Task-Specific Prostate MRI Quality Assessment via Downstream Performance and Feature-Space Novelty

Yu Long[1], Alexander Ng[2], Weixi Yi[1], Pawel Rajwa[2], Natasha Thorley[3], Francesco Giganti[2,5], Aqua Asif[4,2], Shonit Punwani[5,3], Veeru Kasivisvanathan[2,6], Yipeng Hu[1], and Yucheng Tang[1]

[1] UCL Hawkes Institute and Department of Medical Physics and Biomedical Engineering, University College London, UK; yucheng.tang.24@ucl.ac.uk
[2] Division of Surgery and Interventional Science, University College London, UK
[3] Centre for Medical Imaging, University College London, UK
[4] British Urology Researchers in Surgical Training (BURST), UK
[5] Department of Radiology, University College London Hospitals NHS Foundation Trust, UK
[6] Department of Urology, University College London Hospitals NHS Foundation Trust, UK

Abstract. Clinically adequate prostate MRI examinations are abundant, whereas low-quality cases are scarce, diverse, and difficult to label consistently. Visual quality assessment remains valuable, but it does not directly quantify whether a scan supports a particular automated task. We propose a task-specific IQA framework that derives reject, caution, and accept labels from prostate-segmentation performance using patient-level out-of-fold predictions, and uses feature-space novelty to flag unfamiliar cases. We train a 3D DenseNet121 to predict ternary quality labels from high-b-value DWI, ADC, and T2WI. In 213 locked-test patients, the ternary model achieved accuracy 0.577, macro F1 0.545, quadratic-weighted kappa 0.439, and ordinal mean absolute error 0.479. Predicted quality groups showed increasing mean segmentation Dice from reject to caution and accept (0.885, 0.912, and 0.935), while the continuous expected quality score correlated with Dice ($\rho = 0.505$, $p < 0.001$). Mahalanobis novelty distance correlated negatively with Dice ($\rho = -0.322$, $p < 0.001$). Independently trained diagnostic models achieved test AUCs of 0.699 for PI-RADS $\geq$ 4 and 0.680 for Gleason grade group $\geq$ 2. Transfer of prostate-derived quality to diagnostic correctness was not evident for PI-RADS and weak for Gleason classification, indicating that quality is task-specific rather than universal. These results support downstream-performance-based quality triage and motivate clinically calibrated thresholds with multiple task-specific quality heads.

Keywords: Prostate MRI · task amenability · uncertainty and novelty · deep learning. · cancer detection

1 Introduction

Prostate magnetic resonance imaging (MRI) is central to prostate cancer diagnosis, characterization and treatment planning. T2-weighted imaging (T2WI) provides anatomical context, whereas diffusion-weighted imaging (DWI), together with derived apparent diffusion coefficient (ADC) maps, provides functional information for identifying clinically significant cancer. However, prostate DWI is vulnerable to susceptibility artefacts from rectal air, metallic implants, patient motion and geometric distortion, which can misalign DWI with T2WI and reduce downstream diagnostic reliability [6, 20].

These vulnerabilities make systematic prostate MRI image quality assessment (IQA) essential, not only for identifying visually degraded scans but also for providing complementary evidence on whether an examination remains reliable for downstream diagnostic tasks such as PI-RADS [20], Gleason grade group [7] classification and prostate segmentation. The PI-QUAL [8, 11] system provides radiological quality assessment [4, 1], while automated prostate MRI IQA methods can scale quality control [4, 1]. However, visual labels depend on reader experience and may not quantify how quality affects a particular task. Conversely, a scan with a visible imperfection may remain sufficient for one task but not another. Task amenability therefore defines quality operationally as the degree to which an image supports a specified downstream model or decision [18, 21].

This formulation creates a practical supervision problem: adequate scans dominate many real-world clinical image sets, whereas failed scans are scarce and heterogeneous. A model may therefore encounter clinically plausible cases that are poorly represented by its development data. Novelty detection provides a general paradigm for flagging such unfamiliar inputs for additional review. In this study, we investigate feature-space novelty as an implementation of this paradigm by measuring departure from the training feature distribution [13]. Recent work has also shown that prostate MRI quality distributions affect downstream learning and diagnostic performance [19].

To address this gap in task-specific amenability assessment, we propose a task-specific prostate MRI IQA framework. Prostate segmentation provides the primary anatomical quality criterion using expert-annotated prostate masks as reference. The framework predicts reject, caution, and accept categories and a continuous ordinal score. Softmax confidence and entropy describe ambiguity among categories, whereas Mahalanobis distance in the penultimate feature space measures novelty. PI-RADS [20] ≥ 4 and Gleason grade group [7] ≥ 2 classifiers provide secondary tests of whether prostate-segmentation-derived quality transfers to diagnostic tasks.

The contributions are as follows. First, we propose a patient-level out-of-fold strategy to generate segmentation-derived quality labels and use these labels to train a ternary ordinal IQA model. Second, we incorporate uncertainty measures and feature-space novelty to characterize ambiguity among quality categories and identify cases that depart from the training feature distribution. Third, using a real-world multi-institutional biopsy-trial dataset with prostate masks and multiple downstream labels, we evaluate quality classification, task amenability,

uncertainty, and novelty, and assess whether segmentation-derived quality relates to PI-RADS and Gleason grade group classification. Codes are available at `https://github.com/YuLong111/prostate_mri_task_iqa`.

2 Methods

Task-Specific Quality Definition and IQA Model Training. Patients were first split at the patient level into training, development-validation, and locked-test sets. For each multimodal MRI input x_i, the IQA model predicts an ordinal quality label $y_i \in \{0, 1, 2\}$, denoting reject, caution, and accept. These labels must not be generated from a segmentation model that has already been trained on the same patient, because memorization could make the resulting labels optimistically biased. We therefore split only the training patients into five patient-held-out label-generation folds. For each fold, a U-Net prostate-segmentation model [3] was trained on the other four folds and used only on the held-out patients. The models received high-b-value DWI, ADC, and T2WI concatenated as three input channels. Concatenating the five held-out prediction sets produced one segmentation result per eligible training patient from a model that had never been trained on that patient. The resulting Dice and HD95 values were converted to quality labels as follows:

$$y_i = \begin{cases} 2, & \text{Dice}_i \geq 0.93 \text{ and } \text{HD95}_i \leq 3.5, \\ 0, & \text{Dice}_i < 0.88 \text{ or } \text{HD95}_i \geq 5.0, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

The thresholds in Eq. 1 were selected from the empirical distribution of five-fold out-of-fold prostate-segmentation performance in the training set. Each training patient was assigned a quality label using a segmentation model that had not seen that patient during training, reducing label-generation bias. Development-validation and locked-test patients were excluded from out-of-fold label generation. For these patients, Dice and HD95 were computed from predictions made by a final U-Net segmentation model that used the same architecture and training protocol as the fold-specific models but was trained on the full training set, and the same thresholds were applied without modification. These thresholds operationally divide cases into three task-amenability categories and are not proposed as universal clinical cut-offs. Future studies may refine them according to the downstream model, patient cohort, intended clinical action, and consequences of unnecessary rejection or missed failure.

Specifically, we use a three-dimensional DenseNet121 [10] as the backbone of the IQA model, with high-b-value DWI, ADC, and T2WI concatenated as three input channels. The model outputs class probabilities $\mathbf{p}_\theta(x_i) = (p_i^0, p_i^1, p_i^2)$ for the ordinal classes $k \in \{0, 1, 2\}$. In addition to the hard class prediction, we compute an expected ordinal quality score as $s_i = \sum_{k=0}^{2} k p_i^k$. Training minimizes cross-entropy loss [9] with an additional ordinal absolute-error term as $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \, |s_i - y_i|$, where λ controls the weight of the ordinal term. This objective

encourages the expected quality score to remain close to the ordinal target, penalizing distant ordinal errors more strongly than adjacent errors.

Uncertainty and Feature-Space Novelty. For the softmax output $\mathbf{p}_\theta(x_i)$, we compute predictive confidence and entropy as $C(x_i) = \max_k p_i^k$ and $H(x_i) = -\sum_{k=0}^{2} p_i^k \log p_i^k$, respectively. These quantities capture ambiguity among the learned quality classes, but they may not identify inputs that are unfamiliar to the model. We therefore extracted the penultimate feature representation $z_i = h_\theta(x_i)$ and fitted a regularized Gaussian distribution to the training features with empirical mean and regularized covariance as $\mu = \frac{1}{N}\sum_{j=1}^{N} z_j$ and $\Sigma_{\text{reg}} = \hat{\Sigma} + \lambda_{\text{reg}} I$, respectively, where I is the $d \times d$ identity matrix, and d is the dimensionality of the penultimate feature representation. The empirical covariance matrix $\hat{\Sigma}$ of the training features is calculated as $\hat{\Sigma} = \frac{1}{N-1}\sum_{j=1}^{N}(z_j - \mu)(z_j - \mu)^\top$, where N is the number of training feature vectors. Feature-space novelty was then quantified using the Mahalanobis distance [12] as $D_M(x_i) = \sqrt{(z_i - \mu)^\top \Sigma_{\text{reg}}^{-1}(z_i - \mu)}$. A larger $D_M(x_i)$ indicates greater departure from the training feature distribution and is used as a continuous warning signal rather than a calibrated out-of-distribution diagnosis.

Downstream Evaluation Protocols. To evaluate whether the proposed IQA model captured task-relevant image quality rather than generic visual appearance, we assessed its relationship with downstream prostate MRI tasks on the locked-test set. Specifically, we examined whether IQA predictions, continuous ordinal scores, predictive uncertainty, and feature-space novelty were associated with performance in prostate segmentation and diagnostic classification. The prostate segmentation task used the final U-Net described above and served as the primary task-specific evaluation target. For diagnostic classification, two separate 3D DenseNet121 models, with high-b-value DWI, ADC, and T2WI concatenated as three input channels, were trained using the training set: PI-RADS ≥ 4 vs. < 4 and Gleason grade group ≥ 2 vs. < 2. Both models were optimized using cross-entropy loss. Class imbalance was handled only in the training loader using inverse-frequency weighted random sampling, which increased the sampling probability of patients in the minority class. The decision thresholds were selected on the development-validation set to balance sensitivity and specificity, yielding 0.90 for PI-RADS and 0.35 for Gleason classification; both thresholds were then frozen before locked-test evaluation.

3 Experiments and Locked-Test Results

Dataset. We used a pooled multi-institutional prostate biopsy-trial MRI cohort [15] with real-world clinical imaging and complementary downstream-task annotations. Available data included high-b-value DWI ($b = 2000\ \text{s/mm}^2$), ADC, and T2WI, together with expert annotations and labels, including prostate masks, lesion masks, PI-RADS scores, and binary Gleason grade group labels (≥ 2 vs. < 2). For IQA development, novelty analysis, and prostate segmentation, eligible patients required DWI, ADC, T2WI, and a prostate reference mask,

yielding 1016, 217, and 213 patients in the training, development-validation, and locked-test sets, respectively. Diagnostic evaluation additionally required the corresponding PI-RADS or Gleason grade group label, yielding 933, 198, and 194 eligible patients, respectively. All volumes were preprocessed using SimpleITK [14] and MONAI [2]. Each intensity channel was independently clipped to its non-zero first and 99th percentiles and scaled to $[0, 1]$. Segmentation and IQA inputs were centre-cropped to $160 \times 160 \times 64$ voxels, with zero-padding applied only when a volume dimension was smaller than the target size. For the diagnostic classifiers, the input channels were first cropped to the prostate-mask foreground with a margin of $(16, 16, 8)$ voxels and then cropped or padded to the same spatial size. Training augmentation included random flips, affine transformations, Gaussian noise, intensity scaling, and intensity shifting. Spatial transformations were applied synchronously across modalities and masks. Validation and locked-test preprocessing was deterministic.

Implementation Details. All models were implemented using MONAI [2] and PyTorch [17]. For the prostate-segmentation models used for quality-label generation and downstream segmentation evaluation, all U-Net models used encoder channel widths of $(16, 32, 64, 128, 256)$ and were trained for 20 epochs using AdamW with a learning rate of 10^{-4}, a batch size of 1, and a combined Dice and cross-entropy loss. The ternary IQA model was trained for 30 epochs using AdamW with a learning rate of 10^{-4}, batch size of 1, and ordinal-loss weight $\lambda = 0.5$. For downstream diagnostic evaluation, separate 3D DenseNet121 classifiers were trained for PI-RADS ≥ 4 and Gleason grade group ≥ 2. The PI-RADS and Gleason models were trained for 20 and 15 epochs, respectively, using AdamW with learning rate 10^{-4}, batch size 1 and cross-entropy loss.

Evaluation Metrics. Ternary IQA performance was assessed using accuracy (ACC), macro F1 score (m-F1), quadratic weighted kappa (QWK) [5], ordinal mean absolute error (O-MAE), confusion matrix, and class-wise recall. For downstream task evaluations, prostate segmentation performance was evaluated using Dice coefficient, intersection over union (IoU), average surface distance (ASD), and 95th-percentile Hausdorff distance (HD95) [16]. Diagnostic classification performance was evaluated using area under curve (AUC), balanced accuracy (B-ACC), sensitivity (Sens.), and specificity (Spec.). Where reported, 95% confidence intervals (95CI) were estimated from 5,000 patient-level bootstrap resamples, and groupwise diagnostic accuracies were reported with Wilson 95CI. Associations between IQA-derived measures and downstream outcomes were assessed using Spearman rank correlation [22], reported as Spearman's ρ with exploratory, unadjusted p-values and 95CI where available.

Downstream Evaluation Model Performance. We first evaluated the downstream models on the locked test set to establish reference performance for task-amenability and cross-task transfer analyses. The prostate-segmentation model achieved strong performance across 213 eligible patients (mean Dice 0.919), while the PI-RADS and Gleason classifiers showed moderate discrimination across 194 eligible patients (mean AUC 0.699 and 0.680, respectively), providing case-level variation for evaluating cross-task transfer. These results indicate

Table 1. Locked-test downstream performance stratified by predicted IQA group. Segmentation performance was evaluated using Dice, IoU, ASD, and HD95 in 213 eligible locked-test patients, whereas PI-RADS and Gleason classification performance was evaluated using AUC, B-ACC, Sens., and Spec. in 194 locked-test patients with diagnostic labels. Values are reported as mean (standard deviation).

Task	Group	Dice	IoU	ASD	HD95
Segmentation	Reject ($n = 32$)	0.89 (0.06)	0.80 (0.08)	1.40 (0.56)	4.24 (2.76)
	Caution ($n = 80$)	0.91 (0.03)	0.84 (0.05)	1.14 (0.46)	3.31 (1.52)
	Accept ($n = 101$)	0.94 (0.02)	0.88 (0.03)	1.00 (0.27)	2.94 (0.94)
		AUC	B-ACC	Sens.	Spec.
PI-RADS	Reject ($n = 29$)	0.49 (0.11)	0.42 (0.08)	0.67 (0.11)	0.18 (0.12)
	Caution ($n = 69$)	0.75 (0.06)	0.69 (0.05)	0.74 (0.08)	0.63 (0.08)
	Accept ($n = 96$)	0.67 (0.06)	0.66 (0.05)	0.56 (0.09)	0.75 (0.05)
Gleason	Reject ($n = 29$)	0.70 (0.10)	0.60 (0.09)	0.36 (0.15)	0.83 (0.09)
	Caution ($n = 69$)	0.60 (0.07)	0.59 (0.06)	0.41 (0.08)	0.77 (0.08)
	Accept ($n = 96$)	0.65 (0.06)	0.58 (0.05)	0.91 (0.04)	0.25 (0.08)

that the downstream models were sufficiently functional and variable to support subsequent analysis of image-quality amenability and diagnostic transfer.

Ternary IQA Results. We evaluated whether the ternary IQA model, trained on segmentation-derived labels from the training set, could recover segmentation-derived reject (35 cases), caution (82 cases), and accept (96 cases) labels on the locked test set. The trained ternary IQA model achieved ACC 0.577 [95CI 0.512–0.643], m-F1 0.545 [95CI 0.470–0.616], QWK 0.439 [95CI 0.313–0.552], and O-MAE 0.479 [95CI 0.399–0.559]. Class-wise recall was 0.429 for reject, 0.512 for caution, and 0.688 for accept, with the confusion matrix shown in Fig. 1 (a). Although hard-class separation was moderate, the model preserved the ordinal structure of the task. The expected ordinal score increased from 0.834 in true reject cases to 1.136 in caution cases and 1.384 in accept cases, and predicted quality groups showed a monotonic Dice gradient, as shown in Fig. 1 (b,c). These findings suggest that the proposed task-specific IQA can recover an operationally interpretable ordinal quality signal without direct visual quality labels.

Downstream Task Amenability of IQA Model. We evaluated whether IQA-predicted quality groups reflected downstream task amenability for prostate segmentation, PI-RADS classification, and Gleason grade group classification. Specifically, we tested whether downstream performance followed the expected ordinal pattern, with reject cases performing worst, accept cases best, and caution cases in between. For the defining prostate-segmentation task, this expected pattern was clearly observed. Predicted quality groups stratified segmentation performance in the intended direction, with progressively better Dice and IoU and lower ASD and HD95 from reject to caution to accept, as shown in Table 1 and Fig. 1 (b,c). The continuous IQA-predicted quality score correlated positively with Dice ($\rho = 0.505$, 95CI 0.387–0.606, $p < 0.001$). The Spearman

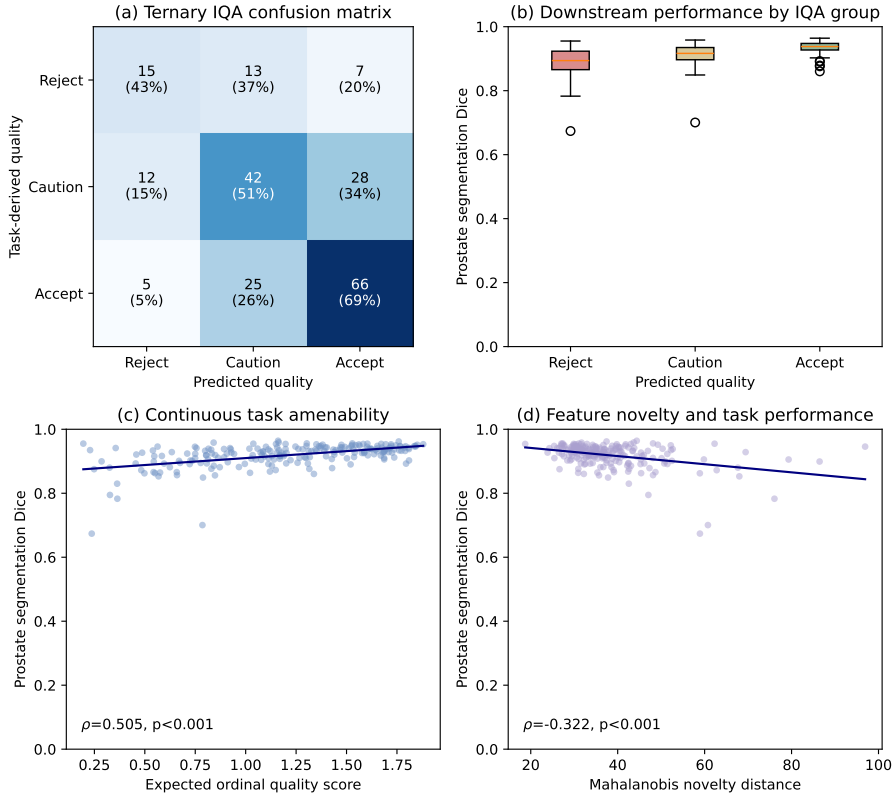

Fig. 1. Locked-test results for the task-specific IQA model ($n = 213$): (a) confusion matrix with row percentages; (b) prostate-segmentation Dice by predicted quality; (c) expected ordinal quality score versus Dice; and (d) Mahalanobis novelty distance versus Dice.

correlation with IoU was identical because case-level Dice and IoU are monotonically related. The quality score correlated negatively with ASD ($\rho = -0.297$, $p < 0.001$) and HD95 ($\rho = -0.208$, $p = 0.002$).

For the two diagnostic tasks, the pattern was less consistent, as shown in Table 1. PI-RADS classification showed a descriptive improvement from reject to the higher-quality groups, but the continuous expected quality score was not significantly associated with PI-RADS correctness ($\rho = 0.092$, 95CI -0.053 - 0.231, $p = 0.200$). Gleason grade group classification showed a weak positive association between expected quality and correctness ($\rho = 0.165$, 95CI 0.038 - 0.299, $p = 0.021$), but the categorical ordering across reject, caution, and accept was not strictly monotonic. We also found that the diagnostic confidence showed a different pattern from correctness: mean confidence was highest in predicted reject cases for both diagnostic classifiers, as shown in Fig. 2(c), suggesting that softmax confidence reflected probability extremity rather than downstream

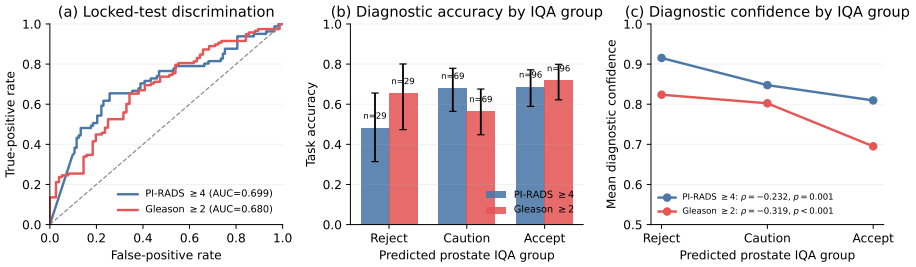

Fig. 2. Locked-test diagnostic-task analysis ($n = 194$): (a) ROC curves for PI-RADS and Gleason classifiers; (b) task accuracy across predicted prostate-IQA groups with Wilson 95% confidence intervals; and (c) mean diagnostic confidence by predicted group, showing that higher confidence did not imply better task amenability.

correctness or task amenability. These findings indicate that the IQA model captured clear task amenability for prostate segmentation, while transfer to PI-RADS and Gleason classification was limited and inconsistent; therefore, the proposed IQA score should be interpreted as task-specific rather than as a universal image-quality scale.

Uncertainty and Novelty. We evaluated whether softmax uncertainty and feature-space novelty provided warning signals beyond predicted quality labels for prostate-segmentation performance. Softmax confidence correlated weakly with Dice ($\rho = 0.205$, 95CI 0.061–0.340, $p = 0.003$), while entropy correlated weakly and negatively ($\rho = -0.171$, 95CI -0.309 to -0.030, $p = 0.013$). Mahalanobis distance correlated negatively with Dice ($\rho = -0.322$, 95CI -0.450 to -0.179, $p < 0.001$) and positively with ASD ($\rho = 0.323$, $p < 0.001$) and HD95 ($\rho = 0.282$, $p < 0.001$). Mean novelty distance was 49.08 in predicted reject cases, 39.28 in caution, and 33.77 in accept. These results suggest that unfamiliar feature representations were associated with worse anatomical task performance, even though novelty was not used to define the quality labels. Feature-space novelty therefore provided a complementary safety signal beyond softmax confidence and entropy.

4 Conclusion

This study demonstrates a task-specific prostate MRI IQA framework that defines image quality by downstream task amenability rather than direct visual ratings. The proposed model recovered an interpretable ordinal quality signal on a locked-test cohort: although three-class classification was moderate, the continuous quality score and predicted reject/caution/accept groups showed a consistent relationship with prostate-segmentation performance. This provides evidence that prostate MRI quality can be operationalized in a model- and task-specific manner, enabling automated identification of cases that are more or less suitable for a defined downstream task. The limited transfer of this prostate-

segmentation-derived quality signal to PI-RADS and Gleason classification highlights an important implication: image quality is not universal, but depends on the clinical task being supported. Feature-space novelty further added a complementary safety signal by identifying unfamiliar cases associated with poorer anatomical performance, beyond softmax confidence alone. Together, these findings support task-specific IQA as a practical foundation for prostate MRI quality control, failure-risk flagging, and future multi-task systems with shared representations but task-specific quality heads.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alis, D., Kartal, M.S., Seker, M.E., Guroz, B., Basar, Y., Arslan, A., Sirolu, S., Kurtcan, S., Denizoglu, N., Tuzun, U., et al.: Deep learning for assessing image quality in bi-parametric prostate mri: A feasibility study. European Journal of Radiology **165**, 110924 (2023)
2. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
4. Cipollari, S., Guarrasi, V., Pecoraro, M., Bicchetti, M., Messina, E., Farina, L., Paci, P., Catalano, C., Panebianco, V.: Convolutional neural networks for automated classification of prostate multiparametric magnetic resonance imaging based on image quality. Journal of Magnetic Resonance Imaging **55**(2), 480–490 (2022)
5. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin **70**(4), 213 (1968)
6. Donato Jr, F., Costa, D.N., Yuan, Q., Rofsky, N.M., Lenkinski, R.E., Pedrosa, I.: Geometric distortion in diffusion-weighted mr imaging of the prostate—contributing factors and strategies for improvement. Academic radiology **21**(6), 817–823 (2014)
7. Epstein, J.I.: An update of the gleason grading system. The Journal of urology **183**(2), 433–440 (2010)
8. Giganti, F., Allen, C., Emberton, M., Moore, C.M., Kasivisvanathan, V., Group, P.S., et al.: Prostate imaging quality (pi-qual): a new quality control scoring system for multiparametric magnetic resonance imaging of the prostate from the precision trial. European urology oncology **3**(5), 615–619 (2020)
9. Ho, Y., Wookey, S.: The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE access **8**, 4806–4813 (2019)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
11. Karanasios, E., Caglic, I., Zawaideh, J.P., Barrett, T.: Prostate mri quality: clinical impact of the pi-qual score in prostate cancer diagnostic work-up. The British Journal of Radiology **95**(1133), 20211372 (2022)

12. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
13. Liu, J., Lou, B., Diallo, M., Meng, T., von Busch, H., Grimm, R., Tian, Y., Comaniciu, D., Kamen, A., Collaborators, P.C., et al.: Detecting out-of-distribution via an unsupervised uncertainty estimation for prostate cancer diagnosis. In: International Conference on Medical Imaging with Deep Learning. pp. 796–807. PMLR (2022)
14. Lowekamp, B.C., Chen, D.T., Ibáñez, L., Blezek, D.: The design of simpleitk. Frontiers in neuroinformatics **7**, 45 (2013)
15. Min, Z., Bianco, F.J., Yang, Q., Yan, W., Shen, Z., Cohen, D., Rodell, R., Barratt, D.C., Hu, Y.: Segmentation versus detection: Development and evaluation of deep learning models for prostate imaging reporting and data system lesions localisation on bi-parametric prostate magnetic resonance imaging. CAAI Transactions on Intelligence Technology **10**(3), 689–702 (2025). https://doi.org/10.1049/cit2.12318
16. Müller, D., Soto-Rey, I., Kramer, F.: Towards a guideline for evaluation metrics in medical image segmentation. BMC research notes **15**(1), 210 (2022)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Saeed, S.U., Fu, Y., Stavrinides, V., Baum, Z.M., Yang, Q., Rusu, M., Fan, R.E., Sonn, G.A., Noble, J.A., Barratt, D.C., et al.: Image quality assessment for machine learning tasks using meta-reinforcement learning. Medical Image Analysis **78**, 102427 (2022)
19. Tang, Y., Rajwa, P., Ng, A., Wang, Y., Yan, W., Thorley, N., Asif, A., Allen, C., Dickinson, L., Giganti, F., et al.: Impact of clinical image quality on efficient foundation model finetuning. In: International Workshop on Efficient Medical Artificial Intelligence. pp. 194–204. Springer (2025)
20. Turkbey, B., Rosenkrantz, A.B., Haider, M.A., Padhani, A.R., Villeirs, G., Macura, K.J., Tempany, C.M., Choyke, P.L., Cornud, F., Margolis, D.J., et al.: Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. European urology **76**(3), 340–351 (2019)
21. Wang, Y., Yang, Q., Drukker, L., Papageorghiou, A., Hu, Y., Noble, J.A.: Task model-specific operator skill assessment in routine fetal ultrasound scanning. International Journal of Computer Assisted Radiology and Surgery **17**(8), 1437–1444 (2022)
22. Zar, J.H.: Spearman rank correlation. Encyclopedia of biostatistics **7** (2005)