Learning-based fully automated prediction of lumbar disc degeneration progression with specified clinical parameters and preliminary validation

Abstract

Background

Lumbar disc degeneration (LDD) may be related to aging, biomechanical and genetic factors. Despite the extensive work on understanding its etiology, there is currently no automated tool for accurate prediction of its progression.

Purpose

We aim to establish a novel deep learning-based pipeline to predict the progression of LDD-related findings using lumbar MRIs.

Materials and Methods

We utilized our dataset with MRIs acquired from 1,343 individual participants (taken at the baseline and the 5-year follow-up timepoint), and progression assessments (the Schneiderman score, disc bulging, and Pfirrmann grading) that were labelled by spine specialists with over ten years clinical experience. Our new pipeline was realized by integrating the MRI-SegFlow and the Visual Geometry Group-Medium (VGG-M) for automated disc region detection and LDD progression prediction correspondingly. The LDD progression was quantified by comparing the Schneiderman score, disc bulging and Pfirrmann grading at the baseline and at follow-up. A 5-fold cross-validation was conducted to assess the predictive performance of the new pipeline.

Results

Our pipeline achieved very good performances on the LDD progression prediction, with high progression prediction accuracy of the Schneiderman score (Accuracy: 90.2 ± 0.9%), disc bulging (Accuracy: 90.4% ± 1.1%), and Pfirrmann grading (Accuracy: 89.9% ± 2.1%).

Conclusion

This is the first attempt of using deep learning to predict LDD progression on a large dataset with 5-year follow-up. Requiring no human interference, our pipeline can potentially achieve similar predictive performances in new settings with minimal efforts.

Keywords: Lumbar disc degeneration; Convolutional neural network; Magnetic resonance imaging; Disease progression prediction

Introduction

Lumbar disc degeneration (LDD) is one of the main potential causes for low back pain and is associated with reduced quality of life, work disability, potential psychological distress, and increased health-care costs [1]. Magnetic resonance imaging (MRI) of the lumbar spine is used to diagnose LDD and to guide clinical management. Assessment of LDD on MRIs often includes characterization of reduced disc signal intensity, high-intensity zones and structural abnormalities [2]. Many known parameters are associated with LDD such as increasing age and body mass index, presence of Modic changes and low pelvic incidence [3, 4]. However, there is less work on LDD progression prediction. Despite the intuitive association with aging, some contradicting evidence from a population-based cohort suggests this association is insignificant [5]. Based on twins data, there may be genetic heritability for longitudinal changes in disc signal intensity and disc bulging [6]. Disc bulging may not progress and may even resolve in some cases [7, 8]. There are no learning-based studies to predict LDD progression.

Machine learning for utilizing longitudinal big data to establish predictive models can be a potential solution [9-11]. Convolutional Neural Network (CNN) has achieved a remarkable performance in MRI analysis tasks including pathology classification [12-15], landmark detection [16, 17], and segmentation [18-21]. In comparison with the conventional machine learning approach, such as support vector machine (SVM) [22, 23], CNN does not rely on the rule-based shallow image features that are often perceptible for humans. By performing a series of convolution operations, CNN models can extract the hierarchical features automatically from the input image. Since the feature extraction is mathematical and does not always conform to human visual patterns [24, 25], CNN can utilize both perceptible and non-perceptible image features.

There is no previous work using CNNs to predict longitudinal changes in LDD. The major obstacle for such studies is the lack of labelled MRI datasets with follow-up for training the model. In this study, we aim to develop and validate a deep learning pipeline for the 5-year progression prediction of LDD. The objectives include 1) mapping the data of a large MRI dataset with labels and follow-up; 2) developing a pipeline for LDD progression prediction; 3) testing the progression prediction accuracy.

Materials and Methods

Dataset

The dataset was constructed from the Hong Kong Disc Degeneration Population-Based Cohort of Southern Chinese participants [3]. Written consent was obtained from all subjects and ethics was approved by the local institutional review board. Subjects who were 18 years or older were recruited by open invitation using newspaper advertisements, posters, and e-mails. Subjects were interviewed for demographic data and underwent MRIs examinations. Participants with prior surgical treatment of the spine, spinal tumors, and marked spinal deformities were excluded from the cohort. The dataset consisted of 1343 participants' sagittal lumbar T2-weighted MRIs at baseline and follow-up timepoint (in total 2686 sets of MRIs). The follow-up images were obtained at 5-year (within 6 months deviation) from the initial image. The images were obtained from three different institutions with the same MRI protocol, which demonstrated the diversity of our dataset. All patients have been previously reported [3]. The prior article dealt with the association between LDD and body weight in adult, whereas in this study we were predicting the long-term progression of the LDD using MR and deep learning technologies.

MRI Protocol

All subjects included in this study underwent 1.5T HD MRI with sagittal imaging at L1-S1. The detailed MRI protocol has been described in the previous study [26], but briefly participants were oriented in supine position. For T2-weighted sagittal scans, the field of view was 28cm×28cm, slice thickness was 5mm, slice spacing was 1mm, and imaging matrix was 448×336. The repetition time for T2-weighted MRI 3320ms, and the echo time was 85ms.

MRI parameters

Three MRI phenotypes (Figure 1) of Schneiderman score, disc bulging, and Pfirrmann grading were examined. For Schneiderman's score [27], the disc signal intensity was divided into 4 grades: grade 0 represents normal disc height and signal intensity; grade 1 represents speckled pattern or heterogenous decreased signal intensity; grade 2 represents diffuse loss of signal; and grade 3 indicates a signal void. Disc bulging was subclassified as: 0 = no disc herniation; 1 = posterior disc bulging (disc displaced beyond a virtual line connecting the posterior edges of two adjacent vertebrae); 2 = disc extrusion (distance between the edge of the protruded disc into the spinal canal was greater than the distance between edges of the base of the disc); 3 = disc sequestration [2]. Disc degeneration was also evaluated using the Pfirrmann grading [28] which assessed disc signal intensity by 5 grades: 1 = homogeneous bright white disc; 2 = inhomogeneous white disc and/or horizontal bands; 3 = inhomogeneous grey disc; 4 = inhomogeneous grey to black disc; 5 = inhomogeneous black disc with probable disc space collapse. All measurements were performed by two spine specialists with over ten years clinical experience, blinded to the participant's demographics. Any deviation in gradings was discussed and a final consensus score was determined. For each grading, progression was labelled when the follow-up grade was more than the initial baseline score.

Prediction Pipeline

The pipeline of our learning-based progression prediction system is summarized in Figure 2. First, the region of each disc was detected and extracted from the lumbar MRI, based on the pixel-wise vertebral masks produced by the MRI-SegFlow [29], a novel unsupervised vertebrae segmentation method published by our group. The disc region was defined as the $1.5w \times 2w \times n$ cuboid between two adjacent vertebrae, where w represents the average width of vertebrae in the MRI, and n represents the slice number of the MRI series. It was followed by resizing the disc regions with different shapes to a standard size and input to a deep learning model using the basic architecture of a CNN model by adopting the framework of Visual Geometry Group-Medium (VGG-M) [30], which has relatively deep network architecture, extracting highly abstract image features. The encoder of the model was trained to extract the features from the disc region. With these features, the classifier produced the probability for each follow-up pathology grade, and the grade with the highest probability was defined as the grade prediction. Referring to the baseline grade, the state of disc pathology progression was determined. Since one model only handled one specific pathology, three models were built for the prediction of Schneiderman score, bulging, and Pfirrmann grading, respectively. The detailed network architecture, training strategy, and implementation details are stated in the Appendix.

Evaluation Metrics

The performance of our pipeline was evaluated by the Accuracy (Acc), Precision (Pre), Recall (Rec) and F_1 of progression prediction for each pathology. They were defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Pre = \frac{TP}{TP + FP}$$

$$Rec = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 \times Pre \times Rec}{Pre + Rec}$$

where TP represented the number of true positive samples which were the samples labeled as progression and predicted correctly by the deep learning method. The FN represented the number of false negative samples which were also progression samples but predicted incorrectly. The TN and FP represented the numbers of true negative and false positive samples, which were the non-progression samples predicted correctly and incorrectly respectively. The Acc represented the overall performance of our method in the progression prediction task, while the Pre, Rec, and F_1 illustrated our method's ability on recognition of progression samples.

Results

The study subjects (39% male and 61% female) had a mean age of 44.8 years (SD 9.7) with the major participants older than 45 years (57.1%), weight of 61.1kg (SD 11.0), height of 1.62m (SD 0.09) and body mass index of 23.1 kg/m2 (SD 3.5). The detailed demographics of dataset is presented in Table 1. The summary of the label and progression distribution are shown in Table 2. We found that the distributions were imbalanced. The percentages of the cases with LDD progression were less than those with no progression. The baseline grade distribution is presented in Table 3, which shows that the discs with the same follow-up grade tend to have similar baseline grades.

Our new progression prediction pipeline was validated according to the implementation details presented in the Appendix. The percentages of the TP, TN, FP and FN samples were calculated first (Table 4). Then the evaluation metrics, including Accuracy, Precision, Recall and F1 were derived (Table 4). Our method achieved remarkable overall accuracy in all predictions of the three LDD clinical parameters (Schneiderman score: 90.2%, Disc Bulging: 90.4%, Pfirrmann grading: 89.9%). For the Schneiderman score, the Precision, Recall and F1 were 89.6%, 96.0% and 92.7% respectively, which illustrated the superior ability of our method on the identification of progression samples. However, due to the imbalanced sample distribution, our method only achieved suboptimal performance on the progression identification for Disc Bulging and Pfirrmann grading (the Precision, Recall and F1 were 80.2%, 76.5% and 78.3% for Disc Bulging, and 64.9%, 60.4% and 62.6% for Pfirrmann grading).

Discussion

We developed the first deep learning embedded pipeline for predicting LDD progression, which integrated the published MRI-SegFlow and the basic network architecture of VGG-M. Compared with other machine learning approaches for MRI analysis, our method can extract the highly abstract features from the raw MRIs automatically without relying on any rule-based feature extraction. Thus, it can learn the information that is non-perceptible for humans by looking at MRIs. Since the prediction process is fully mathematical without any subjective or random factors, the results from our pipeline is consistent, providing accurate detection of LDD progression. Our findings lay the foundations for early detection of progressive diseases whereby preventive measures and interventions may be implemented to potentially reduce the number of surgeries.

The heterogenicity of either the progression group or the non-progression group is large. Each group may have different pathological grades in Schneiderman score, disc bulging and Pfirrmann grading. The difficulties in learning whether the pathology will progress or not based on different baselines and follow-up pathologies is a challenging task for a deep learning method [31]. It can be observed (Table 3) that the discs with the same follow-up pathology grading tend to have similar baseline grades. For instance, for the disc with follow-up Pfirrmann grade of 4, 74.9% of them had baseline grade of 4, 20.3% of them had baseline grade of 3, and 4.0% of them progressed from grade 2. The discs with different follow-up grades usually have a different distribution of baseline grades. Additionally, 85.1% of the discs with follow-up Schneiderman score of 1 progressed from grade 0, while only 23.6% of the discs with follow-up score of 2 had baseline score of 0. Therefore, instead of directly predicting whether the pathology will progress or not, we predicted the follow-up stage of the pathology to reduce the heterogenicity of the groups, and then computed any progression or no progression based on the predicted pathology grading.

It must be acknowledged that the deep learning model is data driven [9-11], which means the performance of the model is depended on the label distribution of the training dataset. Our pipeline achieved remarkable Pre, Rec and F_1 in the progression prediction of the Schneiderman score, which illustrated the excellent ability of this pipeline on the identification of progression samples. It is mainly because the distribution of progression and non-progression samples is balanced for the Schneiderman score. However, for disc bulging and Pfirrmann grading, the sample distributions are highly imbalanced. For disc bulging (Table 2), 76.6% samples did not progress, and for Pfirrmann grading 85.9% samples did not progress. A model trained with this imbalanced data will tend to distinguish an unseen sample as non-progressive. Therefore, our model achieved sub-optimal performance in the progression detection of these two pathologies in comparison with Schneiderman score. With an increase in the data volume, especially with the number of progression samples, our method can produce improved performance in the prediction of disc bulging and Pfirrmann grading. As for the Schneiderman score, our method already provides a reliable progression prediction.

We adopted several data-level and algorithm-level methods to deal with the unbalanced label distribution problem, such as oversampling, undersampling, variable loss weight [32] and SMOTE [33]. However, there is no significant improvement and even reduction in the model performance. This may be because the small number of progression samples cannot fully represent their patterns and lack a clear data structure, thus the model is not able to learn an optimal decision boundary for the identification of the progression samples [32]. Despite this, our dataset is based on a population cohort and with a large sample size and five-year longitudinal follow-up. This dataset should already reflect the true pathology progression distribution. Data skewed towards non-progression in our learned model may reflect the real situation and true progression probability.

There are still some limitations to our pathology progression prediction method. Since our method is based on CNN, it requires a large number of labelled training data, and has high dataset sensitivity. This means that if the method is tested on an MRI with different image quality from the training data, the performance will be reduced. To accommodate a new image quality, a well-trained CNN model still requires several hundreds of labelled follow-up MRI for finetuning. Also, this model was developed using a dataset based on southern Chinese individuals. Whether this is applicable in other ethnicities require further replication. We will further validate our method in other populations. Besides, the prediction process of our method is only based on the MRI findings, and clinical information such as age, sex and weight were not involved. These additional factors may be required to link the imaging findings to real clinical implications. We will continue to enrich our follow-up MRI dataset to improve the performance of our CNN model. Prospective testing at other external institutes will be conducted to validate the robustness of our method, and a similar performance will likely be achieved due to the diversity of our dataset. In addition, the clinical information will be merged into the prediction process, and the network architecture will be further modified to inhibit the dataset sensitivity. Longer term follow-up data is also useful as the progression potential may be higher with more subjects experiencing LDD progression.

Conclusion

We have developed and tested a new pipeline for predicting LDD progression. This is the first deep learning embedded pipeline to be used in the task of pathology prediction. A large labelled MRI dataset with follow-up was utilized for the training and testing of our method. The validation result shows that our method achieved remarkable accuracy in the progression prediction of the Schneiderman score, disc bulging, and Pfirrmann grading. Our method has shown superior ability on the identification of progression samples for the Schneiderman score. With increased training data, the performance of our method can be further improved, and it has significant potential for clinical implementation. Future study will be conducted for interpretation of the model, identifying image features and related underlying pathology.

References

1. Disease GBD, Injury I, Prevalence C (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 392:1789-1858. doi: 10.1016/S0140-6736(18)32279-7
2. Teraguchi M, Cheung JPY, Karppinen J, Bow C, Hashizume H, Luk KDK, Cheung KMC, Samartzis D (2020). Lumbar high-intensity zones on MRI: imaging biomarkers for severe, prolonged low back pain and sciatica in a population-based cohort. The spine journal : official journal of the North American Spine Society 20:1025-1034. doi: 10.1016/j.spinee.2020.02.015
3. Samartzis D, Karppinen J, Chan D, Luk KD, Cheung KM (2012). The association of lumbar intervertebral disc degeneration on magnetic resonance imaging with body mass index in overweight and obese adults: a population-based study. Arthritis Rheum 64:1488-1496. doi: 10.1002/art.33462
4. Zehra U, Cheung JPY, Bow C, Crawford RJ, Luk KDK, Lu W, Samartzis D (2020). Spinopelvic alignment predicts disc calcification, displacement, and Modic changes: Evidence of an evolutionary etiology for clinically-relevant spinal phenotypes. JOR Spine 3:e1083. doi: 10.1002/jsp2.1083
5. Teraguchi M, Yoshimura N, Hashizume H, Yamada H, Oka H, Minamide A, Nagata K, Ishimoto Y, Kagotani R, Kawaguchi H, Tanaka S, Akune T, Nakamura K, Muraki S, Yoshida M (2017). Progression, incidence, and risk factors for intervertebral disc degeneration in a longitudinal population-based cohort: the Wakayama Spine Study. Osteoarthritis Cartilage 25:1122-1131. doi: 10.1016/j.joca.2017.01.001
6. Williams FM, Popham M, Sambrook PN, Jones AF, Spector TD, MacGregor AJ (2011). Progression of lumbar disc degeneration over a decade: a heritability study. Ann Rheum Dis 70:1203-1207. doi: 10.1136/ard.2010.146001
7. Kjaer P, Tunset A, Boyle E, Jensen TS (2016). Progression of lumbar disc herniations over an eight-year period in a group of adult Danes from the general population--a longitudinal MRI study using quantitative measures. BMC Musculoskelet Disord 17:26. doi: 10.1186/s12891-016-0865-6
8. Zhong M, Liu JT, Jiang H, Mo W, Yu PF, Li XC, Xue RR (2017). Incidence of Spontaneous Resorption of Lumbar Disc Herniation: A Meta-Analysis. Pain Physician 20:E45-E52
9. Han SS, Azad TD, Suarez PA, Ratliff JK (2019). A machine learning approach for predictive models of adverse events following spine surgery. The spine journal : official journal of the North American Spine Society 19:1772-1781. doi: 10.1016/j.spinee.2019.06.018
10. Horng MH, Kuok CP, Fu MJ, Lin CJ, Sun YN (2019). Cobb Angle Measurement of Spine from X-Ray Images Using Convolutional Neural Network. Comput Math Methods Med 2019:6357171. doi: 10.1155/2019/6357171
11. Jin R, Luk KD, Cheung JPY, Hu Y (2019). Prognosis of cervical myelopathy based on diffusion tensor imaging with artificial intelligence methods. NMR Biomed 32:e4114. doi: 10.1002/nbm.4114
12. Jamaludin A, Kadir T, Zisserman A (2017). SpineNet: Automated classification and evidence visualization in spinal MRIs. Med Image Anal 41:63-73. doi: 10.1016/j.media.2017.07.002
13. Jamaludin A, Kadir T, Zisserman A (2017). Self-supervised learning for spinal MRIs. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support.

Springer. pp. 294-302.
14. Lootus M, Kadir T, Zisserman A (2015). Automated radiological grading of spinal MRI. In: Recent advances in computational methods and clinical applications for spine imaging. Springer. pp. 119-130.
15. Lu J-T, Pedemonte S, Bizzo B, Doyle S, Andriole KP, Michalski MH, Gonzalez RG, Pomerantz SR (2018). Deepspine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. arXiv preprint arXiv:10215
16. Mader A, Lorenz C, Meyer C (2019). A General Framework for Localizing and Locally Segmenting Correlated Objects: A Case Study on Intervertebral Discs in Multi-modality MR Images. In: Annual Conference on Medical Image Understanding and Analysis. Springer. pp. 364-376.
17. Rouhier L, Romero FP, Cohen JP, Cohen-Adad J (2020). Spine intervertebral disc labeling using a fully convolutional redundant counting model. arXiv preprint arXiv:04387
18. Gros C, De Leener B, Badji A, Maranzano J, Eden D, Dupont SM, Talbott J, Zhuoquiong R, Liu Y, Granberg T (2019). Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. Neuroimage 184:901-915
19. Han Z, Wei B, Mercado A, Leung S, Li S (2018). Spine-GAN: Semantic segmentation of multiple spinal structures. Med Image Anal 50:23-35. doi: 10.1016/j.media.2018.08.005
20. Li X, Dou Q, Chen H, Fu C-W, Qi X, Belavý DL, Armbrecht G, Felsenberg D, Zheng G, Heng P-A (2018). 3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images. Med Image Anal 45:41-54
21. Perone CS, Calabrese E, Cohen-Adad J (2018). Spinal cord gray matter segmentation using deep dilated convolutions. Sci Rep 8:5966. doi: 10.1038/s41598-018-24304-3
22. Beulah A, Sharmila TS (2016). Classification of Intervertebral Disc on Lumbar MR Images using SVM. In: 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). IEEE. pp. 293-297.
23. Huang S-H, Chu Y-H, Lai S-H, Novak CL (2009). Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. IEEE transactions on medical imaging 28:1595-1605
24. Chen L, Wang S, Fan W, Sun J, Naoi S (2015). Beyond human recognition: A CNN-based framework for handwritten character recognition. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE. pp. 695-699.
25. LeCun Y, Bengio Y, Hinton GJn (2015). Deep learning. 521:436-444
26. Cheung JP, Samartzis D, Shigematsu H, Cheung KM (2014). Defining clinically relevant values for developmental spinal stenosis: a large-scale magnetic resonance imaging study. Spine 39:1067-1076. doi: 10.1097/BRS.0000000000000335
27. Schneiderman G, Flannigan B, Kingston S, Thomas J, Dillin WH, Watkins RG (1987). Magnetic resonance imaging in the diagnosis of disc degeneration: correlation with discography. Spine 12:276-281
28. Pfirrmann CW, Metzdorf A, Zanetti M, Hodler J, Boos N (2001). Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine 26:1873-1878
29. Xihe K, Jason P, Cheung, Honghan W, Socrates D, Teng Z (2020). MRI-SegFlow: a novel unsupervised deep learning pipeline enabling accurate vertebral segmentation of MRI images In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in

Medicine and Biology Society (EMBC). Montréal, Canada.
30. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014). Return of the devil in the details: delving deep into convolutional nets; 2014. arXiv preprint arXiv:14053531
31. Bishop CM (2006). Pattern recognition and machine learning. springer
32. Krawczyk B (2016). Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5:221-232. doi: 10.1007/s13748-016-0094-0
33. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002). SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 16:321-357. doi: DOI 10.1613/jair.953

Table 1: Demographics of Dataset

Age Group (years)		**18 to 40**	**40 to 50**	**over 50**
Subject Number		412	512	419
Progression Percentage	**Schneiderman score**	56.4%	67.4%	69.1%
	Disc Bulging	13.7%	24.7%	31.5%
	Pfirrmann grading	10.3%	14.5%	17.5%
Gender		**Male**		**Female**
Subject Number		524		819
BMI (kg/m^2)		**under 18.5**	**18.5 to 25.0**	**over 25.0**
Subject Number		100	920	323

Table 2: Label distribution

Schneiderman score					
Baseline Grade	0	1	2	3	
Percentage	58.6%	16.1%	18.5%	6.8%	
Follow-up Grade	0	1	2	3	
Percentage	6.4%	51.0%	39.2%	3.4%	
Progression State	Progression		Non-progression		
Percentage	64.6%		35.4%		
Disc Bulging					
Baseline Grade	0	1	2	3	
Percentage	80.2%	19.0%	0.8%	0.0%	
Follow-up Grade	0	1	2	3	
Percentage	60.3%	38.0%	1.5%	0.2%	
Progression State	Progression		Non-progression		
Percentage	23.4%		76.6%		
Pfirrmann grading					
Baseline Grade	1	2	3	4	5
Percentage	0.2%	24.2%	44.8%	30.1%	0.7%
Follow-up Grade	1	2	3	4	5
Percentage	0.8%	35.0%	32.9%	29.4%	1.9%
Progression State	Progression		Non-progression		
Percentage	14.1%		85.9%		

Table 3: Baseline grade distribution of samples with different follow-up grades

Schneiderman score					
Follow-up Grade: 0					
Baseline Grade	0	1	2	3	
Percentage	92.8%	4.2%	2.6%	0.4%	
Follow-up Grade: 1					
Baseline Grade	0	1	2	3	
Percentage	85.1%	9.4%	4.6%	0.9%	
Follow-up Grade: 2					
Baseline Grade	0	1	2	3	
Percentage	23.6%	27.8%	38.6%	10.0%	
Follow-up Grade: 3					
Baseline Grade	0	1	2	3	
Percentage	2.2%	3.5%	25.3%	69.0%	
Disc Bulging					
Follow-up Grade: 0					
Baseline Grade	0	1	2	3	
Percentage	95.8%	4.2%	0.0%	0.0%	
Follow-up Grade: 1					
Baseline Grade	0	1	2	3	
Percentage	57.4%	41.1%	1.5%	0.0%	
Follow-up Grade: 2					
Baseline Grade	0	1	2	3	
Percentage	39.1%	54.5%	5.5%	0.9%	
Follow-up Grade: 3					
Baseline Grade	0	1	2	3	
Percentage	10%	50%	40%	0.0%	
Pfirrmann grading					
Follow-up Grade: 1					
Baseline Grade	1	2	3	4	5
Percentage	1.9%	87.0%	11.1%	0.0%	0.0%
Follow-up Grade: 2					
Baseline Grade	1	2	3	4	5
Percentage	0.5%	48.3%	49.3%	1.9%	0.0%
Follow-up Grade: 3					
Baseline Grade	1	2	3	4	5
Percentage	0.0%	16.3%	64.9%	18.8%	0.0%
Follow-up Grade: 4					
Baseline Grade	1	2	3	4	5
Percentage	0.0%	4.0%	20.3%	74.9%	0.8%
Follow-up Grade: 5					
Baseline Grade	1	2	3	4	5
Percentage	0.0%	0.8%	3.8%	73.1%	22.3%

Table 4: Sensitivity and specificity of the prediction pipeline with the evaluation matrix of prediction capabilities

Schneiderman score				
Type	TP	TN	FP	FN
Percentage	62.0% ± 2.6%	28.2% ± 3.9%	7.2% ± 1.0%	2.6% ± 1.0%
Evaluation Matrix	Accuracy	Precision	Recall	$F1$
	90.2% ± 0.9%	89.6% ± 1.1%	96.0% ± 1.3%	92.7% ± 1.3%
Disc Bulging				
Type	TP	TN	FP	FN
Percentage	18.0% ± 3.5%	72.4% ± 2.8%	4.2% ± 0.6%	5.4% ± 0.2%
Evaluation Matrix	*Accuracy*	*Precision*	*Recall*	$F1$
	90.4% ± 1.1%	80.2% ± 4.3%	76.5% ± 3.5%	78.3% ± 4.2%
Pfirrmann grading				
Type	TP	TN	FP	FN
Percentage	8.7% ± 1.4%	81.2% ± 3.8%	4.7% ± 1.1%	5.4% ± 1.5%
Evaluation Matrix	*Accuracy*	*Precision*	*Recall*	$F1$
	89.9% ± 2.1%	64.9% ± 3.7%	60.4% ± 3.4%	62.6% ± 3.5%

Figure 1: An example of T2-weighted sagittal MRI of the L1-S1 discs. L1-2 is described as Schneiderman 1, with no disc herniation and Pfirrmann 2. L2-3 is described as Schneiderman 1, with no disc herniation and Pfirrmann 2. L3-4 is described as Schneiderman 1, with no disc herniation and Pfirrmann 3. L4-5 is described as Schneiderman 3, with disc bulging and Pfirrmann 5. L5-S1 is described as Schneiderman 2, with disc bulging, and Pfirrmann 4.

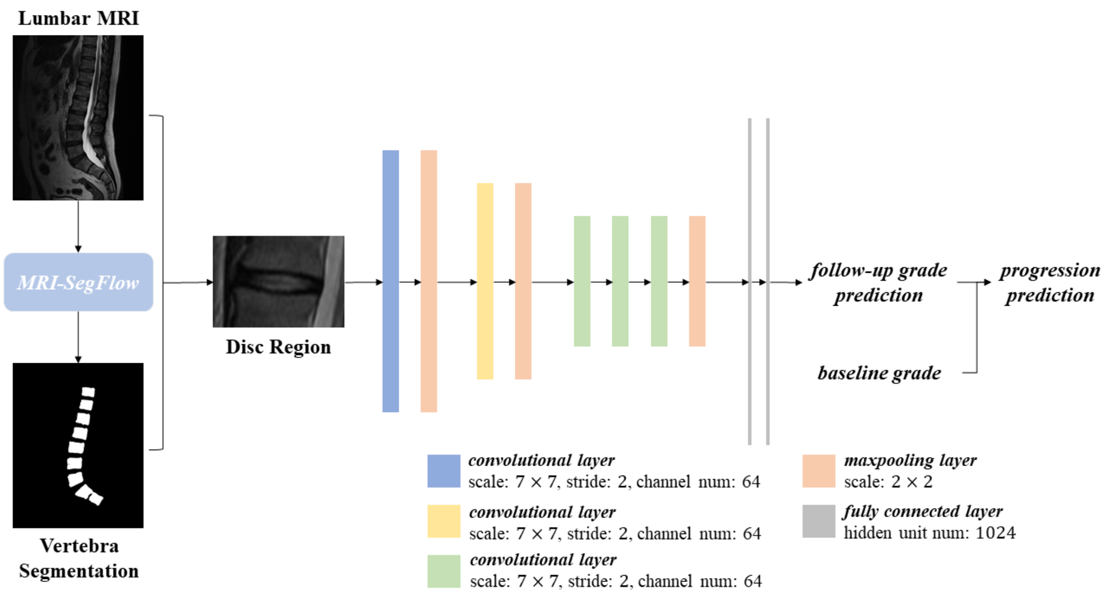

Figure 2: The pipeline of our pathology progression prediction method. MRI-SegFlow was used for the disc region detection using the protocol described in our recent published in prior to the VGG-M. The follow-up grade could be directly predicted from this pipeline. In comparison with the baseline grade, whether the pathology would progress was predicted.