Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-025-01504-8

# Challenges in Using Deep Neural Networks Across Multiple Readers in Delineating Prostate Gland Anatomy

**Shatha Abudalou[1,2] · Jung Choi[3] · Kenneth Gage[3] · Julio Pow-Sang[4] · Yasin Yilmaz[2] · Yoganand Balagurunathan[1,2,3,4]**

Received: 26 December 2024 / Revised: 28 March 2025 / Accepted: 12 April 2025


## Abstract

Deep learning methods provide enormous promise in automating manually intense tasks such as medical image segmentation and provide workflow assistance to clinical experts. Deep neural networks (DNN) require a significant amount of training examples and a variety of expert opinions to capture the nuances and the context, a challenging proposition in oncological studies (H. Wang *et al., Nature*, vol. 620, no. 7972, pp. 47-60, Aug 2023). Inter-reader variability among clinical experts is a real-world problem that severely impacts the generalization of DNN reproducibility. This study proposes quantifying the variability in DNN performance using expert opinions and exploring strategies to train the network and adapt between expert opinions. We address the inter-reader variability problem in the context of prostate gland segmentation using a well-studied DNN, the 3D U-Net model. Reference data includes magnetic resonance imaging (MRI, T2-weighted) with prostate glandular anatomy annotations from two expert readers (R#1, n = 342 and R#2, n = 204). 3D U-Net was trained and tested with individual expert examples (R#1 and R#2) and had an average Dice coefficient of 0.825 (CI, [0.81 0.84]) and 0.85 (CI, [0.82 0.88]), respectively. Combined training with a representative cohort proportion (R#1, n = 100 and R#2, n = 150) yielded enhanced model reproducibility across readers, achieving an average test Dice coefficient of 0.863 (CI, [0.85 0.87]) for R#1 and 0.869 (CI, [0.87 0.88]) for R#2. We re-evaluated the model performance across the gland volumes (large, small) and found improved performance for large gland size with an average Dice coefficient to be at 0.846 [CI, 0.82 0.87] and 0.872 [CI, 0.86 0.89] for R#1 and R#2, respectively, estimated using fivefold cross-validation. Performance for small gland sizes diminished with average Dice of 0.8 [0.79, 0.82] and 0.8 [0.79, 0.83] for R#1 and R#2, respectively.

**Keywords** Reproducibility of deep network · Prostate gland segmentation · Multi-reader variability

## Introduction

Prostate cancer is the second most prevalent cancer among men in the USA [1] and the fourth most common cancer worldwide, with over 18.1 million new cases in 2020 [2]. Current disease diagnosis is dependent on serum-based prostate-specific antigen (PSA), but these assays are limited in their diagnostic ability [3]. The recent development of multi-parametric magnetic resonance imaging (mpMRI) has become the primary modality for assessing prostate disease

✉ Yoganand Balagurunathan
yoganand.balagurunathan@moffitt.org

Shatha Abudalou
shatha.abudalou@moffitt.org; shathaa@usf.edu

Jung Choi
jung.choi@moffitt.org

Kenneth Gage
kenneth.gage@moffitt.org

Julio Pow-Sang
julio.powsang@moffitt.org

Yasin Yilmaz
yasiny@usf.edu

[1] Department of Machine Learning, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA

[2] Department of Electrical Engineering, University of South Florida, Tampa, FL, USA

[3] Department of Diagnostic Radiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA

[4] Department of Genitourinary Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA

Published online: 20 May 2025

conditions [4, 5]. Determining the prostate volume (PV) provides critical information to the oncologist in prostate disease assessment, monitoring, and treatment planning. Prostate volume also provides prostate disease stratification and is often used to standardize prostate-specific antigen (PSA) [6]. MRI provides visualization of the gland and enables the physicians to assess the disease conditions and monitor progression; all of these are useful in treatment planning [7]. In clinical practice, T2-weighted MR imaging (T2-WI) provides structural information on prostate gland anatomy. Additional MR sequences include diffusion-weighted imaging (DWI) and dynamic contrast enhancement (DCE), which are known to provide complementary information and have been shown to improve prostate cancer detection [8].

MRI is the preferred modality for soft tissue–based organs such as the prostate gland, and the volume estimated using this modality provides better accuracy than widely used ultrasound (US) [9]. Measuring the prostate gland volume using manual segmentation is a standard clinical practice. However, it is a time-consuming process [10] that would benefit from automation, with the promise of artificial intelligence (AI)–based models.

There are explicit biases among human experts in manual gland delineation, most often influenced by the physician's experience and glandular differences that exist among the patients. The organ's shape and vasculature are two factors that affect the prostate gland delineation and the glandular volume estimate [11]. DNN models have shown promising results in prostate gland segmentation despite the challenges with variations in the gland shape and texture [12–14]. Training a deep learning model for segmenting the gland anatomy would require a large-scale dataset with clinical annotations [15]. Medical oncological cohorts are often scarce and have many limitations, including unavailable labeled data and limited patient diversity [16, 17]. Developing national databases through the national genome project (The Cancer Genome Archive, TCGA, and The Cancer Imaging Archive, TCIA) has encouraged data sharing and immensely promoted open science [18, 19]. Recent work [20] tested the ability of AI systems to detect clinically significant prostate cancers using MR imaging and compare them with radiologist-provided PI-RADS (v2.1) scores ($n=$ 9207). These models were independently evaluated on a test cohort ($n=1000$).

Advancements in AI have enabled many architectures that are being proposed for clinical imaging applications. Convolutional neural networks (CNNs) have been the first [21] to be widely adopted for medical imaging applications to improve detection, diagnosis, and segmentation [22–25]. These models improve the segmentation quality by utilizing attention mechanisms, incorporating multi-scale learning, and applying various techniques to address shape variability. Transformer models are newer architectures that allow global representation compared to CNN-type models that may improve image-level detection. These models incur significant computational expenses and require robust GPUs to facilitate efficient training. The 3D U-Net, albeit resource-intensive, is particularly well-suited for detailed segmentation, especially in medical imaging.

Multiple studies have examined the inter-reader variability concerning the segmentation of the prostate gland and the zonal regions. In [26], three deep learning models were compared to segment the gland, with annotations provided by four readers. They report that the models' Dice performances range from 0.84 to 0.91. In [27], the research employed a U-Net architecture and utilized five cohorts established through expert reader annotations to examine the impact of label errors. The performance of the models was determined through the consensus of multiple model predictions. The findings indicated that the inter-reader agreement of deep learning networks surpassed that of human readers. In [12], they used DenseNet and U-Net to segment the gland structure. The authors trained their model on a cohort of ($n=141$) participants and tested it on smaller ($n=47$) cohort size, asserting that the optimal models yielded a Dice coefficient of 0.92. In [28], U-resNet was used to segment the glandular architectures. Two readers were utilized to compare the performance of the deep learning models, resulting in a reported best Dice coefficient of 0.88 for the central gland. The limitation of these studies lies in training the deep learning model on a single cohort/reader type dataset. Additionally, utilizing a limited testing dataset may constrain the generalization capacity.

In [29], a framework was proposed to study the clinical annotator's preferential biases and estimate the user-specific segmentation and stochastic errors. The study does outline the inherent differences among human annotators and provides bias estimates. However, it has not presented a numerical evaluation of this assessment. The study's key limitation is modeling each pixel's label distribution independently, ignoring the spatial correlation among pixels, which is crucial for medical image segmentation. In [30], authors study annotator biases in lung CT delineations. They developed a multi-view "divide-and-rule" (MV—DAR) model that effectively learns from both ambiguous and reliable nodule annotations. The model lacks pathological training results, which are essential for reliable malignancy classification. Ambiguous radiologist labels may bias the model, affecting its generalization to new datasets.

In [31], the authors proposed to study the label noise distribution and reduce the impact of noisy labels. The study attempts to estimate the noise transition matrix that enables the estimation of posterior probabilities. The study reveals the importance of human-assessed label noises on the outcome inference. This work's limitation is that its

accuracy may be compromised if the mini-batch inadequately represents classes or exhibits a deficiency in feature variety. In this work, we evaluate the performance of a modified 3D U-Net [32] with varying expert-provided ground truth that delineates the prostate gland anatomy in a T2w-MRI. We explore strategies to train a balanced network to optimize performance across expert readers. The mixed training strategies seem to enhance model performance, significantly improving reproducibility and reducing annotator variability by allowing the model to adapt to inherent biases. The U-Net model was trained using semi-optimal hyperparameters to segment prostate MR annotations. The study overview is described in Fig. 1. Our methodology addresses the urgent challenge of adapting DNN models across diverse annotators' opinions.

The contribution of the presented work can be summarized as follows:

(a) Our study quantifies variability in deep network performance due to differences in human expert opinion.
(b) We explore the best strategies to train the deep network that can adapt across the expert opinions and generalize the network.

## Methods

### Datasets

We curated 342 patients who had magnetic resonance imaging (MRI with T2-weighted sequences) that were obtained from The Cancer Imaging Archive (TCIA) portal under the ProstateX challenge cohort [33]. The imaging data was stored in our research Picture Archiving and Communication System (PACS) (MIM software®) for our clinical review and assessment. The downloaded patient data records were de-identified without any patient health identifiers, and the patients waived the informed consent. The University of South Florida/Moffitt Cancer Center institutional review board (IRB) protocol approved the research study. All research methods used anonymized patient information and were in compliance with relevant ethical use of human subject guidelines and regulations.

### Clinical Readers

The patients' MR image data (T2 W) was reviewed by our institutional clinical/research radiologist, who delineated the prostate gland and zonal regions in 3D using semi-automated digital tools available on the research PACS tools (MIM software). The first clinical expert team marking was denoted as Reader #1 (R#1) (MCC, Moffitt Cancer Center). The MCC team consists of two experts; the first expert possesses over 15 years of clinical experience in magnetic resonance (MR) prostate imaging, while the second expert brings over 12 years of clinical experience in MR prostate imaging. The regional markings in three dimensions were stored in RT (radiotherapy) format on our research PACS, and the cohort was exported for offline model training and validation. We obtained an independent set of delineations from an expert clinical reader, reader #2 (R#2). The findings were previously presented [14], and their annotations were shared through the TCIA (The Cancer Imaging Archive) data portal. Both R#1 and R#2 annotated the same 204 samples in the dataset. In addition, R#1 annotated 138 more samples, resulting in $n = 342$ patient image annotations. The data from the readers are organized as follows: the same 150 images from R#1 and R#2 were used for training, and the remaining 192 images from R#1 and 54 images from R#2 were used for testing.

### Preprocessing

The MR (T2 W) image data were standardized to a uniform resolution across three dimensions and compared with different resampling settings: (1,1,1), (0.5, 0.5, 1), and (0.5, 0.5, 3). The volumetric image data's original size dimension ($256 \times 256 \times 23$) was resized ($128 \times 128$x16) using the 3D cropping volume method before feeding it into the network. The datasets were preprocessed to eliminate the outliers, and the pixel intensities were snipped within the range of three standard deviation values outside the positive and negative standard deviation values. The image data was normalized by scaling the signal intensity between [0,1] following the Z-transformation [34].

The dataset was augmented by generating five times the number of input images for the chosen batch size. The augmentation function generates one random output image that differs from the original image, producing a total of between 25 and 110 augmented images, actual size depends on the chosen batch size. We further used five different functions to augment the data, namely: zoom in the range [0.5, 1.5], rotation of 90°, width shift of 0.4 for the horizontal shift, height shift of 0.4 for the vertical shift, and random horizontal flip. Data augmentation strategies are powerful strategies that aid in discriminative model training and contribute to improving model performance [35].

### DNN Model and Training Parameters

A 3D U-Net convolutional neural network [32] was built sequentially, starting with random weights, allowing us to train the model using the study data with augmentation as described in the prior section. The U-Net architecture has an encoder analysis path, followed by a decoder synthesis

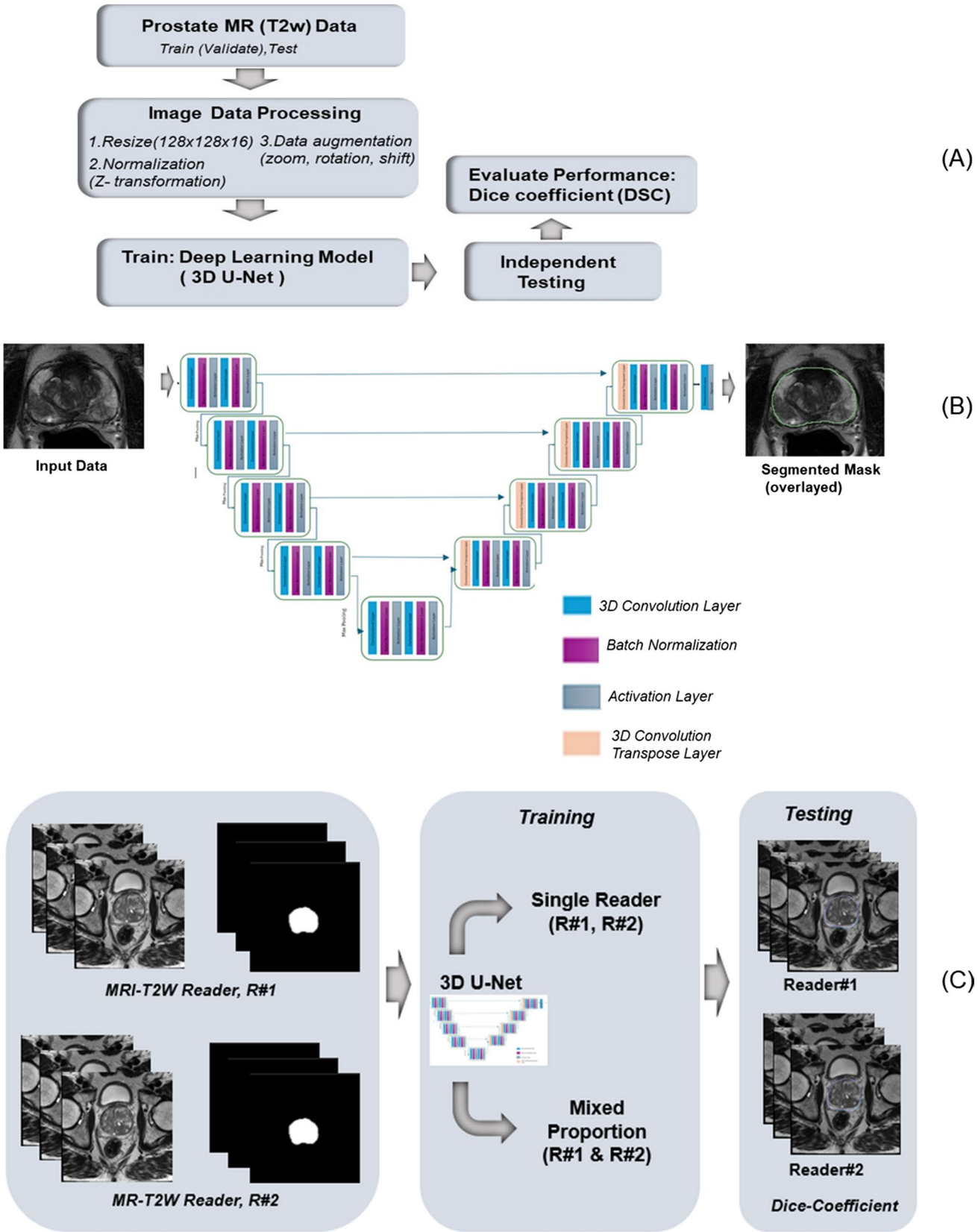


**Fig. 1** Study overview. **A** Process flow diagram. **B** 3D U-Net model architecture. **C** Pictorial representation of the experimental setup

path, widely useful in medical image segmentation [24, 36, 37]. The original architecture has four convolution blocks with complementary arms for up-sampling, often represented in a U shape. Each block consists of two zero-padded convolution layers, followed by the rectified linear unit activation function and max pooling layer for the encoder, which is replaced by an up-sampling layer in the decoder path. The original network architecture [32] was replaced with a 3D convolutional layer ($3 \times 3 \times 3$) to handle the volumetric image dataset. A similar approach was followed for the max-pooling layers. A batch normalization layer was added to the convolution block to standardize the data (zero mean, unit standard deviation) to achieve the training stabilization, following the model architecture recommendations described previously [25, 38, 39]. The most popular U-Net architecture has reported training parameters ranging from 7.85 to 9.6 million [40, 41]. The 3D architecture used in the study has about 90 million training parameters with 23 convolution layers; the number of training parameters depends on the convolutional layers [42] and kernel size [43]. The 3D model needs more significant computational memory than the 2D and 2.5D models. However, 3D models converge 20–40% faster than the 2D and 2.5D models and have been shown to perform better with limited data [44]. In a prior study [45], they applied a modified 3D U-Net with 62 million training parameters and 23 convolution layers. Figure 1(B) shows an overview of the model architecture representing the encoder (left arm) and decoder (right arm).

Several architectural advancements have shown improvements in model performance [46–48]. In our work, the 3D U-Net model was chosen as it allows manual adjustment of the model's parameters, including the number of layers and activation function, and allows architectural flexibility.

In our study, the model was trained with 2000 epochs, using multiple strategies to adjust the learning rate to minimize the loss function in each iteration. Initially, we started with a fixed learning rate ($10^{-5}$). We then tried the cosine annealing technique to reduce the epochs and improve the model's convergence with the learning rate ($\partial$) varied from $10^{-8}$ to 0.01. Nevertheless, we obtained better performance using a fixed learning rate ($\partial$). We illustrate the model efficacy during the training phase for single readers (R#1, R#2), measured by the loss function over several epochs (see Suppl. Figure 1). As the loss functions drop (inverse of Dice coefficient, 1-Dice) over the consistent number of iterations (over 100 epochs), the model weights would be retained.

## Model Performance Evaluation

Dice score (DS) was used to measure the extent of similarity between the predicted region of interest and the ground truth (gland region) [49]. Additionally, the Hausdorff distance (HD) was computed to measure the distance between two mask regions and assess dissimilarity between the masks [50]. These metrics are widely used in medical imaging to measure the similarity or dissimilarity between the predicted region and the expert-provided ground truth.

## Experimental Setup for Multi-reader Training

*The U-Net based* DNN architectures are complex systems with millions of nodes to be trained with smaller sample datasets [51]. We designed our experiments and developed deep network training strategies that are broadly categorized in the following ways.

(i) *Hyperparameter tuning*. To improve parameter convergence, we attempted fixed and cosine annealing techniques. We also experimented with different batch sizes (from 5 to 22) to improve the model performance across batch iterations.

(ii) *Baseline model performance*. We trained a network using reference ground truth provided by individual readers (readers R#1, R#2) and evaluated model performance on an independent cohort across readers ($n$= 192 for R#1 and $n$= 54 for R#2).

(iii) *Mixed proportional training (R#1 and R#2).* We attempted multiple strategies to create a mixed training cohort with proportional data from two expert readers and evaluate the model performance.

(iv) *Small and large gland samples training.* We estimated the average gland volume using delineations provided by two readers (R#1 and R#2). The cohort was divided into small and large gland volumes using the median value as the cut point. The DNN models were trained and tested in each of these sub-cohorts.

**Table 1** Data cohort used for the study

| Groups | Gland annotation (3D) | |
|---|---|---|
| | Reader#1 ($n$ = 342) | Reader#2 ($n$ = 204) |
| Training | 150 | 150 |
| Testing | 192 | 54 |
| Gland volume $cm^3$ ($\mu$, 95% CI), median | | |
| Volume | 63 [59.4 66.5], 55.357 | 67.4 [62.3 72.5], 58.197 |
| Sub-cohort size | | |
| Large (≥ median) | 167 | 109 |
| Small (< median) | 175 | 95 |

## Results

In this study, we assembled a cohort of 342 patients with MRI T2 W scans to train a 3D U-Net architecture from random weights, followed by training approaches previously proposed to segment prostate gland anatomy. We used annotations provided by independent expert readers (R#1: $n = 342$, R#2: $n = 204$); see Table 1. We evaluated different strategies to train the network to obtain the best reproducible performance across multiple expert references (R#1 and R#2). We investigated the key hyperparameters such as the batch size (5, 16, 22), learning rate (fixed $10^{-5}$), and cosine annealing (0.01 to $10^{-8}$) to converge on a baseline network that optimizes the network and provides better convergence. We proposed to use lower batch sizes (range from 5, 16, 22) due to the higher memory requirement. We found that a batch size of 22 and a learning rate of $10^{-5}$ exhibit better network convergence for our study. We evaluated the network model performance with input data resampled at different resolutions (1, 1, 1), (0.5, 0.5, 1), and (0.5, 0.5, 3). The model performance varied across the resolutions with an average Dice coefficient of 0.77, 0.74, and 0.825, respectively, for the respective resolutions (using reader R#1); see Suppl. Table 1.

### Independent Reader Training

The Dice performance using individual readers separately was suboptimal, and details on the experiment are deferred to Suppl. Table 2. Training on R#1 with a 22-batch size, the average Dice coefficient for R#1 test was 0.825, and for R#2, it was 0.68 using R#1 training weights. While using R#2 for training, the average Dice coefficient was 0.707 for R#1 test (using R#2 training weights) and 0.835 for R#2 test (Table 2A). Five-fold cross-validation with individual training (R#1, R#2) was also used to obtain an average Dice of 0.7982 and 0.88, respectively (see Suppl. Table 3B). We also compared our findings with an $n$-fold random selection of train and test cohorts, averaging over five repeats, with $n = 2,3,5$. We find the average Dice coefficient ranging between 0.794 and 0.8204 when trained and tested on R#1, and between 0.8244 and 0.8539 when trained and tested on R#2 (see Suppl. Table 3A).

**Table 2** Models were trained (A) using data from individual readers (R#1, R#2) and (B) using the mixed dataset (R#1 and R#2) compared to cross-validated samples

| (A) Individual cohort (average Dice coefficient ($\mu$, 95% CI) | | |
|---|---|---|
| Training | | |
| | R#1, $n = 150$ | R#2, $n = 150$ |
| Testing | | |
| R#1 ($n = 192$) | 0.825 [0.81 0.84] | 0.707 [0.67 0.75] |
| R#2 ($n = 54$) | 0.68 [0.64 0.72] | 0.835 [0.81 0.86] |
| (B) Mixed training (average Dice coefficient ($\mu$, 95% CI) | | |
| Training | Mixed (R#1, $n = 150$ and R#2, $n = 150$) | Cross-validation (fivefold) (R#1, $n = 150$ and R#2, $n = 150$) |
| Testing | | |
| R#1 ($n = 192$) | 0.811 [0.76 0.86] | 0.826 [0.81 0.84] |
| R#2 ($n = 54$) | 0.851 [0.82 0.88] | 0.875 [0.86 0.89] |

### Mixed Proportions

Using mixed training examples (R#1 and R#2, each with $n = 150$), the average testing Dice coefficient is 0.811 and 0.851 for R#1 and R#2, respectively (see Table 2B). Figure 2 (A and B) shows the sample reader's annotation and the best-performing sample patient (Dice of 0.94 and 0.91 for R#1 and R#2, respectively) obtained by mixed training. In contrast, Fig. 3 (A and B) shows representative suboptimal performance (Dice coefficient ranging from 0.778 to 0.75 for R#1 and R#2, respectively).

We attempted to improve the deep network performance by using different proportions of expert references (R#1 and R#2). While attempting a different number of training examples from each of the experts and fixing the other, the performance was evaluated on an unseen test set. For R#1, we applied different numbers of training samples (R#1, $n = 25$ to 125), considering the entire R#2 dataset. In the mixed training case R#1, n = 100 and R#2, $n = 150$, we find improved performance with an average Dice of 0.862 for R#1 and 0.868 for R#2 with the lowest mean absolute deviation of 0.006 (see Table 3).

We systematically experimented with model training with several proportions of samples from the two readers. In addition to varying R#1 training proportion, we fixed the R#1 example and proportionally varied R#2 training examples. We used fivefold cross-validation to evaluate the model performance (see Suppl. Table 4A&B). The results were not better than the highlighted best result in Table 3.

### Cohorts Based on Gland Size

We estimated the average gland volume using delineations from R#1 and R#2. Using the average median volume of 56.77cm$^3$, we divide the samples into sub-cohorts (large and small glands). For small glands, there are 175 samples in R#1 and 95 in R#2. For large glands, there are 167 samples in R#1 and 109 in R#2 (see Table 1).

We retrained our models in these sub-cohorts using proportionally mixed readers. We used fivefold cross-validation to estimate the performance. We first fixed R#2; we found in the case of a large gland cohort, the best model

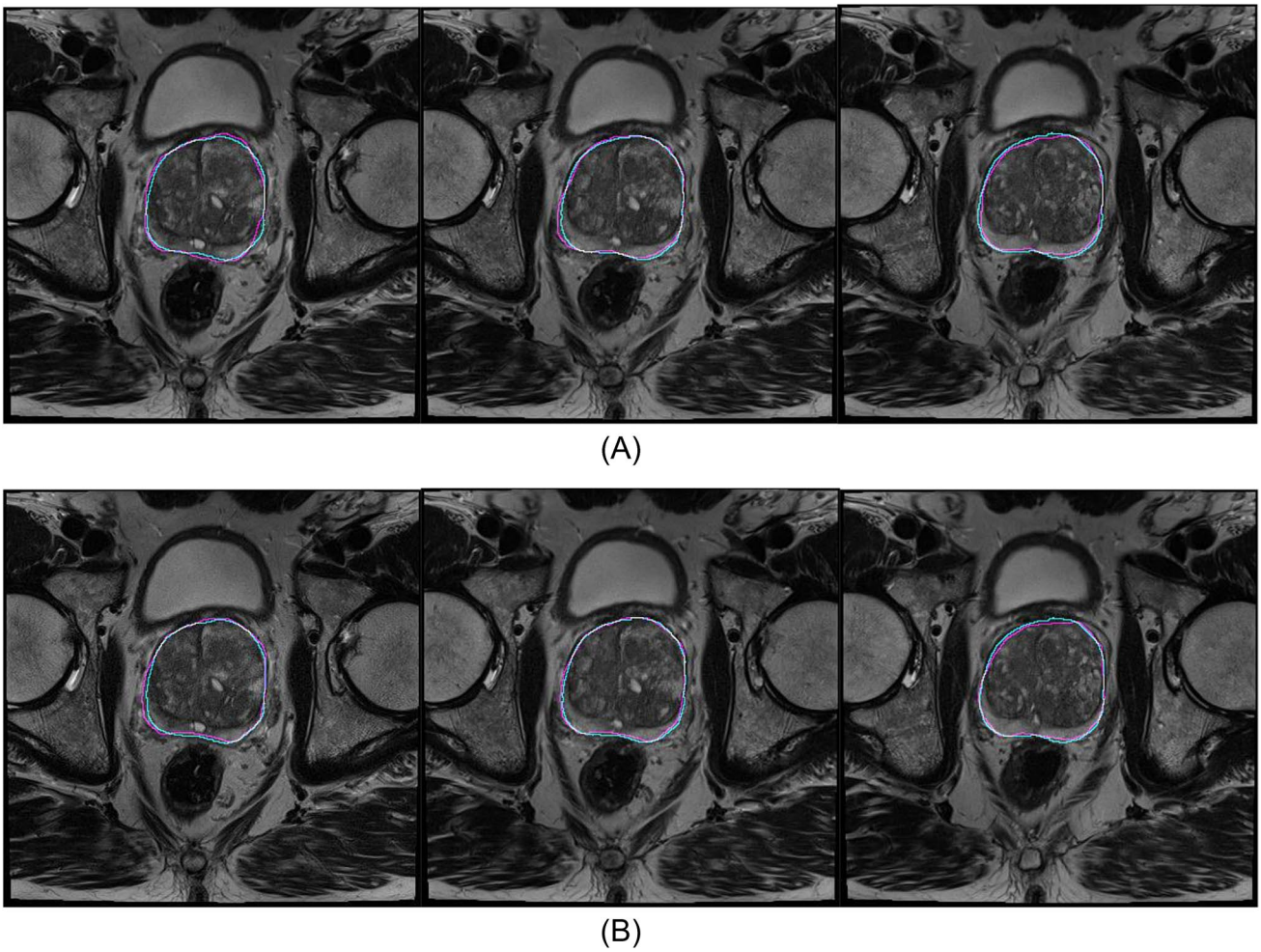


**Fig. 2** Representative slices of a patient's T2-weighted (T2W) MRI showing delineation results by the best-performing network (cyan) trained on the mixed training dataset. (**A**) R#1 annotations (Dice = 0.943). (**B**) R#2 annotations (Dice = 0.914). Purple contours indicate expert references by the respective readers

was obtained for a proportion of 70% of R#1 ($n = 70\%$ of 167) and fixed R#2 ($n = 109$), yielding a Dice of 0.842 for R#1 and a Dice of 0.866 for R#2, and the mean absolute deviation equals 0.024, and results in a reduction in variability, with a mean absolute deviation of 0.002 using (R#1, $n = 34$ and R#2, $n = 87$), achieving an average Dice coefficient of 0.84 (see Table 4A). In the case of a small gland cohort, the best model performance for the proportional mixture was with 85% of the R#1 dataset ($n = 85\%$ of 175 samples) and fixed R#2 ($n = 95$ samples). A Dice of 0.803 for R#1 and 0.829 for R#2, and 0.027 mean absolute deviation was achieved (see Table 4B).

We then fixed R#1 and evaluated the model performance. In a sub-cohort with large gland volumes, the best-performing model's Dice was 0.838 for R#1 and 0.872 for R#2, achieved using 85% of R#2 ($n = 85\%$ of 109), with a mean absolute deviation of 0.034 (see Table 5A). In the small gland volume sub-cohort, we find the best model performance was achieved when selecting 70% of R#2 ($n = 70\%$ of 95) and fixed R#1 ($n = 175$); the model reached an average Dice of 0.804 and 0.837 for R#1 and R#2, respectively, with a mean absolute deviation of 0.033 (see Table 5B).

In addition, applying small and large cohorts improved model reproducibility among readers compared to the single cohort training model performance; Fig. 4 illustrates the Dice coefficients for large- and small-sized glands across different proportions. We noticed that for the small gland volume datasets, the Dice coefficient was approximately 0.8 using 25% of the R#1 and 100% of the R#2 (other possible mixtures for small glands include using 10% of the R#2 with 100% of the R#1). The model performance improved when using 40% of R#1 and 100% of R#2 for the large gland volume group, with an average Dice of $\approx 0.85$.

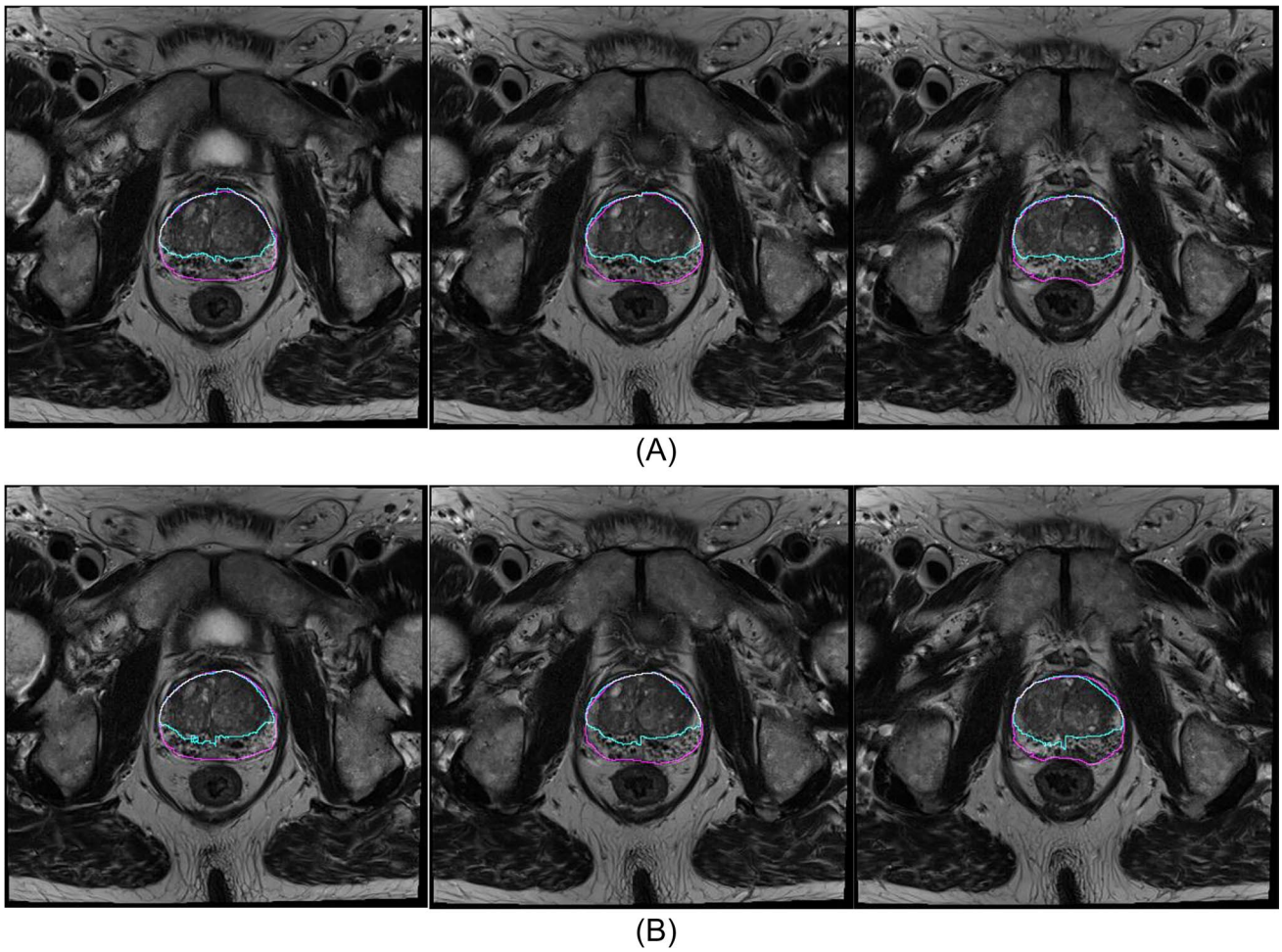


**Fig. 3** Representative slice of a patient's T2 W image with delineation provided by *suboptimal* network performance (in cyan) that was trained on mixed training dataset (R#1, $n = 150$ and R#2, $n = 150$). **A** R#1 annotation with an average Dice coefficient of 0.76. **B** R#2 annotation, with an average Dice coefficient of 0.75. The annotation in purple shows the original reference by the respective reader

## Cohorts Based on Similarity

We followed the investigation by assembling a sub-cohort with high concordance between readers prior to training (high similarity between the ground truth masks) and analyzing the model's performance on the examined dataset for both readers. Of these, 72 patient samples had Dice coefficients of 0.7 or higher from R#1 and R#2. Compared to the previous testing dataset (R#1, $n = 192$; R#2, $n = 54$), the sub-cohort includes 148 samples from both readers, utilizing the cross-validation technique that trained the model, which was then assessed with the testing dataset (R#1, $n = 270$; R#2, $n = 132$). Comparing individual and mixed proportional similarity training examples, the model achieved an average Dice coefficient ranging from 0.826 to 0.848 for R#1 and from 0.869 to 0.875 for R#2. Although using a mixed training dataset enhances the model's performance on the larger testing dataset, the size of the training dataset

**Table 3** Performance comparison of models trained using mixed reader examples with varying R#1(R#2, $n = 150$). Where MAD is mean absolute deviation

| Model performance Dice coefficient ($\mu$, 95% CI) | | | |
|---|---|---|---|
| Training: R1 samples, *R#2 fixed (n = 150)* | Testing | | |
| R#1 | R#1 ($n = 192$) | R#2 ($n = 54$) | MAD (R#1 and R#2) |
| 25 | 0.823 [0.81 0.84] | 0.868 [0.84 0.89] | 0.045 |
| 50 | 0.833 [0.82 0.85] | 0.878 [0.87 0.89] | 0.045 |
| 75 | 0.828 [0.81 0.84] | 0.874 [0.86 0.89] | 0.046 |
| 100 | 0.863 [0.85 0.87] | 0.869 [0.87 0.88] | 0.006 |
| 125 | 0.809 [0.79 0.82] | 0.856 [0.84 0.87] | 0.047 |
| 150 | 0.811 [0.78 0.84] | 0.851 [0.82 0.88] | 0.04 |

**Table 4** Differences in model performance with the proportion of #R1 (fixed R#2), in (A) large-size glands and (B) small-size glands

| (A) Large glands: Dice coefficient ($\mu$, 95% CI): fivefold-CV | | | |
|---|---|---|---|
| Training: R1 (%), R#2 fixed ($n = 87$) | Testing R#1 | R#2 ($n = 22$) | MAD (R#1 and R#2) |
| 100 | 0.835 [0.82 0.85] | 0.859 [0.85 0.87] | 0.022 |
| 85 | 0.843 [0.83 0.85] | 0.862 [0.85 0.88] | 0.019 |
| 70 | 0.842 [0.82 0.86] | 0.866 [0.86 0.87] | 0.024 |
| 55 | 0.84 [0.83 0.85] | 0.849 [0.83 0.86] | 0.009 |
| 40 | 0.846 [0.82 0.87] | 0.849 [0.84 0.86] | 0.003 |
| 25 | 0.838 [0.83 0.85] | 0.836 [0.82 0.86] | 0.002 |
| 10 | 0.833 [0.8 0.87] | 0.838 [0.82 0.85] | 0.005 |
| (B) Small glands: Dice coefficient ($\mu$, 95% CI): fivefold-CV | | | |
| Training: R1 (%), R#2 fixed ($n = 76$) | Testing R#1 | R#2 ($n = 19$) | MAD (R#1 and R#2) |
| 100 | 0.809 [0.8 0.82] | 0.831 [0.82 0.85] | 0.022 |
| 85 | 0.803 [0.78 0.83] | 0.829 [0.81 0.85] | 0.027 |
| 70 | 0.798 [0.77 0.82] | 0.813 [0.79 0.84] | 0.015 |
| 55 | 0.798 [0.78 0.82] | 0.817 [0.81 0.83] | 0.018 |
| 40 | 0.8 [0.79 0.81] | 0.809 [0.79 0.83] | 0.009 |
| 25 | 0.805 [0.79 0.82] | 0.798 [0.78 0.81] | 0.007 |
| 10 | 0.754 [0.74 0.77] | 0.787 [0.75 0.82] | 0.033 |

The results were evaluated using cross-validation using proportional (15% increment of R#1) training between readers (R#1 and R#2) where MAD is mean absolute deviation

remains limited compared to other combination techniques. We explored different proportional mixture strategies to enhance the model's performance and reproducibility, thus minimizing variability among readers (see Suppl. Table 6).

Figure 5 shows how the model performance gradually improves on example patient delineations using different training strategies; Fig. 5(A) applies a single reader training dataset, and Fig. 5(B) shows the mixed training dataset (train R#1 and R#2, $n = 150$ and 150) result. Then, using a network trained on the highly concordant mixed training datasets (R#1, $n = 72$ and R#2, $n = 72$), the result is shown in Fig. 5(C).

Figure 6 illustrates the comparison of model performance utilizing various pairwise training strategies. Figure 6(A) shows that the model performance remained consistent when trained with individual cohorts as opposed to mixed cohort training (readers R#1 and R#2, $n = 150$ samples each). Figure 6(B) delineates the outcomes when the model was trained using individual cohort training compared to mixed training (readers R#1 and R#2, $n = 150$ samples each), incorporating five-fold cross-validation; a notable enhancement in model performance was observed for reader R#2, with an increase in the average Dice coefficient from 0.811 to 0.826. In contrast, Fig. 6(C) illustrates how a proportionately mixed training dataset (reader R#1, $n = 100$ samples and reader R#2, $n = 150$ samples) contributed to improved model repeatability among readers in the test cohort when compared to mixed training (R#1 and R#2, $n = 150$ samples each). Lastly, Fig. 6(D) presents the performance outcomes using similar samples for mixed training (R#1 and R#2).

## Statistical Significance

We compared the model performance between single reader and multiple reader training and computed statistical significance (Wilcoxon rank-sum test) with Bonferroni multiple testing correction. A significant difference was found comparing single cohort to mixed cohort ($p = 0.0058$) and single cohort to proportional mixed cohort with a $p < < 0.001$. Additionally, the mixed cohort with and without cross-validation significantly differed from the proportionally mixed testing ($p < < 0.001$) for R#1. Regarding R#2, significant differences were identified across all training strategy comparisons ($p \leq 0.0038$), except the comparison between the single cohort and the mixed cohort without cross-validation. We summarize the testing pairs in Table 6 and Fig. 7.

In addition, we evaluated the model utilizing three independent, previously unexamined datasets. Each cohort weight was analyzed separately (R#1, $n = 150$ and R#2, $n = 150$) alongside mixed training weights (R#1, $n = 100$ and R#2, $n = 150$). As illustrated in Table 7, the model exhibited suboptimal performance when assessed with the single cohort models; however, there was a marked performance improvement when the mixed training model was applied to the testing cohorts. We show that a proportionally mixed training technique improves reproducibility compared to individual trained models ($p$-value $< < 0.01$ and 0.004 for R#1 and R#2, respectively, R#1, $n = 100$/R#2, $n = 150$).

And we compared the 3D U-Net used in the study with alternate architectures, SegResNet and 3D U-Net (Nvidia's Monai-Zoo models [52, 53]); these models were re-trained for our datasets starting from random weights. We achieved an average Dice coefficient of 0.807 and 0.845 for R#1 ($n = 192$) and R#2 ($n = 54$), respectively. In addition, we compared the SegResNet Monai architecture; test performance yielded an average Dice coefficient of 0.815 and 0.852 for R#1 ($n = 192$) and R2# ($n = 54$), respectively. The model's test performance using the (R1, $n = 150$ and R2, $n = 150$) mixed training dataset approach (for R#1, test) yielded an average Dice of 0.825 and 0.839 for U-Net and SegResNet models, respectively. For R#2 ($n = 54$), the Dice coefficients were 0.851 and 0.873 for U-Net and SegResNet models using the mixed training dataset approach, a notable performance improvement when the mixed training model was used (see Suppl. Table 7).

**Table 5** Differences in model performance with the proportion of #R2 (fixed R#1), for (A) large-size glands and (B) small-size glands. Where MAD is mean absolute deviation

| (A) Large glands: Dice coefficient ($\mu$, 95% CI): fivefold-CV | | | |
|---|---|---|---|
| Training: R2 (% proportion), R#1 fixed ($n = 134$) | Testing | | |
| | R#1 ($n = 33$) | R#2 | MAD (R#1 and R#2) |
| 100 | 0.834 [0.82 0.85] | 0.866 [0.86 0.87] | 0.032 |
| 85 | 0.838 [0.83 0.84] | 0.872 [0.86 0.89] | 0.034 |
| 70 | 0.823 [0.81 0.83] | 0.866 [0.85 0.88] | 0.043 |
| 55 | 0.832 [0.82 0.84] | 0.870 [0.86 0.88] | 0.038 |
| 40 | 0.825 [0.82 0.83] | 0.869 [0.84 0.89] | 0.044 |
| 25 | 0.827 [0.82 0.84] | 0.86 [0.84 0.88] | 0.033 |
| 10 | 0.815 [0.8 0.83] | 0.857 [0.78 0.93] | 0.042 |
| (B) Small glands: Dice coefficient ($\mu$, 95% CI): fivefold-CV | | | |
| Training: R2 (% proportion), R#1 fixed ($n = 140$) | Testing | | |
| | R#1 ($n = 35$) | R#2 | MAD (R#1 and R#2) |
| 100 | 0.808 [0.79 0.83] | 0.826 [0.81 0.84] | 0.018 |
| 85 | 0.803 [0.79 0.82] | 0.817 [0.79 0.84] | 0.014 |
| 70 | 0.804 [0.79 0.81] | 0.837 [0.82 0.85] | 0.033 |
| 55 | 0.792 [0.78 0.81] | 0.824 [0.81 0.84] | 0.032 |
| 40 | 0.795 [0.79 0.81] | 0.841 [0.82 0.86] | 0.045 |
| 25 | 0.802 [0.79 0.81] | 0.824 [0.8 0.84] | 0.023 |
| 10 | 0.794 [0.78 0.81] | 0.808 [0.77 0.85] | 0.014 |

The results were evaluated using cross-validation using proportional (15% increment of R#2) training between readers (R#1 and R#2)

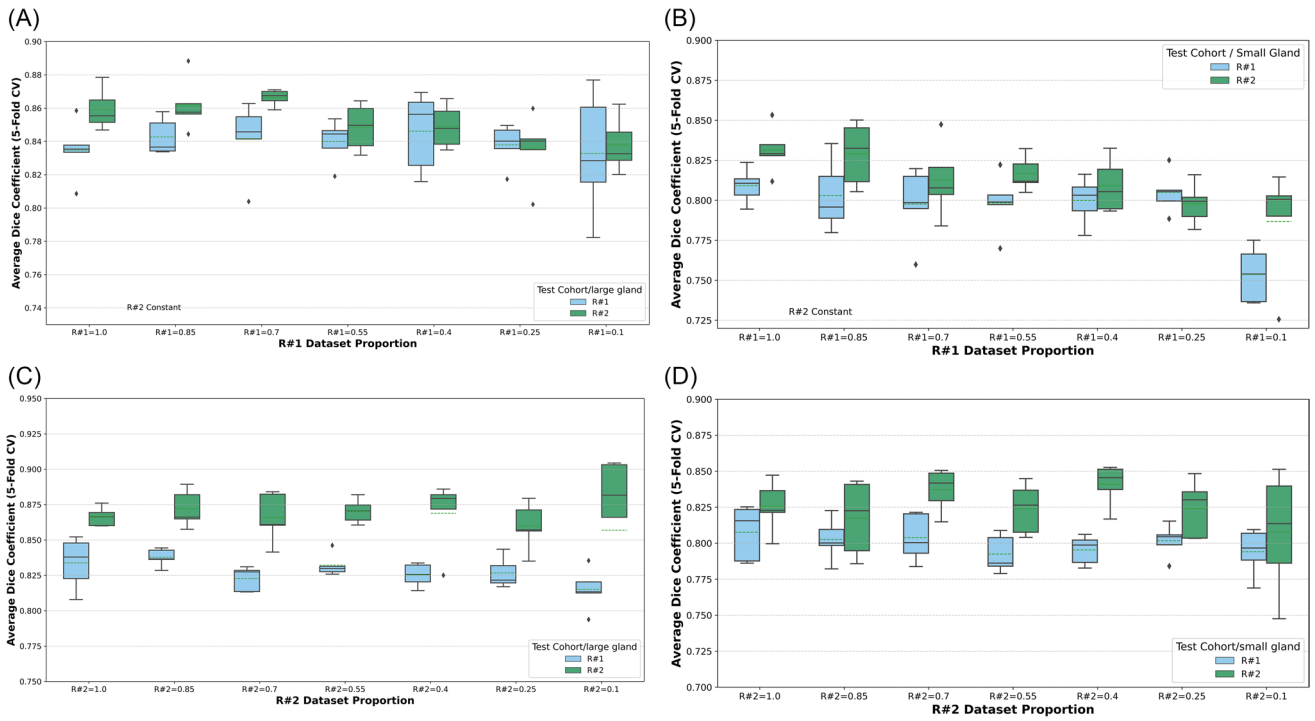


**Fig. 4** Deep network models test performance for mixed proportional training for: **A** large glands (fixed R#2), **B** small glands (fixed R#2), **C** large glands (fixed R#1), and **D** small glands (fixed R#1). The Dice coefficient was estimated by fivefold cross-validation

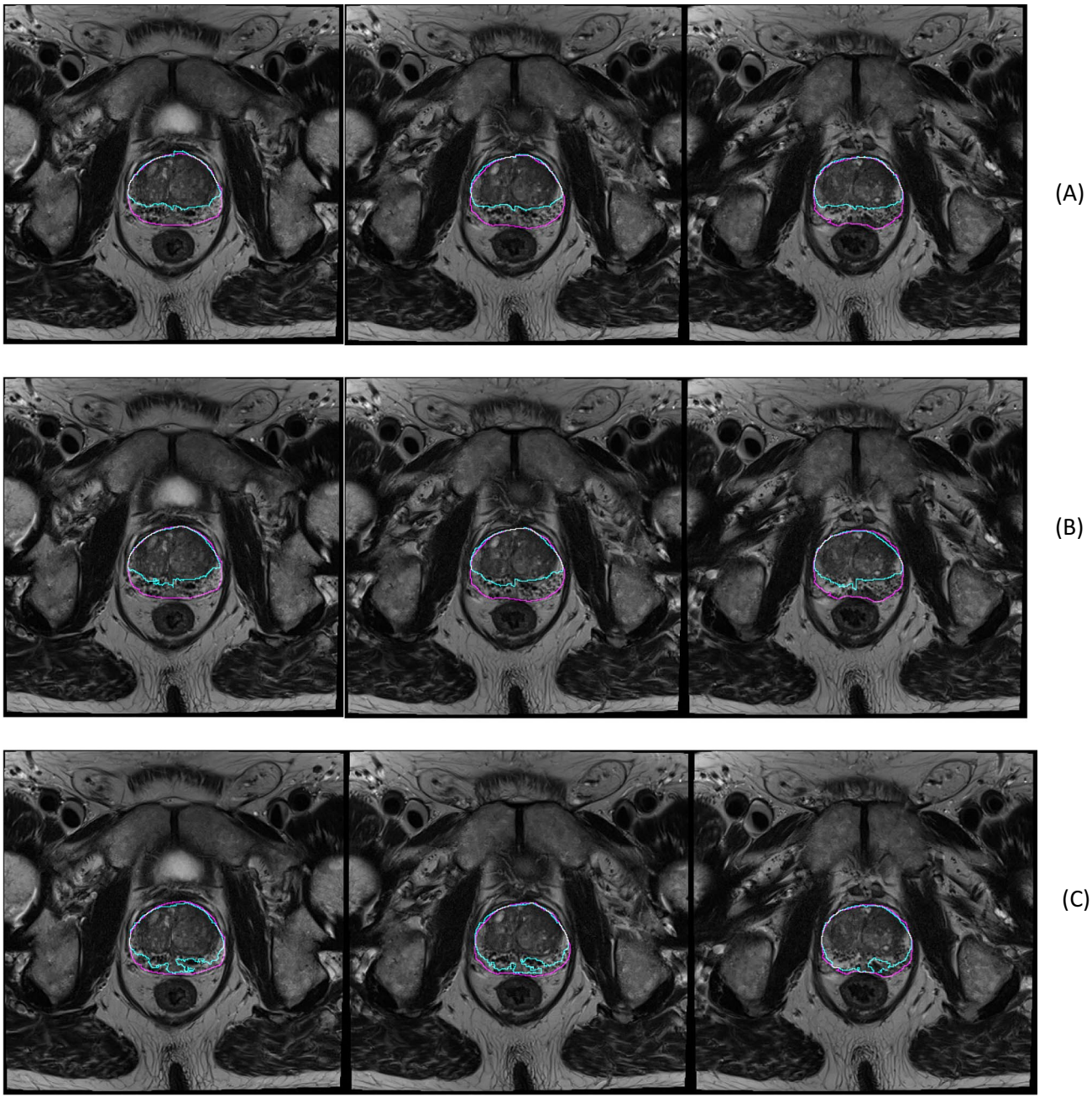


**Fig. 5** Representative slice of a patient's T2 W image with delineation provided by the best network performance (in cyan). The model was trained using examples from **A** Reader R#2, and tested on R#2 (Dice of 0.736), **B** mixed training (R#1, $n = 150$ and R#2, $n = 150$), tested on R#2, yielded a dice of 0.75, **C** samples with similar examples and mixed training (R#1, $n = 72$ and R#2, $n = 72$), tested on R#1 (Dice = 0.7) and R#2 (Dice of 0.81). Annotation in purple is the original reference by respective readers

## Discussion

The study used a 3D U-Net architecture [32] where a convolutional layer is added to a 2D U-Net to segment the prostate gland in three dimensions. We trained the network using annotations obtained from two expert groups for the same patient's T2w-MR imaging. Using these as references, we developed strategies to train deep models to perform at their best and reproducibly across the two readers. We show that proportionally mixed training with expert examples improved the performance and reproducibility across the readers.

The 3D U-Net architecture has been popularly used in segmenting structures in clinical imaging and provides the possibility of replacing manually intense tasks [38, 42]. The architecture has an encoder/decoder arm that complements each other,

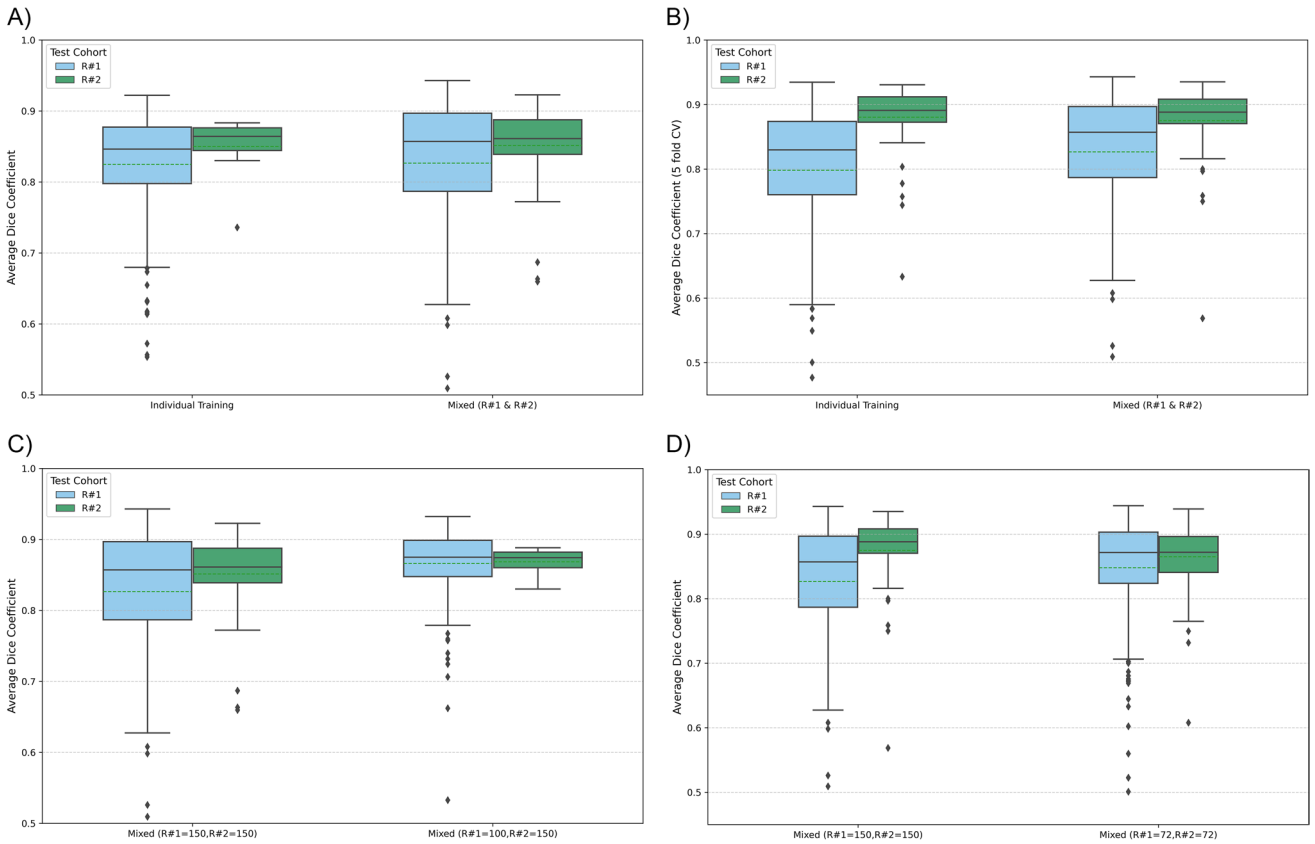


**Fig. 6** Deep network models test performance using different training datasets: **A** individual cohort training (R#1, R#2) and mixed training (R#1 and R#2, $n=$ 150 samples each), **B** individual cohort and mixed training estimated using five-fold cross-validation, **C** mixed training (R#1 and R#2, $n=$ 150 samples each) compared to proportion mixture (R#1, $n=$ 100 samples and R#2, $n=$ 150 samples), **D** mixed training (R#1 and R#2, $n=$ 150 samples each) compared to mixed training of similar examples (R#1 and R#2, $n=$ 72 samples each) with five-fold cross-validation

**Table 6** Comparing the test performance of deep network models using different training strategies, including single cohort training, mixed cohort training (with and without cross-validation), and proportional mixed cohort training

| Training strategy comparison | *p*-value (R#1) | Is Significant # | *p*-value (R#2) | Is Significant # |
|---|---|---|---|---|
| Single cohort (Table 2A) vs Mixed cohort (Table 2B) | 0.0058 | Yes | 0.5905 | No |
| Single cohort (Table 2A) vs Mixed cohort (CV) (Table 2B) | 0.379 | No | 0.00053 | Yes |
| Single cohort (Table 2A) vs Proportional mixed cohort (Table 3) | < 0.001 | Yes | 0.0036 | Yes |
| Mixed training cohort (with CV and without CV) (Table 2B) | 0.1588 | No | < 0.001 | Yes |
| Mixed cohort (Table 2B) vs Proportional mixed cohort (Table 3) | < 0.001 | Yes | 0.0036 | Yes |
| Mixed cohort (CV) (Table 2B) vs Proportional mixed cohort (Table 3) | < 0.001 | Yes | 0.0038 | Yes |

# The Wilcoxon rank-sum test with Bonferroni correction is applied, setting the significance threshold at $p \leq 0.0083$

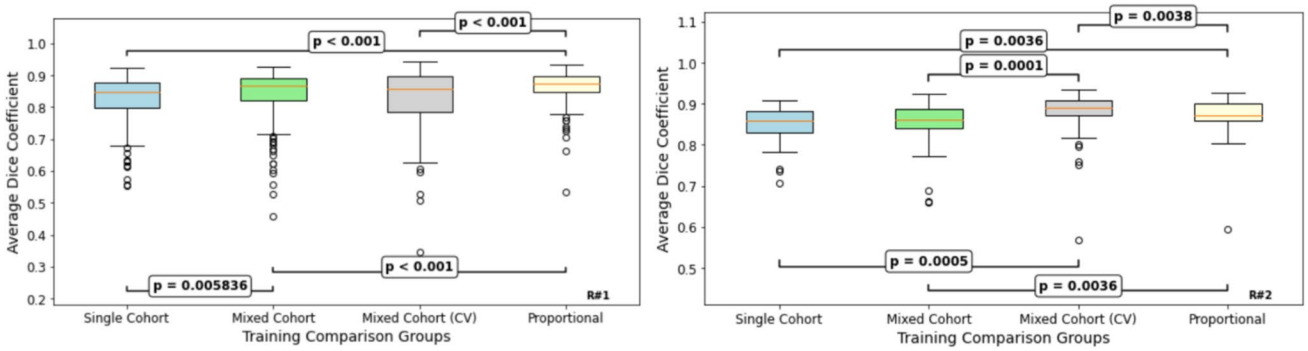


**Fig. 7** Deep network model test performance by applying different training strategies, including single cohort training, mixed cohort training (with and without cross-validation), and proportional mixed cohort training. The Wilcoxon rank-sum test with Bonferroni correction is applied, setting the significance threshold at $p < 0.0083$

allowing better boundary decisions [38, 39]. These architectures have been widely used in segmenting prostate MRI [13, 14, 54]. Prostate gland segmentation poses a challenge due to differences in the organ's shape and texture and the cellular heterogeneity that adds to the complexity of developing automated models [55]. There have been many prior studies that have proposed automated segmentation methods for the gland anatomy [11, 13, 56]. The challenge has been developing an architecture that can work across different population groups.

Our study is one of the first to evaluate network performance and provides strategies to train deep models using two expert reader annotations (in 3D) on T2 W MR images. Previous studies in prostate gland segmentation [56–59] have shown that U-Net and CNN-type architectures have been used on smaller cohorts to test their network performance, which limits reproducibility across clinical centers. Our study has assembled a relatively larger cohort ($n = 345$) that allows better model training and evaluation. Data augmentation techniques were employed to enhance model training efficacy. Prior research [60] have substantiated that data augmentation significantly improves segmentation performance, as quantified by the Dice coefficient, yielding enhancements ranging from 2 to 26%.

In [61], the model was trained on an institutional cohort and tested on two independent datasets, claimed that central gland segmentation achieved Dice score 0.909 and for zonal segmentation performance reached Dice of 0.86 for internal testing group. They concluded that the central gland volume is a crucial factor affecting the model's segmentation performance. In reference to [62], three distinct deep learning models were evaluated to segment the prostate gland. Signal quality and prostate gland volume were considered influencing elements of network performance. It was reported that both prostate volume and inadequate signal quality adversely affect the performance of the model. In [63], which reports the model performance for prostate segmentation, the Dice coefficient score showed improvements for the model with larger glands that were consistent in independent cohorts.

**Table 7** Evaluating model performance on three independent datasets using a single cohort model (R#1, R#2) and a mixed training model (R#1, $n = 100$ and R#2, $n = 150$)

| Unseen testing dataset | Dice coefficient (test) | | |
|---|---|---|---|
| | Train: (R#1, $n = 150$) | Train: (R#2, $n = 150$) | Mixed (R#1, $n = 100$ and R#2, $n = 150$) |
| Public dataset (TCIA—prostatectomy, $n = 25$) | 0.6 | 0.54 | 0.765 |
| MCC#1 dataset ($n = 171$) | 0.665 | 0.545 | 0.77 |

Deep learning models present significant potential for enhancing the accuracy of prostate gland segmentation, which is critical for the standardization of prostate-specific antigen (PSA) density [64–66]. The integration of PSA and segmentation outputs seeks to improve the classification of Gleason grades. Increased precision in tumor boundary detection may enable more targeted therapeutic strategies and signifies a considerable advancement in the detection of prostate cancer, as well as in the formulation of patient treatment plans.

In summary, our study addresses real-world clinical challenges to reproduce the deep network across readers and identify strategies to improve the performance of these networks. In our study, of the several experiments, the proportional mixture-based model training significantly improved model performance and decreased variability among readers.

## Limitations

Deep model reproducibility across cohorts is challenging due to many acquisition level differences, such as the patient cohort, scanner type, and operator setting. Our study uses a large cohort of examples, with multiple

readers from two diverse institutions. Having a much larger cohort with a few more readers would capture the spectrum of variability among the readers and improve the network's generalization. To mitigate the differences in imaging systems, acquisition settings, patient level factors, expert reader training, a much larger cohort size with diverse readers becomes necessary to improve model training/validation.

## Conclusion

We used a 3D U-Net-based DL architecture to segment prostate gland anatomy on MRI-T2 W and quantified variability between independent expert readers. The proposed models use balanced hyperparameters and strategies to improve training to mitigate variability and provide a reproducible network across varying ground truth. We show that proportionally mixed data training is one strategy that improves network translation across expert references.

We find the cohort with large glands (over median value) shows better model performance and improved variability between the readers.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10278-025-01504-8.

**Acknowledgements** We sincerely thank our imaging researchers and clinical/research radiologists at Moffitt Cancer Center (Drs. Qian Li, Jin Qi, and Hong Lu) for their effort to collate consensus readings to delineate prostate glandular structures manually. We also thank the open-source data host, The Cancer Imaging Archive (TCIA), data contributors, and the Prostate X challenge organizers from Radbound University Medical Center (Dr. Huisman et al. and his team) for providing high-quality MR imaging data as part of an open science community for scientific exploration.

**Author Contribution** Concept and problem: SA, YB. Model building: SA, YY, YB. Clinical opinion: JC, KG, JP-S. Radiological overread: JC, KG. Paper write up and edits: SA, JC, KG, JP-S, YY, YB.

**Funding** We acknowledge funding support (5U01 CA200464 -07, Cohon's family funding) that partly supported the investigators' research time.

**Data Availability** The data used in the study can be obtained from the archive accessed through the following URL: https://www.cancerimagingarchive.net/collection/prostatex/an

## Declarations

**Ethics Approval** The downloaded patient data records were de-identified without any patient health identifiers, and the patients waived the informed consent. The University of South Florida/Moffitt Cancer Center approved the research protocol for the study. All research methods that used patient information were performed in compliance with relevant guidelines and regulations.

**Consent to Participate** Patients' clinical records were anonymized before receipt and waived patients' informed consent.

**Consent for Publication** Senior authors of studies bear responsibility for the integrity of the study and approve scholarly publication.

**Competing Interests** The authors declare no competing interests.

## References

1. P. Rawla, "Epidemiology of prostate cancer," *World journal of oncology,* vol. 10, no. 2, p. 63, 2019.
2. H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians,* vol. 71, no. 3, pp. 209-249, 2021.
3. C. J. Magnani *et al.*, "PSA Testing Use and Prostate Cancer Diagnostic Stage After the 2012 U.S. Preventive Services Task Force Guideline Changes," (in eng), *J Natl Compr Canc Netw,* vol. 17, no. 7, pp. 795–803, Jul 1 2019, https://doi.org/10.6004/jnccn.2018.7274.
4. L. M. Johnson, B. Turkbey, W. D. Figg, and P. L. Choyke, "Multiparametric MRI in prostate cancer management," (in eng), *Nature reviews. Clinical oncology,* vol. 11, no. 6, pp. 346-53, Jun 2014, https://doi.org/10.1038/nrclinonc.2014.69.
5. J. Thompson, N. Lawrentschuk, M. Frydenberg, L. Thompson, P. Stricker, and USANZ, "The role of magnetic resonance imaging in the diagnosis and management of prostate cancer," *BJU International,* vol. 112, no. S2, pp. 6-20, 2013, https://doi.org/10.1111/bju.12381.
6. A. Bezinque, A. Moriarity, C. Farrell, H. Peabody, S. L. Noyes, and B. R. Lane, "Determination of prostate volume: a comparison of contemporary methods," *Academic radiology,* vol. 25, no. 12, pp. 1582-1587, 2018.
7. T. Barrett, M. de Rooij, F. Giganti, C. Allen, J. O. Barentsz, and A. R. Padhani, "Quality checkpoints in the MRI-directed prostate cancer diagnostic pathway," *Nature Reviews Urology,* vol. 20, no. 1, pp. 9-22, 2023.
8. A. Firjani, A. Elmaghraby, and A. El-Baz, "MRI-based diagnostic system for early detection of prostate cancer," in *2013 Biomedical Sciences and Engineering Conference (BSEC)*, 2013: IEEE, pp. 1–4.
9. D. R. Christie and C. F. Sharpley, "How accurately can prostate gland imaging measure the prostate gland volume? Results of a systematic review," *Prostate Cancer,* vol. 2019, 2019.
10. W. L. Smith *et al.*, "Prostate volume contouring: a 3D analysis of segmentation using 3DTRUS, CT, and MR," *International Journal of Radiation Oncology* Biology* Physics,* vol. 67, no. 4, pp. 1238-1247, 2007.
11. Z. Khan, N. Yahya, K. Alsaih, M. I. Al-Hiyali, and F. Meriaudeau, "Recent automatic segmentation algorithms of MRI prostate regions: a review," *IEEE Access,* vol. 9, pp. 97878-97905, 2021.
12. N. Aldoj, F. Biavati, F. Michallek, S. Stober, and M. Dewey, "Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net," *Scientific reports,* vol. 10, no. 1, p. 14315, 2020.
13. A. Comelli *et al.*, "Deep learning-based methods for prostate segmentation in magnetic resonance imaging," *Applied Sciences,* vol. 11, no. 2, p. 782, 2021.
14. R. Cuocolo *et al.*, "Deep Learning Whole-Gland and Zonal Prostate Segmentation on a Public MRI Dataset," (in eng), *J Magn Reson Imaging,* vol. 54, no. 2, pp. 452-459, Aug 2021, https://doi.org/10.1002/jmri.27585.
15. K. V. Sarma *et al.*, "Harnessing clinical annotations to improve deep learning performance in prostate segmentation," *Plos one,* vol. 16, no. 6, p. e0253829, 2021.

16. S. Montagne *et al.*, "Challenge of prostate MRI segmentation on T2-weighted images: inter-observer variability and impact of prostate morphology," *Insights into imaging,* vol. 12, no. 1, p. 71, 2021.
17. A. S. Korsager *et al.*, "The use of atlas registration and graph cuts for prostate segmentation in magnetic resonance images," *Medical physics,* vol. 42, no. 4, pp. 1614-1624, 2015.
18. F. W. Prior *et al.*, "TCIA: An information resource to enable open science," (in eng), *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference,* vol. 2013, pp. 1282–5, 2013, https://doi.org/10.1109/embc.2013.6609742.
19. U. R. Chandran *et al.*, "TCGA Expedition: A Data Acquisition and Management System for TCGA Data," (in eng), *PLoS One,* vol. 11, no. 10, p. e0165395, 2016, https://doi.org/10.1371/journal.pone.0165395.
20. A. Saha *et al.*, "Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study," *The Lancet Oncology,* vol. 25, no. 7, pp. 879-887, 2024, https://doi.org/10.1016/S1470-2045(24)00220-1.
21. I. a. H. Alex Krizhevsky; Sutskever, Geoffrey E., *ImageNet Classification with Deep Convolutional Neural Networks* (Advances in Neural Information Processing Systems 25). Curran Associates, Inc., 2012.
22. A. Albayrak and G. Bilgin, "Mitosis detection using convolutional neural network based features," in *2016 IEEE 17th International Symposium on Computational Intelligence and Informatics (CINTI)*, 17–19 Nov. 2016 2016, pp. 000335–000340, https://doi.org/10.1109/CINTI.2016.7846429.
23. R. Cao *et al.*, "Prostate Cancer Detection and Segmentation in Multi-parametric MRI via CNN and Conditional Random Field," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 8–11 April 2019 2019, pp. 1900–1904, https://doi.org/10.1109/ISBI.2019.8759584.
24. M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," (in eng), *Journal of digital imaging,* vol. 32, no. 4, pp. 582-596, Aug 2019, https://doi.org/10.1007/s10278-019-00227-x.
25. H. Mzoughi *et al.*, "Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification," *Journal of Digital Imaging,* vol. 33, pp. 903-915, 2020.
26. R. Cuocolo *et al.*, "Deep Learning Whole-Gland and Zonal Prostate Segmentation on a Public MRI Dataset," *Journal of Magnetic Resonance Imaging,* vol. 54, no. 2, pp. 452-459, 2021, https://doi.org/10.1002/jmri.27585.
27. J. Meglič, M. R. S. Sunoqrot, T. F. Bathen, and M. Elschot, "Label-set impact on deep learning-based prostate segmentation on MRI," *Insights into Imaging,* vol. 14, no. 1, p. 157, 2023/09/25 2023, https://doi.org/10.1186/s13244-023-01502-w.
28. L. C. Adams *et al.*, "Prostate158 - An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection," *Computers in Biology and Medicine,* vol. 148, p. 105817, 2022/09/01/ 2022, https://doi.org/10.1016/j.compbiomed.2022.105817.
29. Z. Liao, S. Hu, Y. Xie, and Y. Xia, "Modeling annotator preference and stochastic annotation error for medical image segmentation," *Medical Image Analysis,* vol. 92, p. 103028, 2024/02/01/ 2024, https://doi.org/10.1016/j.media.2023.103028.
30. Z. Liao, Y. Xie, S. Hu, and Y. Xia, "Learning From Ambiguous Labels for Lung Nodule Malignancy Prediction," *IEEE Transactions on Medical Imaging,* vol. 41, no. 7, pp. 1874-1884, 2022, https://doi.org/10.1109/TMI.2022.3149344.
31. Z. Liao, S. Hu, Y. Xie, and Y. Xia, "Instance-dependent Label Distribution Estimation for Learning with Label Noise," *International Journal of Computer Vision,* 2024/12/02 2024, https://doi.org/10.1007/s11263-024-02299-x.
32. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015// 2015: Springer International Publishing, pp. 234–241.
33. G. Litjens, Debats, O., Barentsz, J., Karssemeijer, N., & Huisman, H. "PROSTATEx | SPIE-AAPM-NCI PROSTATEx Challenges." Cancer Imaging Archive. https://www.cancerimagingarchive.net/collection/prostatex/ (accessed.
34. A. R. Mayer *et al.*, "An evaluation of Z-transform algorithms for identifying subject-specific abnormalities in neuroimaging data," (in eng), *Brain Imaging Behav,* vol. 12, no. 2, pp. 437-448, Apr 2018, https://doi.org/10.1007/s11682-017-9702-2.
35. Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential data augmentation techniques for medical imaging classification tasks," in *AMIA annual symposium proceedings*, 2017, vol. 2017: American Medical Informatics Association, p. 979.
36. S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," (in eng), *J Med Syst,* vol. 42, no. 11, p. 226, Oct 8 2018, https://doi.org/10.1007/s10916-018-1088-1.
37. Y. Zhu *et al.*, "Fully automatic segmentation on prostate MR images based on cascaded fully convolution network," (in eng), *J Magn Reson Imaging,* vol. 49, no. 4, pp. 1149-1156, Apr 2019, https://doi.org/10.1002/jmri.26337.
38. M. H. Asnawi *et al.*, "Lung and Infection CT-Scan-Based Segmentation with 3D UNet Architecture and Its Modification," in *Healthcare*, 2023, vol. 11, no. 2: MDPI, p. 213.
39. F. Abdalbagi, S. Viriri, and M. T. Mohammed, "Bata-unet: Deep learning model for liver segmentation," *Signal & Image Processing: An International Journal (SIPIJ) Vol,* vol. 11, 2020.
40. A. Temenos, N. Temenos, A. Doulamis, and N. Doulamis, "On the exploration of automatic building extraction from RGB satellite images using deep learning architectures based on U-Net," *Technologies,* vol. 10, no. 1, p. 19, 2022.
41. S.-T. Tran, C.-H. Cheng, T.-T. Nguyen, M.-H. Le, and D.-G. Liu, "TMD-Unet: Triple-Unet with Multi-Scale Input Features and Dense Skip Connection for Medical Image Segmentation," *Healthcare,* vol. 9, no. 1, p. 54, 2021. [Online]. Available: https://www.mdpi.com/2227-9032/9/1/54.
42. W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3D-UNet: separable 3D U-Net for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, 2019: Springer, pp. 358-368.
43. L. Ding, K. Zhao, X. Zhang, X. Wang, and J. Zhang, "A lightweight U-Net architecture multi-scale convolutional network for pediatric hand bone segmentation in X-ray image," *IEEE Access,* vol. 7, pp. 68436-68445, 2019.
44. A. Avesta, S. Hossain, M. Lin, M. Aboian, H. M. Krumholz, and S. Aneja, "Comparing 3D, 2.5 D, and 2D approaches to brain image auto-segmentation," *Bioengineering,* vol. 10, no. 2, p. 181, 2023.
45. M. Perslev, E. B. Dam, A. Pai, and C. Igel, "One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 2019: Springer, pp. 30-38.
46. O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint* arXiv:1804.03999*,* 2018.

47. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
48. Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, 2018: Springer, pp. 3–11.
49. K. H. Zou *et al.*, "Statistical validation of image segmentation quality based on a spatial overlap index," (in eng), *Acad Radiol,* vol. 11, no. 2, pp. 178-89, Feb 2004, https://doi.org/10.1016/s1076-6332(03)00671-8.
50. O. U. Aydin *et al.*, "On the usage of average Hausdorff distance for segmentation performance assessment: hidden error when used for ranking," (in eng), *Eur Radiol Exp,* vol. 5, no. 1, p. 4, Jan 21 2021, https://doi.org/10.1186/s41747-020-00200-2.
51. L. Cai, Z. Wang, R. Kulathinal, S. Kumar, and S. Ji, "Deep Low-Shot Learning for Biological Image Classification and Visualization From Limited Training Samples," (in eng), *IEEE Trans Neural Netw Learn Syst,* vol. 34, no. 5, pp. 2528-2538, May 2023, https://doi.org/10.1109/tnnls.2021.3106831.
52. A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, 2019: Springer, pp. 311-320.
53. L. C. Adams *et al.*, "Prostate158-An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection," *Computers in Biology and Medicine,* vol. 148, p. 105817, 2022.
54. D. Karimi, G. Samei, C. Kesch, G. Nir, and S. E. Salcudean, "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models," *International journal of computer assisted radiology and surgery,* vol. 13, pp. 1211-1219, 2018.
55. M. Montazerolghaem, Y. Sun, G. Sasso, and A. Haworth, "U-Net Architecture for Prostate Segmentation: The Impact of Loss Function on System Performance," *Bioengineering,* vol. 10, no. 4, p. 412, 2023. [Online]. Available: https://www.mdpi.com/2306-5354/10/4/412.
56. A. Meyer *et al.*, "Automatic high resolution segmentation of the prostate from multi-planar MRI," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018: IEEE, pp. 177–181.
57. T. Clark, A. Wong, M. A. Haider, and F. Khalvati, "Fully deep convolutional neural networks for segmentation of the prostate gland in diffusion-weighted MR images," in *Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings 14*, 2017: Springer, pp. 97-104.
58. L. Yu, X. Yang, H. Chen, J. Qin, and P. A. Heng, "Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31, no. 1.
59. Q. Zhu, B. Du, B. Turkbey, P. L. Choyke, and P. Yan, "Deeply-supervised CNN for prostate segmentation," in *2017 international joint conference on neural networks (IJCNN)*, 2017: IEEE, pp. 178–184.
60. F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine,* vol. 152, p. 106391, 2023.
61. L. Xu *et al.*, "Development and clinical utility analysis of a prostate zonal segmentation model on T2-weighted imaging: a multicenter study," *Insights into Imaging,* vol. 14, no. 1, p. 44, 2023/03/16 2023, https://doi.org/10.1186/s13244-023-01394-w.
62. L. A. Johnson *et al.*, "Automated prostate gland segmentation in challenging clinical cases: comparison of three artificial intelligence methods," *Abdominal Radiology,* vol. 49, no. 5, pp. 1545–1556, 2024/05/01 2024, https://doi.org/10.1007/s00261-024-04242-7.
63. M. Baldeon-Calisto *et al.*, "A multi-object deep neural network architecture to detect prostate anatomy in T2-weighted MRI: Performance evaluation," (in English), *Frontiers in Nuclear Medicine,* Original Research vol. 2, 2023-February-06 2023, https://doi.org/10.3389/fnume.2022.1083245.
64. O. J. Pellicer-Valero *et al.*, "Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images," *Scientific Reports,* vol. 12, no. 1, p. 2975, 2022/02/22 2022, https://doi.org/10.1038/s41598-022-06730-6.
65. S. Kuanar *et al.*, "Transition-zone PSA-density calculated from MRI deep learning prostate zonal segmentation model for prediction of clinically significant prostate cancer," *Abdominal Radiology,* vol. 49, no. 10, pp. 3722–3734, 2024/10/01 2024, https://doi.org/10.1007/s00261-024-04301-z.
66. D. Li *et al.*, "Deep learning in prostate cancer diagnosis using multiparametric magnetic resonance imaging with whole-mount histopathology referenced delineations," *Frontiers in medicine,* vol. 8, p. 810995, 2022.
67. H. Wang *et al.*, "Scientific discovery in the age of artificial intelligence," (in eng), *Nature,* vol. 620, no. 7972, pp. 47-60, Aug 2023, https://doi.org/10.1038/s41586-023-06221-2.