

# UCSF

## UC San Francisco Previously Published Works

### Title

Automatic detection and voxel-wise mapping of lumbar spine Modic changes with deep learning

### Permalink

<https://escholarship.org/uc/item/6mp610m6>

### Journal

JOR Spine, 5(2)

### ISSN

2572-1143

### Authors

Gao, Kenneth T  
Tibrewala, Radhika  
Hess, Madeline  
[et al.](#)

### Publication Date

2022-06-01


### DOI

10.1002/jsp2.1204

Peer reviewed

## RESEARCH ARTICLE

# Automatic detection and voxel-wise mapping of lumbar spine Modic changes with deep learning

Kenneth T. Gao<sup>1,2</sup>  | Radhika Tibrewala<sup>1</sup> | Madeline Hess<sup>1</sup> |  
Upasana U. Bharadwaj<sup>1</sup> | Gaurav Inamdar<sup>1</sup> | Thomas M. Link<sup>1</sup> |  
Cynthia T. Chin<sup>1</sup> | Valentina Padoia<sup>1</sup> | Sharmila Majumdar<sup>1</sup>

<sup>1</sup>Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, California, USA

<sup>2</sup>Department of Bioengineering, University of California Berkeley–University of California San Francisco Graduate Program in Bioengineering, Berkeley, California, USA

## Correspondence

Kenneth T. Gao, Department of Radiology and Biomedical Imaging, University of California, San Francisco, 1700 4th Street, 203D Byers Hall, San Francisco, CA 94158, USA.  
Email: [kenneth.gao@ucsf.edu](mailto:kenneth.gao@ucsf.edu)

## Funding information

National Institute of Arthritis and Musculoskeletal and Skin Diseases, Grant/Award Number: UH2AR076724; National Institutes of Health, United States

## Abstract

**Background:** Modic changes (MCs) are the most prevalent classification system for describing magnetic resonance imaging (MRI) signal intensity changes in the vertebrae. However, there is a growing need for novel quantitative and standardized methods of characterizing these anomalies, particularly for lesions of transitional or mixed nature, due to the lack of conclusive evidence of their associations with low back pain. This retrospective imaging study aims to develop an interpretable deep learning-based detection tool for voxel-wise mapping of MCs.

**Methods:** Seventy-five lumbar spine MRI exams that presented with acute-to-chronic low back pain, radiculopathy, and other symptoms of the lumbar spine were enrolled. The pipeline consists of two deep convolutional neural networks to generate an interpretable voxel-wise *Modic map*. First, an autoencoder was trained to segment vertebral bodies from T<sub>1</sub>-weighted sagittal lumbar spine images. Next, two radiologists segmented and labeled MCs from a combined T<sub>1</sub>- and T<sub>2</sub>-weighted assessment to serve as ground truth for training a second autoencoder that performs segmentation of MCs. The voxels in the detected regions were then categorized to the appropriate Modic type using a rule-based signal intensity algorithm. Post hoc, three radiologists independently graded a second dataset with the aid of the model predictions in an artificial (AI)-assisted experiment.

**Results:** The model successfully identified the presence of changes in 85.7% of samples in the unseen test set with a sensitivity of 0.71 ( $\pm 0.072$ ), specificity of 0.95 ( $\pm 0.022$ ), and Cohen's kappa score of 0.63. In the AI-assisted experiment, the agreement between the junior radiologist and the senior neuroradiologist significantly improved from Cohen's kappa score of 0.52 to 0.58 ( $p < 0.05$ ).

**Conclusions:** This deep learning-based approach demonstrates substantial agreement with radiologists and may serve as a tool to improve inter-rater reliability in the assessment of MCs.

Kenneth T. Gao and Radhika Tibrewala contributed equally to this study.

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.

© 2022 The Authors. *JOR Spine* published by Wiley Periodicals LLC on behalf of Orthopaedic Research Society.

## KEYWORDS

deep learning, magnetic resonance imaging, Modic changes, vertebral body

## 1 | INTRODUCTION

Low back pain (LBP) is the leading cause of disability globally, accounting for 60.1 million disability-adjusted life-years in 2015.<sup>1,2</sup> While the nociceptive source in the vast majority of LBP cases cannot be identified,<sup>2,3</sup> there has been a growing collection of evidence showing that properties of vertebral endplates are closely linked to intervertebral disc degeneration and LBP.<sup>4-7</sup> Modic changes (MCs) are the most commonly used classification system for describing changes in endplate-adjacent vertebral bone marrow.<sup>8</sup> Despite its prevalence, the association of MCs with LBP is inconsistent.<sup>5,9-11</sup>

Hypothesized to cause LBP through structural and inflammatory changes in the bony structures of the spine,<sup>12-14</sup> MCs are defined as signal variations seen in the combined assessment of T<sub>1</sub>-weighted and T<sub>2</sub>-weighted magnetic resonance imaging (MRI).<sup>8</sup> Bone marrow edema-like changes or fibrovascular changes appear distinctly hypointense on T<sub>1</sub>-weighted images and hyperintense on T<sub>2</sub>-weighted images (Modic type 1).<sup>10,15</sup> Meanwhile, conversion of red hematopoietic bone marrow to yellow fatty marrow is hyperintense on T<sub>1</sub> MRI and iso- to hyperintense in fat saturated T<sub>2</sub> and non-fat saturated T<sub>2</sub> sequences, respectively (Modic type 2). And lastly, sclerotic bone appears hypointense in both sequences (Modic type 3).

Thus, the semiquantitative nature of the MC classification system is highly susceptible to variability in non-standardized imaging. Fields et al., detailed how evaluation of MCs is prone to inter-rater variability through a wide range of factors related to equipment and image acquisition parameters.<sup>15</sup> Magnetic field strength, in particular, has been shown to have significant effects on the prevalence of MCs, with type 2 changes being easily distinguishable in low-field MRI and type 1 changes visualized more easily in high-field MRI.<sup>16</sup> Pulse sequence design and parameters can also effectively influence image quality, signal-to-noise, fat suppression, and, importantly, tissue contrast. Due to a lack of systemic standardization in spine imaging, it is pivotal to adapt grading procedures with objective and quantitative methodologies.

Several quantitative approaches have been recently applied to the assessment of vertebral changes. Specialized pulse sequences, such as chemical shift encoding-based water-fat imaging,<sup>17</sup> magnetic resonance spectroscopy,<sup>18</sup> diffusion, and perfusion,<sup>19</sup> can provide additional information on tissue composition. Post-acquisition, Wang et al. extracted morphological and signal intensity-based metrics from contours of MCs, reporting improved inter- and intra-rater agreement as compared to unassisted MC classification.<sup>20</sup> However, a limitation with these approaches is the need for manual demarcation of MCs, which may be labor-intensive.

Data-driven strategies to address these drawbacks have emerged from the recent surge of development in deep learning (DL) and convolutional neural networks. Notable applications to spinal imaging analysis include automated segmentation of spinal structures,<sup>21-23</sup> detection of spinal anomalies,<sup>24-26</sup> and predictive modeling of spinal

surgery outcomes.<sup>27,28</sup> Automated endplate assessments have seen relative success, as well. Jamaludin et al. have shown that endplate defects can be detected from MRI using convolutional neural networks with approximately 83.7% and 86.9% accuracy in their test set for upper and lower endplates, respectively.<sup>29</sup> While these efforts automate spinal analysis to near human-performance, there remain opportunities to translate such models into clinical utility.

The adoption of a DL model into widespread use to address inconsistencies of the assessment and reporting of MCs hinges on its interpretability. Our study aims to (1) develop a DL-based automatic contouring method to identify MCs in vertebral bodies, (2) classify these changes as Modic types 1, 2, or 3 (MC 1/2/3) on a voxel-wise level, thereby providing granular, quantitative information about the vertebral bodies as a *Modic map*, and (3) use the automatic detection as an aid to radiologists to demonstrate capability to improve agreement and pave the way for more consistent evaluations of the relationship between MCs and LBP.

## 2 | MATERIALS AND METHODS

This retrospective, single-center study was approved by the local Institutional Review Board, and the informed consent requirement was waived.

### 2.1 | Dataset and annotations

Seventy-five exams with the following inclusion and exclusion criteria were sampled at random from lumbar spine MRIs acquired between 2008 and 2019 at our institution. *Inclusion*: patients aged 19 years or older presenting with acute-to-chronic LBP, radiculopathy, and other symptoms of the lumbar spine including numbness, tingling, weakness, dysesthesia, and tightness. *Exclusion*: (1) vertebral fractures, (2) post-operative changes, (3) extensive hardware, (4) primary tumors, (5) metastatic spinal disease, (6) infection, and (7) transitional anatomy. Imaging was performed on GE Signa HDxt 1.5 T and GE Discovery MR750 3.0 T (GE Healthcare, Milwaukee, WI) with acquisition details of the relevant T<sub>1</sub>-weighted sagittal and T<sub>2</sub>-weighted sagittal sequences provided in Table 1. All images were deidentified for this study.

To serve as ground truth for the DL components, vertebral bodies with visible MCs were segmented for these changes (Type 1, 2, and 3) by a board-certified neuroradiologist (C. C. with over 25 years of experience) and a musculoskeletal junior radiologist (U. U. B. with 3 years of experience) after initial adjudication for calibration on 15 exams not included in the study cohort. To promote further standardization between grading assessments, MCs with diameter less than 5 mm were excluded and mixed MCs were annotated as the predominant type. All manual annotations were performed using the medical imaging platform, MD.ai (MD.ai, New York, NY).

## 2.2 | Image analysis

This Modic mapping scheme consists of three stages, as depicted in Figure 1: (1) segmentation and localization of the vertebral bodies, (2) binary detection and segmentation of signal variabilities characteristic of MCs, and (3) voxel-wise classification of the detected regions to classify Modic type.

### 2.2.1 | Image alignment

As MCs are characterized by local signal variations in both  $T_1$ - and  $T_2$ -weighted images, these images were aligned with image position coordinates prior to processing. The rigid alignment was performed by first matching positions of each sagittal slice of the  $T_2$ -weighted images to the  $T_1$ -weighted images in the frontal axis. Then,  $T_2$ -weighted slices were rotated, translated, and scaled to the dimensions of their

**TABLE 1** Summary of the range of acquisition parameters from dataset curated from clinical magnetic resonance imaging (MRI) exams

	$T_1$ -weighted	$T_2$ -weighted
Field strength (T)	1.5, 3.0	1.5, 3.0
Matrix	$256 \times 256$ – $512 \times 512$	$256 \times 256$ – $512 \times 512$
Field-of-view (cm)	24.0–37.0	24.0–37.0
Slice thickness (mm)	3.0–4.0	3.0–4.0
Pixel bandwidth (Hz)	88.8–250.0	81.4–325.5
Repetition time (ms)	377–975	2430–6307
Echo time (ms)	6.8–31.8	26.1–107.8
Flip angle (°)	90–180	90–160

corresponding  $T_1$  counterpart. Finally, each slice was similarly translated and scaled to harmonize in-plane resolution using bicubic interpolation.

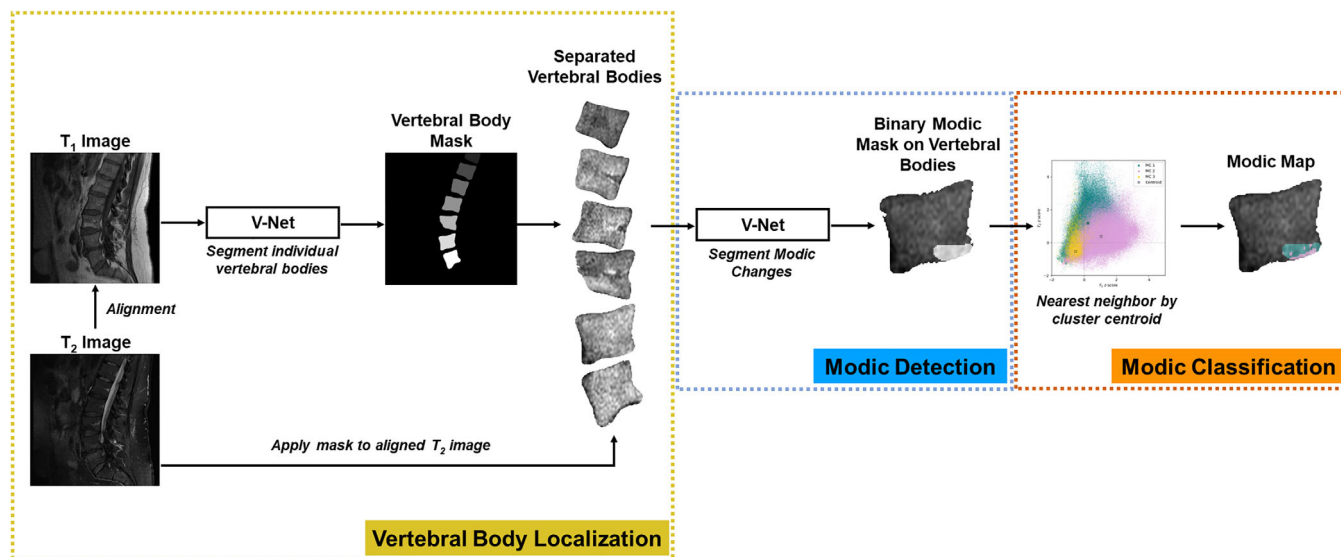
### 2.2.2 | Vertebral body localization

Our first goal was to isolate vertebral bodies to fixate on image features pertaining to the vertebral body and endplates. To achieve this, we developed and trained a preliminary V-Net convolutional neural network<sup>30</sup> for semantic segmentation. A research associate (G. I.) manually segmented vertebral bodies from  $T_1$ -weighted images in a subset of 40 exams. These MRIs were randomly split into training ( $n = 20$ ), validation ( $n = 17$ ), and test ( $n = 3$ ) sets and then separated into 2D slices. The V-Net was trained on a single NVIDIA TITAN X GPU using Tensorflow v1.14 with the following hyperparameters: batch size = 3; optimizer = Adam; learning rate =  $1e-4$ ; loss function = Dice (Equation (1)); dropout rate = 0.8. Post-training, the performance of the segmentation model was assessed using the Dice coefficient overlap between the manual and predicted segmentations. To evaluate inter-rater variability, a second research associate (K. T. G.) manually segmented vertebral bodies from a subset of five exams.

$$\text{Dice loss} = \frac{2 \sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}, \quad (1)$$

where  $N$  is the total number of voxels,  $p_i \in P$  represents voxel values of the prediction, and  $g_i \in G$  represents voxel values of the ground truth.

We utilized this model to segment vertebral bodies of the 75 lumbar spine MRI exams in the dataset. The individual vertebral bodies



**FIGURE 1** Schematic of the full Modic mapping approach. Vertebral bodies are first segmented and extracted from  $T_1$ -weighted magnetic resonance imaging (MRI), allowing extraction of the bodies on the  $T_1$  and aligned  $T_2$  images. Next, a binary segmentation network localizes and detects regions of Modic changes (MCs). Lastly, each voxel of the detected regions is classified to a Modic type using a nearest neighbor algorithm and  $T_1$  and  $T_2$  z-scores to form a Modic map.

in the inferred masks were identified using 3D connected component labeling, in which segmented masks joined within a six-connected neighborhood were given a unique label. The masked vertebral body masks were then zero-padded to a standardized size of  $100 \times 100$ .

### 2.2.3 | Modic detection and segmentation

MC detection was achieved using a second segmentation neural network that utilized these localized vertebral bodies and the radiologist-annotated MCs. In each exam, we used z-score standardization to convert each voxel to the number of standard deviations from the mean signal intensity in the segmented vertebral bodies. Next, the  $100 \times 100$  vertebral body masks were applied to the  $T_1$ -weighted and aligned  $T_2$ -weighted images and these images were stacked, producing input images of dimensions  $100 \times 100 \times 2$ . Binary radiologist-annotated MC segmentations (presence vs. absence of MCs) were similarly masked. The 75 exams, consisting of 1872 vertebral body image-Modic segmentation pairs, were randomly split into training ( $n = 50$ ), validation ( $n = 15$ ), and test ( $n = 10$ ) sets. Figure 2 portrays the demographic distribution of the data splits.

We developed and modified the 2D V-Net for MC segmentation. The network consists of two branches, each with four levels. The encoder branch is responsible for compressing the input to an abstract latent space of representative features. At each level, convolutional layers (1, 2, 3, and 3 layers in the respective levels) extract features with 32 kernels of size  $5 \times 5$  and stride 1 followed by downsampling with a  $2 \times 2$  kernel with stride 2. The subsequent decoder branch deconvolves the latent space back to the input's original dimension and passes the array through a combined cross-entropy and Dice loss layer with sigmoid activation to ultimately produce probabilistic segmentation masks for MCs. Hyperparameters for training include: batch size = 128; optimizer = Adam; learning rate =  $1e-4$ ; loss function = weighted cross entropy and Dice (Equation (2)); loss weights = 20:1 (foreground:background); dropout = 0.2. Training was deemed complete after a designated 15 validation cycles without improvement (500 iterations per cycle).

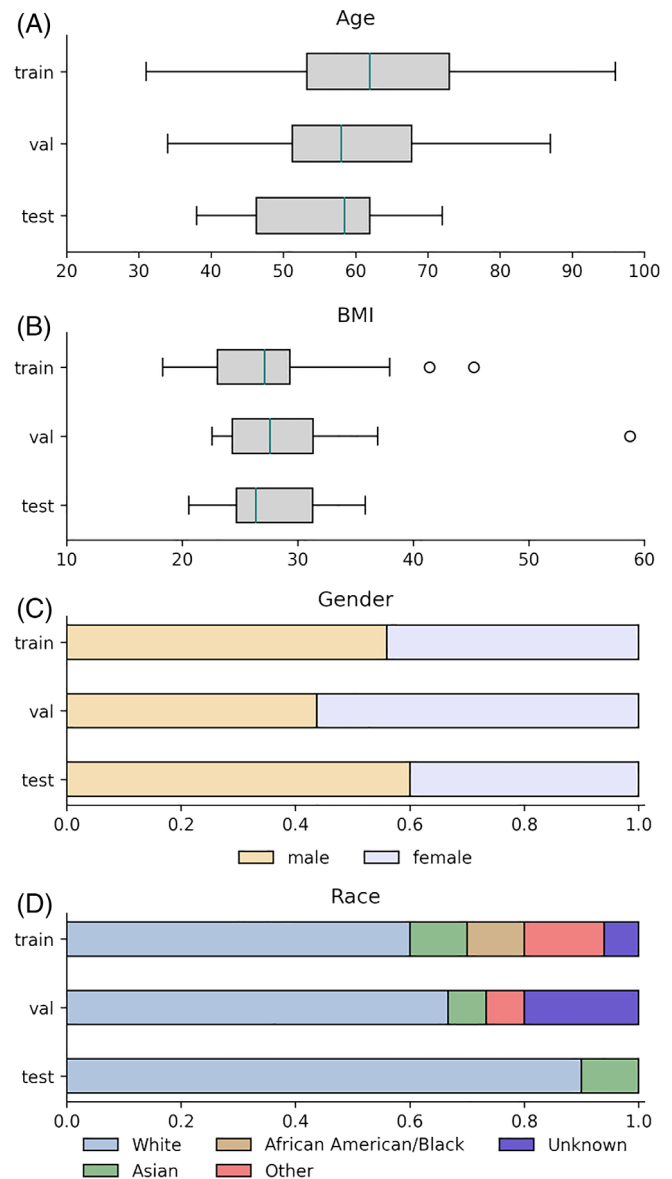
$$\text{Combined loss} = \lambda(\text{Cross entropy loss}) + \text{Dice loss}, \quad (2)$$

where  $\lambda$  is a weighting coefficient set to 0.1, and

$$\text{Cross entropy loss} = - \sum_{i=1}^N g_i \log(p_i) + (1 - g_i) \log(1 - p_i). \quad (3)$$

### 2.2.4 | Voxel-wise Modic change mapping

With a trained model for Modic segmentation, we then utilized a nearest-neighbor algorithm to classify each voxel in the detected MCs into one of three types. Again, we utilized the training set; each voxel in the regions annotated by the radiologist was characterized by its  $T_1$  z-score and  $T_2$  z-score and then grouped into the appropriate MC



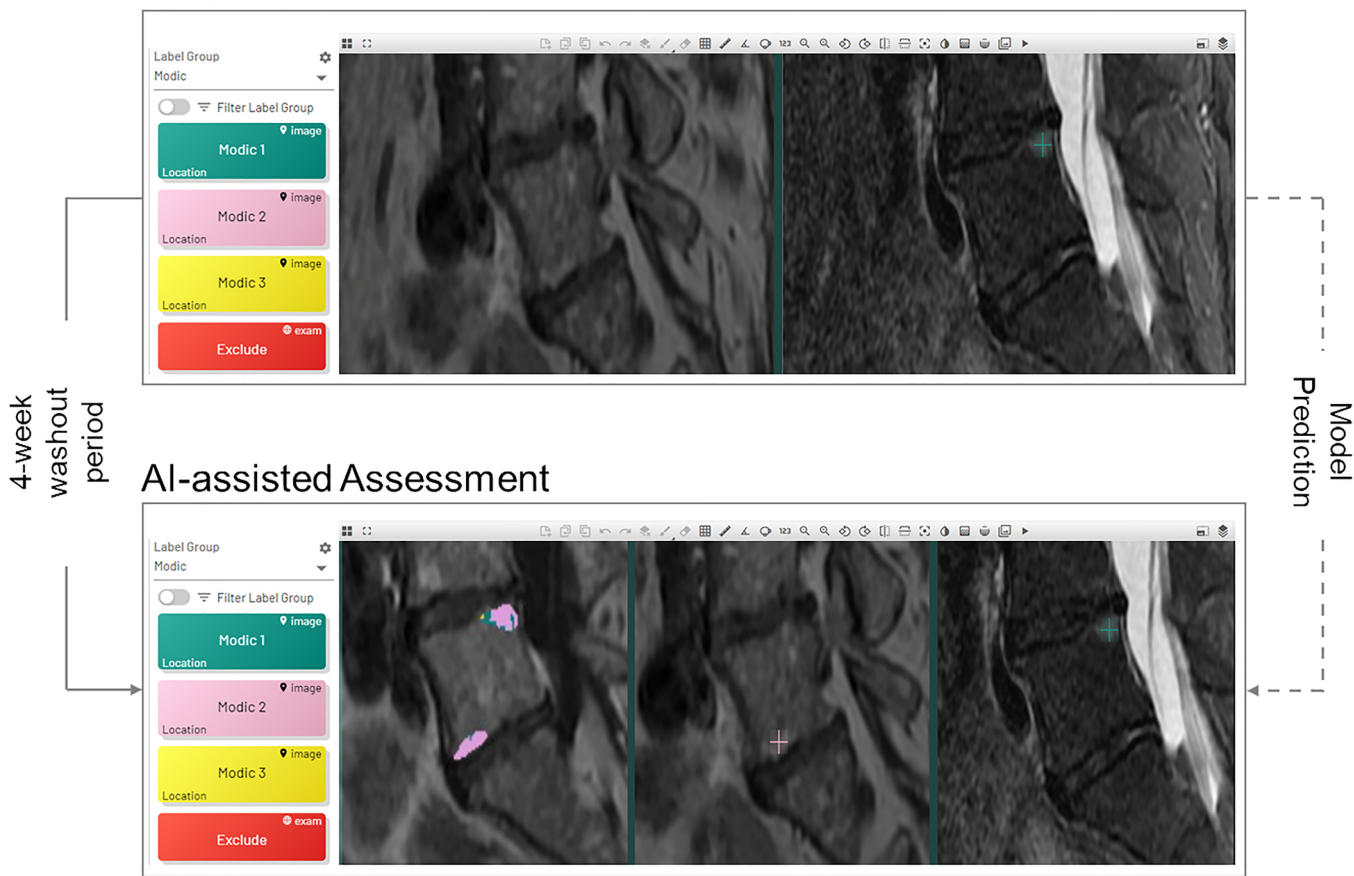
**FIGURE 2** Distribution of subject demographics, including (A) age, (B) BMI, (C) gender, and (D) race, of the 75 magnetic resonance imaging (MRI) exams after randomly splitting into training ( $n = 50$ ), validation ( $n = 15$ ), and test sets ( $n = 10$ )

group. The centroid of the [ $T_1$  z-score,  $T_2$  z-score] clusters was computed. To classify the test set and exams in inference, each voxel in detected MCs was similarly characterized by [ $T_1$  z-score,  $T_2$  z-score] then categorized by the nearest cluster centroid neighbor. This ultimately produced voxel-wise Modic maps.

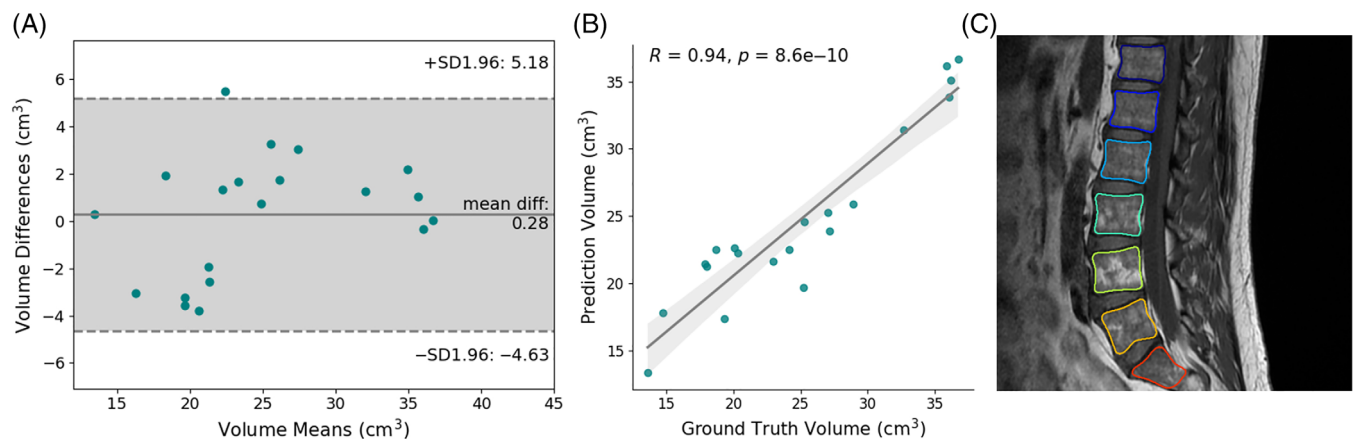
## 2.3 | Statistical analysis

We created a rule-based algorithm that produces binary labels of each MC in upper and lower vertebral bodies to assess the effectiveness of this scheme as compared to human performance and past works. Upper and lower sections were approximated by finding the convex

### Initial Assessment



**FIGURE 3** Experimental setup of the AI-assisted assessments in the labeling platform, MD.ai. Three readers graded an independently curated dataset ( $n = 20$ ). Using the trained Modic mapping schema, predictions for Modic changes (MCs) were generated in the same dataset, and after a 4-week washout period, readers 2 and 3 re-graded these exams with the assistance of the model predictions.



**FIGURE 4** Post hoc analysis of vertebral body segmentation of the test set. (A) Bland-Altman plot indicates the average difference in vertebral body volume between model prediction and ground truth was  $0.28 \text{ cm}^3$ . The gray areas portray the 95% confidence intervals. (B) The correlation plot of vertebral body volume has an intercept of  $14.8 \text{ cm}^3$ , demonstrating a measurement bias, and  $R$ -value of 0.94. (C) Representative example of vertebral body segmentation contours on  $T_1$ -weighted image

hull of the vertebral body mask and bisecting them along the long axis. Thus, each bisection was described with three binary labels, representing the presence or absence of voxels characteristic of Modic

types 1, 2, and 3, respectively. Sensitivity, specificity, and Cohen's kappa score ( $\kappa$ ) were computed to evaluate the overall Modic detection performance, and the subsequent classification.

### 2.3.1 | AI-assisted experiment

A second dataset ( $n = 20$ ) was curated to explore the effect of inter-rater agreement of Modic grading with the aid of this Modic mapping pipeline. A senior neuroradiologist (C. C., over 25 years of experience), a senior musculoskeletal radiologist (T. M. L., over 25 years of experience), and a junior radiologist in-training (U. U. B., 3 years of experience) graded these exams independently. Inter-rater reliability was assessed using Cohen's kappa coefficient. After a 4-week washout period, the musculoskeletal radiologist and junior radiologist re-graded the same dataset, with the aid of Modic maps generated from our developed pipeline. Agreement was reassessed to measure differences with the initial trial using Cohen's kappa score and the McNemar's test, with the neuroradiologist established as the baseline. The experimental setup is summarized in Figure 3.

## 3 | RESULTS

### 3.1 | Vertebral body localization

Training the vertebral body segmentation network was completed in approximately 10 h with 20 000 iterations. Evaluated with the unseen test set, the model achieved  $0.882 \pm 0.018$  Dice overlap with the ground truth segmentations. This performance is comparable to the inter-rater Dice overlap between two research associates, which was reported as  $0.927 \pm 0.011$ .

Post hoc analysis of vertebral body segmentation was performed (Figure 4). The mean volumetric error of the model prediction was  $0.28 \text{ cm}^3$  per vertebral body, approximately 1.1% of the average vertebral volume. Manually segmented ground truth and model predictions were well correlated with an  $R$ -value of 0.94 and  $p$ -value  $< 0.001$  using Pearson correlation.

### 3.2 | Modic detection and segmentation

The Modic detection model, after training for 11 500 iterations, successfully identified the presence or absence of changes in 85.7% of samples in the unseen test set. Sensitivity and specificity of the model were computed and summarized in Table 2, resulting in  $0.71 (\pm 0.072)$  and  $0.95 (\pm 0.022)$ , respectively. Cohen's kappa score was similarly computed against the radiologist-annotated ground truth as 0.63, interpreted as substantial agreement.

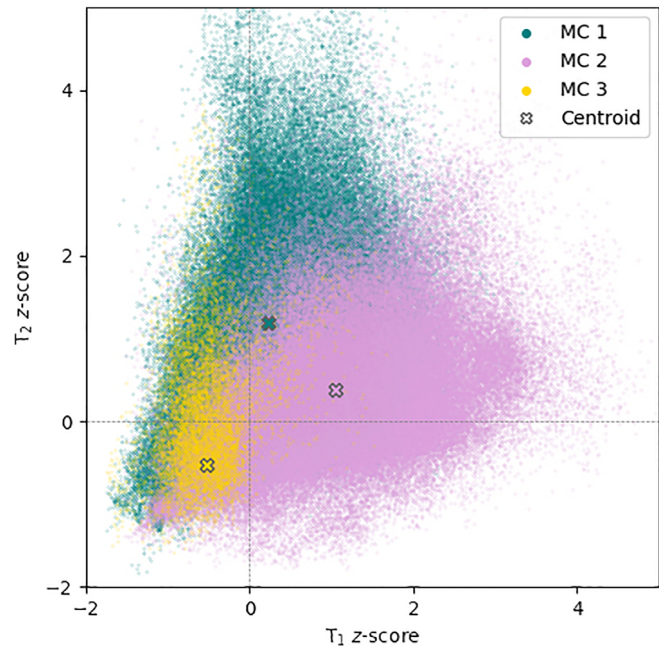
### 3.3 | Voxel-wise Modic change mapping

Figure 5 shows the  $[T_1 \text{ z-score}, T_2 \text{ z-score}]$  voxel-wise characterization of MCs in the training set. Cluster centroids of Modic 1, 2, and 3 were centered at  $[0.23 (\pm 0.73), 1.20 (\pm 1.16)]$ ,  $[1.04 (\pm 1.00), 0.37 (\pm 0.85)]$ , and  $[-0.53 (\pm 0.41), -0.52 (\pm 0.85)]$ , respectively, corresponding well with the qualitative classification system defined by hyper- and

**TABLE 2** Performance of the full pipeline on the unseen test set

	Sensitivity (95% CI)	Specificity (95% CI)
Overall	0.71 ( $\pm 0.072$ )	0.95 ( $\pm 0.022$ )
MC 1	0.67 ( $\pm 0.113$ )	0.87 ( $\pm 0.030$ )
MC 2	0.67 ( $\pm 0.102$ )	0.89 ( $\pm 0.028$ )
MC 3	0.44 ( $\pm 0.324$ )	0.83 ( $\pm 0.032$ )

Abbreviations: CI, 95% confidence interval; MC, Modic change.



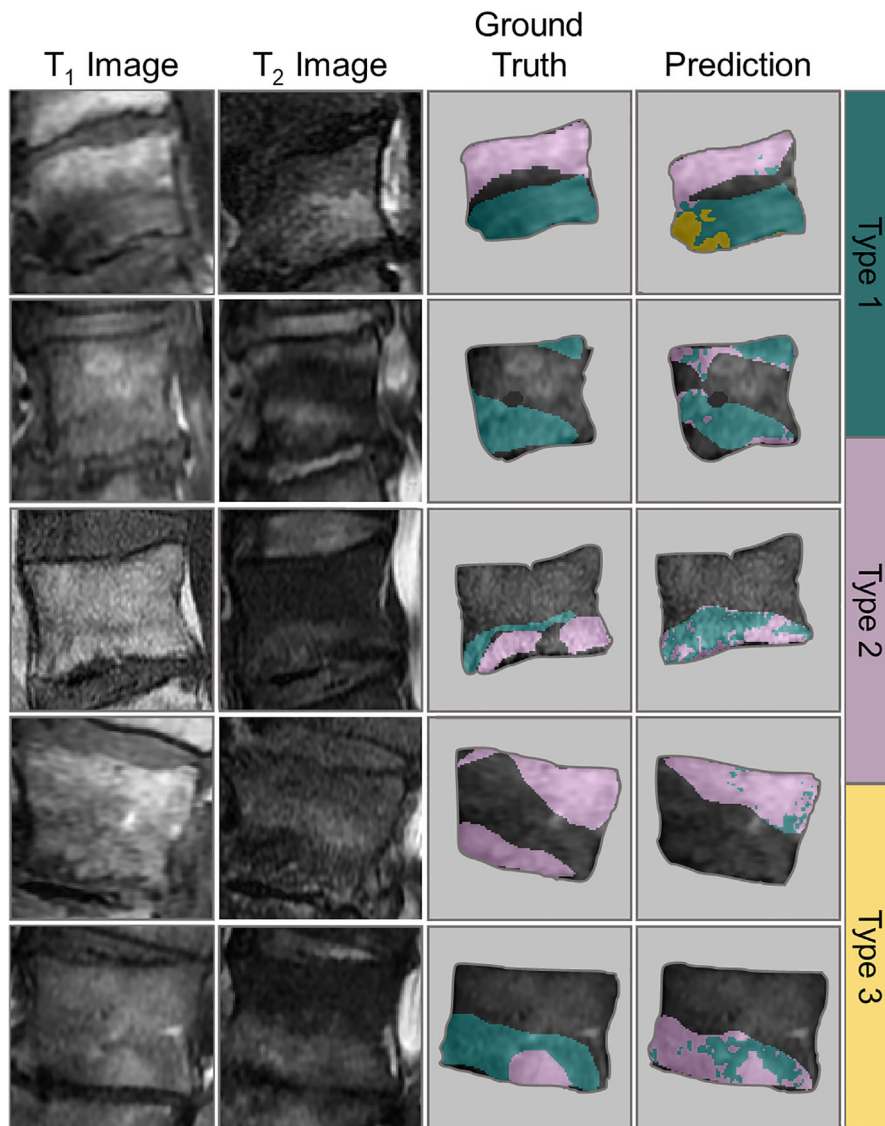
**FIGURE 5** Paired  $T_1$  and  $T_2$  z-score coordinates of each voxel within Modic changes in the training set. These centroid coordinates align well with the qualitative Modic grading system and its corresponding variations in signal intensity (e.g., Modic type 1 is hyperintense in  $T_2$ -weighted imaging, Modic type 2 is hyperintense in  $T_1$ -weighted imaging). Detected Modic changes in the test set were classified on a voxel-by-voxel basis using a nearest neighbor algorithm to these cluster centroids.

hypo-intensities. Labeling of upper and lower vertebral bodies using the rule-based classification system resulted in sensitivities of  $[0.67 (\pm 0.113), 0.67 (\pm 0.102), \text{ and } 0.44 (\pm 0.324)]$  and specificities of  $[0.87 (\pm 0.030), 0.89 (\pm 0.028), \text{ and } 0.83 (\pm 0.032)]$  for Modic types 1, 2, and 3, respectively, as seen in Table 2. The overall prevalence of MCs in the test set was 0.27 in the ground truth and, correspondingly, 0.23 in the model predictions. Further stratification of MC prevalence is described in Figure 6. In Figure 7, representative examples of Modic maps are shown with their corresponding  $T_1$  and  $T_2$  images.

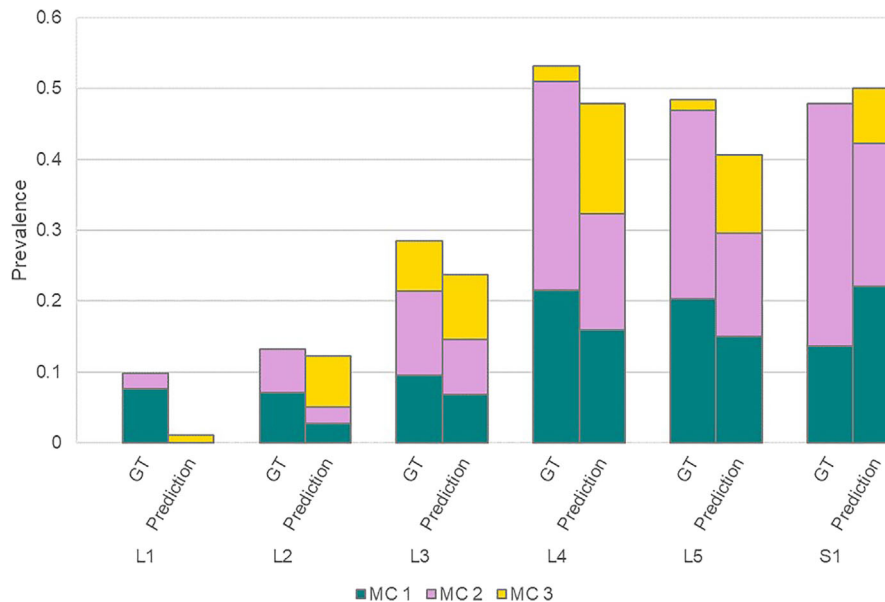
### 3.4 | AI-assisted experiment

Inter-rater agreement was initially assessed with an independently curated dataset ( $n = 20$ ) (Table 3). Between the three radiologists, the

**FIGURE 6** Representative examples of the model inputs ( $T_1$  and  $T_2$  images), radiologist-annotated ground truth segmentations, and the predicted Modic maps. The mapping technique is advantageous for visualizing heterogeneity and transitional pathology. Notably, in the top row, the model detects Modic change (MC) 3-like characteristics in the anterior inferior endplate. In the second row, a small MC 1 region in the anterior superior endplate, unnoticed by the radiologist, was annotated by the automatic model.



**FIGURE 7** Prevalence of Modic changes (MCs) in the ground truth and prediction of the test set, stratified by vertebral body level. The two distributions share similarities, with the highest number of MCs in the lower lumbar region (L4-S1). The prevalence is further apportioned by the relative ratios of each Modic type. The model tends to overestimate MC 3s due to low representation in the ground truth and inductive bias.





	Initial agreement ( $\kappa$ )	Post-AI-assist experiment ( $\kappa$ )	$\Delta\kappa$	$p$ -value
Readers 1 and 2	0.63	0.62	-0.01	NS
Readers 1 and 3	0.52	0.58	+0.06	<0.05
Readers 2 and 3	0.45	0.48	+0.03	NS

**TABLE 3** Cohen's kappa coefficients between three readers in AI-assisted experiment

Abbreviation: NS, not significant.

two senior readers (reader 1 [C. C.] and reader 2 [T. M. L.]) were in the most agreement, with a Cohen's kappa score  $\kappa = 0.63$ . The junior radiologist (reader 3 [U. U. B.]) had moderate agreement,  $\kappa = 0.52$ , with reader 1 and,  $\kappa = 0.45$ , with reader 2.

With the assistance of the model prediction, agreement of reader 3 with reader 1 significantly improved to  $\kappa = 0.58$  ( $p < 0.05$ ). Agreement between readers 3 and 2 increased to  $\kappa = 0.48$ , though this result was insignificant by the McNemar's test. Meanwhile, reliability between readers 1 and 2 decreased slightly to  $\kappa = 0.62$ , again, without statistical significance.

## 4 | DISCUSSION

This study used DL-based models to automatically localize and map MCs in vertebral bodies. Overall, these results demonstrate substantial agreement of the detection model with radiologist-annotated grading and a novel Modic mapping technique that provides grading assistance when incorporated into a radiology workflow. A design goal of this schema is to provide clinical utility through objective and interpretable models. We aimed to achieve this in two ways.

The first pertains to reducing and streamlining the semiquantitative Modic classification system into a data-driven, yet easily understood multistep algorithm. To limit the effective field-of-view to regions of the vertebral bone, rather than confounding structures such as the neighboring intervertebral discs, foramen, or spinal cord, we performed vertebral body segmentation using the V-Net,<sup>30</sup> a widely used encoder-decoder for biomedical image segmentation. This is particularly important when considering intervertebral disc degeneration due to the strong correlation between presence of the two anomalies.<sup>31,32</sup> The performance of this model is consistent with previous works in spinal segmentation<sup>21,23</sup> and conveys to users of this tool which regions were evaluated by the subsequent Modic detection tool. Similarly, the rule-based classification system proposed here, based on  $T_1$  and  $T_2$  z-scores, intuitively follows the semiquantitative blueprint originally proposed by Modic et al.<sup>8</sup> Ultimately, the availability of intermediary results and interfaces for the pipeline's decision-making process may build confidence toward the adoption of such methodologies into clinical settings.

The second strategy adopted in this approach capitalizes on the ability of Modic maps to describe heterogeneous tissues. Systematic reviews of works involving MCs note inconsistencies in reporting procedures.<sup>10,15</sup> In both research studies and in clinical practice, MCs are dictated as isolated, homogeneous lesions when they are often conglomerated and characterized by spatial heterogeneity. Past literature suggested that MRI changes may progress from Modic type 1 to type

2 to type 3 in a linear fashion,<sup>33</sup> though recent studies have demonstrated that pathologies are often reversible.<sup>34</sup> Not only can MCs be transitional, it has been reported that 27.2% of MCs are regarded as mixed, comprising of characteristics of multiple Modic types.<sup>35</sup> Capturing the granularity of mixed MCs is challenging for the human eye, yet neural networks have proven capable of identifying detailed textural and shape features from medical imaging.<sup>36,37</sup> In this work, we chose to implement a voxel-wise MC segmentation method over a classification model due to the key capability of visualizing the heterogeneity of mixed MCs. In addition, the segmentation methodology offers higher degree of supervision, where each voxel in an image is attributed with a label. This granular supervision retains context of the neighboring tissue and improves label specificity. Further works using this approach can unravel attributes of progressive or transitional MCs that may interact with LBP, as heterogeneous tissues are often correlated with degeneration.

Performance of the vertebral body segmentation and MC detection components reached or neared human reliabilities. Error analysis showed predictive inaccuracies in the lateral-most slices where partial volume effects tend to impact the delineation of bone from surrounding tissues. The performance metric is artificially deflated as the research associate manually segmented complete vertebral bodies while the model would be apt to predict all instances of bone, some of which were only partially visible in the prescribed field of view. In the MC detection component, the distribution of predicted MCs across the lumbar vertebrae was predominantly in the L4-S1 range (74.4%), which matches well with the radiologist annotations (78.8%) and past work (75.5%).<sup>35</sup> Detection of MCs in L1 was notably underestimated by the model. We speculate this is due to signal loss at the periphery of the coil. Voxel-wise classification of MCs yielded high predictive value of Modic types 1 and 2, arguably the two groups most important to classify due to their prevalence<sup>35,38</sup> and the strong association of MC 1 with nonspecific LBP.<sup>39,40</sup> Notably, the models are trained and evaluated on a dataset with a wide arrangement of acquisition parameters to capture the variability in non-standardized imaging procedures.

In the pilot AI-assisted experiment, we found that the additional utility of the model predictions improved agreement of the junior radiologist with the senior radiologists ( $\Delta\kappa = +0.06$  and  $\Delta\kappa = +0.03$  with reader 1 and reader 2, respectively). However, agreement did not improve, but rather slightly decreased ( $\Delta\kappa = -0.01$  with reader 1), for reassessment by reader 2. This is likely explained by the differences in training and preferences between neuroradiology and musculoskeletal radiology. The participating readers reported that a key advantage of the tool was its utility as "attention focuses," which may have contributed to boosting agreement between reader 3 with reader 1.

The technologies developed in this study can be applied in various ways. With further development, this tool could potentially assist training efforts of junior radiologists by highlighting complex cases which depict the nuances of heterogeneous spinal pathologies. Furthermore, because this model was trained using non-standardized clinical data, the AI-assist tool can be adapted to a continuous learning paradigm to improve model generalizability and utility without the need for additional data curation. Specifically, this model demonstrates the capability to predict transitional and heterogeneous MCs which have been hypothesized to be associated with LBP. Using this tool, more data can be gathered on these changes to make consistent associations with LBP and help pave the path to elucidate the mechanisms of nonspecific LBP.

While our results demonstrate that DL-based approaches can contribute to identifying MCs, there are several notable limitations. First, despite the quantitative nature of this methodology, data-driven techniques are still biased by its training data and annotators. Two participants of the AI-assisted experiment were responsible for labeling the training data, which may have biased the agreement metrics against other readers. For these reasons, this algorithm is not intended to be a standalone fully diagnostic tool. Second, relatedly, we acknowledge that the exams used in this study are from a single institution, and the model is not validated with multi-institutional testing. Lastly, our results are limited by the small sample size with poor representation of Modic type 3. Modic type 3 is described by signal void in both T<sub>1</sub>- and T<sub>2</sub>-weighted images, which makes it difficult to grade and susceptible to errors in cases with low signal-to-noise ratio. This is impactful in the nearest neighbor component of the pipeline, which is notably sensitive. Fortunately, several collaborative efforts are in progress to amass additional data from other institutions with wider variability in imaging equipment and acquisition parameters. We also aim to extend this work by exploring domain adaptation strategies to improve generalizability and performing longitudinal analysis to further investigate transitional pathologies.

## 5 | CONCLUSION

In this work, we present a novel DL-based approach to localize and segment MCs, with results that demonstrate high agreement with radiologist grading. The introduction of this fully automatic, quantitative mapping technique may increase inter-rater reliability and ultimately improve robustness in understanding the associations of MCs with LBP and spinal degeneration.

### AUTHOR CONTRIBUTIONS

Kenneth T. Gao, Radhika Tibrewala, Upasana U. Bharadwaj, Cynthia T. Chin, Valentina Pedita, and Sharmila Majumdar contributed to study design. Upasana U. Bharadwaj, Gaurav Inamdar, and Cynthia T. Chin manually annotated the MRI exams. Kenneth T. Gao, Radhika Tibrewala, and Madeline Hess developed and trained the deep learning models. Upasana U. Bharadwaj, Thomas M. Link, and Cynthia T. Chin performed the AI-assisted experiment and interpreted the

results. All authors provided critical feedback and approved the final submitted manuscript.

### ACKNOWLEDGMENTS

This work is fully supported by the National Institutes of Health and National Institute of Arthritis and Musculoskeletal and Skin Diseases (Project #: UH2AR076724).

### CONFLICT OF INTEREST

The authors declare no conflicts of interest.

### ORCID

Kenneth T. Gao  <https://orcid.org/0000-0002-5975-0127>

### REFERENCES

- Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *Lancet*. 2016;388(10053):1545-1602.
- Hartvigsen J, Hancock MJ, Kongsted A, et al. What low back pain is and why we need to pay attention. *Lancet*. 2018;391(10137):2356-2367.
- Maher C, Underwood M, Buchbinder R. Non-specific low back pain. *Lancet*. 2017;389(10070):736-747.
- Liu J, Hao L, Suyou L, et al. Biomechanical properties of lumbar endplates and their correlation with MRI findings of lumbar degeneration. *J Biomech*. 2016;49(4):586-593.
- Brinjikji W, Diehn FE, Jarvik JG, et al. MRI findings of disc degeneration are more prevalent in adults with low Back pain than in asymptomatic controls: a systematic review and meta-analysis. *Am J Neuroradiol*. 2015;36(12):2394-2399.
- Weishaupt D, Zanetti M, Hodler J, et al. Painful lumbar disk derangement: relevance of endplate abnormalities at MR imaging. *Radiology*. 2001;218(2):420-427.
- Kjaer P, Korsholm L, Bendix T, Sorensen JS, Leboeuf-Yde C. Modic changes and their associations with clinical findings. *Eur Spine J*. 2006; 15(9):1312-1319.
- Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. *Radiology*. 1988;166(1 Pt 1):193-199.
- Herlin C, Kjaer P, Espeland A, et al. Modic changes—their associations with low back pain and activity limitation: a systematic literature review and meta-analysis. *PLoS One*. 2018;13(8):e0200677.
- Zhang Y-H, Zhao C-Q, Jiang L-S, Chen X-D, Dai L-Y. Modic changes: a systematic review of the literature. *Eur Spine J*. 2008;17(10):1289-1299.
- Din RU, Cheng X, Yang H. Diagnostic role of magnetic resonance imaging in low Back pain caused by vertebral endplate degeneration. *J Magn Reson Imaging*. 2022;55:755-771.
- Fields AJ, Liebenberg EC, Lotz JC. Innervation of pathologies in the lumbar vertebral end plate and intervertebral disc. *Spine J*. 2014; 14(3):513-521.
- Ohtori S, Inoue G, Ito T, et al. Tumor necrosis factor-immunoreactive cells and PGP 9.5-immunoreactive nerve fibers in vertebral endplates of patients with discogenic low back pain and Modic type 1 or type 2 changes on MRI. *Spine*. 2006;31(9):1026-1031.
- Farshad-Amacker NA, Hughes A, Herzog RJ, Seifert B, Farshad M. The intervertebral disc, the endplates and the vertebral bone marrow as a unit in the process of degeneration. *Eur Radiol*. 2017;27(6):2507-2520.
- Fields AJ, Battie MC, Herzog RJ, et al. Measuring and reporting of vertebral endplate bone marrow lesions as seen on MRI (Modic changes): recommendations from the ISSLS degenerative spinal phenotypes group. *Eur Spine J*. 2019;28(10):2266-2274.

16. Bendix T, Sorensen JS, Henriksson GAC, Bolstad JE, Narvestad EK, Jensen TS. Lumbar Modic changes—a comparison between findings at low- and high-field magnetic resonance imaging. *Spine*. 2012; 37(20):1756-1762.
17. Fields AJ, Ballatori A, Han M, et al. Measurement of vertebral endplate bone marrow lesion (Modic change) composition with water-fat MRI and relationship to patient-reported outcome measures. *Eur Spine J*. 2021;30(9):2549-2556.
18. Karampinos DC, Melkus G, Baum T, Bauer JS, Rummeny EJ, Krug R. Bone marrow fat quantification in the presence of trabecular bone: initial comparison between water-fat imaging and single-voxel MRS. *Magn Reson Med*. 2014;71(3):1158-1165.
19. Biffar A, Dietrich O, Sourbron S, Duerr H-R, Reiser MF, Baur-Melnyk A. Diffusion and perfusion imaging of bone marrow. *Eur J Radiol*. 2010;76(3):323-328.
20. Wang Y, Videman T, Niemeläinen R, Battié MC. Quantitative measures of Modic changes in lumbar spine magnetic resonance imaging: intra- and inter-rater reliability. *Spine*. 2011;36(15):1236-1243.
21. Lessmann N, van Ginneken B, de Jong PA, Išgum I. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Med Image Anal*. 2019;53:142-155.
22. Weber KA, Abbott R, Bojilov V, et al. Multi-muscle deep learning segmentation to automate the quantification of muscle fat infiltration in cervical spine conditions. *Sci Rep*. 2021;11(1):16567.
23. Han Z, Wei B, Mercado A, Leung S, Li S. Spine-GAN: semantic segmentation of multiple spinal structures. *Med Image Anal*. 2018;50: 23-35.
24. Merali Z, Wang JZ, Badhiwala JH, Witiw CD, Wilson JR, Fehlings MG. A deep learning model for detection of cervical spinal cord compression in MRI scans. *Sci Rep*. 2021;11(1):10473.
25. Lu J-T, Pedemonte S, Bizzo B, et al. DeepSPINE: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. ArXiv180710215 Cs [Internet]. July 26, 2018 [cited October 1, 2020]. <http://arxiv.org/abs/1807.10215>
26. Hallinan JTPD, Zhu L, Yang K, et al. Deep learning model for automated detection and classification of central canal, lateral recess, and neural foraminal stenosis at lumbar spine MRI. *Radiology*. 2021; 300(1):130-138.
27. Wang KY, Ikwuezunma I, Puvanesarajah V, et al. Using predictive modeling and supervised machine learning to identify patients at risk for venous thromboembolism following posterior lumbar fusion. *Glob Spine J*. 2021;21925682211019360.
28. Fatemi P, Zhang Y, Han SS, et al. External validation of a predictive model of adverse events following spine surgery. *Spine J*. 2022;22: 104-112.
29. Jamaludin A, Kadir T, Zisserman A. SpineNet: automated classification and evidence visualization in spinal MRIs. *Med Image Anal*. 2017;41: 63-73.
30. Milletari F, Navab N, Ahmadi S-A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. ArXiv160604797 Cs [Internet]. June 15, 2016 [cited February 5, 2021]. <http://arxiv.org/abs/1606.04797>
31. Albert HB, Briggs AM, Kent P, Byrhagen A, Hansen C, Kjaergaard K. The prevalence of MRI-defined spinal pathoanatomies and their association with Modic changes in individuals seeking care for low back pain. *Eur Spine J*. 2011;20(8):1355-1362.
32. Ekşi MŞ, Özcan-Ekşi EE, Özmen BB, et al. Lumbar intervertebral disc degeneration, end-plates and paraspinal muscle changes in children and adolescents with low-back pain. *J Pediatr Orthop B*. 2020;31:93-102.
33. Mitra D, Cassar-Pullicino VN, McCall IW. Longitudinal study of vertebral type-1 end-plate changes on MR of the lumbar spine. *Eur Radiol*. 2004;14(9):1574-1581.
34. Hutton MJ, Bayer JH, Powell JM. Modic vertebral body changes: the natural history as assessed by consecutive magnetic resonance imaging. *Spine*. 2011;36(26):2304-2307.
35. Xu L, Chu B, Feng Y, Xu F, Zou Y-F. Modic changes in lumbar spine: prevalence and distribution patterns of end plate oedema and end plate sclerosis. *Br J Radiol*. 2016;89(1060):20150650.
36. Andrearczyk V, Whelan PF. Using filter banks in convolutional neural networks for texture classification. *Pattern Recognit Lett*. 2016;84:63-69.
37. Hattikatti P. Texture based interstitial lung disease detection using convolutional neural network. In: 2017 International Conference on Big Data, IoT and Data Science (BIGD); 2017, pp. 18-22.
38. Jensen TS, Karppinen J, Sorensen JS, Niinimäki J, Leboeuf-Yde C. Vertebral endplate signal changes (Modic change): a systematic literature review of prevalence and association with non-specific low back pain. *Eur Spine J*. 2008;17(11):1407-1422.
39. Albert HB, Manniche C. Modic changes following lumbar disc herniation. *Eur Spine J*. 2007;16(7):977-982.
40. Kjaer P, Leboeuf-Yde C, Korsholm L, Sorensen JS, Bendix T. Magnetic resonance imaging and low Back pain in adults: a diagnostic imaging study of 40-year-old men and women. *Spine*. 2005;30(10): 1173-1180.

**How to cite this article:** Gao, K. T., Tibrewala, R., Hess, M., Bharadwaj, U. U., Inamdar, G., Link, T. M., Chin, C. T., Padoia, V., & Majumdar, S. (2022). Automatic detection and voxel-wise mapping of lumbar spine Modic changes with deep learning. *JOR Spine*, 5(2), e1204. <https://doi.org/10.1002/jsp2.1204>