Enhancing Reporting Efficiency and Accuracy for Degenerative Cervical Spine MRI with Deep Learning

Dr Aric Lee[1], Mr Junran Wu[2], Dr Changshuo Liu[2], Dr Andrew Makmur[1,2], Dr Yonghan Ting[1,2], Dr Shannon Lee[1], Dr Matthew Chan[1], Dr Desmond Lim[1], Dr Vanessa Khoo[1], Dr Jonathan Sng[1], Dr Han Yang Ong[1], Dr Amos Tan[1], Dr Shuliang Ge[1], Dr Faimee Erwan Muhamat Nor[1], Dr Yi Ting Lim[1], Dr Joey Beh[3], Ms Qai Ven Yap[2], Dr Jiong Hao Tan[1], Dr Naresh Kumar[1], Prof Beng Chin Ooi[2], Dr James TPD Hallinan[1,2]
[1]National University Hospital, [2]National University of Singapore and [3]Ng Teng Fong General Hospital, Singapore

Background

Cervical spine MRI plays a crucial role in evaluating degenerative cervical spondylosis (DCS). However, it is **time-consuming to interpret and prone to interobserver variability**. Deep learning models (DLMs) could mitigate these issues by allowing more consistent and efficient assessment. Recently, a DLM was developed to grade spinal canal and neural foraminal stenosis at each disc level, similar to radiologist evaluations.

The performance of DLMs in real-world clinical settings varies, with some studies indicating significant enhancements in radiologist efficiency and accuracy, while others report mixed outcomes. Therefore, evaluating these models in practical scenarios is essential.

This study aims to assess whether implementing a previously developed DLM enhances radiologists' efficiency and accuracy in reporting DCS MRIs.

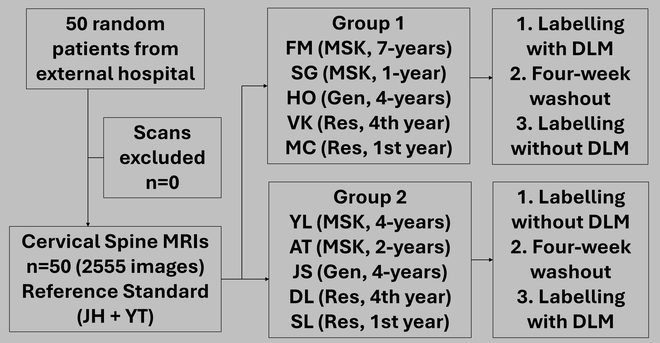
Materials and Methods

The test dataset included 50 MRIs from 2015-2018 from Ng Teng Fong General Hospital, Singapore. **Two musculoskeletal radiologists (MSKs) (JH, YT; both 12 years-post-board-certification) provided reference standard labels** in consensus, using established grading systems. Ten readers were split into two groups and prospectively interpreted anonymized MRIs at using a crossover study design (Figure 1).

The DLM was trained on 454 adult cervical spine MRIs from National University Hospital, Singapore. It comprises a convolutional neural network (CNN) and transformer (Figure 2) allowing **detection and grading of spinal canal and neural foramina stenosis at each disc level, and achieved high agreement on intitial validation**. No further model training was performed.

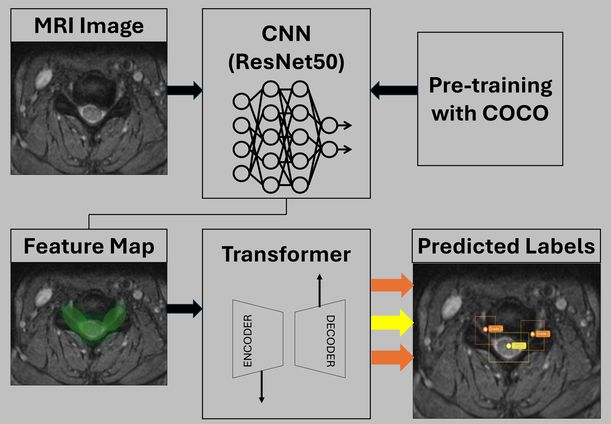
The primary outcomes were interpretation time (seconds per MRI, Linear Mixed Model analysis) and interobserver agreement (Gwet's kappa) with the reference standard. Assuming a medium effect size (Cohen's d=0.5), 80% power and 0.05 significance, 50 MRIs were included.

Figure 1: Using a crossover design, **ten readers with varying experience** labeled cases **with and without deep learning model use**. The experience of each reader post-board certification is in parentheses (for residents, the year of training is displayed).



Note- MSK = musculoskeletal radiologist, Gen = general radiologist, Res = resident, DLM = deep learning model

Figure 2: The two-component deep learning model uses a **convolutional neural network** (ResNet50, pre-trained with COCO) backbone for feature extraction. The 2D feature map is combined with positional encoding before being fed into a **transformer**. This then uses a multi-attention head mechanism and feed forward networks to generate predicted class labels, overlaid for review by readers. Readers could modify gradings as required.



Results

The test dataset comprised 50 MRIs (2555 images) from 50 patients (mean age=60, SD14; 13(26%) women). The distribution of reference standard gradings showed a slightly higher proportion of grade 0/1 labels.

DLM use significantly improved interpretation time by 69-308 seconds (p<0.001) (Table 1). There was no difference between groups (Bonferroni corrected p=1.000)

DLM use improved interobserver agreement across all stenosis gradings compared to baseline (Table 2), although not all metrics reached statistical significance. Notably, DLM performance alone was equivalent or superior to all readers.

Table 1: Timings taken for readers to label the test dataset at baseline and with deep learning model use.

Readers	Baseline	With DLM Use	Difference in time (95% CI)*	p value
Within group#				
MSK	209 (101)	119 (57)	-90 (-102- -79)	<0.001
Gen	159 (27)	90 (42)	-69 (-80- -58)	<0.001
Res	490 (649)	182 (218)	-308 (-380- -235)	<0.001

Note- data are times in seconds with standard deviation or 95% confidence intervals (*) in parentheses, or p values. # paired T tests

Table 2: Interobserver variability with the reference standard for all-class and dichotomous (grade 1/2 vs 2/3) grading

Readers	Baseline	With DLM Use	p value for Difference
All-Class Grading			
Spinal canal			
MSK	0.47 (0.43-0.51)	0.59 (0.56-0.63)	<0.001
Gen	0.48 (0.42-0.53)	0.69 (0.64-0.74)	<0.001
Res	0.46 (0.42-0.50)	0.68 (0.65-0.72)	<0.001
DLM	N/A	0.72 (0.66-0.79)	N/A
Neural foramina			
MSK	0.44 (0.41-0.47)	0.52 (0.49-0.55)	<0.001
Gen	0.43 (0.39-0.47)	0.55 (0.51-0.59)	<0.001
Res	0.41 (0.38-0.44)	0.48 (0.45-0.51)	0.003
DLM	N/A	0.56 (0.50-0.62)	N/A
Dichotomous Grading			
Spinal canal			
MSK	0.60 (0.56-0.65)	0.67 (0.62-0.71)	0.052
Gen	0.65 (0.59-0.72)	0.77 (0.71-0.82)	0.008
Res	0.63 (0.58-0.68)	0.77 (0.73-0.81)	<0.001
DLM	N/A	0.81 (0.74-0.88)	N/A
Neural foramina			
MSK	0.60 (0.56-0.63)	0.72 (0.69-0.75)	<0.001
Gen	0.66 (0.62-0.71)	0.73 (0.69-0.77)	0.003
Res	0.67 (0.64-0.70)	0.70 (0.67-0.73)	0.230
DLM	N/A	0.73 (0.67-0.79)	N/A

Note- data are Gwet's kappas with 95% confidence interval in parentheses, or p values. MSK = musculoskeletal radiologist, Gen = general radiologist, Res = resident, DLM = deep learning model

Discussion

DLM use enhanced efficiency and accuracy in DCS MRI assessment across experience levels, potentially addressing conventional limitations like interpretation tedium and interobserver variability.

Time-savings and accuracy gains for spinal canal assessement correlated inversely with experience level, echoing similar studies on lumbar spine MRI. Conversely, improvement in accuracy levels for neural foraminal assessment did not follow this trend. Other factors like radiologist's adaptability and personality likely also play a part.

The DLM had equivalent or better performance than all readers. However, comprehensive radiologist assessment remains necessary for evaluating findings beyond the model's training, such as marrow signal abnormalities. Our DLM implementation focuses on supporting radiologists, integrating with our PACS to overlay bounding boxes on DICOMs, prioritising patient safety and data security.

Study limitations include a relatively small sample size, variability in institutional imaging protocols and grading scales. Future studies could include other findings important to treatment like alignment or cord signal.

Conclusion

Using a DLM improved radiologists' interpretation time and interobserver agreement, regardless of experience level.

References
1. Theodore N. Degenerative cervical spondylosis. N Engl J Med. 2020;383(2):159-168. doi:10.1056/NEJMra2003558
2. Karpova A, Arun R, Davis AM, et al. Reliability of quantitative magnetic resonance imaging methods in the assessment of spinal canal stenosis and cord compression in cervical myelopathy. Spine (Phila Pa 1976). 2013;38(3):245-252. doi:10.1097/BRS.0b013e3182672307
3. Kim S, Lee JW, Chai JW, et al. A new MRI grading system for cervical foraminal stenosis based on axial T2-weighted images. Korean J Radiol. 2015;16(6):1294-1302. doi:10.3348/kjr.2015.16.6.1294
4. Grochmal JK, Lozen AM, Klein AP, Mark LP, Li J, Wang MC. Interobserver reliability of magnetic resonance imaging predictors of outcome in cervical spine degenerative conditions. World Neurosurg. 2018;117:e215-e220. doi:10.1016/j.wneu.2018.05.242
5. Lee A, C Liu, J Wu, et al. A140: Deep learning model for detection and classification of central canal and neural foraminal stenosis combined with vertebral segmentation on cervical spine MRI. In: Oral presentations. Global Spine J. 2024;14(4_suppl):50S-51S. doi:10.1177/21925682241239498
6. Lim DSW, Makmur A, Zhu L, et al. Improved productivity using deep learning-assisted reporting for lumbar spine MRI. Radiology. 2022;305(1):160-166. doi:10.1148/radiol.220076
7. Muhle C, Metzner J, Weinert D, et al. Classification system based on kinematic MR imaging in cervical spondylitic myelopathy. AJNR Am J Neuroradiol. 1998;19(9):1763-1771.

Correspondence:
Dr Aric Lee, Senior Resident, Department of Diagnostic Imaging, National University Hospital
Email: aricleewz@gmail.com