





## RESEARCH ARTICLE

# Neural network segmentation of disc volume from magnetic resonance images and the effect of degeneration and spinal level

Milad I. Markhali | John M. Peloquin  | Kyle D. Meadows  |  
Harrah R. Newman  | Dawn M. Elliott 

Department of Biomedical Engineering, University of Delaware, Newark, Delaware, USA

## Correspondence

Dawn M. Elliott, Blue and Gold Distinguished Professor of Biomedical Engineering, Biomedical Engineering, University of Delaware, 201 STAR Health Sciences Center, Newark, DE 19716, USA.  
Email: [delliott@udel.edu](mailto:delliott@udel.edu)

## Funding information

National Institute of Arthritis and Musculoskeletal and Skin Diseases, Grant/Award Numbers: F31AR081687, R01AR050052; National Institute of General Medical Sciences, Grant/Award Number: P20GM139760

## Abstract

**Background:** Magnetic resonance imaging (MRI) noninvasively quantifies disc structure but requires segmentation that is both time intensive and susceptible to human error. Recent advances in neural networks can improve on manual segmentation. The aim of this study was to establish a method for automatic slice-wise segmentation of 3D disc volumes from subjects with a wide range of age and degrees of disc degeneration. A U-Net convolutional neural network was trained to segment 3D T1-weighted spine MRI.

**Methods:** Lumbar spine MRIs were acquired from 43 subjects (23–83 years old) and manually segmented. A U-Net architecture was trained using the TensorFlow framework. Two rounds of model tuning were performed. The performance of the model was measured using a validation set that did not cross over from the training set. The model version with the best Dice similarity coefficient (DSC) was selected in each tuning round. After model development was complete and a final U-Net model was selected, performance of this model was compared between disc levels and degeneration grades.

**Results:** Performance of the final model was equivalent to manual segmentation, with a mean DSC =  $0.935 \pm 0.014$  for degeneration grades I–IV. Neither the manual segmentation nor the U-Net model performed as well for grade V disc segmentation. Compared with the baseline model at the beginning of round 1, the best model had fewer filters/parameters (75%), was trained using only slices with at least one disc-labeled pixel, applied contrast stretching to its input images, and used a greater dropout rate.

**Conclusion:** This study successfully trained a U-Net model for automatic slice-wise segmentation of 3D disc volumes from populations with a wide range of ages and disc degeneration. The final trained model is available to support scientific use.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *JOR Spine* published by Wiley Periodicals LLC on behalf of Orthopaedic Research Society.

## KEYWORDS

aging, degeneration, imaging, intervertebral disc

## 1 | INTRODUCTION

Disc degeneration, structural abnormalities, and pathology are implicated in low back pain (LBP) and noninvasive measurements of these often require segmenting the disc regions from spine magnetic resonance images (MRI). Trained human readers can produce disc segmentations, but this process is time intensive and interpretation of ambiguous disc borders are susceptible to human error and bias.<sup>1,2</sup> Furthermore, the time it takes to segment the MRI causes a gap between MRI acquisition and MRI evaluation for obtaining useful outputs. Although numerous auto- or semi-automatic segmentation approaches have been developed over the past decade, recent advances in neural networks (deep learning) may improve segmentation by reducing processing time and increasing segmentation consistency.<sup>1,3-8</sup>

Convolutional neural networks (CNNs) are a powerful and convenient method for automatic segmentation, provided a pre-labeled dataset exists to train the CNN model. Because the derivative of a CNN can be computed efficiently, the model parameters can be automatically optimized (trained) by gradient descent to classify each pixel in the image as *disc* or *not disc* according to a set of pre-labeled data, the training set.<sup>9,10</sup> During training, the model *learns* to extract and use image features automatically. The features used for segmentation are thus not chosen by a human-driven design process and are not necessarily even human-perceptible. The U-Net architecture is a widely used encoder-decoder CNN for medical image segmentation and has remarkable performance across various imaging modalities, including challenging low-contrast images and small image datasets.<sup>11</sup> Therefore, the U-Net architecture may be ideal for disc segmentation of spine MRI and was evaluated in this study.

CNNs, such as U-Net or others, have been applied to the problem of disc segmentation, with typical accuracy, as measured using the Dice similarity coefficient (DSC), of 0.89–0.94 for segmentations of 2D (mid-sagittal) and 3D MRI.<sup>1,3-8,12,13</sup> Studies with 181–4075 subjects had mean DSC  $\geq 0.93$ ,<sup>3,4,8,12</sup> reflecting the utility of a large training set. Although large datasets are readily achieved for qualitative assessments of spine clinical phenotypes (e.g., degenerative grade, herniation, Modic changes) in a single mid-sagittal slice,<sup>9,14-16</sup> large training sets of volume segmentations from 3D MRI are not readily available. Fortunately, good model performance can also be obtained with moderately sized training sets, with a DSC of 0.96 on a single 2D mid-sagittal MRI slice reported by a 50-subject study using a U-Net model.<sup>5</sup> However, one of the reasons to segment the disc is to study disc degeneration, and disc degeneration causes shape and boundary irregularities that may adversely impact segmentation accuracy. Similarly, shape differences with spinal level, particularly L5-S1, which is highly variable among the population, may affect model performance. Unfortunately, with the exception of the 2D Huang et al.<sup>5</sup> study, prior

studies have not reported the degree of disc degeneration in their training set. Thus, the accuracy of 3D volume segmentation when the training set includes degenerated discs remains unknown, limiting our ability to accurately use automatic segmentation models on discs of higher degeneration grades, which are more likely to be clinically relevant.

The aim of this study was to establish a method for automatic 3D slice-wise segmentation of disc volumes from populations with a wide range of age and degrees of disc degeneration. A U-Net CNN was trained to segment 3D T1w spine MRI and its performance was tuned by comparing several versions of the training process and model structure. The performance of the final U-Net model was assessed by disc level and degeneration grade.

## 2 | METHODS

### 2.1 | Subjects and magnetic resonance image acquisition

Lumbar spine MRIs were acquired from 43 subjects across a large age range (23–83 years old, 24 female/19 male) under an approved IRB protocol. T1-weighted (T1w) FLASH (Fast Low Angle Shot) images (Repetition Time [TR] = 9.6 ms, Echo time [TE] = 3.7 ms, resolution =  $0.52 \times 0.52 \times 3.00$  mm, sagittal slices, run time 11 min) were acquired on a 3T Siemens scanner.<sup>17</sup> The T1w FLASH sequence provided sufficient contrast between the disc and surrounding structures. Each subject was imaged four times, in different postures and times of day.<sup>17</sup> T2-weighted TSE (turbo spin echo) was also acquired to grade disc degeneration with the Pfirrmann scale, where low grade (I) is healthy and high grade (V) is considered degenerated.<sup>18</sup> Disc grading was performed by 3 trained graders who reached consensus.<sup>17</sup>

### 2.2 | Manual segmentation and human performance assessment

Each FLASH MRI was manually segmented for all slices and lumbar levels, L1–L2 to L5–S1, by trained readers using the open-source ITK-SNAP software.<sup>19</sup> These segmentations were considered ground truth for U-Net model training and validation. To set a target for the expected performance of CNN segmentation, inter-reader segmentation accuracy was measured for six readers across 9 subjects (7 young: age < 60 y, 2 older: age  $\geq 60$  y) with grade I–IV discs. Segmentation of grade V discs was avoided because they are often collapsed and partially ossified. The hypothesis that inter-reader DSC differed across discs ( $N = 45$ ) by grade and level was tested using a one-factor

Kruskal–Wallis test with post-hoc Wilcoxon signed rank tests with Holm correction for multiple comparisons. In this study, all statistical analyses were performed in R version 4.3.0.<sup>20</sup>

## 2.3 | Image split

To train and evaluate the U-Net architecture, MR images and the masks from 35 subjects were divided into a training set and a validation set (Table 1). Because degenerated discs have reduced MR contrast and variable disc boundaries, the 35 subjects were split semi-randomly to achieve a balanced distribution of degeneration grades (Table 1). The subjects with grade V discs were assigned exclusively to the training set due to their scarcity. After the model training was completed, an expanded validation set was used to test for the effect of degeneration and spinal level. To do so, 8 additional subjects (containing at least one grade V disc) were added to the original validation set (Table 1).

For each subject, MRIs of all four postures/times were used for training, with the repetition serving as data augmentation, but only the first image acquired was used for validation so that each example would be independent. There was no crossover of subjects between the training and validation sets.

## 2.4 | U-Net model

Based on its proven architecture and highly competitive performance in various applications, including spine MRI, this study used U-Net architecture,<sup>11</sup> set up to operate on sagittal slices from each spine image recruited as TIFF stacks. The U-Net accepts square images as input, so the 352 × 512 pixel sagittal MRI slices in the current study were resized to 512 × 512 pixels prior to processing by the U-Net (Figure 1).

The model was trained using the TensorFlow framework (version 2.8.2)<sup>21,22</sup> on a personal computer with an NVIDIA GeForce RTX 3080 using the training set and ground truth labels described above. The optimization algorithm was Adaptive Moment Estimation (Adam)

and the loss function was the sum of cross-entropy and Dice loss. Dropout was applied at the U-Net bottleneck and parameters were randomly initialized (Figure 1). The U-Net model produces 388 × 388 pixel segmentations, so these were resized to match the input 352 × 512 pixel sagittal image slices. Lastly, the original MRI header information, including image origin, voxel spacing, and image orientation, was added to the resized output segmentation using the Convert3d.

## 2.5 | U-net model tuning

### 2.5.1 | Overview

Two rounds of model tuning were performed to choose specific hyperparameters and modifications to the model's convolutional filters. Initial hyperparameter values were found by a grid search in a pilot study, which is common practice, and were learning rate = 0.001, first moment decay rate ( $\beta_1$ ) = 0.95, second moment decay rate ( $\beta_2$ ) = 0.999, dropout rate = 0.2, weight decay = 0.05, batch size = 6, and number of epochs = 2000. This was the starting point for round 1 tuning.

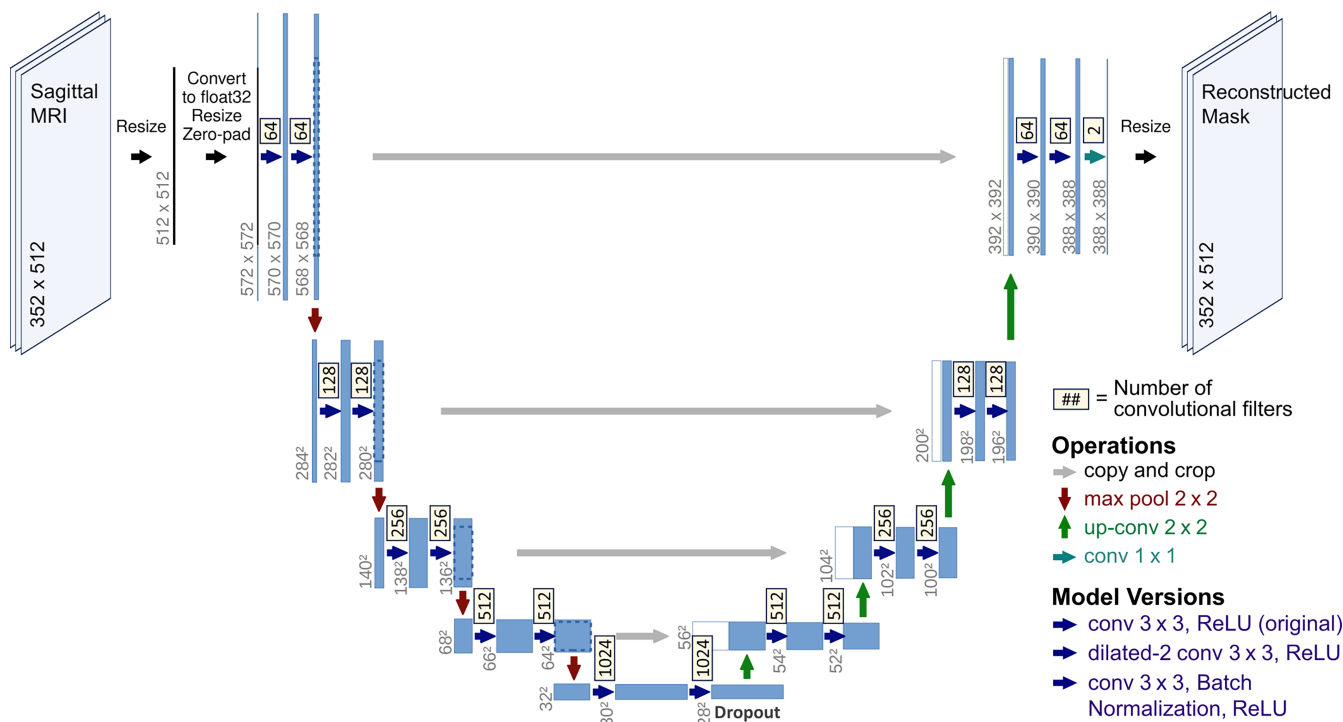
In each round of tuning, all candidate model versions under evaluation were trained 8 times (e.g., 8 runs per version) to account for randomness in model training. The predictive performance of the model resulting from each run was measured using the validation set, where DSC was calculated using Convert3D<sup>19</sup> (see Appendix A). The model version with the best average performance (DSC) across runs was selected for subsequent use.

### 2.5.2 | Round 1 tuning

The objective of the first round of tuning was to (a) choose the number of convolutional filters (i.e., parameters) in the model, and (b) choose the image preprocessing method. The starting point was

**TABLE 1** Split of images into training and validation datasets.

Age group	Number of subjects	Number of discs	Number (percent) of discs for each degeneration grade				
			I	II	III	IV	V
Training set							
Young	17	85	23 (27)	45 (53)	7 (8)	10 (12)	0 (0)
Older	10	50	0 (0)	3 (6)	18 (36)	24 (48)	5 (10)
All	27	135	23 (17)	48 (36)	25 (19)	34 (25)	5 (4)
Validation set							
Young	4	20	8 (40)	7 (35)	2 (10)	3 (15)	0 (0)
Older	4	20	0 (0)	2 (10)	8 (40)	10 (50)	0 (0)
All	8	40	8 (20)	9 (22)	10 (25)	13 (33)	0 (0)
Expanded validation set							
Added	8	40	1 (2)	3 (8)	8 (20)	11 (28)	17 (42)
Full set	16	80	9 (11)	12 (15)	18 (22)	24 (30)	17 (21)



**FIGURE 1** U-Net model architecture used in this study,<sup>11</sup> showing processing of one MRI slice. Several model versions with different convolution operations were created and compared during model tuning.

“model 1” in Table 2. Nine additional model versions were evaluated with 100%, 75%, or 50% of the original number of parameters and three different image preprocessing methods (Table 2, models 2–10). Since reducing the number of parameters also reduced the memory required for training, the batch size was increased to 8 with 75% parameters and 14 with 50% parameters. Adjusting the number of trainable parameters was intended to detect overfit or underfit of the training data.

Four image inputs were evaluated: raw data and three candidate image preprocessing steps. These preprocessing methods were applied to images prior to using them as input to the U-Net model. Input 1, “Raw Input,” had no modification to the images (Figure 2A,C). Input 2, “Deleted Blanks,” removed MRI slices that had zero disc-labeled pixels from the training set, reducing the number of training slices from 2434 to 1918. No slices were removed from the validation set, as it is meant to estimate model performance on unlabeled images. Input 3, “Deleted Blanks + Stretched Contrast,” modified input 2 by scaling the intensity such the 2.5th percentile (“input minimum”) became 0 and the 97.5th percentile (“input maximum”) became 255 (Figure 2B,D). Input 4, “Deleted Blanks + Augmentation,” modified input 3 by adding a second copy of each slice scaled using an input minimum randomly selected from 0% to 10% with input maximum = 100% – input minimum. This contrast stretching was intended as training data augmentation set to make the model more robust to intensity variation.

The hypothesis that varying the number of model parameters and the input image preprocessing method affected model performance was tested using a linear mixed model with the number of parameters and the preprocessing method as fixed parameters, and subject as a

**TABLE 2** Model versions used in round 1 tuning.

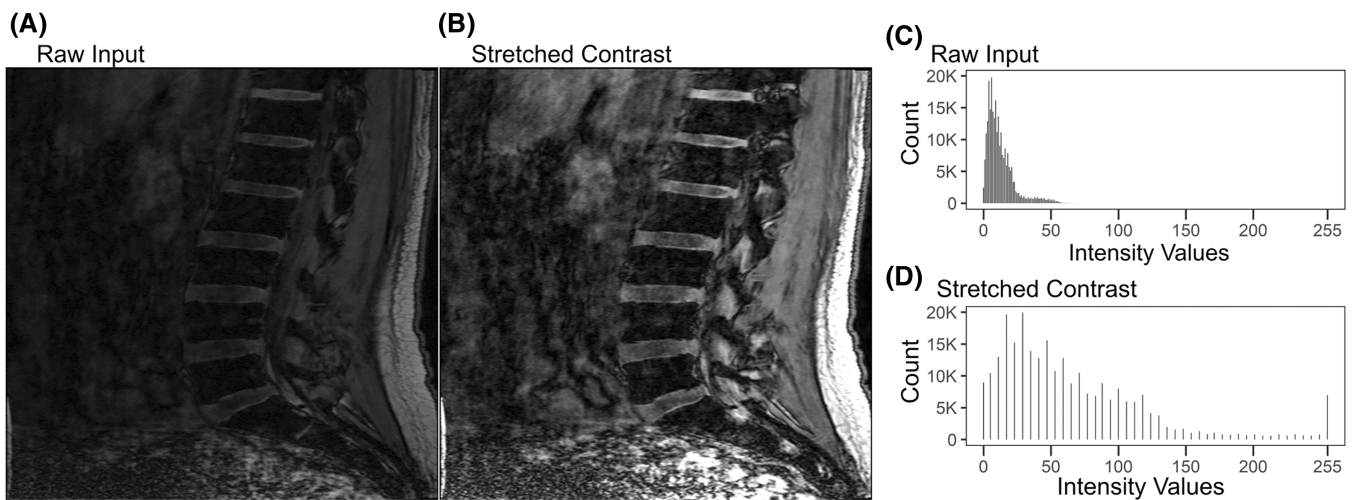
Round 1 model version	Number of U-net parameters	Pre-processed input image
Model 1	100% of Original	Input 1: Raw Input
Model 2	100% of Original	Input 2: Deleted Blanks
Model 3	100% of Original	Input 3: Deleted Blanks + Stretched Contrast
Model 4	100% of Original	Input 4: Deleted Blanks + Augmentation
Model 5	75% of Original	Input 2: Deleted Blanks
Model 6	75% of Original	Input 3: Deleted Blanks + Stretched Contrast
Model 7	75% of Original	Input 4: Deleted Blanks + Augmentation
Model 8	50% of Original	Input 2: Deleted Blanks
Model 9	50% of Original	Input 3: Deleted Blanks + Stretched Contrast
Model 10	50% of Original	Input 4: Deleted Blanks + Augmentation

random intercept. This accounted for correlation of each subject's results across the 8 runs of each model.

### 2.5.3 | Round 2 tuning

The objective of the second round of tuning was (a) to optimize the regularization-related parts of the training process, and (b) to test a





**FIGURE 2** Representative preprocessed input images with (A) original contrast, referred to as “Raw Input,” and (B) stretched contrast with the 95th percentile intensity treated as the maximum, referred to as “Stretched Contrast.” The images have additional brightening for clearer visualization in print. (C) Histogram of the original intensity values in the raw input image. (D) Histogram of modified intensity values in the stretched contrast image. (Note that images in A and B have already been resized for input to the U-Net as described in section 2.4).

**TABLE 3** Model versions used in round 2 tuning.

Round 2 model version	Learning rate	Decay rate $\beta 1$	Dropout at bottleneck	Weight decay	Batch normalization	Dilation rate
Baseline (Round 1, Model 6)	1e-03	0.95	0.2	0.05	N	1
Batch Normalization	1e-03	0.95	0.2	0.05	Y	1
Increased Dropout Rate	5e-04	0.90	0.8	0.05	N	1
Dilated Convolution	1e-03	0.95	0.2	0.05	N	2

structural modification of the U-Net architecture, in which the filter convolutions were replaced by dilated convolutions (Figure 1, “Model Versions”). Three model versions, based on round 1 model 6 (see results), were compared in round 2 (Table 3). In the first version, batch normalization was added to the hidden layer activations, which often stabilizes and speeds convergence. In the second, the dropout rate was increased from 0.2 to 0.8. A higher dropout rate can force the network to learn more robust features, which is especially helpful when the training set is relatively small. In the third, dilated convolution, with corresponding padding of the layer’s input to maintain the size of its output, was used in place of the original hidden layer filter convolutions to increase the receptive field of the convolutions. Dilated convolution has been shown to improve performance.<sup>23</sup> Similar to round 1, segmentation accuracy was compared between the candidate models using a linear mixed model with the model version as a fixed parameter and subject as a random intercept to account for potential correlation of each subject’s results across model runs.

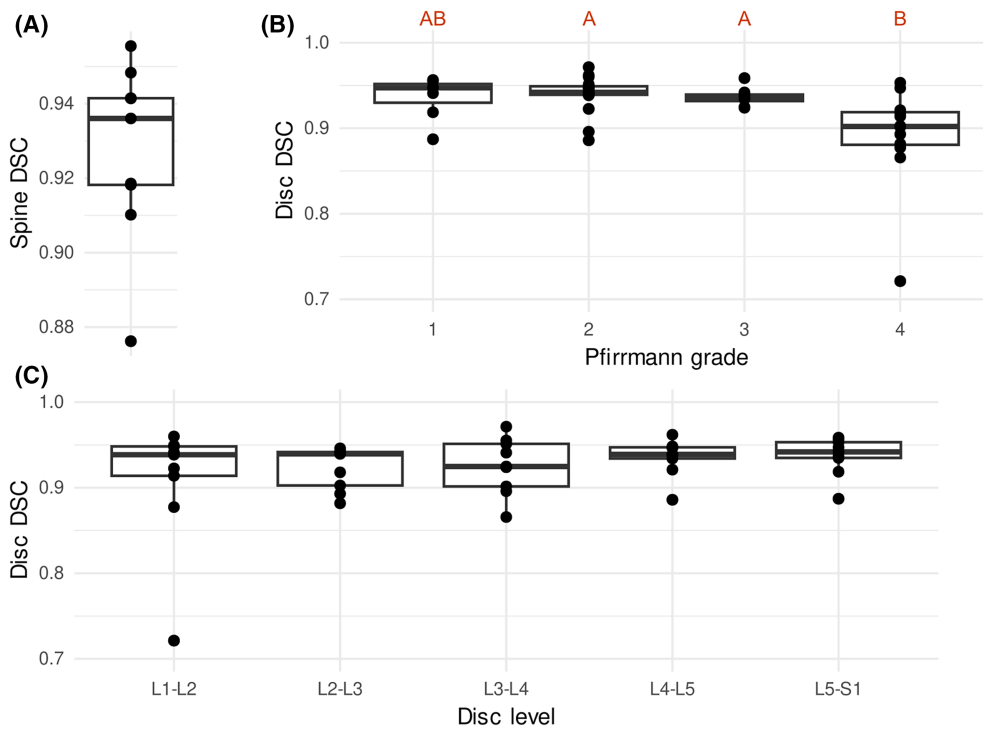
## 2.6 | Orphan removal

Small groups of voxels that are incorrectly labeled as disc and are separated from the main part of the segmentation, or “orphans,” can often be removed by postprocessing. Here, orphans were identified

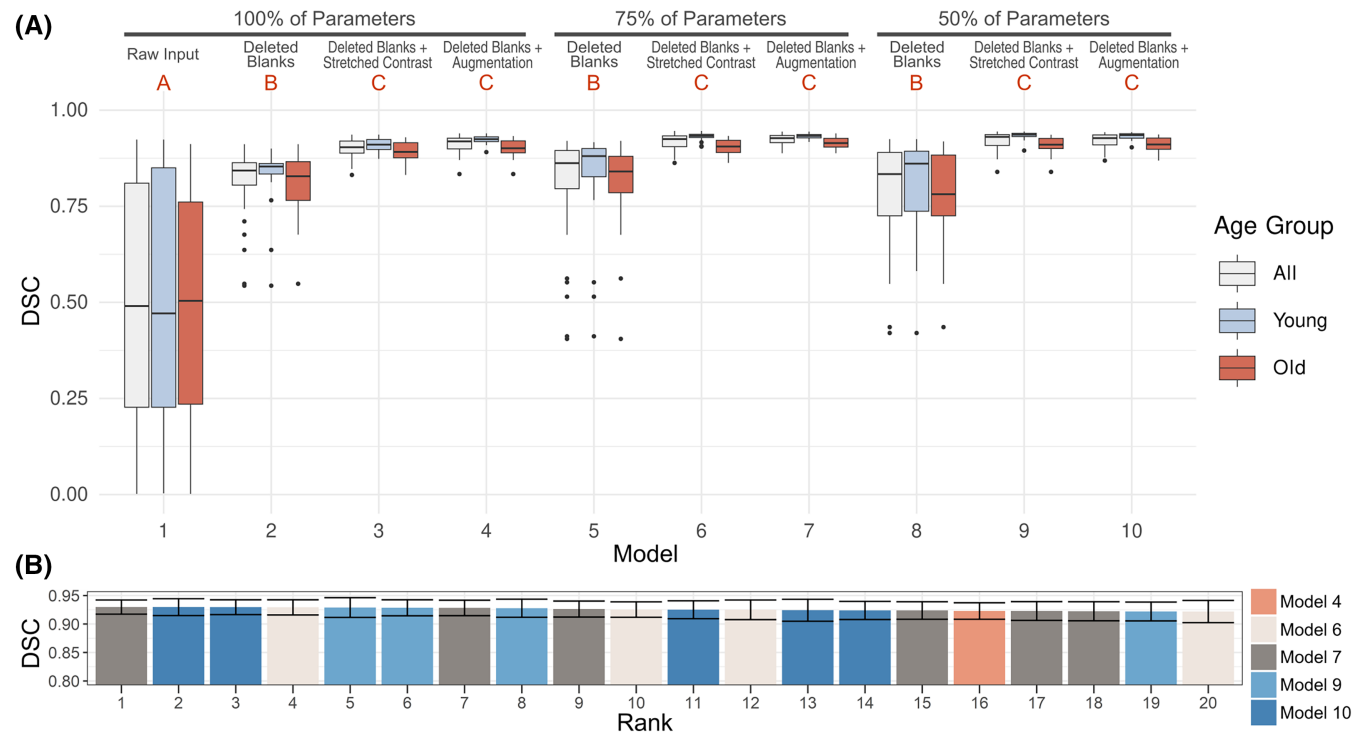
within each slice as any separate labeled area that either (1) had size  $<10$  pixels, (2) did not intersect any disc-labeled pixels in the central slice when superimposed, or (3) did not intersect any disc-labeled pixel in a neighboring slice when superimposed. This was done using a Matlab script. To focus on comparison of the intrinsic performance of the various U-Net model versions, orphan removal was applied only as an extra evaluation step after the second round of model tuning.

## 2.7 | Spine level-based and degenerative grade-based analysis

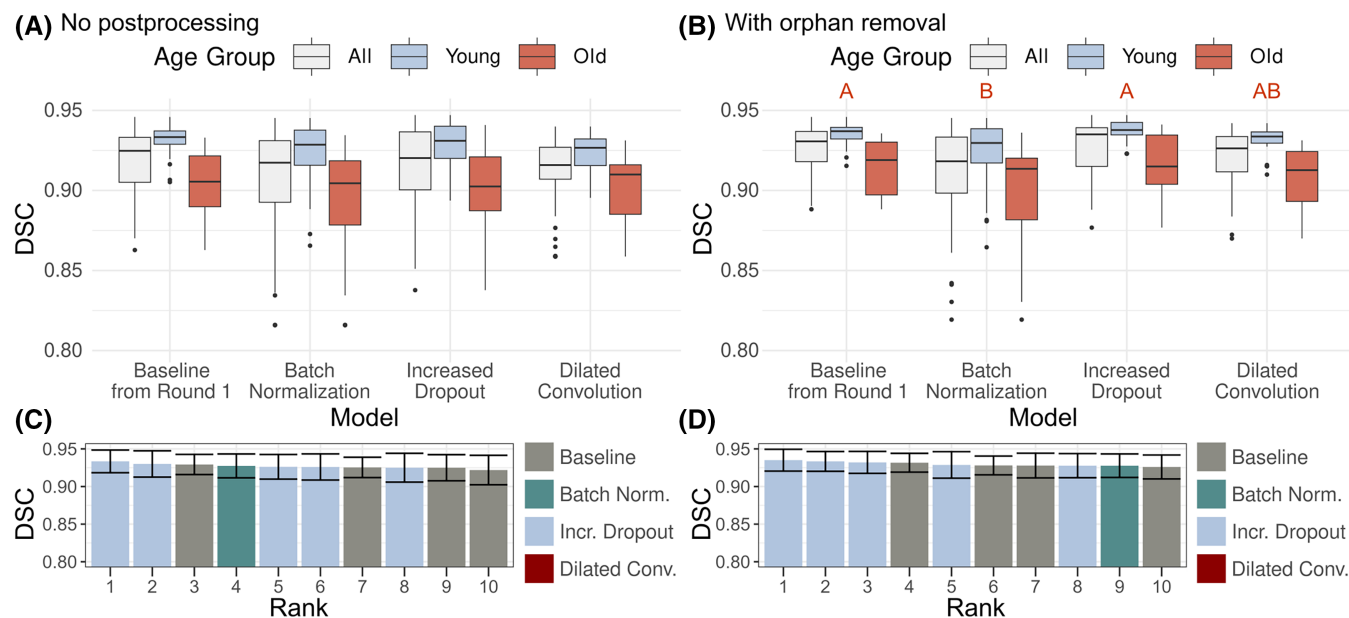
After model development (i.e., training) was complete and a final U-Net model was selected, performance of this “best” model, including orphan removal, was compared between disc levels and degeneration grades. This comparison used data from the 8 subjects that comprised the validation set used in model development as well as 8 additional subjects with grade V discs, with the resulting 16-subject set subsequently referred to as the “expanded validation set” (Table 1). For this final assessment, two similarity metrics, DSC and volume ratio (Appendix A) were used to compare between disc levels and degeneration grades using a one-factor Kruskal-Wallis test with post-hoc Wilcoxon signed rank tests with Holm correction for multiple comparisons ( $N = 80$  discs).



**FIGURE 3** (A) Inter-reader segmentation Dice similarity coefficient (DSC) for 9 subjects. (B) Inter-reader DSC by grade, with a significant overall effect of grade ( $p = 0.006$ ). (C) Inter-reader DSC by disc level. In B and C, groups that do not share a letter on the top margin have significantly different mean DSC ( $p < 0.05$  by pairwise Wilcoxon rank sum test with Holm correction).



**FIGURE 4** Validation performance from round 1 model tuning. (A) Dice similarity coefficient (DSC) (median, quartiles, and range, with outlier points classified using Tukey's definition) including data points for each of the 8 training runs in each model version. Letters A, B, C on the top margin of (A) show significant differences, where models that do not share a letter have significantly different mean DSC ( $p < 0.05$ , pairwise  $t$ -test with Holm correction). (B) The leaderboard of top 20 model runs (of 80 total), ranked from highest mean DSC, shows which model versions occur in the best 20 runs.



**FIGURE 5** Validation Dice similarity coefficient (DSC) from round 2 model tuning. (A) DSC from each model version without postprocessing including data points for each of the 8 training runs in each model version. (B) DSC for each model version with orphan removal. The box plots in A & B show median, quartiles, and range, with outlier points classified using Tukey's definition. Letters on the top margin of A & B show pairwise significant differences, where models that do not share a letter have significantly different mean DSC ( $p < 0.05$ , pairwise  $t$ -test with Holm correction). The leaderboard of top 10 model runs (of 36 total), ranked from highest mean DSC, shows which model versions occur in the best 10 runs for (C) without postprocessing, and (D) with orphan removal.

### 3 | RESULTS

#### 3.1 | Human performance assessment

Inter-reader segmentation similarity metrics for 9 subjects are shown in Figure 3. The inter-reader DSC was  $0.927 \pm 0.024$ , with a range from 0.876 to 0.955 for full spines (Figure 3A). Inter-reader DSC was lower for more degenerated discs (Figure 3B), meaning those segmentations were less reliable. There was no significant effect of disc level on inter-reader DSC (Figure 3C).

#### 3.2 | Round 1 tuning

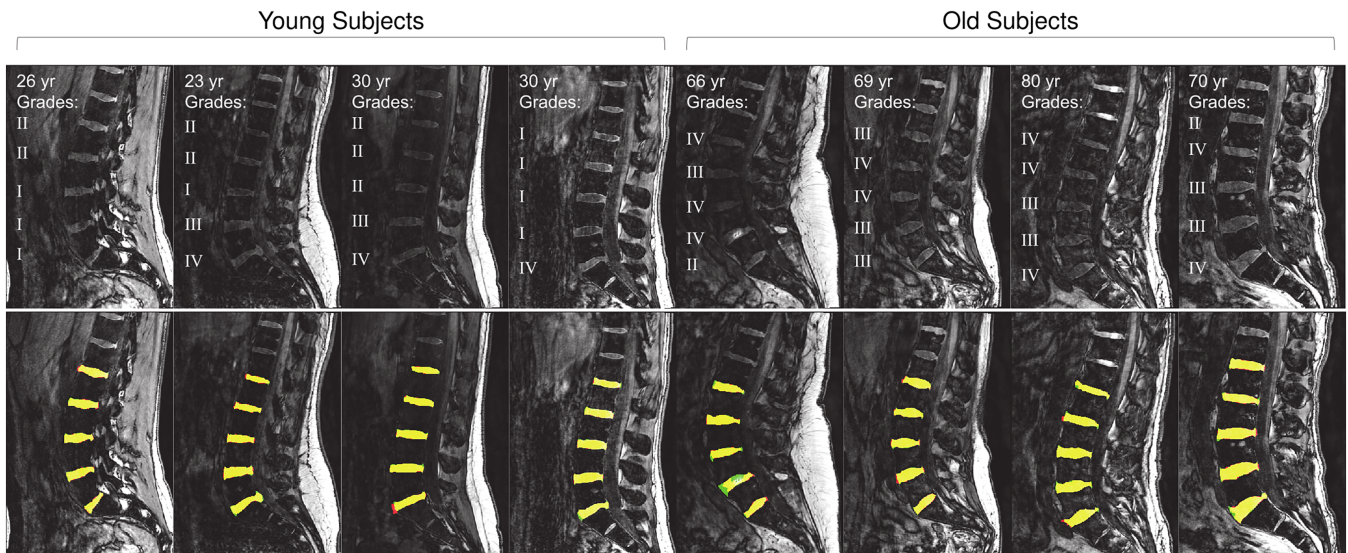
Runs of the baseline model (Model 1, Table 2), with no image preprocessing, had poor performance, with a mean DSC  $\approx 0.5$  on the validation set (Figure 4A). The linear mixed model indicated that the image preprocessing method had a significant effect on model performance ( $p < 0.001$ ) but that the number of model parameters had negligible effect ( $p = 0.6$ ). Post-hoc pairwise model comparisons indicated that all models for which blank MRI slices were removed from the training data set ("Deleted Blanks") had significantly greater DSC on the validation set than the baseline model (Figure 4A). Either form of contrast adjustment ("Stretched Contrast" or "Augmentation") further improved validation DSC (Figure 4A). The leaderboard shows model versions of DSC of the best 20 runs and consisted solely of models with contrast adjustment (Figure 4B). Model 6 (DSC =  $0.918 \pm 0.020$ ), with 75% parameters and contrast stretching, was chosen to

carry forward to round 2 tuning on the basis that (a) addition of synthetic training images through augmentation increased training time with no benefit and (b) use of an intermediate number of parameters minimizes risk of under and over-fitting.

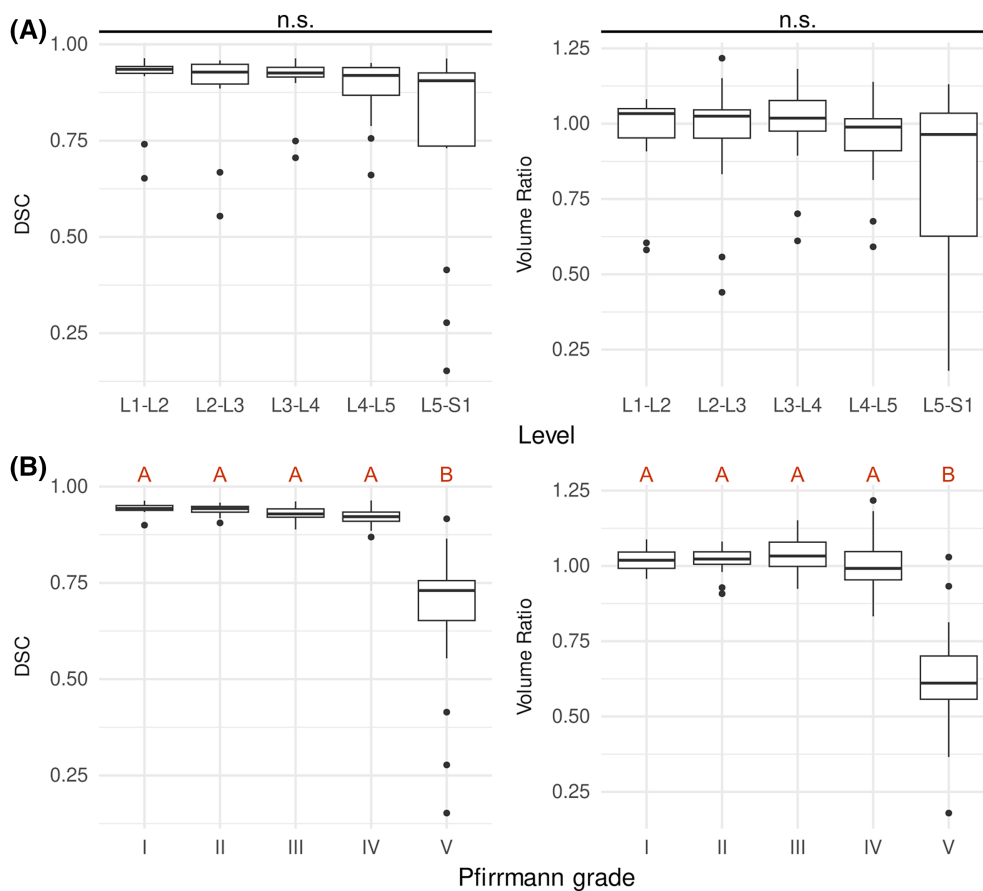
#### 3.3 | Round 2 tuning

Round 2 used model 6 from round 1 (Table 2; Figure 4) as a starting point for further adjustments targeted at regularization-related hyperparameters used during training and increasing the receptive field of the model's filters through use of dilated convolution (Table 3). Without postprocessing, all round 2 model versions, including the baseline, had median DSC above 0.90 (Figure 5A). The linear mixed model indicated that model version had a significant effect ( $p = 0.007$ ), but no post-hoc pairwise comparison was significant. The increased dropout rate model had the best individual run and contributed 5 of the top 10 runs (Figure 5C), even though it had a lower mean validation DSC ( $0.916 \pm 0.025$ ) than the baseline model ( $0.918 \pm 0.020$ ) (Figure 5A).

Applying orphan removal as a postprocessing step slightly improved performance, with the increased dropout model's DSC increasing the most ( $0.011 \pm 0.005$  across runs) and the batch normalization model's DSC increasing the least ( $0.003 \pm 0.003$  across runs). The increased dropout rate model consequently had significantly greater mean validation DSC across runs,  $0.920 \pm 0.018$ , than the other models ( $p < 0.001$ ) (Figure 5B). It also had the best individual run performance (Figure 5D), with a mean DSC =  $0.935 \pm 0.014$



**FIGURE 6** Validation set segmentations from the best run of the increased dropout rate model version, with orphan removal postprocessing. Yellow = correct label, green = false negative, red = false positive.



**FIGURE 7** Performance (mean and standard deviation) of the final U-Net model on the expanded 16-subject validation set including grade V discs by (A) disc level and (B) degeneration grade. Degeneration grade had significant overall effects on both dice similarity coefficient (DSC) and volume ratio. Letters A, B on the top margin of (B) show significant differences, where models that do not share a letter have significantly different mean DSC or volume ratio ( $p < 0.05$ ).

across subjects. The increased dropout rate model was therefore selected as the final model, and its best run (best specific set of trained parameter values) was subsequently examined for performance variation with respect to disc level and degeneration grade. Segmentations from this run are shown in Figure 6.

### 3.4 | U-net performance by disc level and degeneration grade

To test the hypotheses that the U-Net model was sensitive to disc level and degeneration grade, the final model's best run was evaluated



using an expanded validation set with eight additional subjects that contained at least one grade V disc. Disc level had no statistically significant effect on DSC or volume ratio ( $0.05 < p < 0.1$ ; Figure 7A). L5–S1 discs had noticeably greater variance, as expected given they are the most variably sized and shaped lumbar disc. Disc degeneration had a significant overall effect on both DSC ( $p < 1 \times 10^{-8}$ ) and volume ratio ( $p < 1 \times 10^{-6}$ ) (Figure 7B), with grade V discs having smaller DSC and volume ratio. Therefore, the U-Net model has similar performance for grades I–IV, but a major performance loss in the form of under-segmentation for grade V discs.

## 4 | DISCUSSION

### 4.1 | Summary

Automatic segmentation of disc volumes from T1-weighted FLASH MRIs was achieved in this study with a U-Net CNN trained on a small-to-medium size dataset. The model was developed in two rounds of tuning by comparison between multiple candidate models. Performance of the final model, with a mean validation DSC = 0.935 for degeneration grades I–IV, was similar to that of human readers, mean DSC = 0.927. Compared with the baseline model at the beginning of round 1, the best model had fewer filters/parameters (75%), was trained using only slices with at least one disc-labeled pixel, applied contrast stretching to its input images, and used a greater dropout rate. Postprocessing of the output segmentations by removal groups of voxels disconnected from the largest connected component (orphan removal) was also beneficial. The U-Net CNN did not generalize well to grade V discs, with highly variable segmentation quality, reflecting the challenge of developing automatic procedures that are robust to the collapsed and irregular disc shape that characterizes severe disc degeneration.

### 4.2 | Training dataset and comparison to prior work

The size and quality of the training set influence the accuracy of a CNN model. The dataset used in this study for model development consisted of 43 subjects split into a 27-subject training set and an 8-subject validation set, with an additional 8 subjects added to the validation set for evaluation of the final model. This is larger than the 27- and 12-subject datasets used in prior CNN-based 3D segmentation of discs from T1w MRI.<sup>1,7</sup> Importantly, in the present work we quantified the accuracy of the manual segmentations used in our training set, with mean inter-reader DSC =  $0.927 \pm 0.024$  (Figure 3A), a typical level of quality for a “ground truth.”<sup>24–27</sup> Across all candidate models, performance saturated at DSC  $\approx 0.93$ , equivalent to the intrinsic accuracy of the training set. Models with good performance in prior work also have DSCs clustered in the range 0.92–0.94.<sup>3,4,12,28</sup> It is likely that the accuracy of manual segmentation, and thus validation set accuracy, limits the field's ability to detect any improvements in model performance beyond this level. The one study with a greater

reported DSC (0.96) invested more resources in manual segmentation than is typical, using a detailed formal segmentation procedure and requiring consensus between two orthopedic residents and a spine surgeon.<sup>5</sup> This intensive cross-checking was presumably feasible because those segmentations were not fully 3D, with only three MRI slices labeled per subject. Measurement of CNN model performance beyond the DSC = 0.92–0.94 level may require significantly greater investment in preparation of validation data (e.g., multiple segmenters) or use of different evaluation methods (e.g., paired high and low contrast/resolution MRIs).

### 4.3 | Segmentation quality and degeneration grade

The present study examined how CNN segmentation accuracy changed with disc degeneration, which has not been previously reported. Greater disc degeneration was associated with decreases in both manual and automatic segmentation accuracy. The decrease in DSC was slight ( $\sim 0.05$ ) for grade IV but severe for grade V (Figures 3 and 7). The inability of the CNN to accurately segment grade V discs is not particularly surprising. These discs are often collapsed, fragmented, calcified, or have other abnormalities that introduce ambiguity regarding what should be considered disc, bone, or non-disc soft tissue. They are also relatively rare, limiting the number of training examples. This issue extends to application areas that would use automatic 3D segmentation such as finite element modeling, in which it is typical to exclude abnormal disc shapes associated with severe degeneration.<sup>29–32</sup>

### 4.4 | Interpretation of performance differences between models

The main improvements in U-Net model performance compared with baseline were due to (1) deleting slices with blank segmentations from the training set, (2) stretching image contrast, and (3) increasing dropout in combination with orphan removal in postprocessing. Each MRI had 3–6 slices to the left or right of the spine, which are “blank” in the sense that they contain no disc voxels. Deleting these blank slices may have helped by reducing the potential class imbalance in each batch (the tested batch sizes were between 6 and 14, and an MRI may have up to 5 slices on its left and right with no discs visible). Class imbalance tends to distort the cost function gradient. The form of contrast stretching used here is sometimes called winsorization, and is sometimes used in image registration to suppress intensity outliers<sup>33–36</sup>; here, it may serve to normalize image intensity. Increasing dropout is interesting because it was slightly detrimental when used alone, and only beneficial in combination with orphan removal. Dropout may have forced the CNN to learn more robust features as intended, but also decreased its ability to encode large-scale features that indicate where the disc can be located.

Other changes to the model had no significant effect. Decreasing the number of parameters very slightly improved performance, but not to a statistically significant degree. This suggests the CNN is close to optimal complexity. The use of batch normalization did

not have a significant effect, which has been reported previously for U-Net models.<sup>37,38</sup> Use of dilated convolution was intended to expand the image area accessed by each convolution operation, which was previously reported to be beneficial.<sup>23,39</sup> Here, it had no detectable benefit, possibly because the disc is a relatively simple structure and, at MRI resolution, two successive  $3 \times 3$  convolutions (each  $1.5 \times 1.5$  mm) are sufficient to capture all informative local features.

#### 4.5 | Model availability for scientific use

The final trained model is available in the supplemental information of this article to support scientific use. It should be noted that this model may not necessarily generalize, in that images with different properties than the training set used in this study may not be accurately assessed by the model. For example, this study used T1-weighted MRI to achieve contrast between the disc and the vertebral body, whereas many studies use T2-weighted MRI. To use this model, it may be necessary to fine-tune it, using the existing weights of the current model as a starting point for continued training on new data. Often, a model can be fine-tuned to a specific application with little additional training data.<sup>40–42</sup>

#### 4.6 | Conclusion

The study successfully trained a U-Net model for automatic slice-wise segmentation of 3D disc volumes in a population with a wide range of age and disc degeneration and measured a small decrease in segmentation accuracy at grade IV and a large decrease in accuracy at grade V for both manual segmentation and the U-Net. This trained model is publicly available for use in routine disc segmentation. This decrease in accuracy with advanced degeneration is an important finding for applications of the U-Net model.

#### AUTHOR CONTRIBUTIONS

The project concept was developed by MIM, JMP, and DME. Data collection was conducted by HRN, KDM, MIM, JMP. Data interpretation and manuscript preparation was done by MIM, JMP, HRN, KDM, and DME.

#### ACKNOWLEDGMENTS

This study was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (grant numbers: R01AR050052 and F31AR081687) and the National Institute of General Medical Sciences (grant number: P20GM139760).

#### ORCID

John M. Peloquin  <https://orcid.org/0000-0001-7145-6476>

Kyle D. Meadows  <https://orcid.org/0000-0002-1113-8633>

Harrah R. Newman  <https://orcid.org/0000-0001-5555-4308>

Dawn M. Elliott  <https://orcid.org/0000-0003-4792-1029>

#### REFERENCES

- Hess M, Allaire B, Gao KT, et al. Deep learning for multi-tissue segmentation and fully automatic personalized biomechanical models from BACPAC clinical lumbar spine MRI. *Pain Med.* 2023;24(Suppl 1): S139–S148. doi:10.1093/pm/pnac142
- Li I, Cook K, Le M, Gaonkar BK, Macyszyn L. Multi-resolution deep network ensembles for cervical intervertebral disc segmentation are biased by trainer. *Medical Imaging 2021: Computer-Aided Diagnosis.* SPIE; 2021.
- Sáenz-Gamboa JJ, Domenech J, Alonso-Manjarrés A, Gómez JA, de la Iglesia-Vayá M. Automatic semantic segmentation of the lumbar spine: clinical applicability in a multi-parametric and multi-center study on magnetic resonance images. *Artif Intell Med.* 2023;140: 102559. doi:10.1016/j.artmed.2023.102559
- Suri A, Jones BC, Ng G, et al. A deep learning system for automated, multi-modality 2D segmentation of vertebral bodies and intervertebral discs. *Bone.* 2021;149:115972. doi:10.1016/j.bone.2021.115972
- Huang J, Shen H, Wu J, et al. Spine explorer: a deep learning based fully automated program for efficient and reliable quantifications of the vertebrae and discs on sagittal lumbar spine MR images. *Spine J.* 2020;20(4):590–599. doi:10.1016/j.spinee.2019.11.010
- Zheng G, Chu C, Belavý DL, et al. Evaluation and comparison of 3D intervertebral disc localization and segmentation methods for 3D T2 MR data: a grand challenge. *Med Image Anal.* 2017;35:327–344. doi: 10.1016/j.media.2016.08.005
- Li X, Dou Q, Chen H, et al. 3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images. *Med Image Anal.* 2018;45:41–54. doi:10.1016/j.media.2018.01.004
- Sekuboyina A, Hussein ME, Bayat A, et al. VerSe: a vertebrae labeling and segmentation benchmark for multi-detector CT images. *Med Image Anal.* 2021;73:102166. doi:10.1016/j.media.2021.102166
- Kuang X, Cheung JPY, Wong K-YK, et al. Spine-GFlow: a hybrid learning framework for robust multi-tissue segmentation in lumbar MRI without manual annotation. *Comput Med Imaging Graph.* 2022;99: 102091.
- Chen L, Wang S, Fan W, Sun J, Naoi S. Beyond human recognition: a CNN-based framework for handwritten character recognition. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR).* IEEE; 2015.
- Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18.* Springer; 2015.
- Lu J-T, Pedemonte S, Bizzo B, et al. Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. *Proceedings of the 3rd Machine Learning for Healthcare Conference (PMLR); PMLR;2018:403–419.*
- Ji X, Zheng G, Belavy D, Ni D. Automated intervertebral disc segmentation using deep convolutional neural networks. *4th International workshop and challenge, CSI 2016, Held in Conjunction with MICCAI 2016.* springer; October 17, 2016. <https://link.springer.com/book/10.1007/978-3-319-55050-3>
- Jamaludin A, Kadir T, Zisserman A. SpineNet: automated classification and evidence visualization in spinal MRIs. *Med Image Anal.* 2017;41: 63–73. doi:10.1016/j.media.2017.07.002
- Lewandrowski KU. Retrospective analysis of accuracy and positive predictive value of preoperative lumbar MRI grading after successful outcome following outpatient endoscopic decompression for lumbar foraminal and lateral recess stenosis. *Clin Neurol Neurosurg.* 2019; 179:74–80. doi:10.1016/j.clineuro.2019.02.019
- Castro-Mateos I, Hua R, Pozo JM, Lazary A, Frangi AF. Intervertebral disc classification by its degree of degeneration from T2-weighted



- magnetic resonance images. *Eur Spine J.* 2016;25(9):2721-2727. doi:[10.1007/s00586-016-4654-6](https://doi.org/10.1007/s00586-016-4654-6)
17. Meadows KD, Peloquin JM, Newman HR, Cauchy PJ, Vresilovic EJ, Elliott DM. MRI-based measurement of in vivo disc mechanics in a young population due to flexion, extension, and diurnal loading. *JOR Spine.* 2023;6:e1243.
  18. Pfirrmann CW, Metzendorf A, Zanetti M, Hodler J, Boos N. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine.* 2001;26(17):1873-1878.
  19. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage.* 2006;31(3):1116-1128.
  20. R Core Team. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing; 2021. <https://www.R-project.org/>
  21. Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016.
  22. Sergeev A, Del Balso M. Horovod: fast and easy distributed deep learning in TensorFlow. arXiv preprint arXiv:180205799. 2018.
  23. Sha G, Wu J, Yu B. A robust segmentation method based on improved U-net. *Neural Process Lett.* 2021;53:2947-2965. doi:[10.1007/s11063-021-10531-9](https://doi.org/10.1007/s11063-021-10531-9)
  24. Molière S, Hamzaoui D, Granger B, et al. Reference standard for the evaluation of automatic segmentation algorithms: quantification of inter observer variability of manual delineation of prostate contour on MRI. *Diagn Interv Imaging.* 2024;105(2):65-73. doi:[10.1016/j.diii.2023.08.001](https://doi.org/10.1016/j.diii.2023.08.001)
  25. Willers C, Bauman G, Andermatt S, et al. The impact of segmentation on whole-lung functional MRI quantification: repeatability and reproducibility from multiple human observers and an artificial neural network. *Magn Reson Med.* 2021;85(2):1079-1092. doi:[10.1002/mrm.28476](https://doi.org/10.1002/mrm.28476)
  26. Granzier RWY, Verbakel NMH, Ibrahim A, et al. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. *Sci Rep.* 2020;10(1):14163. doi:[10.1038/s41598-020-70940-z](https://doi.org/10.1038/s41598-020-70940-z)
  27. Guo J, Bouaou K, Houriez-Gombaudo-Saintonge S, et al. Deep learning-based analysis of aortic morphology from three-dimensional MRI. *J Magn Reson Imaging.* 2024. Epub ahead of print. PMID: 38216546. doi:[10.1002/jmri.29236](https://doi.org/10.1002/jmri.29236)
  28. Korez R, Ibragimov B, Likar B, Pernuš F, Vrtovc T. Intervertebral disc segmentation in MR images with 3D convolutional networks. *Proc SPIE 10133, Medical Imaging 2017: Image Processing.* 24 February 2017;1013306. doi:[10.1117/12.2254069](https://doi.org/10.1117/12.2254069)
  29. Newman HR, DeLucca JF, Peloquin JM, Vresilovic EJ, Elliott DM. Multiaxial validation of a finite element model of the intervertebral disc with multigenerational fibers to establish residual strain. *JOR Spine.* 2021;4(2):e1145. doi:[10.1002/jsp2.1145](https://doi.org/10.1002/jsp2.1145)
  30. Sharabi M, Levi-Sasson A, Wolfson R, et al. The Mechanical role of the radial fiber network within the annulus fibrosus of the lumbar intervertebral disc: a finite elements study. *J Biomech Eng.* 2019 February 1; 141(2):021006. doi:[10.1115/1.4041769](https://doi.org/10.1115/1.4041769). Erratum in: *J Biomech Eng.* 2019 April 1;141(4). doi:[10.1115/1.4042685](https://doi.org/10.1115/1.4042685). PMID: 30347039.
  31. Yang B, O'Connell GD. Intervertebral disc swelling maintains strain homeostasis throughout the annulus fibrosus: a finite element analysis of healthy and degenerated discs. *Acta Biomater.* 2019;100:61-74. doi:[10.1016/j.actbio.2019.09.035](https://doi.org/10.1016/j.actbio.2019.09.035)
  32. Fleps I, Newman HR, Elliott DM, Morgan EF. Geometric determinants of the mechanical behavior of image-based finite element models of the intervertebral disc. *J Orthop Res.* 2024;42:1355. doi:[10.1002/jor.25788](https://doi.org/10.1002/jor.25788)
  33. Kaur A, Kaur L, Singh A. FP-MMR: a framework for the preprocessing of multimodal MR images. In: Senjyu T, Mahalle P, Perumal T, Joshi A, eds. *Information and Communication Technology for Intelligent Systems Smart Innovation, Systems and Technologies.* Springer; 2021:363-375.
  34. Tustison NJ, Avants BB. Explicit B-spline regularization in diffeomorphic image registration. *Front Neuroinform.* 2013;7:39. doi:[10.3389/fninf.2013.00039](https://doi.org/10.3389/fninf.2013.00039)
  35. Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC. An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics.* 2011;9(4):381-400. doi:[10.1007/s12021-011-9109-y](https://doi.org/10.1007/s12021-011-9109-y)
  36. Bianciardi M, Strong C, Toschi N, et al. A probabilistic template of human mesopontine tegmental nuclei from in vivo 7T MRI. *Neuroimage.* 2018;170:222-230. doi:[10.1016/j.neuroimage.2017.04.070](https://doi.org/10.1016/j.neuroimage.2017.04.070)
  37. Saidu IC, Csató L. Active learning with bayesian UNet for efficient semantic image segmentation. *J Imaging.* 2021;7(2):37.
  38. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19.* Springer; 2016.
  39. Ibtihaz N, Rahman MS. MultiResUNet: rethinking the U-net architecture for multimodal biomedical image segmentation. *Neural Netw.* 2020;121:74-87. doi:[10.1016/j.neunet.2019.08.025](https://doi.org/10.1016/j.neunet.2019.08.025)
  40. Dar SUH, Özbey M, Çatlı AB, Çukur T. A transfer-learning approach for accelerated MRI using deep neural networks. *Magn Reson Med.* 2020;84(2):663-685. doi:[10.1002/mrm.28148](https://doi.org/10.1002/mrm.28148)
  41. Kaoutar BA, Hall LO, Goldgof DB, Liu R, Robert A. Gatenby "Fine-tuning convolutional deep features for MRI based brain tumor classification". *Proc. SPIE 10134, Medical Imaging 2017: computer-aided diagnosis.* 3 March 2017;101342E. doi:[10.1117/12.2253982](https://doi.org/10.1117/12.2253982)
  42. Wang G, Li W, Zuluaga MA, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Trans Med Imaging.* 2018;37(7):1562-1573. doi:[10.1109/TMI.2018.2791721](https://doi.org/10.1109/TMI.2018.2791721)

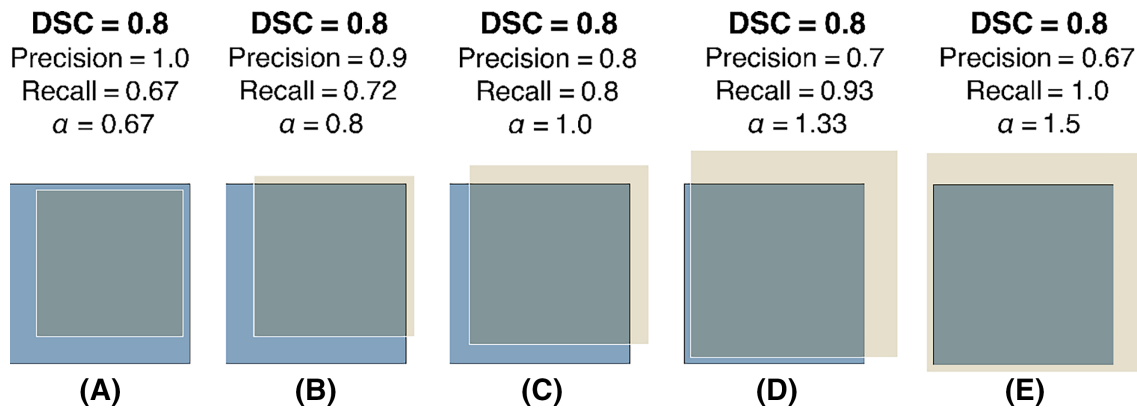
## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Markhali MI, Peloquin JM, Meadows KD, Newman HR, Elliott DM. Neural network segmentation of disc volume from magnetic resonance images and the effect of degeneration and spinal level. *JOR Spine.* 2024;7(3):e70000. doi:[10.1002/jsp2.70000](https://doi.org/10.1002/jsp2.70000)

## APPENDIX A: Similarity Metrics

In this study, the similarity between two segmentation masks was described by two volumetric similarity metrics: Dice Similarity Coefficient (DSC) and Volume Ratio. These metrics are calculated from the true positive (TP), false positive (FP), and false negative (FN) rates and related to the commonly-used Precision and Recall metrics as follows:



**FIGURE A1** A schematic example illustrates situations where different predictions can yield the same DSC (blue is the example ground truth and yellow is the example prediction). It is useful to report at least one metric in addition to DSC to detect bias towards over- or under-segmentation.

$$\left\{ \begin{array}{l} \text{DSC} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \\ \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Volume ratio} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FN}} = \frac{\text{Rec}}{\text{Pre}}. \end{array} \right.$$

Multiple metrics are useful because DSC is symmetric with respect to the two input images, and does not detect a systemic tendency for over- or under-segmentation with respect to a ground truth image. Figure A1 illustrates how other similarity metrics provide additional information in such cases.