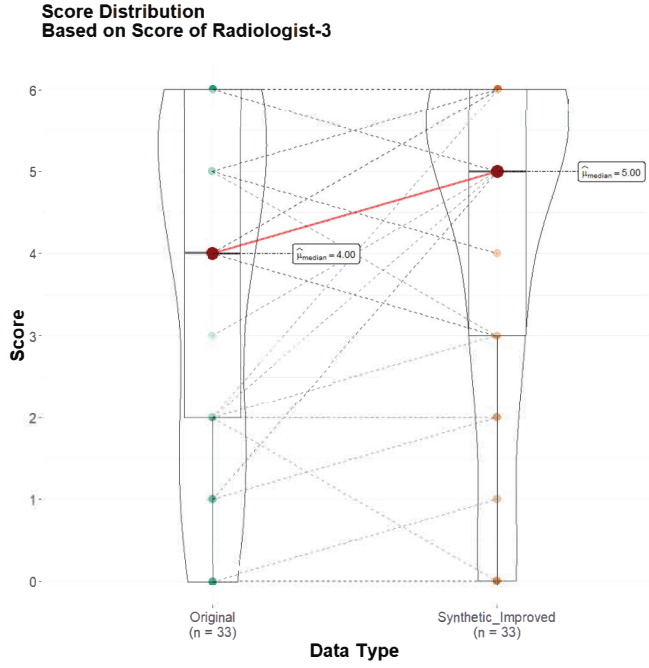


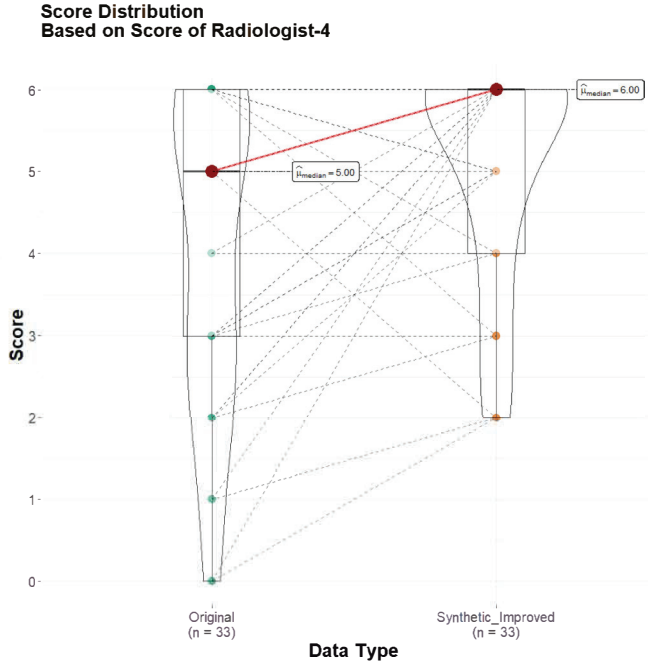
Comparison of radiologists' score for original and synthetic improved prostate MRI images from external test set.



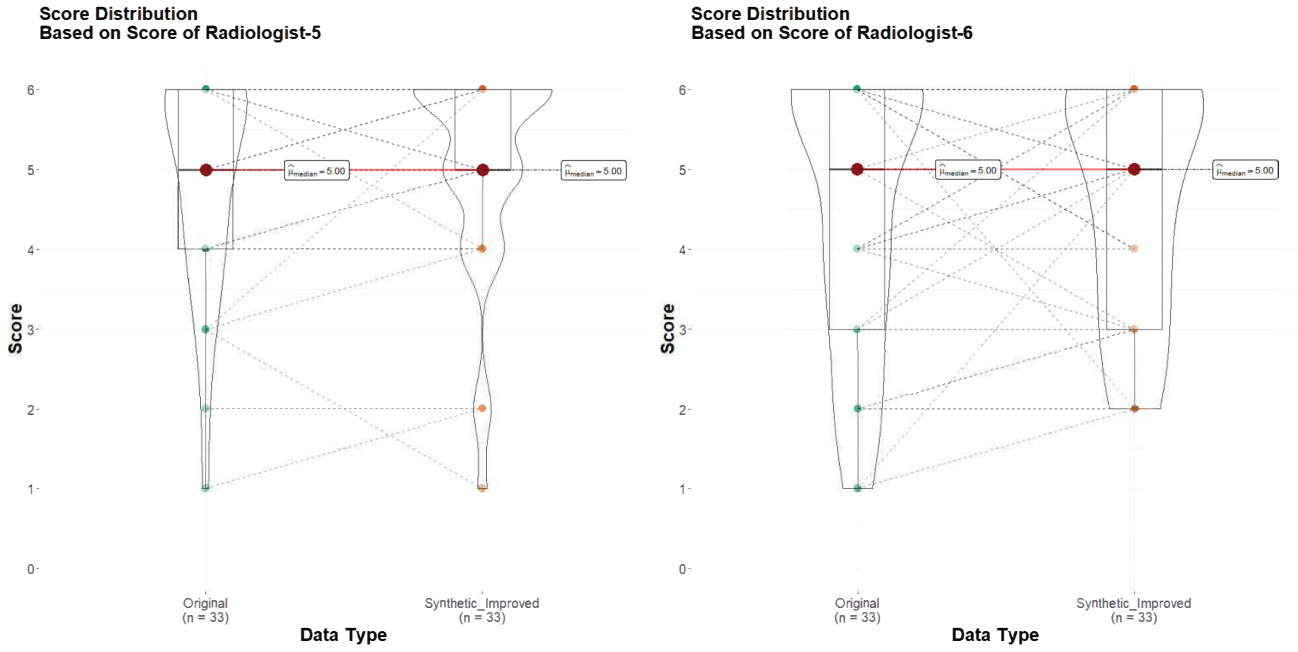


Comparison of radiologists' score for original and synthetic improved prostate MRI images from external test set.

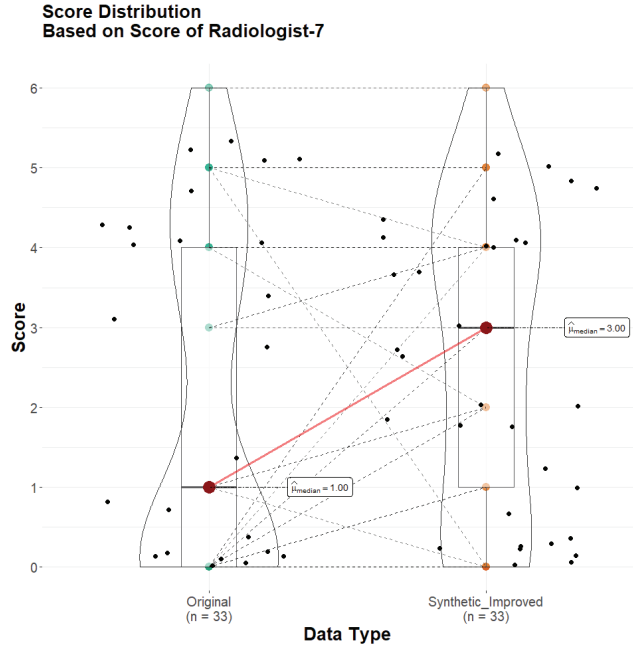




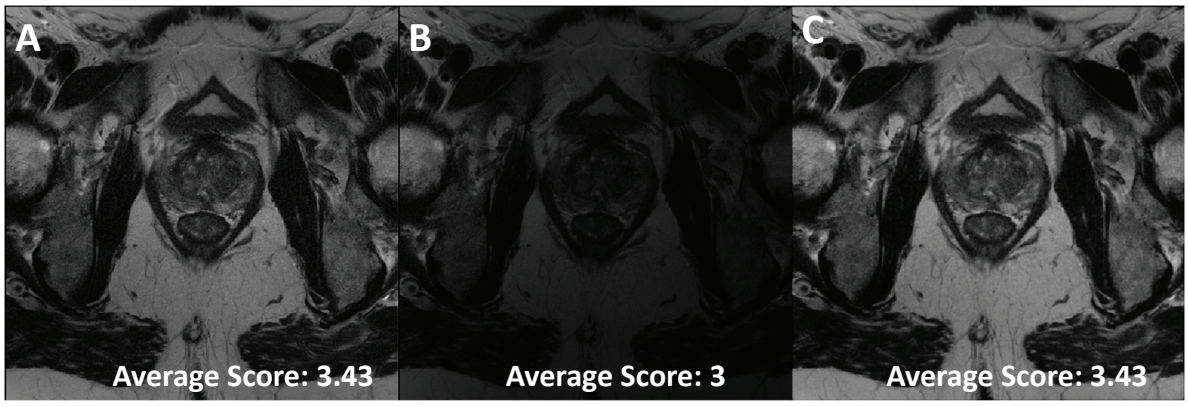
Comparison of radiologists' score for original and synthetic improved prostate MRI images from external test set.



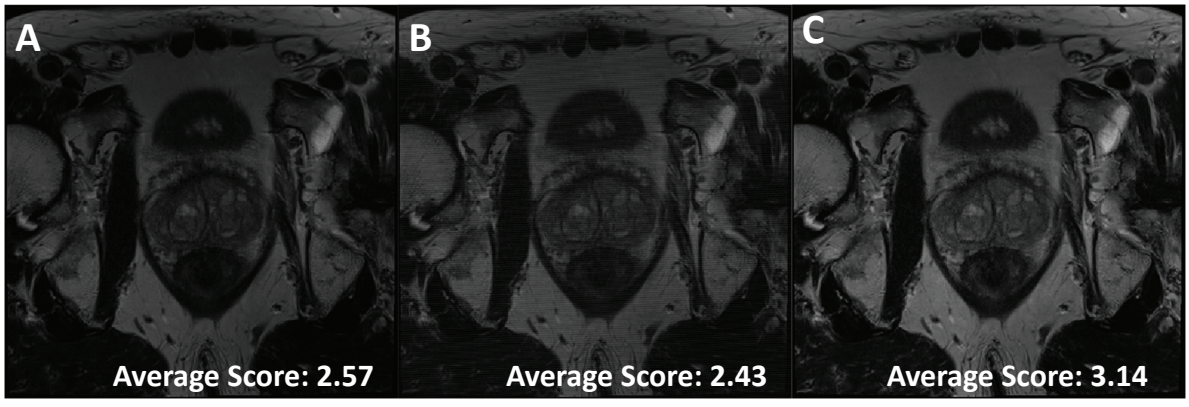
Comparison of radiologists' score for original and synthetic improved prostate MRI images from external test set.



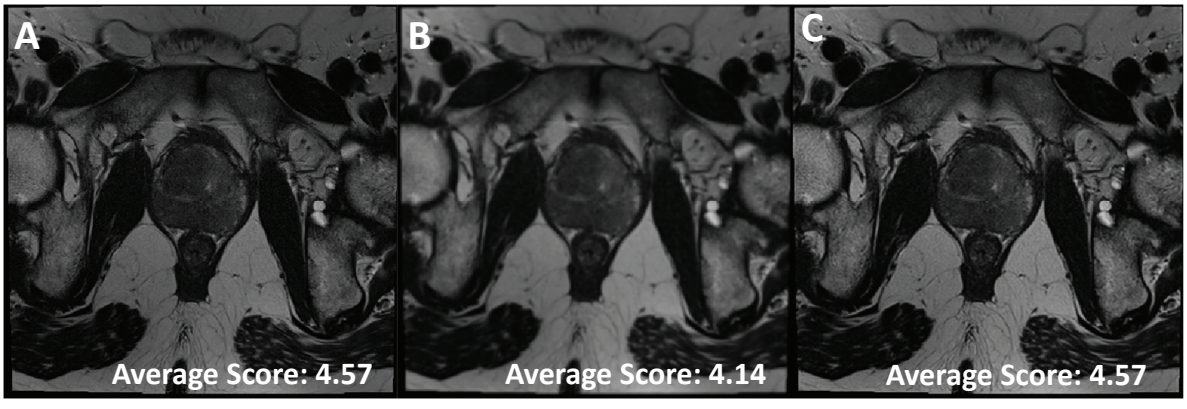
A representative prostate MRI scan (A) original image, (B) synthetic degraded image, (C) synthetic improved image.



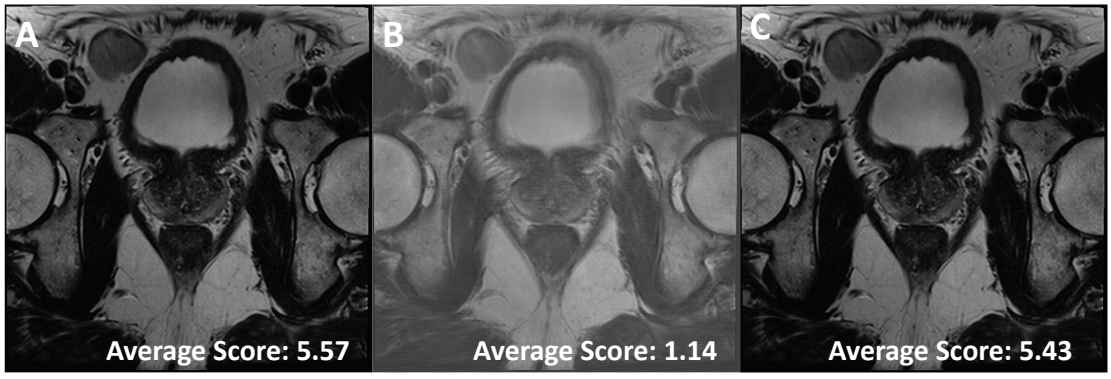
A representative prostate MRI scan (A) original image, (B) synthetic degraded image, (C) synthetic improved image.



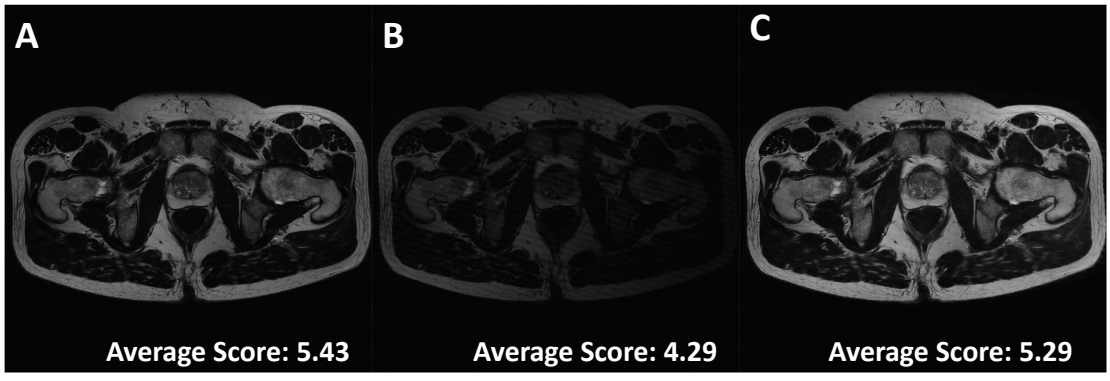
A representative prostate MRI scan (A) original image, (B) synthetic degraded image, (C) synthetic improved image.



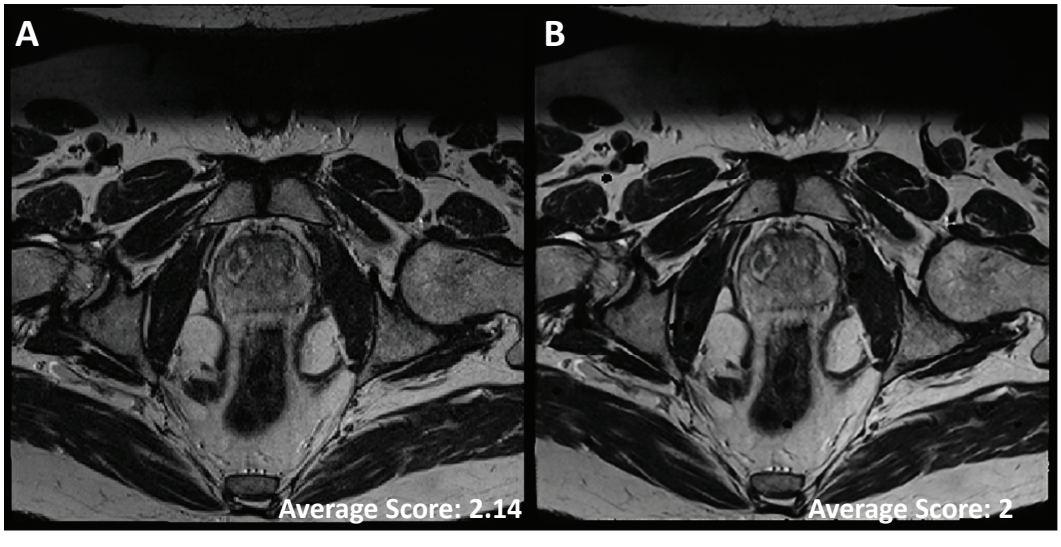
A representative prostate MRI scan (A) original image, (B) synthetic degraded image, (C) synthetic improved image.



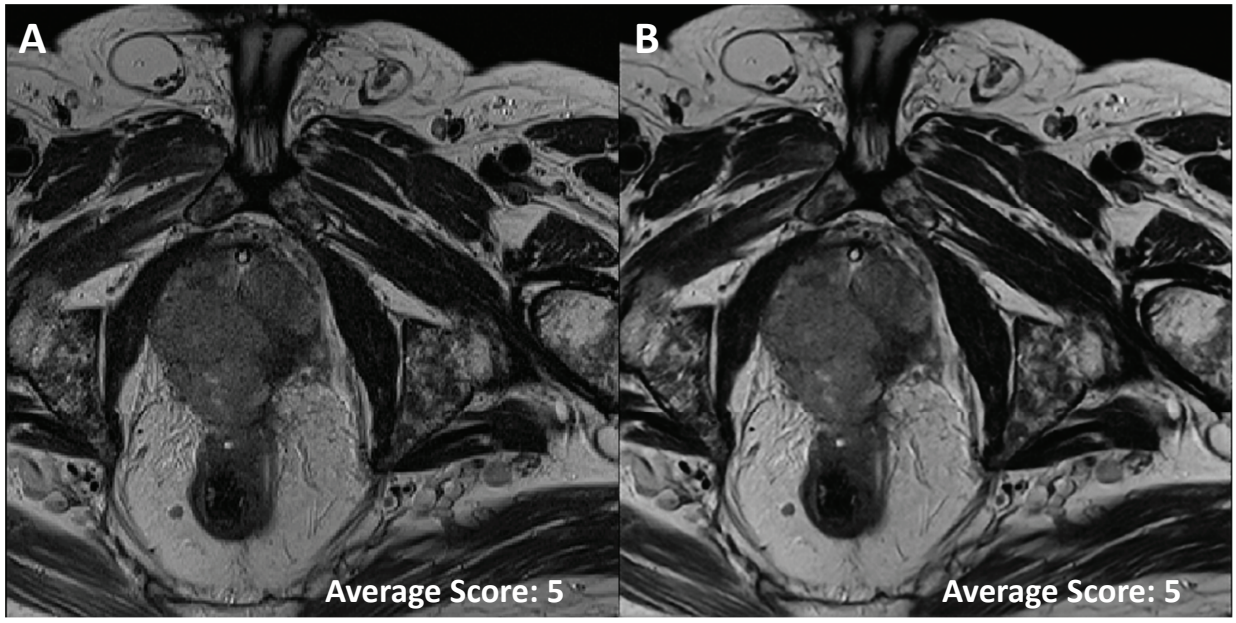
A representative prostate MRI scan (A) original image, (B) synthetic degraded image, (C) synthetic improved image.



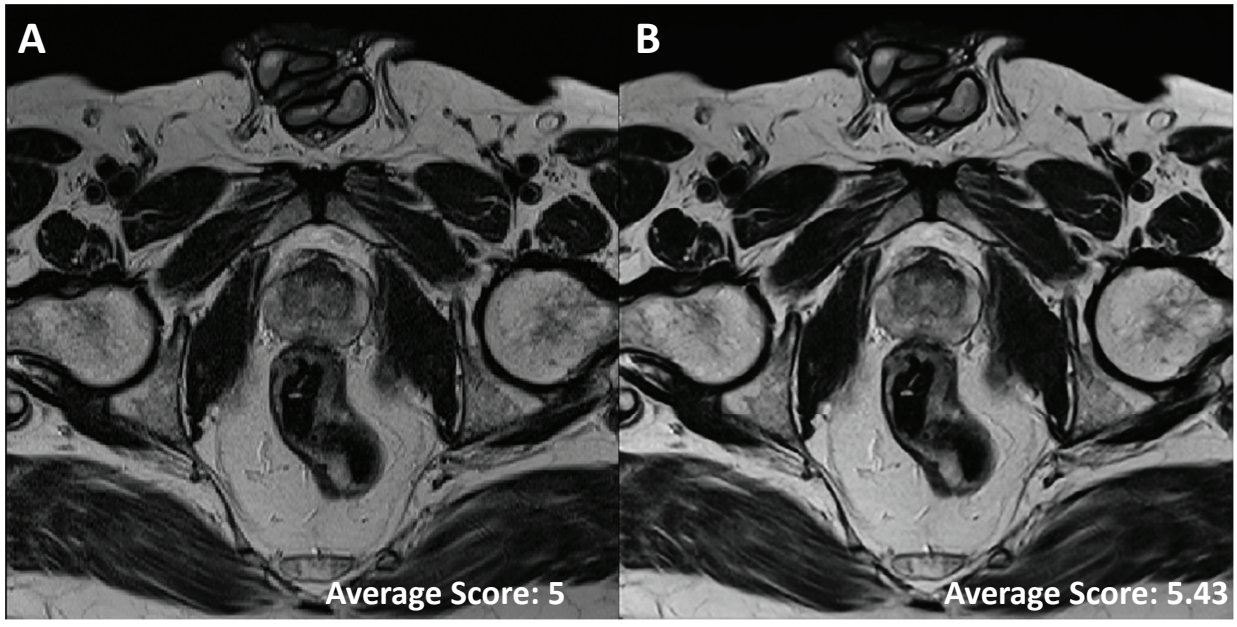
A representative prostate MRI scan from external test set (A) original image, (B) synthetic improved image.



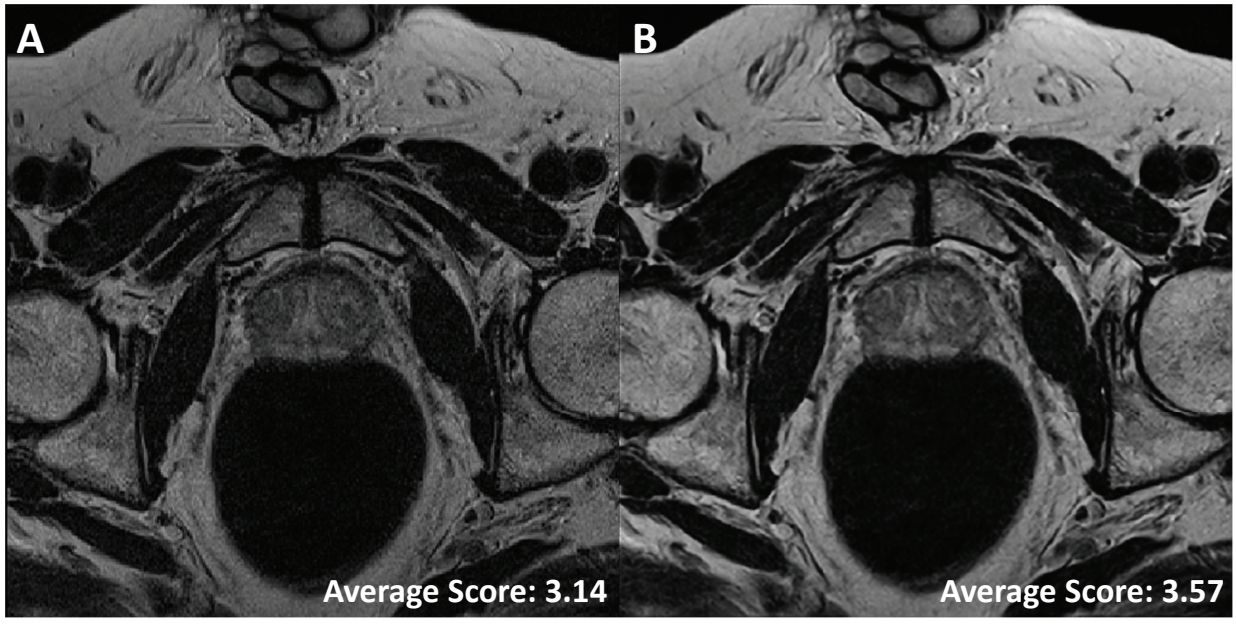
A representative prostate MRI scan from external test set (A) original image, (B) synthetic improved image.



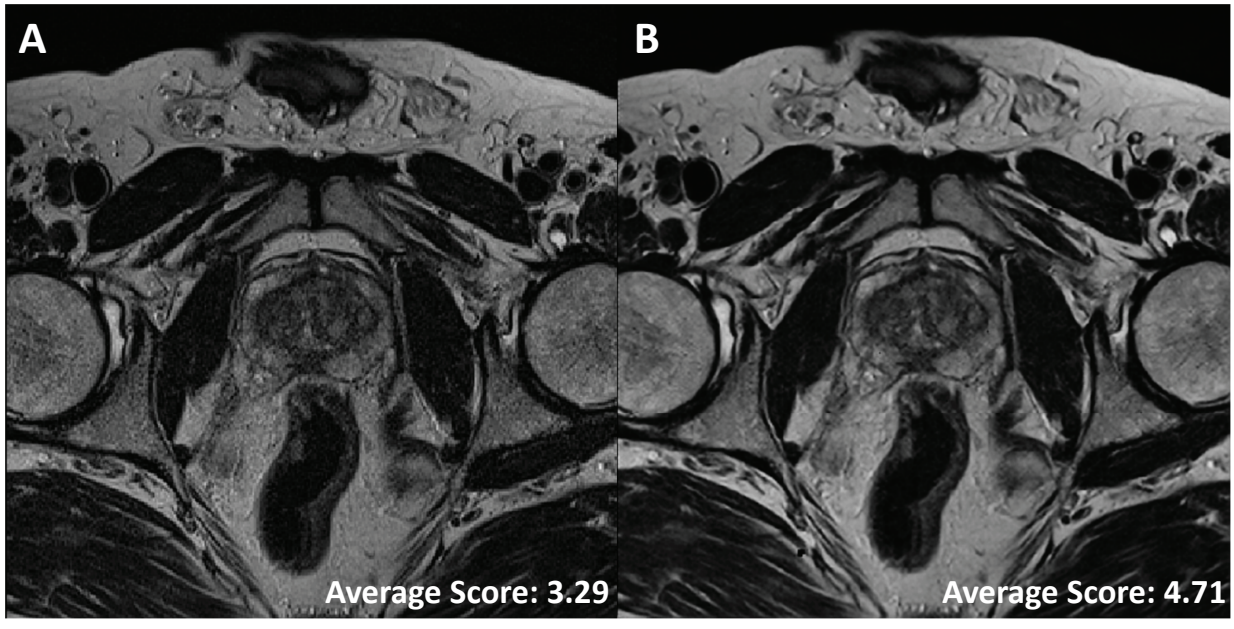
A representative prostate MRI scan from external test set (A) original image, (B) synthetic improved image.



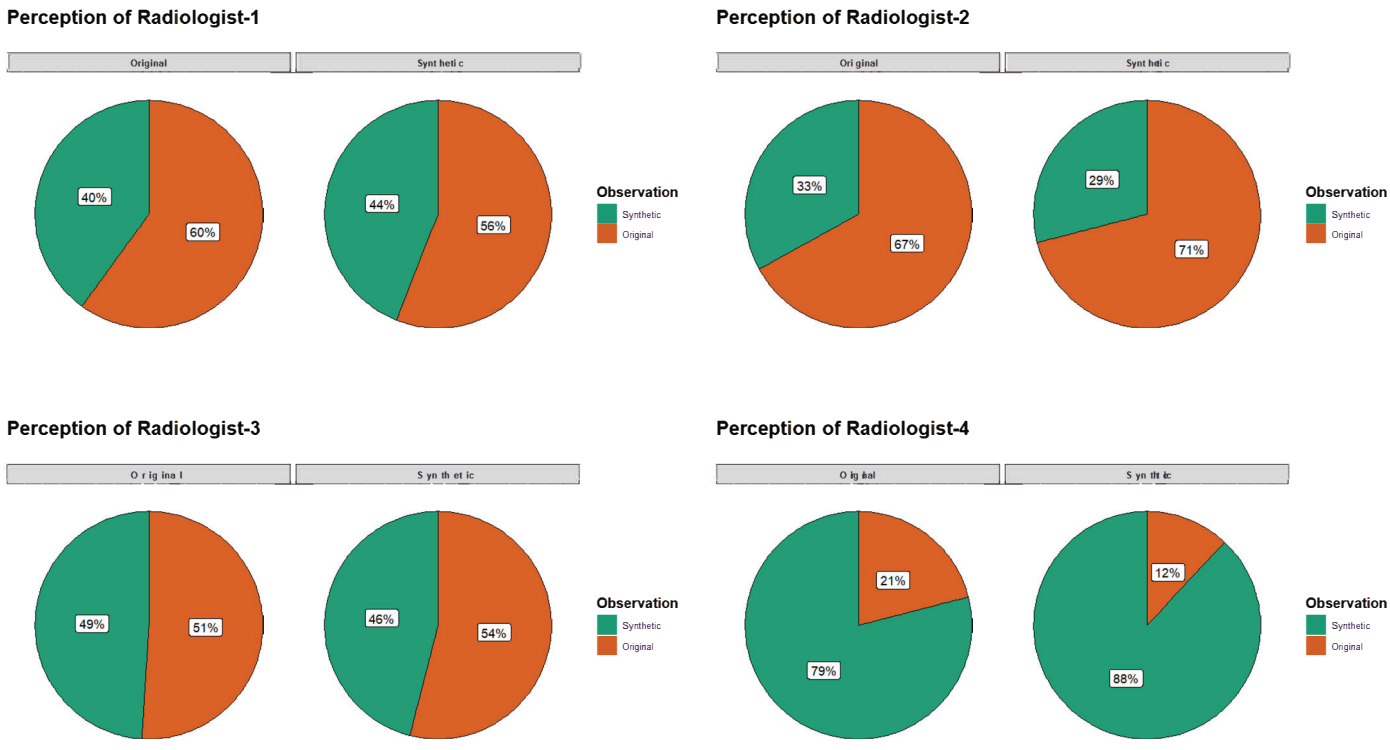
A representative prostate MRI scan from external test set (A) original image, (B) synthetic improved image.



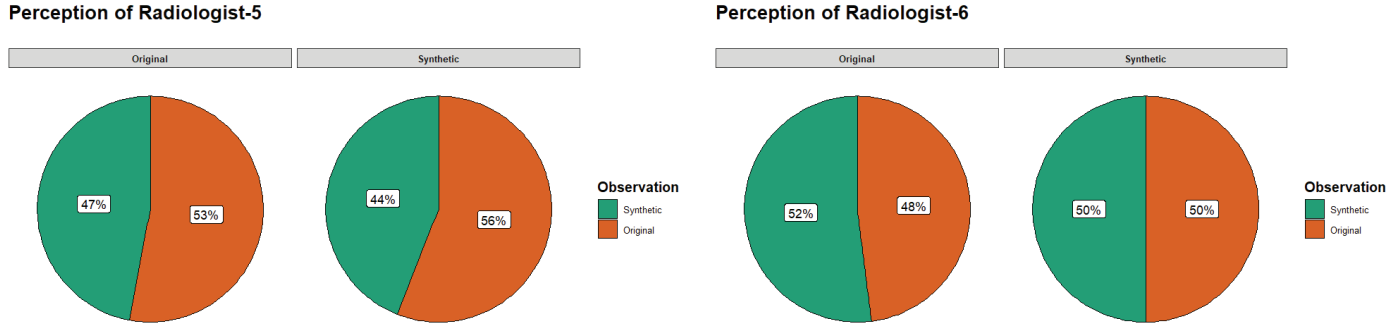
A representative prostate MRI scan from external test set (A) original image, (B) synthetic improved image.



Comparison of radiologists' assessment on whether the image is original or synthetic.



Comparison of radiologists' assessment on whether the image is original or synthetic.



Comparison of radiologists' assessment on whether the image is original or synthetic.