Title: Automated classification of intramedullary spinal cord tumors and inflammatory demyelinating lesions using deep learning

Running title: Deep learning classification of spinal cord lesions

Manuscript type: AI in Brief

Summary Statement

A deep learning pipeline for segmentation and classification of spinal cord lesions was established to support an accurate radiological diagnosis, which sometimes outperforms experienced neuroradiologists.

Key Points

Dice scores of 0.77, 0.80, 0.50 and 0.58 were obtained based on the segmentation of spinal cord lesions for astrocytoma, ependymoma, multiple sclerosis and neuromyelitis optica spectrum disorders (NMOSD), respectively, against manual labels.

Accuracies of 96%, 82% and 79% were obtained for the classifications of tumor vs. demyelinating lesion, astrocytoma vs. ependymoma, and multiple sclerosis vs. NMOSD, respectively.

In radiologically difficult cases, an accuracy of 79-95% was still achieved by the classifier.

Abstract

Accurate and robust differentiation of intramedullary spinal cord tumors and inflammatory demyelinating lesions and their subtypes are warranted, since they have overlapping MRI characteristics but different treatments and prognosis. We aimed to develop a pipeline for spinal cord lesion segmentation and classification using 2D MultiResUNet and DenseNet121 networks based on T2-weighted images. A retrospective cohort of 490 patients (118 astrocytoma, 130 ependymoma, 101 multiple sclerosis (MS) and 141 neuromyelitis optica spectrum disorders (NMOSD)) was used for model development, and an additional prospective cohort of 157 patients (34 astrocytoma, 45 ependymoma, 33 MS and 45 NMOSD) was used for model testing. In the test cohort, dice scores of 0.77, 0.80, 0.50 and 0.58 were obtained based on the segmentation of spinal cord lesions for astrocytoma, ependymoma, MS and NMOSD, respectively, against manual labels. Accuracies of 96% (area under the curve (AUC)=0.99), 82% (AUC=0.90) and 79% (AUC=0.85) were achieved for the classifications of tumor vs. demyelinating lesion, astrocytoma vs. ependymoma, and MS vs. NMOSD, respectively. In a subset of radiologically difficult cases, an accuracy of 79-95% (AUC=0.78-0.97) was still obtained by the classifier. The established deep learning pipeline for segmentation and classification of spinal cord lesions can support an accurate radiological diagnosis.

Keywords: spinal cord MRI; astrocytoma; ependymoma, multiple sclerosis; neuromyelitis optica spectrum disorder; deep learning.

Introduction

Intramedullary spinal cord tumors and inflammatory demyelinating lesions share several MRI characteristics (e.g., localization, shape, signal intensity and contrast enhancement) [1-3], posing a clinical challenge for accurate diagnosis. It is essential to accurately differentiate spinal cord tumors (including astrocytoma and ependymoma) from demyelinating lesions (including multiple sclerosis (MS) and neuromyelitis optica spectrum disorders (NMOSD)), as well as accurate classification of these subtypes, as this implies fundamentally different treatments and prognosis.

Substantial progress has been made in applying deep learning (DL) to diagnosing brain disorders [4-6], but only a few DL studies have focused on spinal cord diseases [7, 8]. The limited evidence to date suggests that DL can be utilized to characterize and segment spinal cord tumors or demyelinating lesions [7, 8], but no study has addressed the differential diagnosis of spinal cord tumors and demyelinating lesions or their subtypes. While automated pipelines for clinical diagnosis integrating lesion segmentation and differential diagnosis by DL have been reported for supratentorial lesions (e.g., gliomas and white matter hyperintensities) [5, 6, 9], they have not yet been reported for intramedullary spinal cord lesions.

We hypothesized that a DL pipeline for the accurate classification of intramedullary lesions, notably spinal cord tumors (astrocytoma and ependymoma) and inflammatory demyelinating lesions (MS and NMOSD), as well as their subtypes, could be achieved

using MR images. Therefore, we conducted this study to validate the above hypothesis using T2-weighted (T2w) images. We deliberately chose basic T2w images as they are generally clinically available in most cases.

Materials and Methods

Authors who are not employees of or consultants for BioMind, Neusoft, Bayer-Schering, Biogen-Idec, GeNeuro, Ixico, Merck-Serono, Novartis or Roche had control of image and clinical data that might present a conflict of interest for authors ZH. L, XP. G, XD. G and F.B.

The aim of this study was to develop a DL pipeline for assisting clinical diagnosis by integrating the segmentation and classification of spinal cord tumors versus demyelinating lesions and their subtypes (3 two-classification models including tumor vs. demyelinating lesion [Model 1], astrocytoma vs. ependymoma [Model 2], MS vs. NMOSD [Model 3]) based on sagittal T2w images (**eTable 1** and **eMethods**) using 2D MultiResUNet [10, 11] and DenseNet121 networks [12]. The code is available at https://github.com/Leezhaohui/spinalcord_classification.

From Jan 2012 to Dec 2018, we retrospectively identified 494 patients based on their first clinical diagnosis and prior to their clinical treatments to train (n=392, 80%) and validate (n=98, 20%) the segmentation and classification models (**Table 1**, **eMethods** and **eResults**). For independent testing, 157 patients were prospectively and consecutively enrolled from Jan 2019 to Dec 2020 (**Table 1**, **eMethods** and **eResults**). Refer to **Figure 1** and **eMethods** for inclusion and exclusion criteria. Radiological assessments, including lesion characteristics, manual lesion segmentation and classification, were performed by neuroradiologists (**eMethods**, **eResults** and **eTable**

2) with reference to other available modalities (e.g., T1w, contrast-enhanced T1w (cT1w) and axial T2w images). Difficult cases were determined as those with disagreement in the most likely diagnoses by the neuroradiologists (**eMethods**). Details of image preparation for DL and model development can be found in **eMethods**. A pipeline including segmentation and classification of spinal cord lesions is shown in Figure 2A. The Dice score was used to evaluate segmentation performance (see **eMethods** for additional statistical analyses). Accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), precision, recall and area under the curve (AUC) were calculated to evaluate classification performance. Additionally, model explanations were conducted by gradient-weighted class activation mapping (Grad-CAM), subgroup analyses according to patient age and sex, and additional combinations with available cT1w images (**eMethods**).

This study was in accordance with the Declaration of Helsinki and approved by the Animal and Human Ethics Committee of the local institute.

Results

DL segmentation of spinal cord tumors and demyelinating lesions

In the independent test cohort, the mean Dice scores were 0.77, 0.80, 0.50 and 0.58 for astrocytoma, ependymoma, MS and NMOSD, respectively (**eTable 3** and **Figure 2B** show representative cases). A subset of DL segmentations (7%-10% tumors and 29%-33% demyelinating lesions) needed further manual review and correction (see **eMethods** and **eResults**).

DL classification of spinal cord tumors and demyelinating lesions

Based on Model 1, an accuracy of 96% (150/157), sensitivity of 97% (76/78), specificity of 94% (74/79) and AUC of 0.99 were achieved on the independent test cohort for the classification of tumor versus demyelination (additional statistics are found in **Table 2, eFigure 1 and eFigure 2**), which is comparable to the neuroradiologists' performance (accuracy of 97% [152/157], **eResults** and **eTable 2)**. An accuracy of 95% (38/40), sensitivity of 95% (21/22) and specificity of 94% (17/18) were achieved for the classification of difficult cases.

Based on Model 2, an accuracy of 82% (65/79), sensitivity of 76% (34/45), specificity of 91% (31/34), and AUC of 0.90 were achieved on the independent test cohort for the classification of astrocytoma versus ependymoma, which is superior to the neuroradiologists' performance (accuracy 72% [57/79]). This performance was maintained for difficult cases, where an accuracy of 83% (15/18), sensitivity of 86%

(6/7) and specificity of 82% (9/11) were achieved.

Based on Model 3, an accuracy of 79% (62/78), sensitivity of 80% (36/45), specificity of 79% (26/33) and AUC of 0.85 were achieved on the independent test cohort for the classification of MS and NMOSD lesions, which is superior to the neuroradiologists' performance (accuracy of 67% [52/78]). This performance was maintained for difficult cases, where an accuracy of 82% (18/22), sensitivity of 87% (13/15) and specificity of 71% (5/7) were achieved.

Model explanation

The Grad-CAM showed that the main activation areas were the lesion and perilesional areas in patients with tumors or demyelinating lesions (**eResults** and **eTable 4**). The model performance only degraded in the pediatric and male subgroups, with decreased sensitivity in the classification of astrocytoma versus ependymoma (**eTable 5**). The cT1w images had no additional contribution to whole lesion segmentation and only improved the classification accuracy of MS versus NMOSD (**eTable 6**).

Discussion

In this study, a DL pipeline for spinal cord lesion segmentation and classification was first developed using the most widely available T2w images, with a manual verification/adjustment step for segmentation in up to 30% of cases. This pipeline could benefit patients without available cT1w images and facilitate fast clinical translation with robust performance across different subpopulations. For the differentiation of demyelinating lesions, cT1w imaging is recommended to achieve a better classification performance.

Few studies have focused on spinal cord lesion segmentation by DL [8]. Spinal cord tumor segmentation benefits from a relatively high tumor intensity compared to surrounding normal spinal cord tissue [13]. Our DL model showed promising segmentation performance (Dice score>0.75) that is comparable to a previous report, where a Dice score of 0.77 was reported [8]. For demyelinating lesions, DL segmentation achieved a slightly lower performance (Dice score$\leqslant$0.6, even combining cT1w images) due to the smaller volume of disseminated lesions and lower contrast of the lesion and surrounding tissue [7], which also poses a challenge in manual delineation (mean Dice score<0.75). Although the current automatic segmentation of demyelinating lesions requires manual review and frequent modification (approximately 30%), it may still aid efficient lesion segmentation.

The novelty of our study is the classification of spinal cord tumors and demyelinating

lesions and their subtypes, a clinically relevant and sometimes challenging task, using DL. Our model showed an excellent differentiation (accuracy of 96%) of spinal cord tumors versus demyelinating lesions using only T2w images, which is comparable to that (mean accuracy of 97%) by neuroradiologists. Our model may benefit from the different intensity contrast and morphological characteristics (e.g., orientation, shape, size and count as shown in Grad-CAM) [14, 15]. In addition, cysts, necrosis, cavities and hemorrhages, which are specific to tumors and typically absent in demyelinating lesions, may also contribute to the final classification [1, 14-16]. Even though the differentiation of different brain tumors has been widely reported in previous studies with accuracies above 80% [5, 17], studies on the differentiation of spinal cord tumors are lacking. The differentiation within spinal cord tumors (accuracy of 82%) using DL in the current study was superior to neuroradiologists' diagnostic performance (mean accuracy<0.75) and comparable to those in previous brain tumor studies [5, 17]. An accuracy of 79% was achieved for the differentiation of demyelinating lesions (MS versus NMOSD) using DL; this performance could be further improved by combining cT1w images (accuracy of 90%, but resulted from a relatively small sample), higher than that by neuroradiologists (mean accuracy<0.7). The contribution of the entire demyelination lesion/lesion central area and perilesional areas along the lesion margin revealed by Grad-CAM indicated potential distinct underlying pathologies within the entire lesion/lesion central area and perilesional areas, which has potential value for radiological diagnosis (**eTable 4**). A good to excellent performance (accuracies from 79% to 95%) was achieved using DL for clinically difficult cases (i.e., conflicting

diagnoses from neuroradiologists), which offers a potential use in solving clinical problems of difficult spinal cord cases.

There are some limitations in this study. First, only spinal cord T2w images were used in this study, and multimodal spinal cord MR and available brain MR images, which would provide complementary profiles, could be considered in further studies. Second, lesion segmentation by DL may have been suboptimal, particularly for demyelinating lesions. Additionally, the whole lesion on the T2w image was segmented, and different tumor components (e.g., cyst, edema and hemorrhage) may improve the classification model performance. Third, a prospective study with more types of spinal cord lesions (e.g., spinal cord infarction) and external validation is warranted to validate the established pipeline and extend the model to other spinal cord diseases.

Conclusion

A DL framework for the segmentation and classification of spinal cord lesions, including tumors (astrocytoma and ependymoma) and demyelinating diseases (MS and NMOSD), was developed and validated, with performance sometimes outperforming that of radiologists.

References

1. Abul-Kasim K, Thurnher MM, McKeever P, Sundgren PC. Intradural spinal tumors: current classification and MRI features. *Neuroradiology*. 2008;50(4):301-314. doi:10.1007/s00234-007-0345-7

2. Karussis D. The diagnosis of multiple sclerosis and the various related demyelinating syndromes: a critical review. *J Autoimmun*. 2014;48-49:134-142. doi:10.1016/j.jaut.2014.01.022

3. Kim HJ, Paul F, Lana-Peixoto MA, et al. MRI characteristics of neuromyelitis optica spectrum disorder: an international update. *Neurology*. 2015;84(11):1165-1173. doi:10.1212/WNL.0000000000001367

4. Zeng C, Gu L, Liu Z, Zhao S. Review of Deep Learning Approaches for the Segmentation of Multiple Sclerosis Lesions on Brain MRI. *Front Neuroinform*. 2020;14:610967. doi:10.3389/fninf.2020.610967

5. Shaver MM, Kohanteb PA, Chiou C, et al. Optimizing Neuro-Oncology Imaging: A Review of Deep Learning Approaches for Glioma Imaging. *Cancers (Basel)*. 2019;11(6):829. doi:10.3390/cancers11060829

6. Zlochower A, Chow DS, Chang P, Khatri D, Boockvar JA, Filippi CG. Deep Learning AI Applications in the Imaging of Glioma. *Top Magn Reson Imaging*. 2020;29(2):115-0. doi:10.1097/RMR.0000000000000237

7. Gros C, De Leener B, Badji A, et al. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. *Neuroimage*. 2019;184:901-915. doi:10.1016/j.neuroimage.2018.09.081

8. Lemay A, Gros C, Zhuo Z, et al. Automatic multiclass intramedullary spinal cord tumor segmentation on MRI with deep learning. *Neuroimage Clin*. 2021;31:102766. doi:10.1016/j.nicl.2021.102766

9. Ye Z, George A, Wu AT, et al. Deep learning with diffusion basis spectrum imaging for classification of multiple sclerosis lesions. *Ann Clin Transl Neurol*. 2020;7(5):695-706. doi:10.1002/acn3.51037

10. Ibtehaz 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

11. Ramamurthy M, Krishnamurthi I, Vimal S, Robinson YH. Deep learning based genome analysis and NGS-RNA LL identification with a novel hybrid model. *Biosystems*. 2020;197:104211. doi:10.1016/j.biosystems.2020.104211

12. Zhang X, Hu Y, Chen W, Huang G, Nie S. 3D brain glioma segmentation in MRI through integrating multiple densely connected 2D convolutional neural networks. *J Zhejiang Univ Sci B*. 2021;22(6):462-475. doi:10.1631/jzus.B2000381

13. Jung JS, Choi YS, Ahn SS, Yi S, Kim SH, Lee SK. Differentiation between spinal cord diffuse midline glioma with histone H3 K27M mutation and wild type: comparative magnetic resonance imaging. *Neuroradiology*. 2019;61(3):313-322. doi:10.1007/s00234-019-02154-8

14. Ogunlade J, Wiginton JGt, Elia C, Odell T, Rao SC. Primary Spinal Astrocytomas: A Literature Review. *Cureus*. 2019;11(7):e5247. doi:10.7759/cureus.5247

15. Wu J, Armstrong TS, Gilbert MR. Biology and management of ependymomas.

Neuro Oncol. 2016;18(7):902-913. doi:10.1093/neuonc/now016

16. Dauleac C, Messerer R, Obadia-Andre N, Afathi M, Barrey CY. Cysts associated with intramedullary ependymomas of the spinal cord: clinical, MRI and oncological features. *J Neurooncol*. 2019;144(2):385-391. doi:10.1007/s11060-019-03241-9

17. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J Clin*. 2019;69(2):127-157. doi:10.3322/caac.21552

Tables

Table 1. Demographics, clinical information and conventional MR features.

	Training cohort					Validation cohort					Testing cohort				
	Astrocytoma n=94	Ependymoma n=104	MS n=81	NMOSD n=113	P value	Astrocytoma n=24	Ependymoma n=26	MS n=20	NMOSD n=28	P value	Astrocytoma n=34	Ependymoma n=45	MS n=33	NMOSD n=45	P value
Female ratio, No. (%)	35 (37)	48 (46)	59 (73)[12]	94 (83)[12]	<.001[a]	10 (42)	13 (50)	12 (60)	25 (89)[123]	.004[a]	11 (32)	20 (44)	23 (70)[12]	40 (89)[123]	<.001[a]
Age, mean (SD), years	31.54(16.48)	42.66(12.61)[1]	34.93(12.16)[2]	39.17(14.67)[1]	<.001[c]	33.33(14.36)	43.19(13.03)[1]	31.2(10.29)[2]	45.46(12.65)[13]	<.001[c]	31.94(14.27)	42.42(15.28)[1]	38.12(11.91)	42.31(13.35)[1]	.003[c]
Lesion location															
Oblongata-cervical, No. (%)	0 (0)	0 (0)	6 (7)[12]	0 (0)[3]	<.001[b]	0 (0)	0 (0)	2 (10)	0 (0)	.04[b]	0 (0)	0 (0)	4 (12)[12]	0 (0)[3]	.002[b]
Cervical, No. (%)	33 (35)	68 (65)[1]	61 (75)[1]	75 (66)[1]	<.001[a]	13 (54)	16 (62)	13 (65)	17 (61)	.91[a]	10 (29)	29 (64)[1]	24 (73)[1]	29 (64)[1]	.001[a]
Cervical-thoracic, No. (%)	15 (16)	1 (1)[1]	11 (14)[2]	19 (17)[2]	<.001[a]	5 (21)	0 (0)	3 (15)	6 (21)	.049[b]	4 (12)	0 (0)	5 (15)	5 (11)	.97[b]
Thoracic, No. (%)	31 (33)	20 (19)[1]	3 (4)[12]	17 (15)[13]	<.001[a]	5 (21)	1 (4)	2 (10)	5 (18)	.26[b]	15 (44)	4 (9)[1]	0 (0)[1]	11 (24)[123]	<.001[a]
Thoracic-lumbar, No. (%)	14 (15)	0 (0)[1]	0 (0)[1]	2 (2)[1]	<.001[a]	1 (4)	0 (0)	0 (0)	0 (0)	.99[b]	5 (15)	1 (2)[1]	0 (0)[1]	0 (0)[1]	.004[b]
Lumbar, No. (%)	1 (1)	15 (14)[1]	0 (0)[2]	0 (0)[2]	<.001[b]	0 (0)	9 (35)[1]	0 (0)[2]	0 (0)[2]	.008[b]	0 (0)	11 (24)[1]	0 (0)[2]	0 (0)[2]	<.001[b]
Lesion count, median IQR)	1 (1,1)	1 (1,1)	2 (1,3)[12]	1 (1,2)[123]	<.001[d]	1 (1,1)	1 (1,1)	2 (1,3.5)[12]	1 (1,1)[3]	<.001[d]	1 (1,1)	1 (1,1)	2 (1,4)[12]	1 (1,1)[3]	<.001[d]
Lesion-associated extension,	4 (3,6)	3 (2,4)[1]	3 (2,5)[1]	4 (2.5,6)[23]	<.001[d]	4 (3,6)	2 (1.5,3)[1]	3 (2,5)	4 (2,7)[2]	.003[d]	4 (2,6)	2 (2,4)[1]	3 (2,5)	3 (2,5)[2]	.008[d]

median (IQR), vertebra count															
Total lesion volume, mean (SD), ml	12.64(10.77)	15.08(11.71)	1.32(1.64)[12]	2.99(2.95)[12]	<.001[c]	11.61(11.66)	10.55(7.65)	8.04(0.63)[12]	2.73(2.14)[12]	<.001[c]	15.05(12.26)	15.30(10.05)	0.99(1.18)[12]	2.31(2.41)[12]	<.001[c]
Contrast enhanced lesion, NO./total case NO. (%)	60/79 (76)	71/99 (72)	7/54 (13)[12]	10/61 (16)[12]	<.001[a]	18/22 (82)	19/24 (79)	0/15 (0)[12]	4/16 (25)[12]	<.001[a]	24/29 (83)	33/42 (79)	2/21 (10)[12]	6/30 (20)[12]	<.001[a]

Note: MS, multiple sclerosis; NMOSD, neuromyelitis optica spectrum disorders; ANOVA, analysis of variance; n, number; SD, standard deviation; IQR, interquartile range; cT1w, contrast-enhanced T1w; The samples in the training cohort were appropriate for model development (e.g., model overfitting was prevented), as indicated by the model performance evaluation with different training sample sizes (eFigure 3).

[a] Pearson’s Chi-squared test between groups.

[b] Fisher's exact test between groups

[c] ANOVA followed by post hoc multiple comparison with Bonferroni correction.

[d] Kruskal–Wallis test followed by post hoc multiple comparison with Bonferroni correction.

[1] Statistically significant compared to astrocytoma.

[2] Statistically significant compared to ependymoma.

Table 2. The differentiation of spinal cord lesions by DL models.

Classification	**Tumor vs. Demyelinating lesion (Model 1)**	**Astrocytoma vs. Ependymoma (Model 2)**	**MS vs. NMOSD (Model 3)**
Validation			
Accuracy (%)	96 (94/98)	80 (40/50)	88 (42/48)
Sensitivity (%)	98 (47/48)	77 (20/26)	86 (24/28)
Specificity (%)	94 (47/50)	83 (20/24)	90 (18/20)
PPV (%)	94 (47/50)	83 (20/24)	92 (24/26)
NPV (%)	98 (47/48)	77 (20/26)	82 (18/22)
Precision (%)	94 (47/50)	83 (20/24)	92 (24/26)
Recall (%)	98 (47/48)	77 (20/26)	86 (24/28)
AUC	0.99	0.85	0.94
Testing			
Accuracy (%)	96 (150/157)	82 (65/79)	79 (62/78)
Sensitivity (%)	97 (76/78)	76 (34/45)	80 (36/45)
Specificity (%)	94 (74/79)	91 (31/34)	79 (26/33)
PPV (%)	94 (76/81)	92 (34/37)	84 (36/43)
NPV (%)	97 (74/76)	74 (31/42)	74 (26/35)
Precision (%)	94 (76/81)	92 (34/37)	84 (36/43)
Recall (%)	97 (76/78)	76 (34/45)	80 (36/45)
AUC	0.99	0.90	0.85
Difficult cases			
Accuracy (%)	95 (38/40)	83 (15/18)	82 (18/22)
Sensitivity (%)	95 (21/22)	86 (6/7)	87 (13/15)
Specificity (%)	94 (17/18)	82 (9/11)	71 (5/7)
PPV (%)	95 (21/22)	75 (6/8)	87 (13/15)
NPV (%)	94 (17/18)	90 (9/10)	71 (5/7)
Precision (%)	95 (21/22)	75 (6/8)	87 (13/15)
Recall (%)	95 (21/22)	86 (6/7)	87 (13/15)
AUC	0.97	0.91	0.78

Note: MS, multiple sclerosis; NMOSD, neuromyelitis optica spectrum disorders; DL, deep learning; PPV, positive predictive value; NPV; negative predictive value; AUC, area under the curve.

Figure legends

Figure 1. A flowchart of the patient selection. MS, multiple sclerosis; NMOSD, neuromyelitis optica spectrum disorders.

Figure 2. A: A DL pipeline established for segmentation and classification of spinal cord lesions. First, the T2w images (all slices, e.g., slices 1-11) were used to segment the lesion, and a manual interaction was conducted to correct the poorly segmented lesions. Then, the slices (e.g., slices 5-8) of T2w images and lesion masks involving lesions were used as the network input for classification tasks. B: Representative cases of segmentation and classification for spinal cord tumors and demyelinating lesions in the test cohort. Diagnosis by DL and four raters (D. C, X. X, C.F and X.H) are also presented. Red areas indicate the deep learning segmentation. DL, deep learning; MS, multiple sclerosis; NMOSD, neuromyelitis optica spectrum disorders.