Deep learning-based detection of lumbar spinal canal stenosis using convolutional neural networks

ABSTRACT[1]

BACKGROUND CONTEXT: Lumbar spinal canal stenosis (LSCS) is the most common spinal degenerative disorder in elderly people and usually first seen by primary care physicians or orthopedic surgeons who are not spine surgery specialists. Magnetic resonance imaging (MRI) is useful in the diagnosis of LSCS, but the equipment is often not available or difficult to read. LSCS patients with progressive neurologic deficits have difficulty with recovery if surgical treatment is delayed. So, early diagnosis and determination of appropriate surgical indications are crucial in the treatment of LSCS. Convolutional neural networks (CNNs), a type of deep learning, offers significant advantages for image recognition and classification, and work well with radiographs, which can be easily taken at any facility.

PURPOSE: Our purpose was to develop an algorithm to diagnose the presence or absence of LSCS requiring surgery from plain radiographs using CNNs.

STUDY DESIGN: Retrospective analysis of consecutive, nonrandomized series of patients at a single institution.

PATIENT SAMPLE: Data of 150 patients who underwent surgery for LSCS, including degenerative spondylolisthesis, at a single institution from January 2022 to August 2022, were

[1]Abbreviations

AUC, area under the curve; CNN, convolutional neural network; LSCS, lumbar spinal canal stenosis; MRI, magnetic resonance imaging; NLR, negative likelihood ratio; NPV, negative predictive value; PLR, positive likelihood ratio; PNG, portable network graphics; PPV, positive predictive value; ROC, receiver operating characteristic; ROI, region of interest

collected. Additionally, 25 patients who underwent surgery at two other hospitals were included for extra external validation.

OUTCOME MEASURES: In annotation 1, the area under the curve (AUC) computed from the receiver operating characteristic (ROC) curve, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, positive likelihood ratio (PLR), and negative likelihood ratio (NLR) were calculated. In annotation 2, correlation coefficients were used.

METHODS: Four intervertebral levels from L1/2 to L4/5 were extracted as region of interest from lateral plain lumbar spine radiographs totaling 600 images were obtained. Based on the date of surgery, 500 images derived from the first 125 cases were used for internal validation, and 100 images from the subsequent 25 cases used for external validation. Additionally, 100 images from other hospitals were used for extra external validation. In annotation 1, binary classification of operative and nonoperative levels was used, and in annotation 2, the spinal canal area measured on axial MRI was labeled as the output layer. For internal validation, the 500 images were divided into each 5 dataset on per-patient basis and five-fold cross-validation was performed. Five trained models were registered in the external validation prediction performance. Grad-CAM was used to visualize area with the high features extracted by CNNs.

RESULTS: In internal validation, the AUC and accuracy for annotation 1 ranged between 0.85–0.89 and 79–83%, respectively, and the correlation coefficients for annotation 2 ranged between 0.53–0.64 (all $P < 0.01$). In external validation, the AUC and accuracy for annotation 1 were 0.90 and 82%, respectively, and the correlation coefficient for annotation 2 was 0.69, using five trained CNN models. In the extra external validation, the AUC and accuracy for annotation 1 were 0.89 and 84%, respectively, and the correlation coefficient for annotation 2 was 0.56. Grad-CAM showed high feature density in the intervertebral joints and posterior intervertebral

discs.

CONCLUSIONS: This technology automatically detects LSCS from plain lumbar spine radiographs, making it possible for medical facilities without MRI or non-specialists to diagnose LSCS, suggesting the possibility of eliminating delays in the diagnosis and treatment of LSCS that require early treatment.

KEY WORDS: Lumbar spinal canal stenosis; Deep learning; Convolutional neural network; Artificial intelligence; Plain lumbar spine radiograph; Magnetic Resonance Imaging.

INTRODUCTION

Lumbar spinal canal stenosis (LSCS) stands as the predominant spinal degenerative ailment among the elderly [1-5]. This condition induces nerve paralysis and impedes ambulation, posing a significant challenge to independent and healthy living. Given the escalating incidence of LSCS in an aging society, overcoming this ailment assumes paramount importance. Surgical interventions, such as spinal decompression and fusion, offer tangible benefits for treating LSCS patients, with a heightened prevalence of surgery observed in those aged 65 and older [1,6]. Nonetheless, individuals with progressive neurologic deficits stemming from LSCS encounter hurdles in recovery if surgical intervention is delayed. Hence, early diagnosis and the judicious determination of appropriate surgical indications emerge as pivotal aspects in the therapeutic approach to LSCS.

Typically, patients with LSCS initially consult primary care physicians or orthopedic surgeons lacking specialization in spine surgery for diagnosis and conservative treatment. A salient

challenge in managing LSCS lies in the intricate task of identifying the optimal moment to refer patients necessitating early surgery to a spine surgeon. Magnetic resonance imaging (MRI), characterized by low invasiveness and absence of radiation exposure, serves as the gold standard for diagnosing LSCS [7,8]. A spinal canal area of less than 50 mm^2 on MRI is resistant to conservative treatment and requires early surgical treatment [9]. Nevertheless, discerning whether an MRI scan is warranted and interpreting LSCS findings on MRI for surgical necessity remains challenging for non-specialists. Moreover, the absence of MRI equipment in numerous primary care medical facilities, particularly in rural areas, compounds the challenge of timely MRI scans. Against this backdrop, the prospect of automating the determination of lumbar spine disorders requiring meticulous management from plain radiographic images assumes significance, given their accessibility across most medical facilities.

Deep learning, a subset of machine learning adept at discerning intricate patterns in extensive datasets [10], becomes particularly pertinent in this context. Convolutional neural networks (CNNs), with their focus on learning features through convolution layers, pooling layers, and fully-connected layers [11,12], offer significant advantages in image recognition and classification [11]. A meta-analysis assessing the diagnostic accuracy of deep learning in medical imaging has underscored the capacity of deep learning algorithms to identify disease with clinically acceptable precision [13]. While CNNs for detecting cervical spondylotic myelopathy from plain radiographs of the cervical spine have been reported [14-16], scant literature exists on CNNs for detecting LSCS from plain radiographs of the lumbar spine.

Our hypothesis posits that deep learning can discern the diagnosis of spinal canal stenosis from plain lumbar spine radiographs. In this study, our objective is to develop algorithms employing CNNs to diagnose the presence or absence of LSCS necessitating surgery from plain lumbar

spine radiographs. The development of an algorithm capable of automatically detecting LSCS from lumbar plain radiographs holds the promise of enabling diagnosis in medical facilities lacking MRI, potentially mitigating delays in the diagnosis and treatment of LSCS requiring early intervention.

MATERIALS AND METHODS

Subjects

We conducted a retrospective collection of image data and corresponding medical records with approval from the institutional review boards of our hospitals. Our study focused on a cohort of 150 patients who underwent surgery for LSCS, including cases of degenerative spondylolisthesis, between January 2022 and August 2022 at a single hospital, as well as 25 patients who underwent surgery at two other hospitals. The exclusion criteria comprised (i) presence of vertebral fractures or lumbar spondylolysis spanning from L1 to L5; (ii) spinal bone metastasis and infection; (iii) degenerative scoliosis with a Cobb angle of 15^{o} or greater and malformations such as transitional lumbosacral anomalies; (iv) absence of lumbar MRI due to a cardiac pacemaker device. Based on the surgical dates, the cohort of 150 cases was stratified into 125 cases from January 2022 to June 2022 for internal validation and 25 cases from July to August 2022 for external validation, aligning with the chronological sequence of medical visits. Additionally, 25 cases from other hospitals were used for extra external validation. During the internal validation phase, CNN models were trained using image data to predict the presence of LSCS. Subsequently, in the external validation phase, the performance of these trained CNN models was evaluated using independent data derived from the internal validation set.

Input image data

The region of interest (ROI) in this study comprised four intervertebral levels spanning from L1/2 to L4/5, which were extracted from lateral radiographs images of the lumbar spine based on the delineation outlined in **Fig. 1a, b**. Initially, each intervertebral level underwent rotation to align parallel to the intervertebral disc. The cranio-caudal segmental region was defined by the lower margin of the upper and lower pedicles, respectively. Meanwhile, the ventrodorsal segmental region was demarcated by the posterior margin of the facet joint and the anterior margin of the intervertebral disc. In total, 500 images were acquired for internal validation, and an additional 100 images were reserved for external validation. Additionally, 100 images from other hospitals were reserved for extra external validation. Because the radiographic images from the other hospitals were taken in a different left–right orientation compared with those from ours, the left and right sides of the images were reversed when extracting the ROI. Subsequently, these images were converted into portable network graphics (PNG) files with dimensions of 224 × 224 pixels, employing zero padding for standardization.

Labeling information for each image

Each image underwent annotation for two types of information: (1) binary classification designating operative and nonoperative levels; (2) the spinal canal area measured on axial MRI images (**Fig. 1c**). In annotation 1, the CNN classified operative levels, encompassing spinal decompression and fusion, as 1, and nonoperative levels as 0, predicting the likelihood of an operative level. For annotation 2, the CNN learned the spinal canal area as a continuous variable. For internal validation, a five-fold cross-validation strategy was employed to mitigate selection

bias and assess the CNN model's generalizability. Based on the surgical day, the 500 images were randomly divided into five datasets, each containing 100 images (datasets 1, 2, 3, 4, and 5) to ensure the inclusion of all four images on a per-patient basis. One dataset served as the validation set, with the remaining four datasets combined as the training set. This validation process was repeated five times by exchanging the datasets, with each serving as the validation set once. Five trained models predicted the likelihood of operative levels and the spinal canal area in the external validation dataset. The prediction performance was evaluated using the average values derived from these five models. Subsequently, multitask learning was performed to confirm that those annotations had common features related with LSCS [17,18]. Finally, the external validation dataset (200 images) and the internal validation dataset (500 images) underwent learning to verify the absence of significant deviations compared with the prediction performance observed in the external validation (internal dataset-external dataset validation).

Convolutional neural network models

The construction of CNN models was facilitated using the neural network console (Sony Network Communications, Tokyo, Japan), employing the architecture outlined in **Fig. 2**. The console featured the "ImageAugmentation" layer, enhancing the robustness of the training network through image processing on training datasets at each epoch. Random variations were introduced to each image during training, involving rotation within the range of -0.1 to 0.1 radians, along with cropping after zooming within folds of 0.8 to 1.2.

For annotation 1 and 2, the architecture comprised three convolution layers, three MaxPooling layers, two DepthwiseConvolution layers, and two AveragePooling layers, following the ImageAugmentation layer. The output layer, employing BinaryCrossEntropy in annotation 1 and

Huber loss in annotation 2, was selected after the fully-connected layers. Multi-task learning had an architecture with a common intermediate layer between annotations 1 and 2 and their respective output layers. A batch size of 64 and the Adam optimizer [19] were configured, with training spanning 2000 epochs due to the slight alterations introduced by the ImageAugmentation layer to the training datasets at each epoch.

To visualize areas with high features extracted by the last convolutional layer of CNNs, Grad-CAM was employed. The computational setup utilized for image learning and output data involved a central processing unit of Core i9-10920X (Intel), a graphics processing unit of Quadro RTX 4000 (NVIDIA), and a random-access memory of 128 GB.

Statistical Analysis

In annotation 1, the computation of the area under the curve (AUC) was derived from the receiver operating characteristic (ROC) curve. Employing a cut-off value of 0.5 with respect to the possibility for an operative level, various indicators for prediction performance, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, positive likelihood ratio (PLR), and negative likelihood ratio (NLR), were calculated. For annotation 2, the relationship between the actual spinal canal area and the predicted spinal canal area was assessed using Pearson's correlation coefficient.

Data analysis was conducted using JMP statistical software for Windows (version 14; SAS, Inc., Cary, NC, USA), and a significance level of $P < 0.05$ was considered statistically significant.

RESULTS

The cohort of 175 subjects comprised 65 females and 110 males, with a mean age of 68 ± 11

years. Annotation 1 encompassed 278 operative levels (40%) and 422 nonoperative levels (60%). Demographic and measurement data of each dataset are shown in **Table 1**.

Internal validation

In the context of annotation 1, the range of the AUC and accuracy during the five-fold cross-validation spanned from 0.85–0.89 and 79–83%, respectively (**Fig. 3** and **Table 2**). Across all five datasets, the performance metrics were comparable, with an AUC of 0.87, sensitivity of 0.74, specificity of 0.86, PPV of 0.80, NPV of 0.81, accuracy of 81%, PLR of 5.39, and NLR of 0.30 (**Fig. 3** and **Table 2**). For annotation 2, the five-fold cross-validation revealed correlation coefficients ranging from 0.53–0.64 (All $P < 0.01$) (**Fig. 4**), and the correlation coefficient across all datasets was 0.58 (**Fig. 4**).

External validation

Utilizing five CNN models for annotation 1 and 2, predictions were made for the possibility of an operative level and the spinal canal area in the external validation dataset consisting of 100 images. In annotation 1, the range of the AUC and accuracy across the five trained CNN models were 0.88–0.90 and 78–83%, respectively. The average values for the possibility indicated an AUC of 0.90 and an accuracy of 82% (**Fig. 5** and **Table 3**). For annotation 2, the correlation coefficient ranged from 0.54–0.71 across the five trained CNN models, and the average value of the area rate demonstrated a correlation coefficient of 0.69 (**Fig. 6**). Grad-CAM visualization highlighted high feature density in the intervertebral joints and posterior intervertebral discs (**Fig. 7**). Additionally, extra external validations were conducted with 100 images from other two hospitals. In annotation 1, the range of the AUC and accuracy across the five trained CNN

models were 0.84–0.91 and 76–88%, respectively. The average values for the possibility indicated an AUC of 0.89 and an accuracy of 84% (**Fig. 8** and **Table 4**). For annotation 2, the correlation coefficient ranged from 0.39–0.63 across the five trained CNN models, and the average value of the area rate demonstrated a correlation coefficient of 0.56 (**Fig. 9**).

Validation by Multitask learning

The results for internal validation were as follows. For binary classification designating operative and nonoperative levels, the range of the AUC and accuracy during the five-fold cross-validation spanned from 0.86–0.88 and 76–84%, respectively (**Fig. 10** and **Table 5**). Across all five datasets, the performance metrics were comparable, with an AUC of 0.86, sensitivity of 0.77, specificity of 0.82, PPV of 0.77, NPV of 0.82, accuracy of 80%, PLR of 4.28, and NLR of 0.28 (**Fig. 10** and **Table 5**). The correlation coefficients for predicting the spinal canal area ranged from 0.57–0.67 (All $P < 0.01$) in the five-fold cross-validation, with an overall correlation coefficient of 0.60 across all datasets (**Fig. 11**). External validation of the trained CNN models for annotation 1 showed an AUC range of 0.86–0.89 and an accuracy range of 77–79%. The average values indicated an AUC of 0.88 and accuracy of 79% (**Fig. 12** and **Table 6**). For annotation 2, the correlation coefficient ranged from 0.67–0.70 across the five trained CNN models, with an average correlation coefficient of 0.71 (**Fig. 13**). Extra external validation for annotation 1 showed an AUC range of 0.85–0.91 and an accuracy range between 77–87% across the five trained CNN models. The average values indicated an AUC of 0.90 and accuracy of 84% (**Fig. 14** and **Table 7**). For annotation 2, the correlation coefficient ranged from 0.49–0.63 across the five trained CNN models, with an average correlation coefficient of 0.58 (**Fig. 15**).

Internal-external cross-validation

During the learning process, with the external validation dataset serving as the validation set and the internal validation dataset as the training set, the performance metrics for annotation 1 were as follows: AUC of 0.89, sensitivity of 0.78, specificity of 0.89, PPV of 0.77, NPV of 0.90, accuracy of 86%, PLR of 7.10, and NLR of 0.25 (**Fig. 16a**). For annotation 2, the correlation coefficient was 0.60 (**Fig. 16b**).

DISCUSSION

In this study, we aimed to assess the capability of deep learning algorithms utilizing CNNs in detecting LSCS from lumbar spine radiographs. In the external dataset, the average probability generated by five models yielded an AUC of 0.90 and an accuracy of 82%. Furthermore, the correlation coefficient between the actual area at the disc level and the predicted area using the models developed in the internal validation was 0.69, indicating a robust relationship. These results suggest that deep learning algorithms hold promise in predicting the presence of LSCS from plain lumbar radiographs. While the etiologies of LSCS have been associated with congenital factors, such as short pedicle length, and degenerative changes in the intervertebral disc, ligamentum flavum, and facet joints [20-23], lumbar spine radiographs may contain features related to these congenital and degenerative factors that surpass the human eye's discernment. Indeed, Grad-CAM analysis revealed high feature density in the inter-vertebral joints and posterior intervertebral discs. Furthermore, the multitask learning results were typically at the same level for the external validation compared to the validation with annotations 1 and 2 separated. These results may confirm that the two items, the presence of the operative level and spinal canal area, had common features to diagnose LSCS. In the extra external

validation using data from the two other hospitals, the AUC, an accuracy and correlation were 0.89, 84% and 0.56, respectively for annotation 1 and 2, and 0.90, 84% and 0.58, respectively, for multitask learning. These results indicate that the models have generalizability and applicability in varied clinical settings, despite slightly lower correlations compared to external validation.

The existing literature includes only one article on a CNN-based transfer learning algorithm designed to detect LSCS from lumbar radiographs using binary classification [24]. This study reported a predicted performance with an AUC of 0.89 and accuracy of 81.8% in external validation. Despite our study having a smaller sample size than the aforementioned research (500 images versus 12,442 images in internal validation), the performance is comparable, demonstrating an AUC of 0.90 and accuracy of 82% in external validation using dataset of the same facility and an AUC of 0.89 and accuracy of 84% in extra external validation using dataset of two other hospitals. This similarity in performance suggests that the ROI extraction in our study, focusing on high-feature areas such as intervertebral joints and posterior intervertebral discs while excluding low-feature areas, contributes to improved learning for deep learning algorithms with CNNs. Feature selection is a valuable technique in deep learning to enhance performance by reducing model complexity. Eliminating tautological and irrelevant features reduces input dimensionality and aids in training the fundamental mechanism that connects effective features [25]. Focusing on the minimum necessary ROI in our study likely achieved a similar effect to feature selection. In the future, we plan to develop an algorithm for extracting the ROI, as illustrated in **Fig. 1**, utilizing semantic segmentation with deep learning.

In the internal-external cross-validation of the current study, the AUC and accuracy reached 0.90 and 83%, respectively, for annotation 1, while the correlation coefficient was 0.73 for annotation

2. These results indicate that the performance in the internal dataset-external dataset validation is comparable to the average possibility observed in internal validation. This suggests that the external validation dataset exhibited a level of dissimilarity with the internal validation dataset, and the trained models did not achieve high performance merely by coincidentally fitting the external validation dataset, thereby mitigating concerns of overfitting.

When the external dataset closely resembles the developed models, the assessment primarily serves for reproducibility rather than transportability [26]. Ensuring the similarity of the validation to the development sets is crucial for the interpretation of an external validation study [27]. While external validation at the time of model development is typically recommended, testing the heterogeneity of predictor effects directly may be more informative than withholding parts of the data using internal-external validation procedures [27]. Such assessments during model development can aid in tempering the overoptimistic expectations of prediction model performance in independent data [27].

This study has several limitations that warrant consideration. Firstly, the exclusion of the L5/S level could potentially introduce bias, as this level often lacks stenosis and exhibits a different shape compared with other intervertebral levels [28, 29]. The models' performance might be overestimated by including L5/S, which could be assessed without stenosis based on vertebral geometry. Secondly, in annotation 1, the determination of operative and nonoperative levels was subjective, relying on surgeons from a single hospital. The inclusion of intervertebral levels with mild and moderate stenosis as surgical indications is at the discretion of the surgeon and may vary. Thirdly, the ROI extraction for training images using CNNs was designed for simplicity; however, for broader societal implementation, an alternative algorithm employing semantic segmentation may be necessary to extract the ROI more accurately. Lastly, Lastly, although we

evaluated prediction performance using data from two other hospitals, it is uncertain whether these models will perform similarly with data from other hospitals. Therefore, further validation using data from multiple institutions is necessary. However, the models can be adapted to each radiographic device by retraining with small datasets using techniques such as fine-tuning. Despite these limitations, we envision that the deep learning algorithms developed in this study could find practical application in clinics lacking MRI facilities and in rural areas with limited access to orthopedic surgeons.

CONCLUSION

We have successfully developed deep learning algorithms capable of diagnosing LSCS based on plain radiographs of the lumbar spine. Leveraging CNN, this algorithm demonstrates a notably high diagnostic accuracy for LSCS. The intended application of this technology is to aid non-specialists and non-orthopedic surgeons in effectively distinguishing cases of LSCS. By automatically detecting LSCS from plain lumbar spine radiographs, this algorithm holds the potential to facilitate diagnosis in medical facilities without access to MRI or lacking specialist expertise. The implementation of this technology raises the prospect of mitigating delays in the diagnosis and treatment of LSCS, particularly in cases requiring early intervention.

REFERENCES

1. Djurasovic, M.; Glassman, S. D.; Carreon, L. Y.; Dimar, J. R. Contemporary management of symptomatic lumbar spinal stenosis. Orthop Clin North Am 2010, 41 (2), 183-191. DOI: 10.1016/j.ocl.2009.12.003.

2. Chen, M.; Carass, A.; Oh, J.; Nair, G.; Pham, D. L.; Reich, D. S.; Prince, J. L. Automatic magnetic resonance spinal cord segmentation with topology constraints for variable fields of view. Neuroimage 2013, 83, 1051-1062. DOI: 10.1016/j.neuroimage.2013.07.060.
3. Laiwalla, A. N.; Ratnaparkhi, A.; Zarrin, D.; Cook, K.; Li, I.; Wilson, B.; Florence, T. J.; Yoo, B.; Salehi, B.; Gaonkar, B.; et al. Lumbar Spinal Canal Segmentation in Cases with Lumbar Stenosis Using Deep-U-Net Ensembles. World Neurosurg 2023, 178, e135-e140. DOI: 10.1016/j.wneu.2023.07.009.
4. Buckwalter, J. A. Aging and degeneration of the human intervertebral disc. Spine (Phila Pa 1976) 1995, 20 (11), 1307-1314. DOI: 10.1097/00007632-199506000-00022.
5. Verbiest, H. Pathomorphologic aspects of developmental lumbar stenosis. Orthop Clin North Am 1975, 6 (1), 177-196.
6. Weinstein, J. N.; Lurie, J. D.; Tosteson, T. D.; Hanscom, B.; Tosteson, A. N.; Blood, E. A.; Birkmeyer, N. J.; Hilibrand, A. S.; Herkowitz, H.; Cammisa, F. P.; et al. Surgical versus nonsurgical treatment for lumbar degenerative spondylolisthesis. N Engl J Med 2007, 356 (22), 2257-2270. DOI: 10.1056/NEJMoa070302.
7. de Schepper, E. I.; Overdevest, G. M.; Suri, P.; Peul, W. C.; Oei, E. H.; Koes, B. W.; Bierma-Zeinstra, S. M.; Luijsterburg, P. A. Diagnosis of lumbar spinal stenosis: an updated systematic review of the accuracy of diagnostic tests. Spine (Phila Pa 1976) 2013, 38 (8), E469-481. DOI:

10.1097/BRS.0b013e31828935ac.

8. McDonald, R. J.; Schwartz, K. M.; Eckel, L. J.; Diehn, F. E.; Hunt, C. H.; Bartholmai, B. J.; Erickson, B. J.; Kallmes, D. F. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 2015, 22 (9), 1191-1198. DOI: 10.1016/j.acra.2015.05.007.
9. Minamide A, Yoshida M, Maio K. The natural clinical course of lumbar spinal stenosis: a longitudinal cohort study over a minimum of 10 years. J Orthop Sci. 2013 Sep;18(5):693-8. doi: 10.1007/s00776-013-0435-9.
10. Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A primer on deep learning in genomics. Nat Genet 2019, 51 (1), 12-18. DOI: 10.1038/s41588-018-0295-5.
11. Alex K.; Ilya S. and Geoffrey E. H. ImageNet classification with deep convolutional neural networks. Commun. ACM 2017, 60, 6, 84–90. DOI: 10.1145/3065386
12. Yamashita, R.; Nishio, M.; Do, R. K. G.; Togashi, K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018, 9 (4), 611-629. DOI: 10.1007/s13244-018-0639-9.
13. Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D. S. W.; Karthikesalingam, A.; King, D.; Ashrafian, H.; Darzi, A. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021, 4 (1), 65. DOI: 10.1038/s41746-021-00438-

z.

14. Maki, S.; Furuya, T.; Horikoshi, T.; Yokota, H.; Mori, Y.; Ota, J.; Kawasaki, Y.; Miyamoto, T.; Norimoto, M.; Okimatsu, S.; et al. A Deep Convolutional Neural Network With Performance Comparable to Radiologists for Differentiating Between Spinal Schwannoma and Meningioma. Spine (Phila Pa 1976) 2020, 45 (10), 694-700. DOI: 10.1097/BRS.0000000000003353.
15. Lee, G. W.; Shin, H.; Chang, M. C. Deep learning algorithm to evaluate cervical spondylotic myelopathy using lateral cervical spine radiograph. BMC Neurol 2022, 22 (1), 147. DOI: 10.1186/s12883-022-02670-w.
16. Tamai, K.; Terai, H.; Hoshino, M.; Tabuchi, H.; Kato, M.; Toyoda, H.; Suzuki, A.; Takahashi, S.; Yabu, A.; Sawada, Y.; et al. Deep Learning Algorithm for Identifying Cervical Cord Compression Due to Degenerative Canal Stenosis on Radiography. Spine (Phila Pa 1976) 2023, 48 (8), 519-525. DOI: 10.1097/BRS.0000000000004595.
17. Ye Z, Yang Y, Li X, Cao D, Ouyang D. An Integrated Transfer Learning and Multitask Learning Approach for Pharmacokinetic Parameter Prediction. Mol Pharm. 2019 Feb 4;16(2):533-541. Doi: 10.1021/acs.molpharmaceut.8b00816.
18. Yang M, Huang W, Tu W, Qu Q, Shen Y, Lei K. Multitask Learning and Reinforcement Learning for Personalized Dialog Generation: An Empirical Study. IEEE Trans Neural Netw Learn Syst. 2021 Jan;32(1):49-62. doi: 10.1109/TNNLS.2020.2975035.

19. Cornell University Library. Adam: a method for stochastic optimization, https://arxiv.org/pdf/1412.6980.pdf. [accessed 13 June 2020].
20. Saifuddin, A. The imaging of lumbar spinal stenosis. Clin Radiol 2000, 55 (8), 581-594. DOI: 10.1053/crad.2000.0223.
21. Gibson, J. N.; Waddell, G. Surgery for degenerative lumbar spondylosis. Cochrane Database Syst Rev 2005, 2005 (4), CD001352. DOI: 10.1002/14651858.CD001352.pub3.
22. Melancia, J. L.; Francisco, A. F.; Antunes, J. L. Spinal stenosis. Handb Clin Neurol 2014, 119, 541-549. DOI: 10.1016/B978-0-7020-4086-3.00035-7.
23. Akar, E.; Somay, H. Comparative morphometric analysis of congenital and acquired lumbar spinal stenosis. J Clin Neurosci 2019, 68, 256-261. DOI: 10.1016/j.jocn.2019.07.015.
24. Kim, T.; Kim, Y. G.; Park, S.; Lee, J. K.; Lee, C. H.; Hyun, S. J.; Kim, C. H.; Kim, K. J.; Chung, C. K. Diagnostic triage in patients with central lumbar spinal stenosis using a deep learning system of radiographs. J Neurosurg Spine 2022, 1-8. DOI: 10.3171/2021.11.SPINE211136.
25. Lee, S. H.; Han, P.; Hales, R. K.; Voong, K. R.; Noro, K.; Sugiyama, S.; Haller, J. W.; McNutt, T. R.; Lee, J. Multi-view radiomics and dosiomics analysis with machine learning for predicting acute-phase weight loss in lung cancer patients treated with radiotherapy. Phys Med Biol 2020, 65 (19), 195015. DOI: 10.1088/1361-6560/ab8531.
26. Justice, A. C.; Covinsky, K. E.; Berlin, J. A. Assessing the generalizability of prognostic

information. Ann Intern Med 1999, 130 (6), 515-524. DOI: 10.7326/0003-4819-130-6-199903160-00016.

27. Steyerberg, E. W.; Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016, 69, 245-247. DOI: 10.1016/j.jclinepi.2015.04.005.

28. Melcher, C.; Paulus, A. C.; Roßbach, B. P.; Gülecyüz,, M. F.; Birkenmaier, C.; Schulze-Pellengahr, C. V.; Teske, W.; Wegener, B. Lumbar spinal stenosis - surgical outcome and the odds of revision-surgery: Is it all due to the surgeon? Technol Health Care 2022, 30 (6), 1423-1434. DOI: 10.3233/THC-223389.

29. Park, D. K.; An, H. S.; Lurie, J. D.; Zhao, W.; Tosteson, A.; Tosteson, T. D.; Herkowitz, H.; Errico, T.; Weinstein, J. N. Does multilevel lumbar stenosis lead to poorer outcomes?: a subanalysis of the Spine Patient Outcomes Research Trial (SPORT) lumbar stenosis study. Spine (Phila Pa 1976) 2010, 35 (4), 439-446. DOI: 10.1097/BRS.0b013e3181bdafb9.

FIGURE LEGENDS

Figure 1. Representative images of radiographs and MRI. (a) Each intervertebral disc level is extracted as a region of interest from the lumbar lateral radiographs. (b) The image is adjusted,

such that the intervertebral disc is horizontal. The cranio-caudal and ventral-dorsal segment regions are the lower margins of the upper and lower pedicles, the posterior margin of the facet joint, and the anterior margin of the intervertebral disc, respectively. (c) Spinal canal area is measured along the line of the dura mater in the MRI axial images.

Figure 2. Schematic of the convolutional neural network architecture utilized in the proposed framework for detecting lumbar spinal canal stenosis. ReLU = rectified linear unit.

Figure 3. Comparison of the area under the curve (AUC) of the five-fold cross-validation (A–E) and all data sets (F) of annotation 1. AUC ranges from 0.85 to 0.89.

Figure 4. Comparison of correlation coefficients of the five-fold cross-validation (A-E) and all data sets (F) of annotation 2. Correlation coefficients range from 0.53 to 0.64 (All $P < 0.01$).

Figure 5. Comparison of the area under the curve (AUC) for external validation using five convolutional neural network models (A–E), and the average value of possibility (F) in annotation 1. AUC ranges from 0.88 to 0.90. The average value of possibility shows an AUC of 0.90.

Figure 6. Comparison of correlation coefficients for the external validation using five convolutional neural network models (A–E), and the average value of the area (F) in annotation 2. Correlation coefficient ranges from 0.54 to 0.71. Average value of the area demonstrates a correlation coefficient of 0.69.

Figure 7. Examples of Grad-CAM showed high feature density in the intervertebral joints and posterior intervertebral discs in both annotation 1 and 2.

Figure 8. Comparison of the area under the curve (AUC) for extra external validation using five convolutional neural network models (A–E), and the average value of possibility (F) in annotation 1. AUC ranges from 0.84 to 0.91. The average value of possibility shows an AUC of 0.89.

Figure 9. Comparison of correlation coefficients for extra external validation using five convolutional neural network models (A–E), and the average value of the area (F) in annotation 2. Correlation coefficient ranges from 0.39 to 0.63. Average value of the area demonstrates a correlation coefficient of 0.56.

Figure 10. Comparison of the area under the curve (AUC) of internal validation by multitask learning for annotation 1. Among the five-fold cross-validation (A–E) and all data sets (F), AUC ranges from 0.86 to 0.88.

Figure 11. Comparison of correlation coefficients of internal validation by multitask learning for annotation 2. Among the five-fold cross-validation (A-E) and all data sets (F), correlation coefficients range from 0.57 to 0.67 (All $P < 0.01$).

Figure 12. Comparison of the area under the curve (AUC) of external validation for

annotation 1 using five convolutional neural network models (A–E) produced by multitask learning, and the average value of possibility (F). AUC ranges from 0.86 to 0.89. The average value of possibility shows an AUC of 0.88.

Figure 13. Comparison of correlation coefficients of external validation for annotation 2 using five convolutional neural network models (A–E) produced by multitask learning, and the average value of the area (F). Correlation coefficient ranges from 0.67 to 0.70. Average value of the area rate demonstrates a correlation coefficient of 0.71.

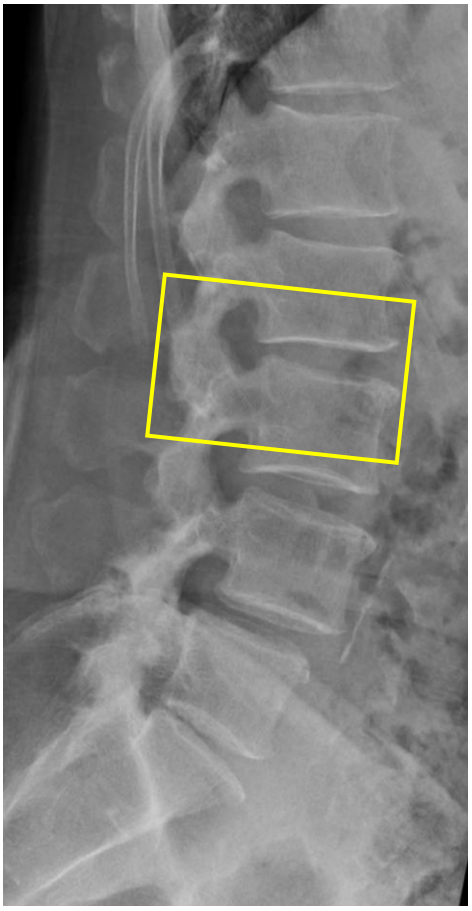
Figure 14. Comparison of the area under the curve (AUC) of extra external validation for annotation 1 using five convolutional neural network models (A–E) produced by multitask learning, and the average value of possibility (F). AUC ranges from 0.85 to 0.91. The average value of possibility shows an AUC of 0.90.

Figure 15. Comparison of correlation coefficients of extra external validation for annotation 2 using five convolutional neural network models (A–E) produced by multitask learning, and the average value of the area (F). Correlation coefficient ranges from 0.49 to 0.63. Average value of the area rate demonstrates a correlation coefficient of 0.58.

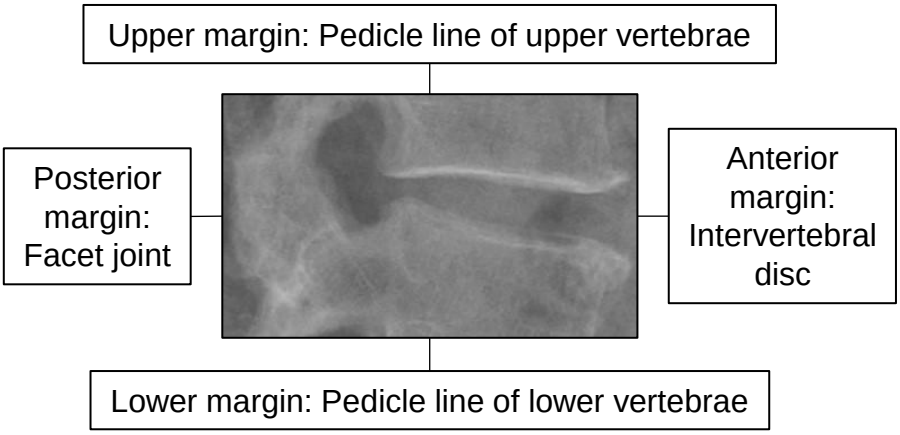
Figure 16. Results of internal-external-cross validation. The area under the curve of annotation 1 is 0.89 (a) and correlation coefficient of annotation 2 is 0.60 (b).

Figure 1.

a



b


Upper margin: Pedicle line of upper vertebrae
Posterior margin: Facet joint
Anterior margin: Intervertebral disc
Lower margin: Pedicle line of lower vertebrae

c

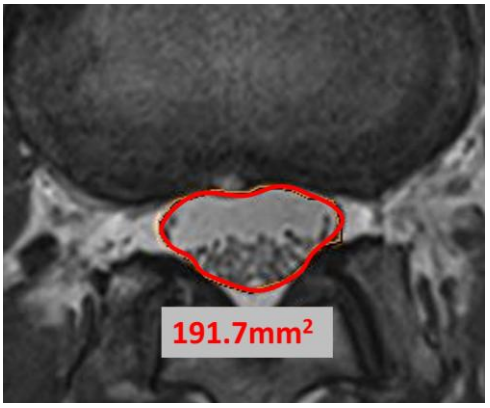

191.7mm²

Figure 2.

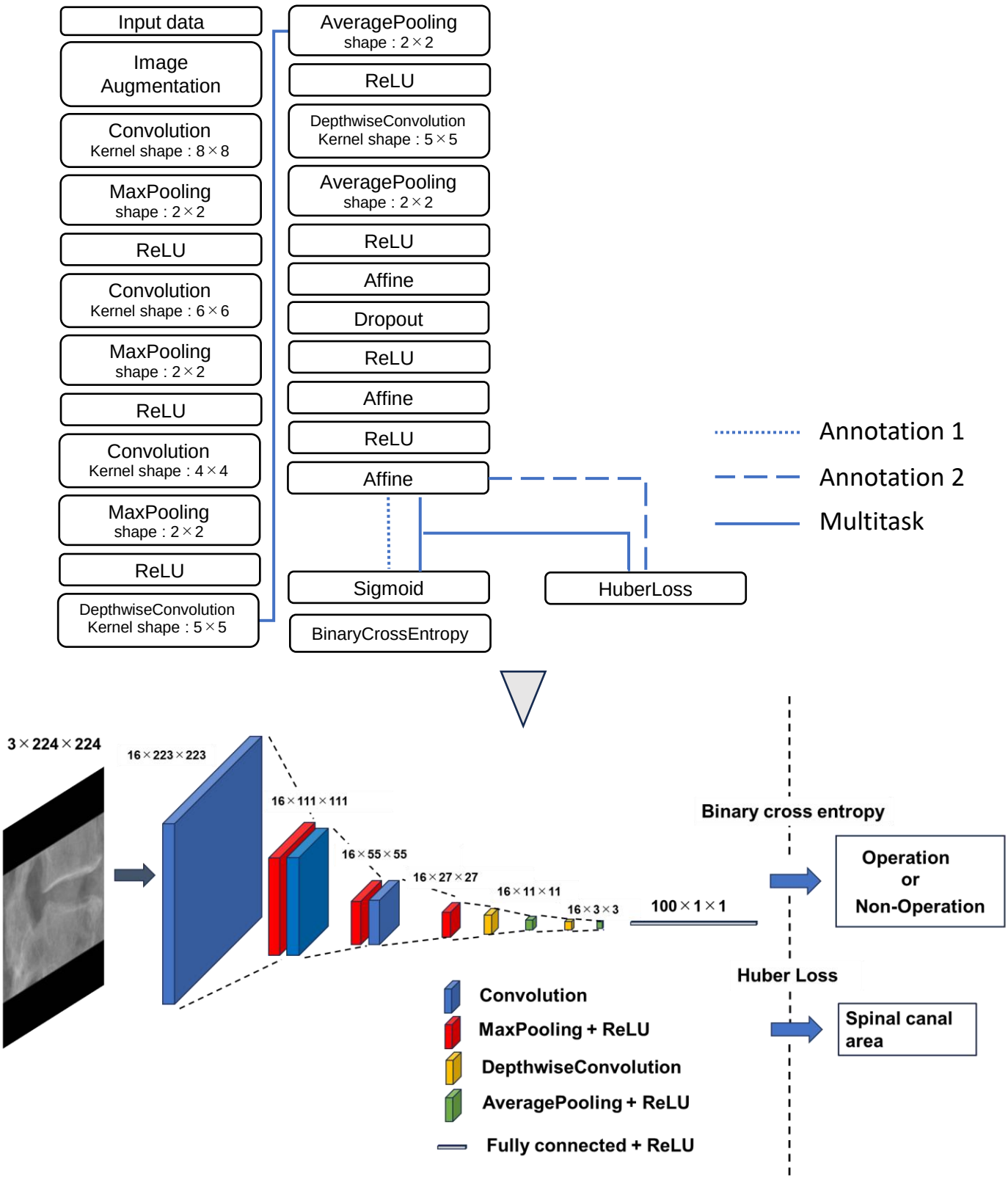

Input data
Image Augmentation
Convolution Kernel shape : 8×8
MaxPooling shape : 2×2
ReLU
Convolution Kernel shape : 6×6
MaxPooling shape : 2×2
ReLU
Convolution Kernel shape : 4×4
MaxPooling shape : 2×2
ReLU
DepthwiseConvolution Kernel shape : 5×5
AveragePooling shape : 2×2
ReLU
DepthwiseConvolution Kernel shape : 5×5
AveragePooling shape : 2×2
ReLU
Affine
Dropout
ReLU
Affine
ReLU
Affine
Sigmoid
BinaryCrossEntropy
HuberLoss
Annotation 1
Annotation 2
Multitask
3×224×224
16×223×223
16×111×111
16×55×55
16×27×27
16×11×11
16×3×3
100×1×1
Binary cross entropy
Operation or Non-Operation
Huber Loss
Spinal canal area
Convolution
MaxPooling + ReLU
DepthwiseConvolution
AveragePooling + ReLU
Fully connected + ReLU

Figure 3.

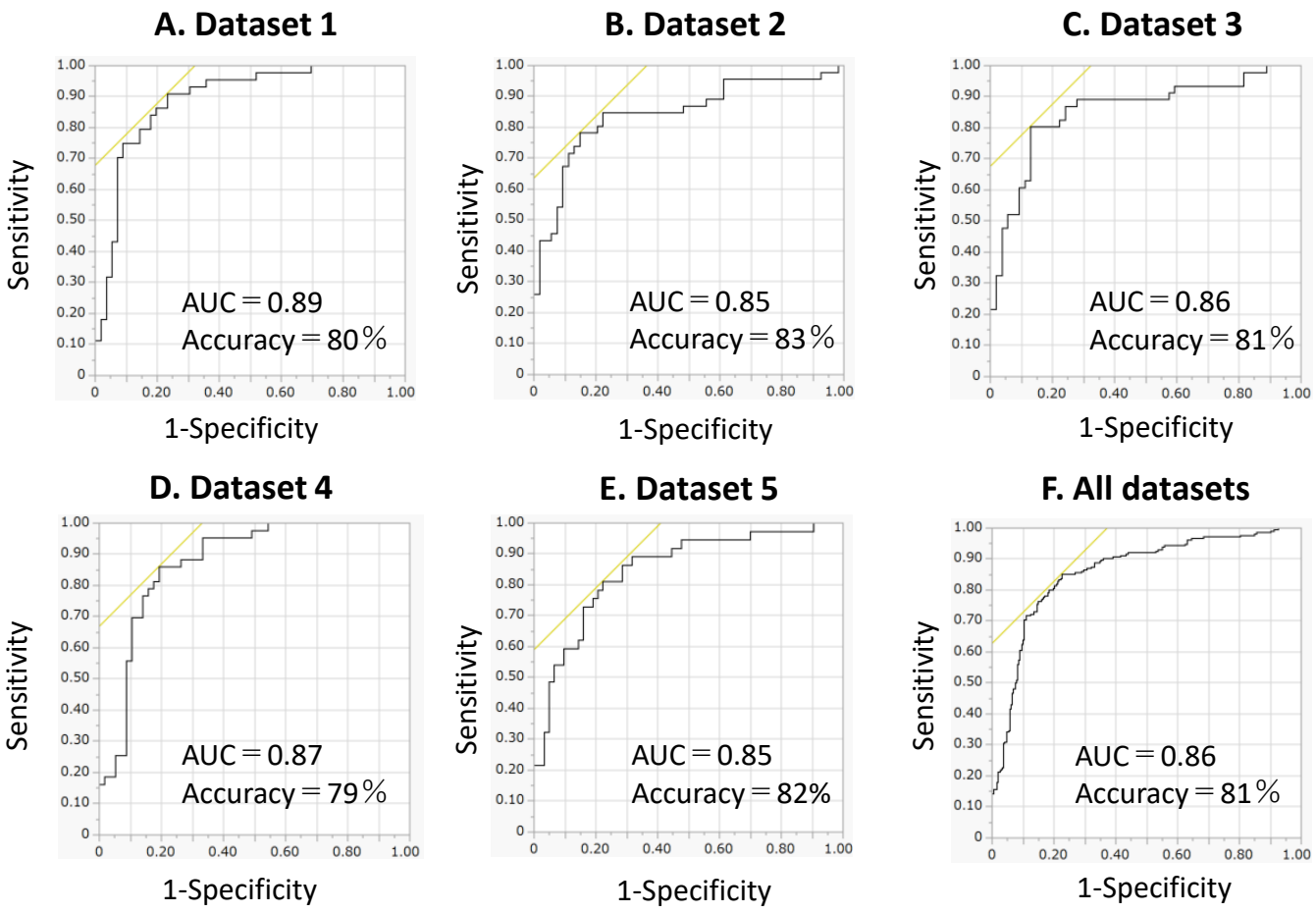

A. Dataset 1
AUC = 0.89
Accuracy = 80%
B. Dataset 2
AUC = 0.85
Accuracy = 83%
C. Dataset 3
AUC = 0.86
Accuracy = 81%
D. Dataset 4
AUC = 0.87
Accuracy = 79%
E. Dataset 5
AUC = 0.85
Accuracy = 82%
F. All datasets
AUC = 0.86
Accuracy = 81%
Sensitivity
1-Specificity

Figure 4.

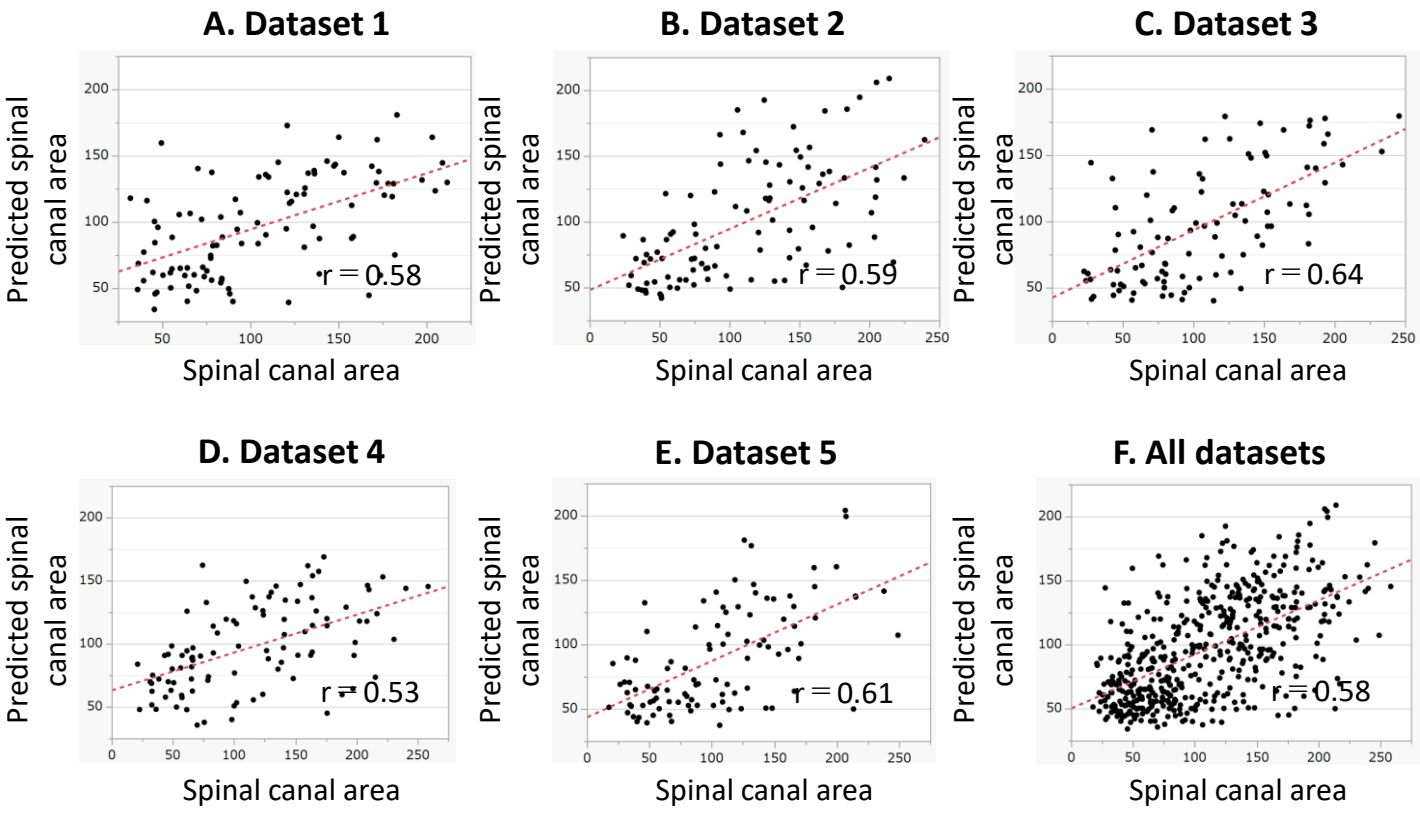

A. Dataset 1
B. Dataset 2
C. Dataset 3
D. Dataset 4
E. Dataset 5
F. All datasets
Predicted spinal canal area
Spinal canal area
r = 0.58
r = 0.59
r = 0.64
r = 0.53
r = 0.61
r = 0.58

Figure 5.

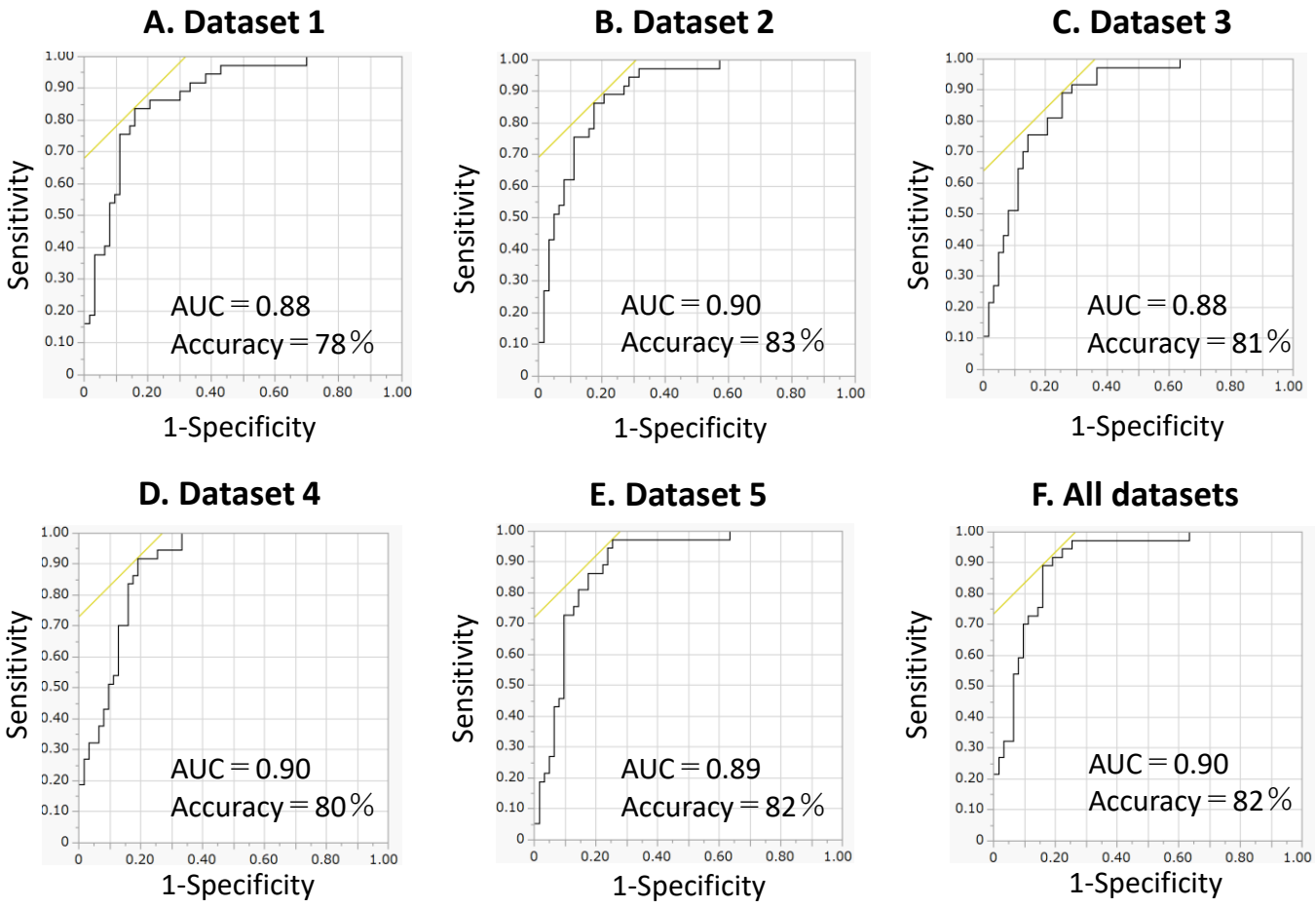

A. Dataset 1
AUC = 0.88
Accuracy = 78%
B. Dataset 2
AUC = 0.90
Accuracy = 83%
C. Dataset 3
AUC = 0.88
Accuracy = 81%
D. Dataset 4
AUC = 0.90
Accuracy = 80%
E. Dataset 5
AUC = 0.89
Accuracy = 82%
F. All datasets
AUC = 0.90
Accuracy = 82%
Sensitivity
1-Specificity

Figure 6.

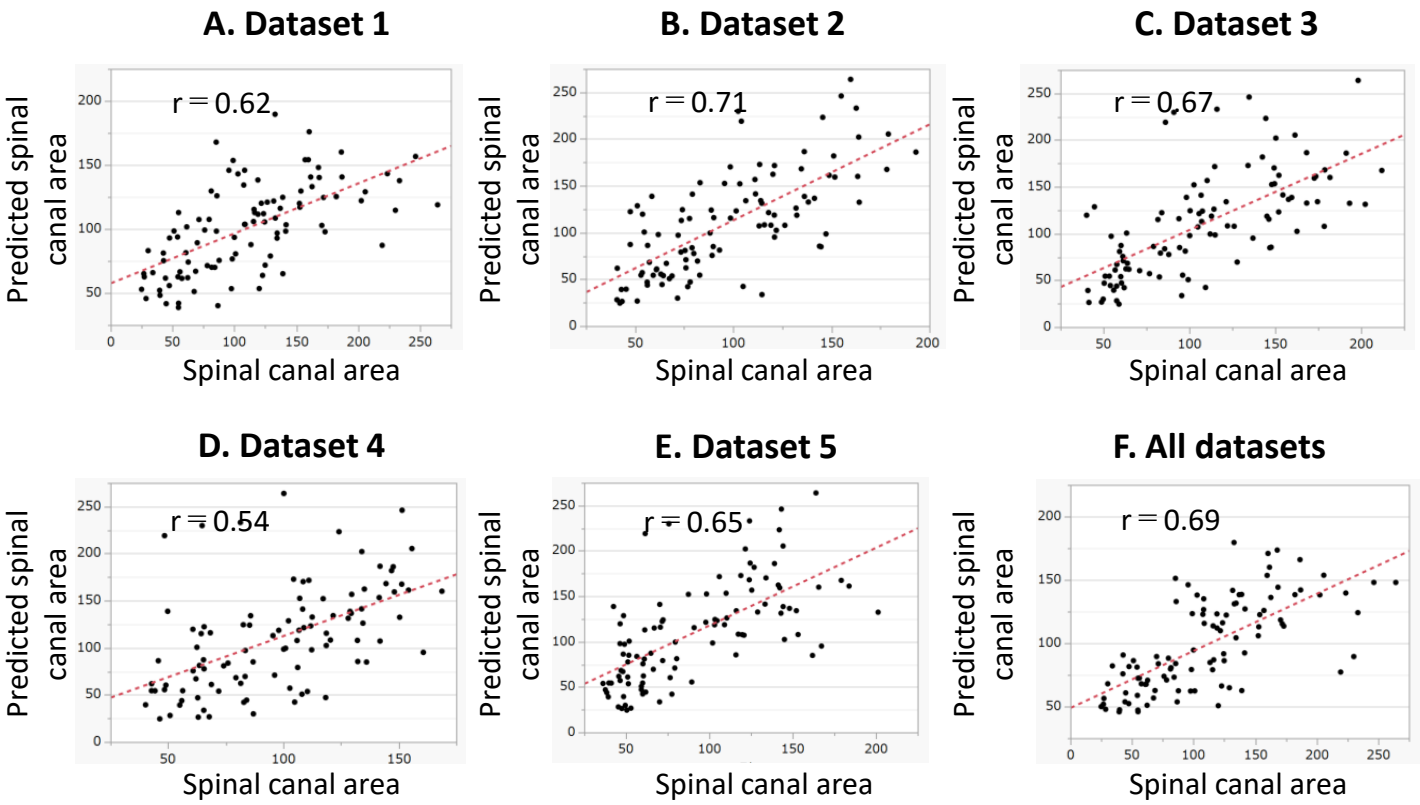

A. Dataset 1
r = 0.62
B. Dataset 2
r = 0.71
C. Dataset 3
r = 0.67
D. Dataset 4
r = 0.54
E. Dataset 5
r = 0.65
F. All datasets
r = 0.69
Predicted spinal canal area
Spinal canal area

Figure 7.

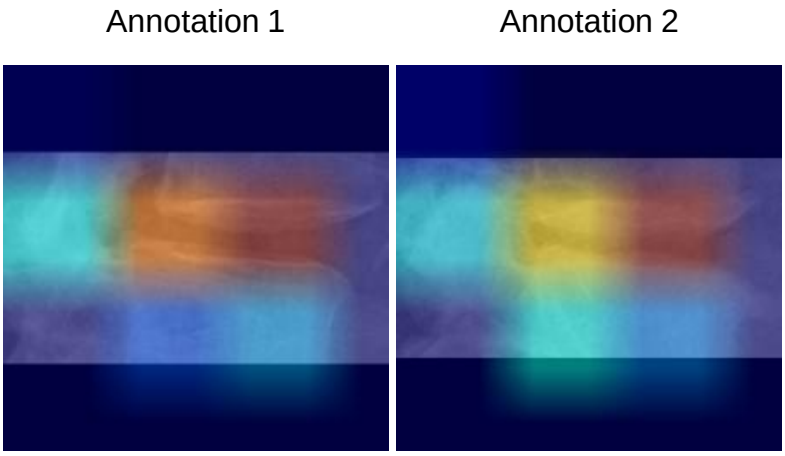


Figure 8.

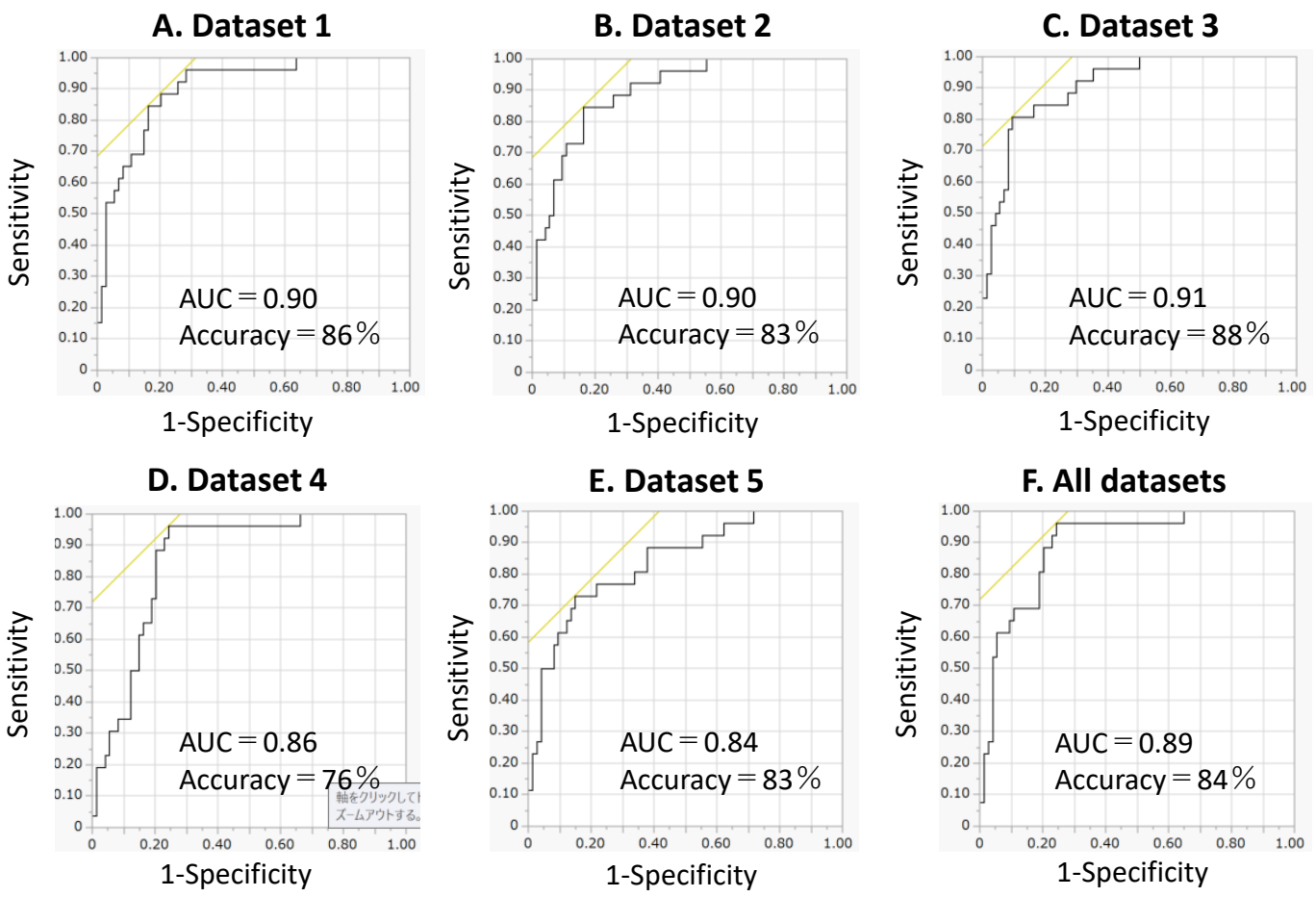

A. Dataset 1
AUC = 0.90
Accuracy = 86%
B. Dataset 2
AUC = 0.90
Accuracy = 83%
C. Dataset 3
AUC = 0.91
Accuracy = 88%
D. Dataset 4
AUC = 0.86
Accuracy = 76%
E. Dataset 5
AUC = 0.84
Accuracy = 83%
F. All datasets
AUC = 0.89
Accuracy = 84%
Sensitivity
1-Specificity

Figure 9.

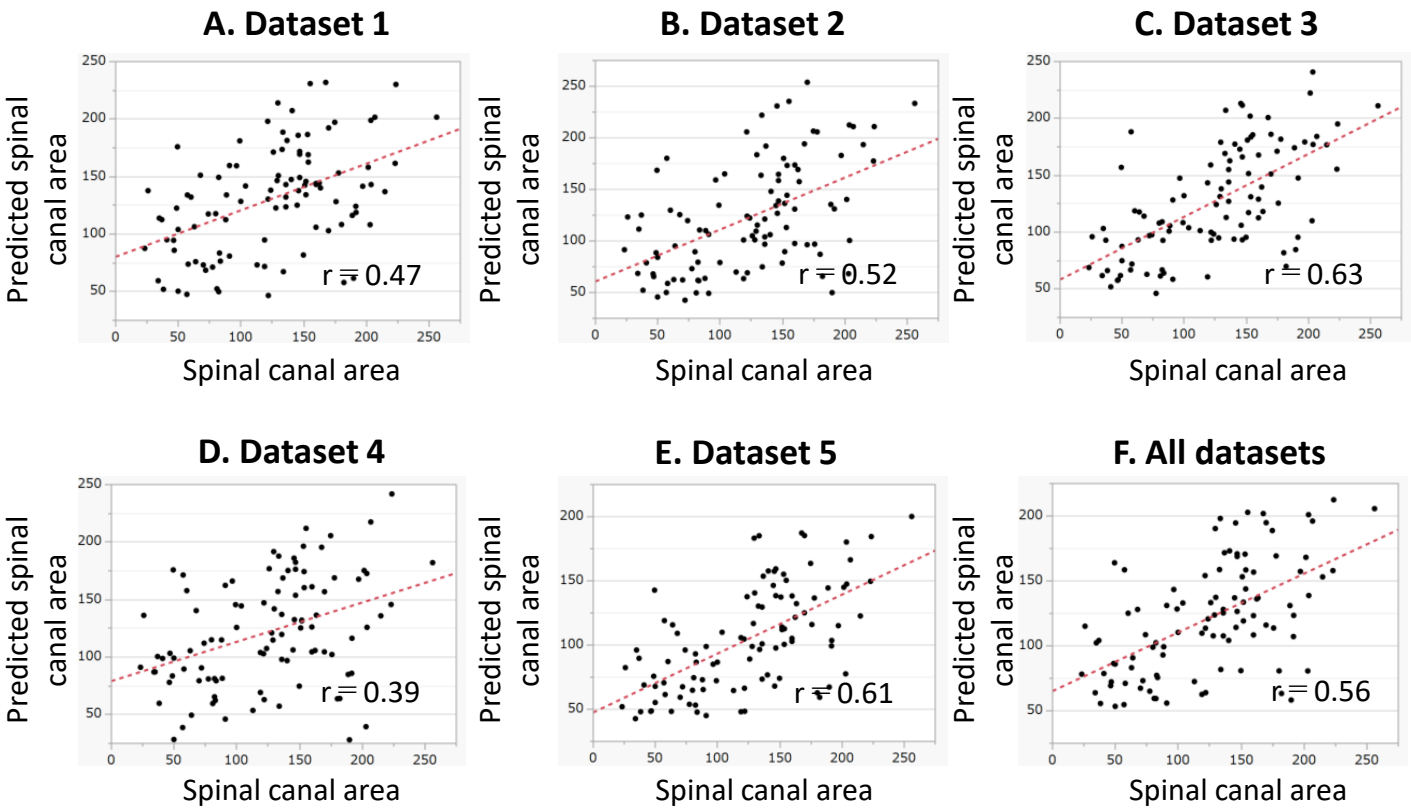

A. Dataset 1
B. Dataset 2
C. Dataset 3
D. Dataset 4
E. Dataset 5
F. All datasets
Predicted spinal canal area
Spinal canal area
r = 0.47
r = 0.52
r = 0.63
r = 0.39
r = 0.61
r = 0.56

Figure 10.

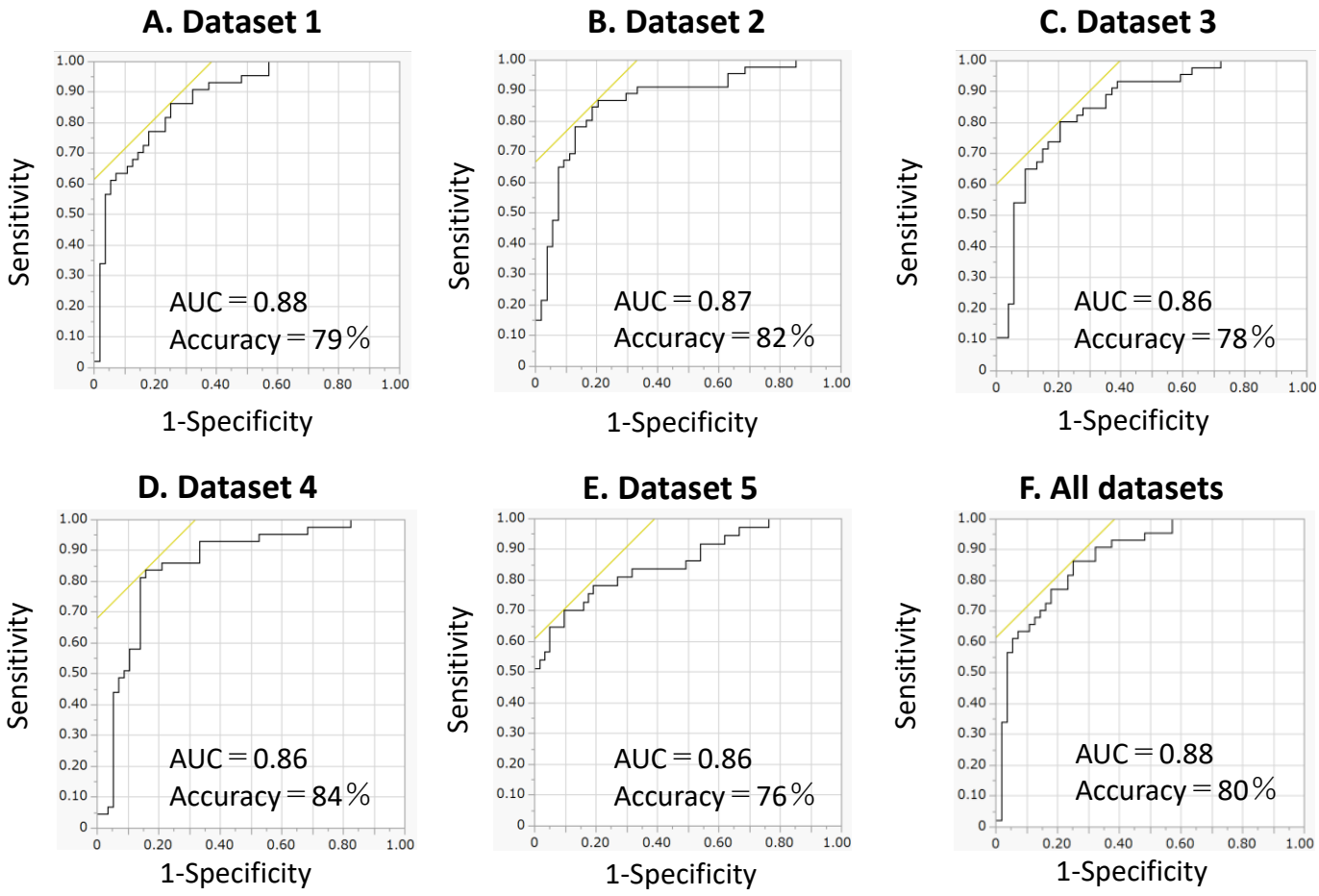

A. Dataset 1
AUC = 0.88
Accuracy = 79%
B. Dataset 2
AUC = 0.87
Accuracy = 82%
C. Dataset 3
AUC = 0.86
Accuracy = 78%
D. Dataset 4
AUC = 0.86
Accuracy = 84%
E. Dataset 5
AUC = 0.86
Accuracy = 76%
F. All datasets
AUC = 0.88
Accuracy = 80%
Sensitivity
1-Specificity

Figure 11.

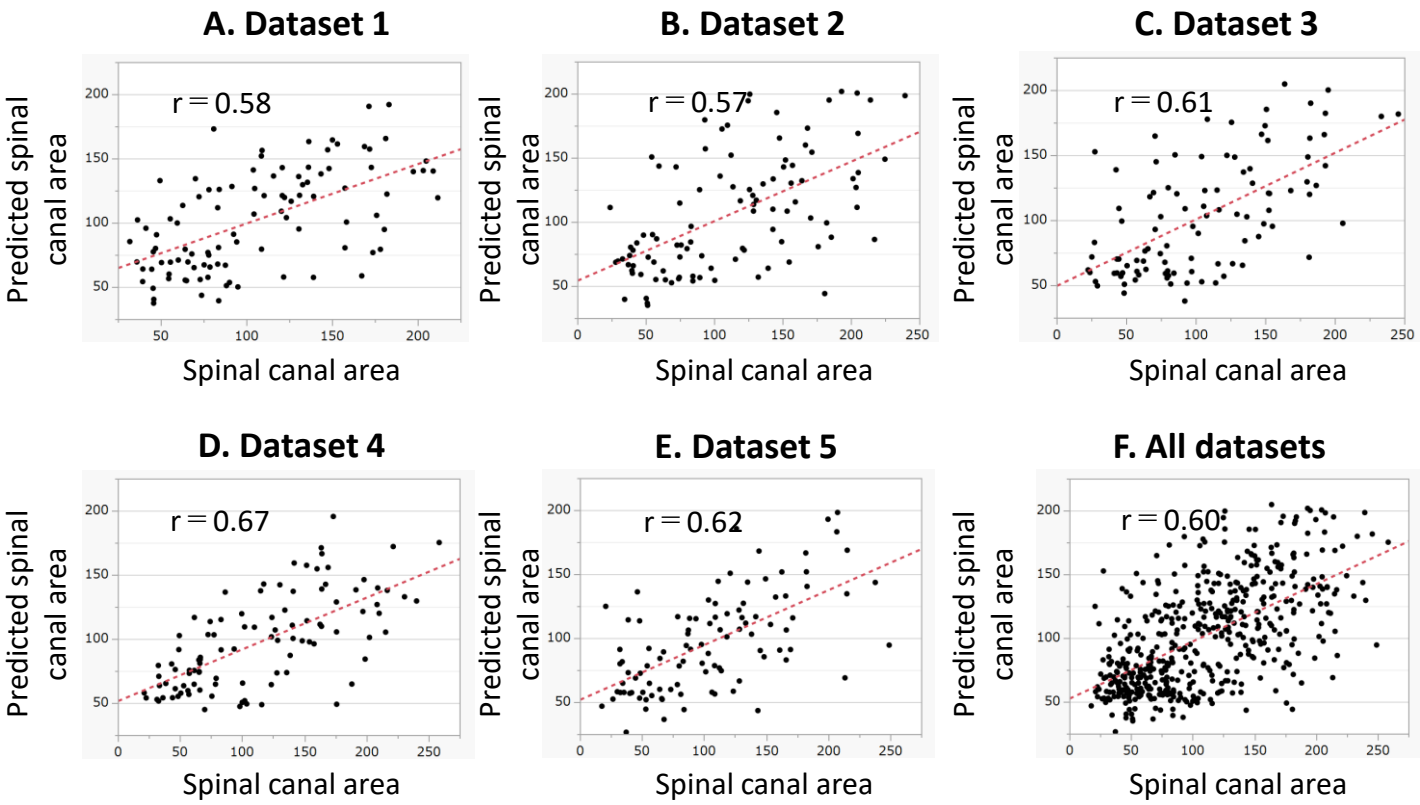

A. Dataset 1
r = 0.58
B. Dataset 2
r = 0.57
C. Dataset 3
r = 0.61
D. Dataset 4
r = 0.67
E. Dataset 5
r = 0.62
F. All datasets
r = 0.60
Predicted spinal canal area
Spinal canal area

Figure 12.

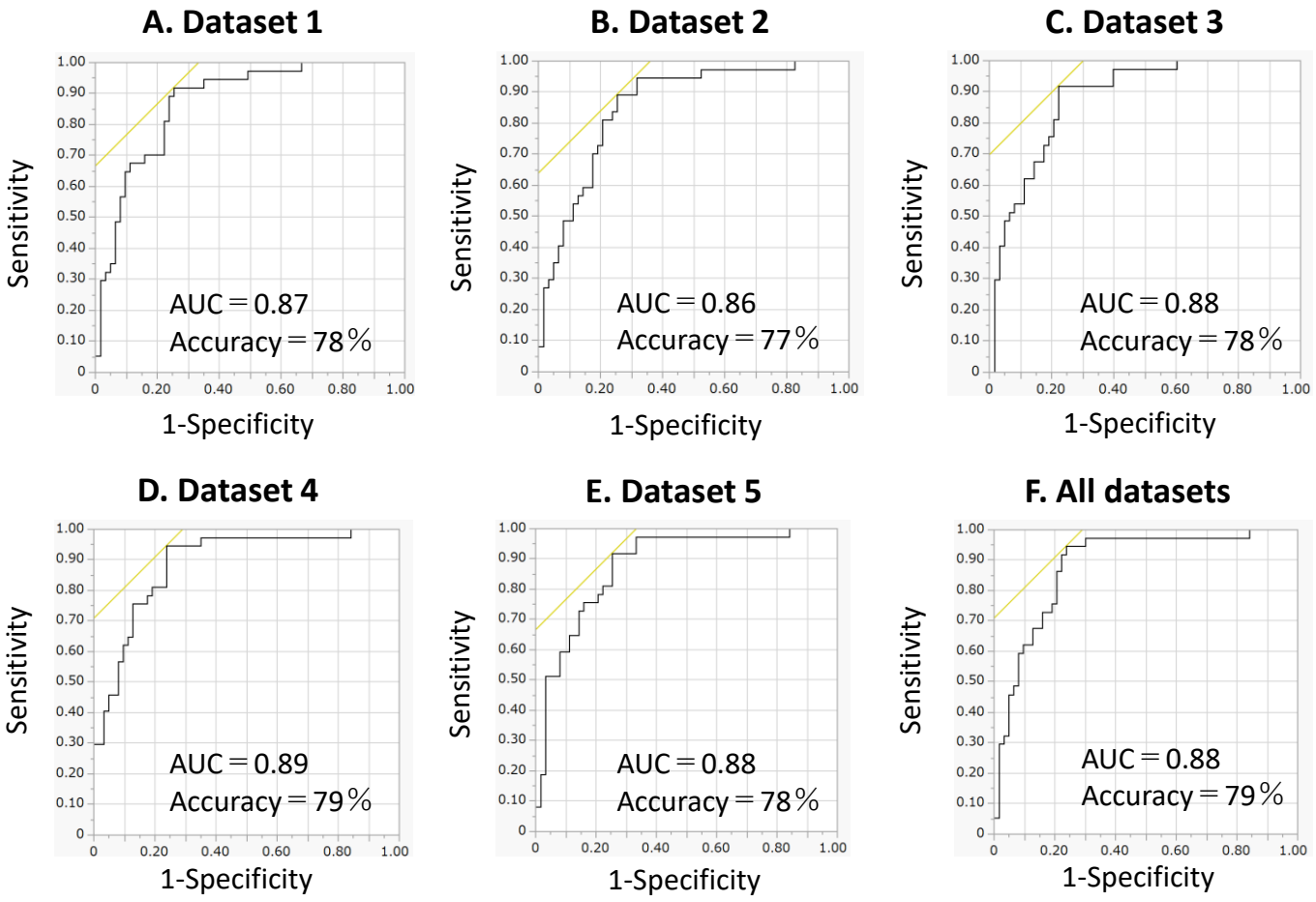

A. Dataset 1
AUC = 0.87
Accuracy = 78%
B. Dataset 2
AUC = 0.86
Accuracy = 77%
C. Dataset 3
AUC = 0.88
Accuracy = 78%
D. Dataset 4
AUC = 0.89
Accuracy = 79%
E. Dataset 5
AUC = 0.88
Accuracy = 78%
F. All datasets
AUC = 0.88
Accuracy = 79%
Sensitivity
1-Specificity

Figure 13.

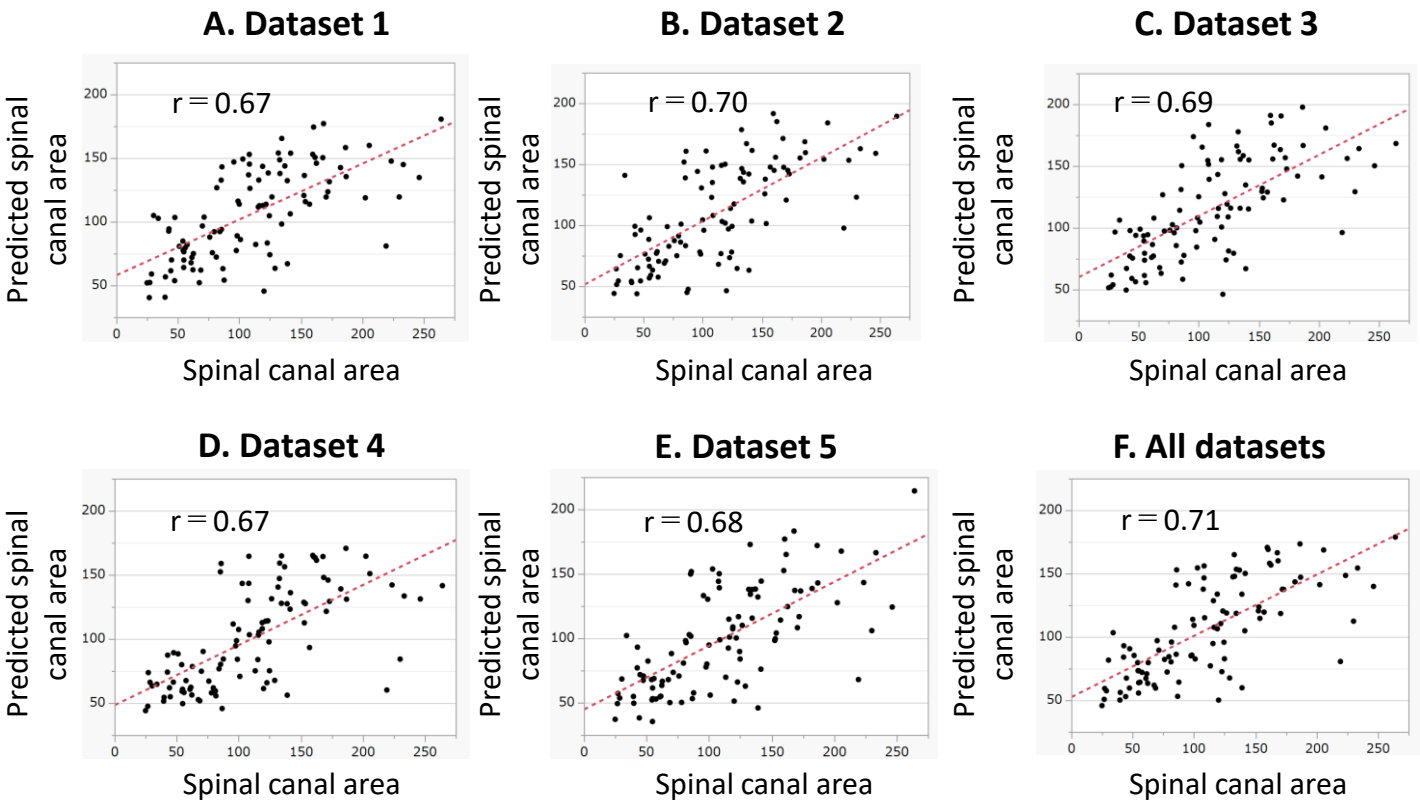

A. Dataset 1
r = 0.67
B. Dataset 2
r = 0.70
C. Dataset 3
r = 0.69
D. Dataset 4
r = 0.67
E. Dataset 5
r = 0.68
F. All datasets
r = 0.71
Predicted spinal canal area
Spinal canal area
0
50
100
150
200
250

Figure 14.

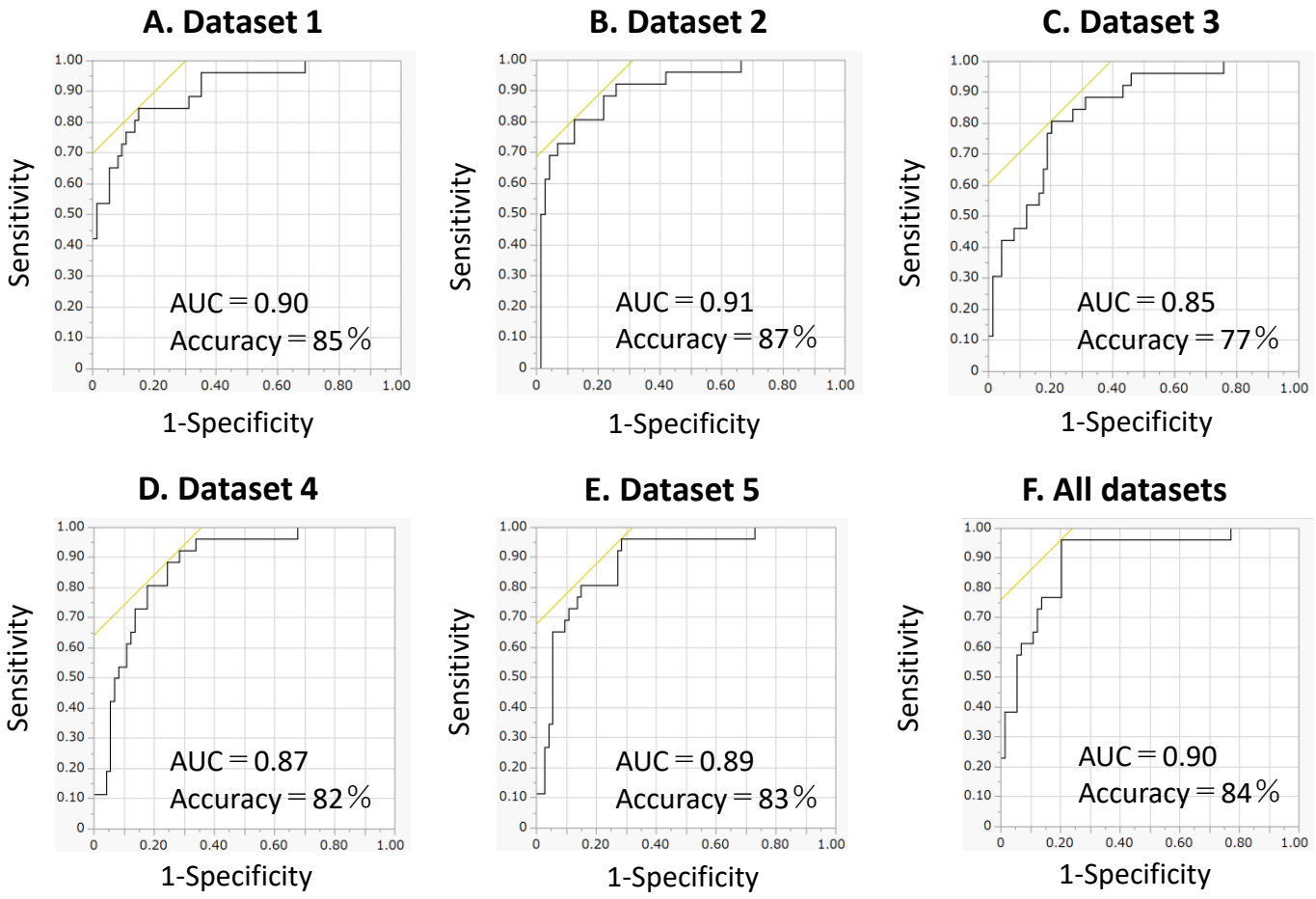

A. Dataset 1
AUC = 0.90
Accuracy = 85%
B. Dataset 2
AUC = 0.91
Accuracy = 87%
C. Dataset 3
AUC = 0.85
Accuracy = 77%
D. Dataset 4
AUC = 0.87
Accuracy = 82%
E. Dataset 5
AUC = 0.89
Accuracy = 83%
F. All datasets
AUC = 0.90
Accuracy = 84%
Sensitivity
1-Specificity

Figure 15.

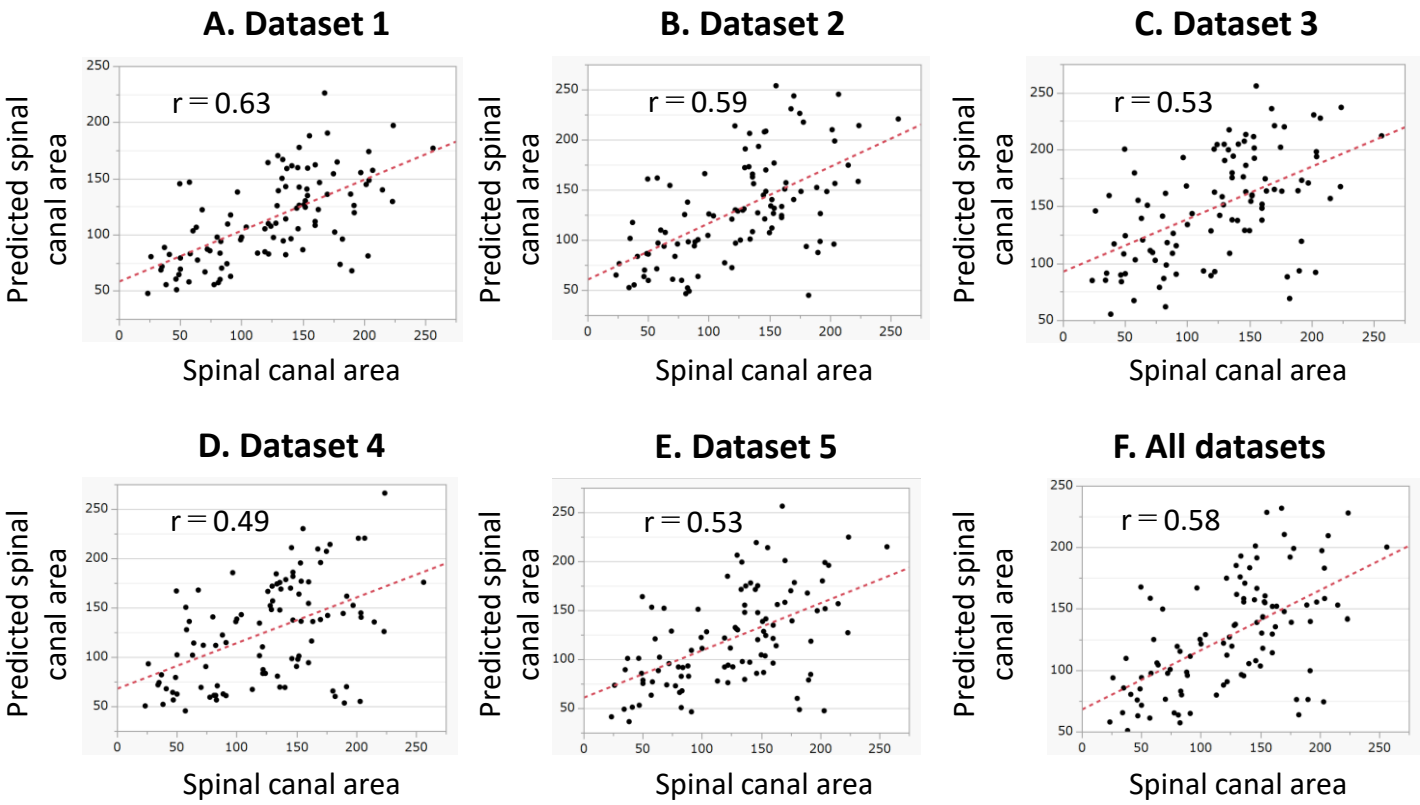

A. Dataset 1
r = 0.63
B. Dataset 2
r = 0.59
C. Dataset 3
r = 0.53
D. Dataset 4
r = 0.49
E. Dataset 5
r = 0.53
F. All datasets
r = 0.58
Predicted spinal canal area
Spinal canal area
0
50
100
150
200
250

Figure 16.

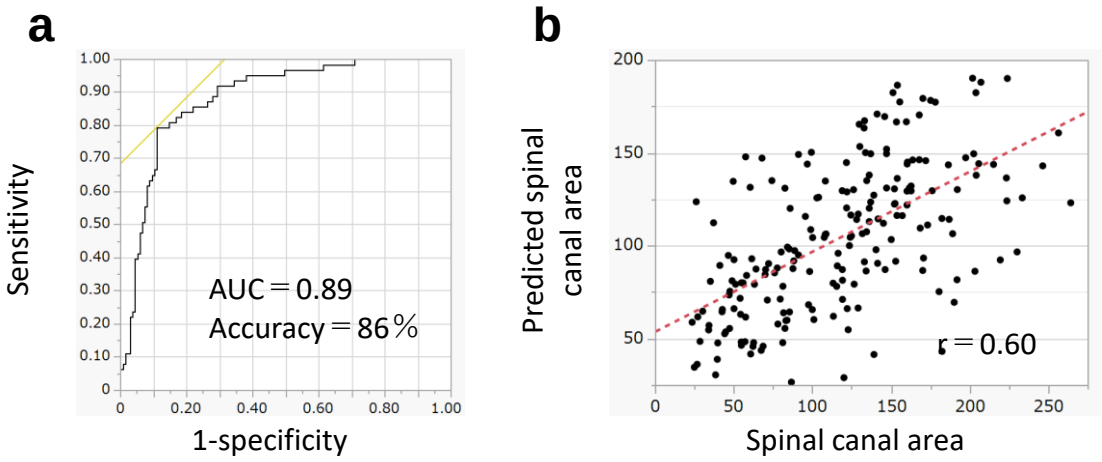

a
Sensitivity
1-specificity
AUC＝0.89
Accuracy＝86％
b
Predicted spinal canal area
Spinal canal area
r＝0.60

Table 1. Demographic and measurement data of each dataset.

	Dataset 1 (n=25)	Dataset 2 (n=25)	Dataset 3 (n=25)	Dataset 4 (n=25)	Dataset 5 (n=25)	External validation (n=25)	Extra external validation (n=25)
Males (Females)	17 (8)	14 (11)	20 (5)	12 (13)	15 (10)	16 (9)	16 (9)
Age	69.7 ± 12.7	69.4 ± 11.3	68.5 ± 12.1	68.8 ± 10.1	69.8 ± 8.8	66.0 ± 11.2	62.7 ± 13.1
Number of operation levels	45 (45.0%)	50 (50.0%)	39 (39.0%)	38 (38.0%)	42 (42.0%)	38 (38.0%)	26 (26.0%)
Spinel canal area (mm^2)	113.5 ± 56.1	100.2 ± 54.3	117.0 ± 54.9	118.4 ± 56.6	100.1 ± 50.1	110.8 ± 57.0	124.1± 53.9

Table 2. Experimental indicators for detecting surgical indications for LSCS in internal validation.

Internal validation	Sensitivity	Specificity	PPV	NPV	Accuracy	PLR	NLR
Dataset 1	0.71	0.88	0.82	0.79	80%	5.64	0.34
Dataset 2	0.83	0.83	0.81	0.85	83%	4.96	0.21
Dataset 3	0.78	0.83	0.80	0.82	81%	4.70	0.26
Dataset 4	0.63	0.91	0.84	0.77	79%	7.16	0.41
Dataset 5	0.76	0.86	0.76	0.86	82%	5.30	0.28
All datasets	0.74	0.86	0.80	0.81	81%	5.39	0.30

PPV= positive predictive value, NPV= negative predictive value,

PLR= positive likelihood ratio, NLR= negative likelihood ratio

Table 3. Experimental indicators for detecting surgical indications for LSCS in external validation.

External validation	Sensitivity	Specificity	PPV	NPV	Accuracy	PLR	NLR
Dataset 1	0.60	0.89	0.76	0.79	78%	5.35	0.46
Dataset 2	0.89	0.79	0.72	0.93	83%	4.32	0.14
Dataset 3	0.70	0.87	0.77	0.83	81%	5.53	0.34
Dataset 4	0.73	0.84	0.73	0.84	80%	4.60	0.32
Dataset 5	0.73	0.87	0.77	0.85	82%	5.75	0.31
All datasets	0.76	0.86	0.76	0.86	82%	5.30	0.28

PPV= positive predictive value, NPV= negative predictive value,

PLR= positive likelihood ratio, NLR= negative likelihood ratio

Table 4. Experimental indicators for detecting surgical indications for LSCS in extra external validation.

External validation	Sensitivity	Specificity	PPV	NPV	Accuracy	PLR	NLR
Dataset 1	0.68	0.92	0.74	0.90	86%	8.50	0.35
Dataset 2	0.73	0.87	0.66	0.90	83%	5.41	0.31
Dataset 3	0.77	0.92	0.77	0.92	88%	9.49	0.25
Dataset 4	0.50	0.85	0.54	0.83	76%	3.36	0.59
Dataset 5	0.58	0.92	0.71	0.86	83%	7.12	0.46
All datasets	0.65	0.91	0.71	0.88	84%	6.91	0.20

PPV= positive predictive value, NPV= negative predictive value,

PLR= positive likelihood ratio, NLR= negative likelihood ratio

Table 5. Experimental indicators for detecting surgical indications for LSCS in internal validation by multitask learning.

Internal validation	Sensitivity	Specificity	PPV	NPV	Accuracy	PLR	NLR
Dataset 1	0.75	0.82	0.77	0.81	79%	4.20	0.30
Dataset 2	0.76	0.87	0.83	0.81	82%	5.87	0.28
Dataset 3	0.72	0.83	0.79	0.78	78%	4.30	0.34
Dataset 4	0.84	0.84	0.80	0.87	84%	5.30	0.19
Dataset 5	0.78	0.75	0.64	0.86	76%	3.09	0.29
All datasets	0.77	0.82	0.77	0.82	80%	4.28	0.28

PPV= positive predictive value, NPV= negative predictive value,

PLR= positive likelihood ratio, NLR= negative likelihood ratio

Table 6. Experimental indicators for detecting surgical indications for LSCS in external validation by multitask learning.

External validation	Sensitivity	Specificity	PPV	NPV	Accuracy	PLR	NLR
Dataset 1	0.70	0.83	0.70	0.83	78%	4.03	0.36
Dataset 2	0.68	0.83	0.69	0.81	77%	3.87	0.39
Dataset 3	0.54	0.92	0.80	0.77	78%	6.81	0.50
Dataset 4	0.84	0.76	0.67	0.89	79%	3.52	0.21
Dataset 5	0.78	0.78	0.67	0.86	78%	3.53	0.28
All datasets	0.76	0.81	0.70	0.85	79%	3.97	0.30

PPV= positive predictive value, NPV= negative predictive value,

PLR= positive likelihood ratio, NLR= negative likelihood ratio

Table 7. Experimental indicators for detecting surgical indications for LSCS in extra external validation by multitask learning.

External validation	Sensitivity	Specificity	PPV	NPV	Accuracy	PLR	NLR
Dataset 1	0.67	0.94	0.85	0.85	85%	11.33	0.35
Dataset 2	0.84	0.88	0.62	0.96	87%	6.82	0.18
Dataset 3	0.80	0.77	0.15	0.97	77%	3.46	0.26
Dataset 4	0.63	0.90	0.73	0.85	82%	6.33	0.41
Dataset 5	0.66	0.90	0.73	0.87	83%	6.65	0.38
All datasets	0.62	0.92	0.73	0.87	84%	7.59	0.42

PPV= positive predictive value, NPV= negative predictive value,

PLR= positive likelihood ratio, NLR= negative likelihood ratio