Using A Convolutional Neural Network Model To Derive Imaging Landmarks For Lumbar Spine Numbering On Axial MR Images

Yongwon Cho
Korea University Anam Hospital, Korea University College of Medicine
Kyung-Sik Ahn (glassesik@gmail.com)
Korea University Anam Hospital, Korea University College of Medicine
Chang Ho Kang
Korea University Anam Hospital, Korea University College of Medicine
Beom Jin Park
Korea University Anam Hospital, Korea University College of Medicine

Research Article

Keywords: Neural networks, computer, MRI, Lumbar vertebrae, Back muscles

Posted Date: February 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1323975/v1

Abstract

Background: Understanding the axial lumbar spine anatomy, including knowledge of the relationship between the lumbar spine level and other paraspinal structures, is important in diagnosing and treating disease. The purpose of this study was to validate the accuracy of a convolutional neural network (CNN) model in lumbar spine level numbering on axial MR images and to find the appropriate anatomic landmarks for numbering using a class activation map (CAM).

Methods: A total of 6055 axial MR images of the lumbar spine from L1–2 to L5–S1 disc levels were obtained to train and validate the CNN model. The MR images were acquired using three 3-Tesla machines. The algorithm was developed with three models, and the best-performing model was selected. The external validation set (n = 493) was obtained from other institutions using various machines. The accuracy of the numbering was analyzed using a confusion matrix and receiver operating characteristic curves. The CAMs were reviewed, and the identified anatomic structures were investigated. A reader study was performed by three radiologists, and their accuracy was compared with that of the model.

Results: The overall accuracy of the best-performing model for lumbar spine numbering was 0.98 on internal validation and 0.95 on external validation. For the CAM review, mappings concentrated on both paraspinal areas, including the kidney, back muscles, and ilium according to the level. Top-1 and top-2 accuracies of the reviewers ranged between 0.56–0.75, and 0.84–0.93, respectively. After reviewing the CAMs, the accuracy increased to 0.75–0.78 and 0.93–0.98, respectively.

Conclusion: A CNN model can accurately determine the level of the lumbar spine on axial MR images, and the configuration of muscles can be used to determine the lumbar level.

Introduction

For accurate diagnoses, proper treatment planning, and good communication among physicians, an understanding of the axial lumbar spine anatomy along with knowledge on how it relates with other paraspinal structures is important. Many anatomic landmarks have been reported for determining the level of the lumbar spine, including the aortic bifurcation, right renal artery, conus medullaris, and iliolumbar ligaments [1, 2]. Although axial images derived from magnetic resonance imaging (MRI) are provided with a navigator, which informs the level or section of the corresponding axial images over the sagittal or coronal plane, many experienced physicians can make an approximate estimation of the level of the axial images based on the neighboring structures such as the liver, kidneys, prevertebral vessels, paravertebral muscles, and iliac bone. However, on a lumbar MRI, the field of view (FOV) of the axial images is focused on the central canal and disc, which often leads to the FOV not including enough solid organ landmarks; this is especially true in patients with a larger body size.

With the advent of deep learning techniques and their application in medical imaging, promising results have been reported in radiologic tasks, including detection, segmentation, and classification [3]. In spine research, the application of this technique has been reported in many studies, and they cover topics such

as localization and labeling of spinal structures, segmentation of the spine, diagnosis of pathology, outcome prediction, clinical decision support, and analysis of biomechanics [4-6]. Although the unaccountability of this technique is a drawback, the class activation map (CAM) can be used to provide hints to obtain the result. By investigating the CAMs, researchers can confirm that the algorithm is working as expected, or inversely, they may discover new avenues to solve the task if the accuracy of the algorithm is guaranteed.

The purpose of this study was to validate the accuracy of a convolutional neural network (CNN) model in lumbar spine numbering on axial MR images and to find the applicable anatomic landmarks for numbering by investigating CAMs.

Materials And Methods

Training and Internal Validation Set

The institutional review board of the institution approved this retrospective study. The axial images of lumbar spine MRIs for disc levels L1–2 to L5–S1 that were performed from January 2017 to December 2019 were obtained from three 3-Tesla machines. The following exclusion criteria were applied:

- patients under 20 or over 60 years;
- images that show transition vertebra where the exact level could not be identified;
- anatomic distortion due to a tumor, infection, or trauma;
- images that show severe central canal stenosis of any kind (disc herniation, ligamentum flavum thickening, etc.).

A total of 6055 axial images (3179 male and 2870 female, mean age 48.5 ± 12 years) were selected: 978 images of L1–2, 1276 images of L2–3, 1344 images of L3–4, 1254 images of L4–5, and 1203 images of L5–S1. Image inclusion and exclusion were performed by a musculoskeletal radiologist with 13 years' experience. The images were in DICOM format and were anonymized. The data set was randomly split into 70 percent for training, 10 percent for tuning, and 20 percent for the final evaluation.

Test Set (External Validation)

A separate test set was obtained from lumbar spine MRIs from other institutions using various machines with various magnetic strengths after the anonymization of the DICOM data. The same inclusion and exclusion criteria were applied as for the training set. The test set comprised 493 axial images (L1–2: 83, L2–3: 112, L3–4: 101, L4–5: 89, and L5–S1: 108). There were 42 images obtained with a 3-T machine, 337 images obtained with a 1.5-T machine, and 74 images obtained with machine below 1.5-T. The number of images used in this study is presented in Table 1. For the reader study, three radiologists (two board-certified musculoskeletal radiologists and one third-year radiology resident) blindly labeled 100 randomly selected cases (20 cases for each level) before and after reviewing the CAM results.

Model Implementation

We classified five classes (from L1–2 to L5–S1) on spine MR images using transfer learning with three pre-trained models. First, DenseNet [7], which was configured with three dense blocks involving four batch normalization rectified linear unit convolution layers. Second, InceptionResNetV2 [8, 9] had an Inception structure and residual block connections. Convolutional filters of various sizes were connected to a residual block defined as the Inception-ResNet block. Finally, Xception [10], which has a stronger hypothesis that underlies the Inception architecture, was proposed as a CNN architecture based on depth wise separable convolution layers. The mapping of cross-channel and spatial correlations in the feature maps of this network can be entirely distributed. It had 36 convolutional layers, including feature extraction of the network. External validation was performed using the model that delivered the highest accuracy during testing. Classification results and localization using Grad-CAM for lumbar spine numbering [11] were visualized after inference. The architectures of the different networks are shown in Figure 1.

The geometric augmentation including zoom, rotation, and shift, on the edges of the images were used to help alleviate scanner-specific biases and improve the robustness of the CNNs against additional sources of variability. The input images were rescaled to 512 × 512 pixels and converted into Numpy arrays in Python 3.6. These datasets were loaded on a GPU server with Ubuntu 20.04, CUDA 11.2, four 24 GB Titan RTXs and Quadro graphics cards, and cuDNN 11.2 (NVIDIA Corporation) with the Keras with backend Tensorflow framework 1.4.1. We used the ADAM optimizer with an initial learning rate of 0.001 for the classification of five classes (from L1–2 to L5–S1). The tuning errors for the selection of optimized models were minimized by running the back-propagation algorithm over 25 training epochs with a batch size of 8.

Statistical Analysis

The performance of each model was evaluated using a confusion matrix. Multiple classifications of lumbar spine level were evaluated in terms of recall, precision, and accuracy using the scikit-learn Python 3.0 library. The differences between the three models were compared using the Delong test with R software (version 1.1.456). ROC curves and AUC were evaluated, and the performances of the models were compared with those of the human reviewers.

Results

Internal and External Validation

Among the three trained models, the Xception model showed the highest accuracy (0.977) compared to InceptionResNetV2 (0.971) and DensNet169 (0.970%), although the difference was not significant (p = 0.304 and 0.386, respectively). For the Xception model, the AUC for each level (L1–2 to L5–S1) was 0.9998, 0.9985, 0.9990, 0.9984, and 0.9990, respectively. The accuracy of the Xception model for external validation was 0.947, with an AUC of 0.9987, 0.9978, 0.9933, 0.9969, and 0.9997 for each level,

respectively. The recall was 0.952, 0.938, 0.941, 0.921, and 0.981, and the precision was 0.975, 0.981, 0.888, 0.911, and 0.981 for each level, respectively. The performance of the model is presented in Table 2.

CAM Evaluation

The CAM results for all trained classes were visualized independently, and the CAMs of the ROIs were extracted individually. Most of the CAMs were located at both paramedian spaces, and included the kidney, retroperitoneal fat, psoas (Ps), quadratus lumborum (QL), erector spinae (ES), iliacus, and ilium according to the respective levels. The CAM results according to the respective levels are shown in Figure 2.

Based on the CAM reviews, a template for muscle configuration was derived. At L1–2, Ps is small and triangular; QL shows a thin and acute lateral margin; the lateral margin of the QL is far medial to that of the ES; the right kidney presents along with the renal pelvis. At L2–3, Ps is slightly enlarged, but still triangular; the lateral margin of the QL is similar or slightly medial to that of the ES; the inferior pole of the right kidney is revealed. At L3–4, Ps changes to a round shape; the QL is thicker, and the lateral margin is lateral to the lateral margin of the ES. At L4–5, Ps is large and round. The lateral margin of the QL is far lateral to that of the ES. At L5–S1, the iliac bone and iliacus are seen. Illustrations of each lumbar level are shown in Figure 3.

Accuracy of Human Reviewers

The Top-1 and top-2 accuracies of the three human reviewers in 100 randomly selected cases ranged between 0.56–0.75 and 0.84–0.93, respectively, in the first session. The human reviewers referred to the kidneys, renal vein, aortic bifurcation, Ps, and iliac bone as landmarks for estimating the lumbar level. The labeling was performed again after the revision of the CAM results and the recognition that muscles could be an indicator for numbering of the spine level. In the second session, the top-1 and top-2 accuracies were increased to 0.75–0.78 and 0.93–0.98, respectively (Table 3). Among the level analyses, the analysis of L5–S1 level showed the highest accuracy among all the reviewers. The ROC curves of the model and human reviewers are presented in Figure 4.

Discussion

In this study, we confirmed that a CNN model could accurately determine the level of axial lumbar spine MR images, and the muscle configuration of the lumbar spine can be used to determine the lumbar level in addition to the solid organs, vessels, and ligaments usually referred to. Our results might be a natural outcome in a way; however, recognition of muscle configuration is an important factor in increasing the accuracy of level estimation on axial MR images.

With the advent of deep learning techniques in spine imaging, automatic segmentation using CNNs has been developed [4]. However, most segmentations are performed on the sagittal plane. Spine numbering is a basic and important issue in medical communication to convey disease localization and treatment

planning. Sometimes, there may be limited information available for spine level recognition (i.e., ultrasound-guided procedures) [12, 13]. In such cases, knowledge of anatomic landmarks is helpful in determining the right level. In addition, with the wide use of axial images such as MRI and CT, the necessity of anatomic landmarks to determine spine level on axial images has arisen. Many anatomic landmarks have been reported in the literature for defining the level of the lumbar spine on axial images, including the aortic bifurcation, inferior vena cava confluence, renal artery, conus medullaris, superior mesenteric artery, psoas muscles, pedicles, neural foramen, and iliolumbar ligaments [14-17]. Because of the convenience of automatic cross-link functions in PACS viewers, physicians usually do not have to determine the level of the spine on axial images themselves. In this regard, knowledge of anatomic landmarks that can indicate the spine level might be helpful in understanding the anatomy and analyzing the images. The axial anatomy of the back muscles according to the spinal level has been described in several studies [18-20]. Considering that intermuscular spaces are well demarcated on MRI and CT images and show differences according to the spinal level [20], the configuration of muscles might be a useful landmark for level determination. To the best of our knowledge, there have been no reports suggesting the use of muscle configuration as a landmark for lumbar level determination.

In this study, we compared three state-of-the-art deep learning networks (DenseNet169, InceptionResNetV2, and Xception), and chose the best performing model. Notably, the difference in accuracy among the models were not significant. All three networks similarly classified five classes (from L1–2 to L5–S1) on spine MR images and showed good performance. In a neural network model, despite some arguments, CAM shows evidence of the decision-making of the algorithm [21]. If the CAM tags are in inappropriate locations, there might be an error in the algorithm or bias in the training set. Conversely, if the accuracy of the algorithm is guaranteed, researchers can use CAMs to find new perspectives on solving a task. In this study, we focused on muscle configuration and showed that the accuracy of human reviewers increased after CAM reviews.

This study had several limitations. The first is the limited resolution of the retroperitoneal space and the organs in the lumbar MR image. This may have influenced our results. In patients with a large body size or those with rapid respiration, image quality of the upper portion of the axial MR images was low. This might decrease the importance of the aorta or renal vessels, which were suggested as landmarks in a CT-based study. In addition, there could be a minor mismatch of the lower lumbar level muscles and discs according to lumbar lordosis. Because axial lumbar MR images are obtained in a plane parallel to the lumbar disc, the muscles in the lower lumbar axial images are slightly higher than the disc. Second, we did not include the transition vertebra, which may comprise 4%–30% of the normal population [22]. Because the exact count of the spine from C2 was not possible in some cases with the ambiguous morphology of L5 or S1 due to a lack of the whole spinal sagittal image, we excluded cases with transitional vertebra. This may have influenced our results, leading to the high accuracy. Finally, although we found that muscle configuration could be the key to determining the lumbar spine level, further detailed quantification analysis for paraspinal muscles (i.e., area dimension or ratio) will be required.

Conclusion

A CNN model can accurately determine the level of the lumbar spine on axial MR images, and the muscle configuration can be used for accurate lumbar level determination.

Abbreviations

AUC: area under curve; CAM: class activation map; CNN: convolutional neural network; CT: computed tomography; DICOM: Digital Imaging and Communications in Medicine; ES: erector spinae; FOV: field of view; MRI: magnetic resonance imaging; PACS: picture archiving communication system; Ps: psoas; QL: quadratus lumborum; ROC: receiver operating characteristic; ROI: regions of interest.

Declarations

Acknowledgements

We thank Hyun Ki Ko for his contribution to this research in terms of data acquisition and preparation.

Author contributions

KSA conceived the study. KSA and YC developed the method and collected the data. YC implemented the algorithms. KSA and YC performed the data analyses and wrote the manuscript. CHK and BJP reviewed the manuscript. All authors have read and approved the manuscript.

Funding

This research was supported by a grant from Korea University Anam Hospital, Seoul, Republic of Korea (Grant No. O1801021)

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to the hospital's regulations, but are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

This retrospective study was approved by the Ethics Committee of Korea University Anam Hospital. Informed consent for the study was waived. All methods were performed in accordance with the relevant guidelines and regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Lee CH, Seo BK, Choi YC, Shin HJ, Park JH, Jeon HJ, et al. Using MRI to evaluate anatomic significance of aortic bifurcation, right renal artery, and conus medullaris when locating lumbar vertebral segments. AJR Am J Roentgenol. 2004;182:1295-300.

2. Hughes RJ, Saifuddin A. Numbering of lumbosacral transitional vertebrae on MRI: role of the iliolumbar ligaments. AJR Am J Roentgenol. 2006;187:W59-W65.

3. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol. 2017;18:570-84.

4. Galbusera F, Casaroli G, Bassani T. Artificial intelligence and machine learning in spine research. JOR Spine. 2019;2:e1044.

5. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Davidson JM, Leslie WD. Identification of Vertebral Fractures by Convolutional Neural Networks to Predict Nonvertebral and Hip Fractures: A Registry-based Cohort Study of Dual X-ray Absorptiometry. Radiology. 2019;293:405-11.

6. Karhade AV, Thio Q, Ogink PT, Shah AA, Bono CM, Oh KS, et al. Development of Machine Learning Algorithms for Prediction of 30-Day Mortality After Surgery for Spinal Metastasis. Neurosurgery. 2019;85:E83-E91.

7. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE. p. 2261-9.

8. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE. p. 770-8.

9. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)2017. p. 4278-84.

10. Chollet F. Xception: deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE. p. 1800-7.

11. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE. p. 618-26.

12. Gupta A, Sondekoppam R, Kalagara H. Quadratus Lumborum Block: a Technical Review. Curr Anesthesiol Rep. 2019;9:257-62.

13. Gofeld M, Bristow SJ, Chiu SC, McQueen CK, Bollag L. Ultrasound-guided lumbar transforaminal injections: feasibility and validation study. Spine (Phila Pa 1976). 2012;37:808-12.

14. Chithriki M, Jaibaji M, Steele RD. The anatomical relationship of the aortic bifurcation to the lumbar vertebrae: a MRI study. Surg Radiol Anat. 2002;24:308-12.

15. Carrino JA, Campbell PD, Jr., Lin DC, Morrison WB, Schweitzer ME, Flanders AE, et al. Effect of spinal segment variants on numbering vertebral levels at lumbar MR imaging. Radiology. 2011;259:196-202.

16. Kornreich L, Hadar H, Sulkes J, Gornish M, Ackerman J, Gadoth N. Effect of normal ageing on the sites of aortic bifurcation and inferior vena cava confluence: a CT study. Surg Radiol Anat. 1998;20:63-8.

17. Beregi JP, Mauroy B, Willoteaux S, Mounier-Vehier C, Rémy-Jardin M, Francke J. Anatomic variation in the origin of the main renal arteries: spiral CTA evaluation. Eur Radiol. 1999;9:1330-4.

18. Osborn AG, Koehler PR. Computed tomography of the paraspinal musculature: normal and pathologic anatomy. AJR Am J Roentgenol. 1982;138:93-8.

19. Crawford RJ, Cornwall J, Abbott R, Elliott JM. Manually defining regions of interest when quantifying paravertebral muscles fatty infiltration from axial magnetic resonance imaging: a proposed method for the lumbar spine with anatomical cross-reference. BMC Musculoskelet Disord. 2017;18:25.

20. Deng X, Zhu Y, Wang S, Zhang Y, Han H, Zheng D, et al. CT and MRI Determination of Intermuscular Space within Lumbar Paraspinal Muscles at Different Intervertebral Disc Levels. PLoS One. 2015;10:e0140315.

21. Philbrick KA, Yoshida K, Inoue D, Akkus Z, Kline TL, Weston AD, et al. What Does Deep Learning See? Insights From a Classifier Trained to Predict Contrast Enhancement Phase From CT Images. AJR Am J Roentgenol. 2018;211:1184-93.

22. Konin GP, Walz DM. Lumbosacral transitional vertebrae: classification, imaging findings, and clinical relevance. AJNR Am J Neuroradiol. 2010;31:1778-86.

Tables

Table 1 The Number of Images for Each Lumbar Disc Level

	Number of Images		
	Training set (tuning set)	Test set	
Level		Internal validation	External validation
L1-2	684 (98)	196	83
L2-3	893 (128)	255	112
L3-4	940 (135)	269	101
L4-5	877 (126)	251	89
L5-S1	842 (121)	240	108
Total	4236 (608)	1211	493

Table 2 Internal and External Validation Results of the Model for Lumbar Disc Level Classification

	Internal Validation			External Validation		
Level	Recall	Precision	AUC	Recall	Precision	AUC
L1-2	0.974	0.989	0.9998	0.952	0.975	0.9987
L2-3	0.980	0.962	0.9985	0.938	0.981	0.9978
L3-4	0.967	0.974	0.9990	0.941	0.888	0.9933
L4-5	0.976	0.972	0.9984	0.921	0.911	0.9969
L5-S1	0.988	0.992	0.9990	0.981	0.981	0.9997
Acc	0.977			0.947		

Table 3 Top-1 and Top-2 Accuracy of Human Reviewers

	1st session		2nd session	
	Top-1 Accuracy	Top-2 Accuracy	Top-1 Accuracy	Top-2 Accuracy
Reviewer 1				
Total	0.56	0.84	0.75	0.93
L1-2	0.55	0.80	0.90	1.00
L2-3	0.50	0.85	0.50	0.75
L3-4	0.55	0.95	0.75	0.90
L4-5	0.45	0.65	0.75	1.00
L5-S1	0.75	0.95	0.85	1.00
Reviewer 2				
Total	0.75	0.93	0.78	0.94
L1-2	0.90	1.00	0.80	0.95
L2-3	0.50	0.75	0.75	0.95
L3-4	0.75	0.90	0.85	1.00
L4-5	0.75	1.00	0.70	0.90
L5-S1	0.85	1.00	0.80	0.90
Reviewer 3				
Total	0.58	0.84	0.76	0.98
L1-2	0.40	0.50	0.90	1.00
L2-3	0.50	0.80	0.70	1.00
L3-4	0.85	0.90	0.80	0.90
L4-5	0.70	0.95	0.40	1.00
L5-S1	0.95	1.00	1.00	1.00

Figures

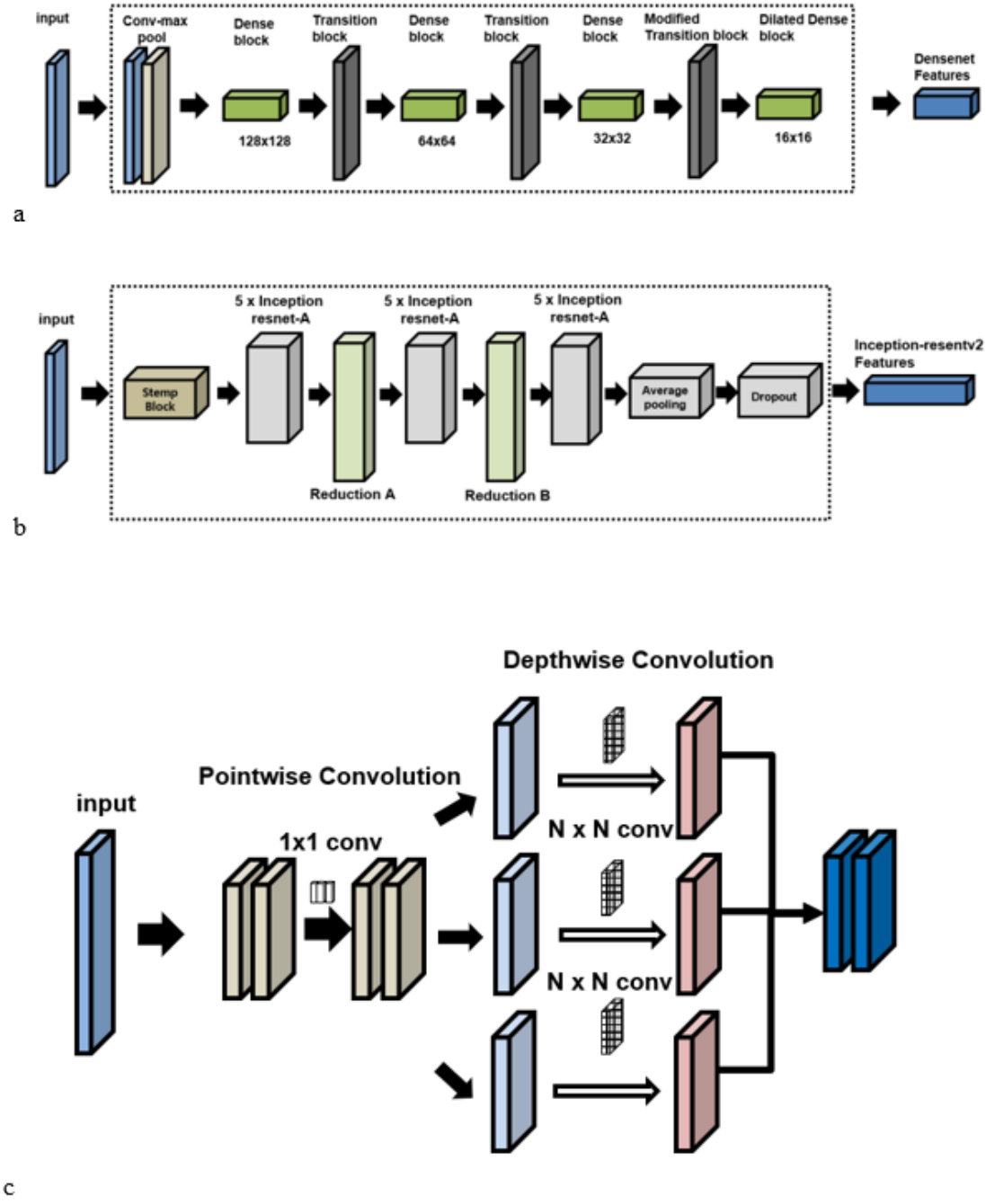

Figure 1

Network architecture of **a** DenseNet, **b** InceptionResNetV2, and **c** Xception.

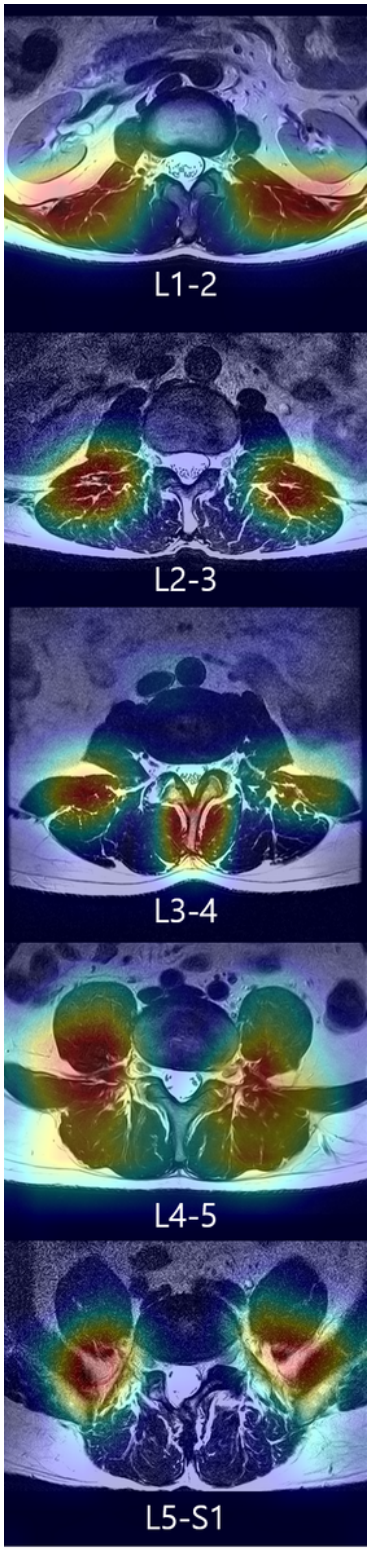

Figure 2

Examples of CAMs at each lumbar level. Note the color mapping around the paraspinal space where the kidney, retroperitoneal fat, psoas, quadratus lumborum, erector spinae, iliacus, and ilium are included according to the level.

Figure 3

Illustrations for lumbar muscle configurations for each level (Green: Psoas, Orange: Quadratus lumborum, Orange dotted line: Lateral border of quadratus lumborum, Blue: Erector spinae, Blue dotted line: Lateral border of erector spinae, Grey: Multifidus, Navy: Iliacus, and Black: Ilium).

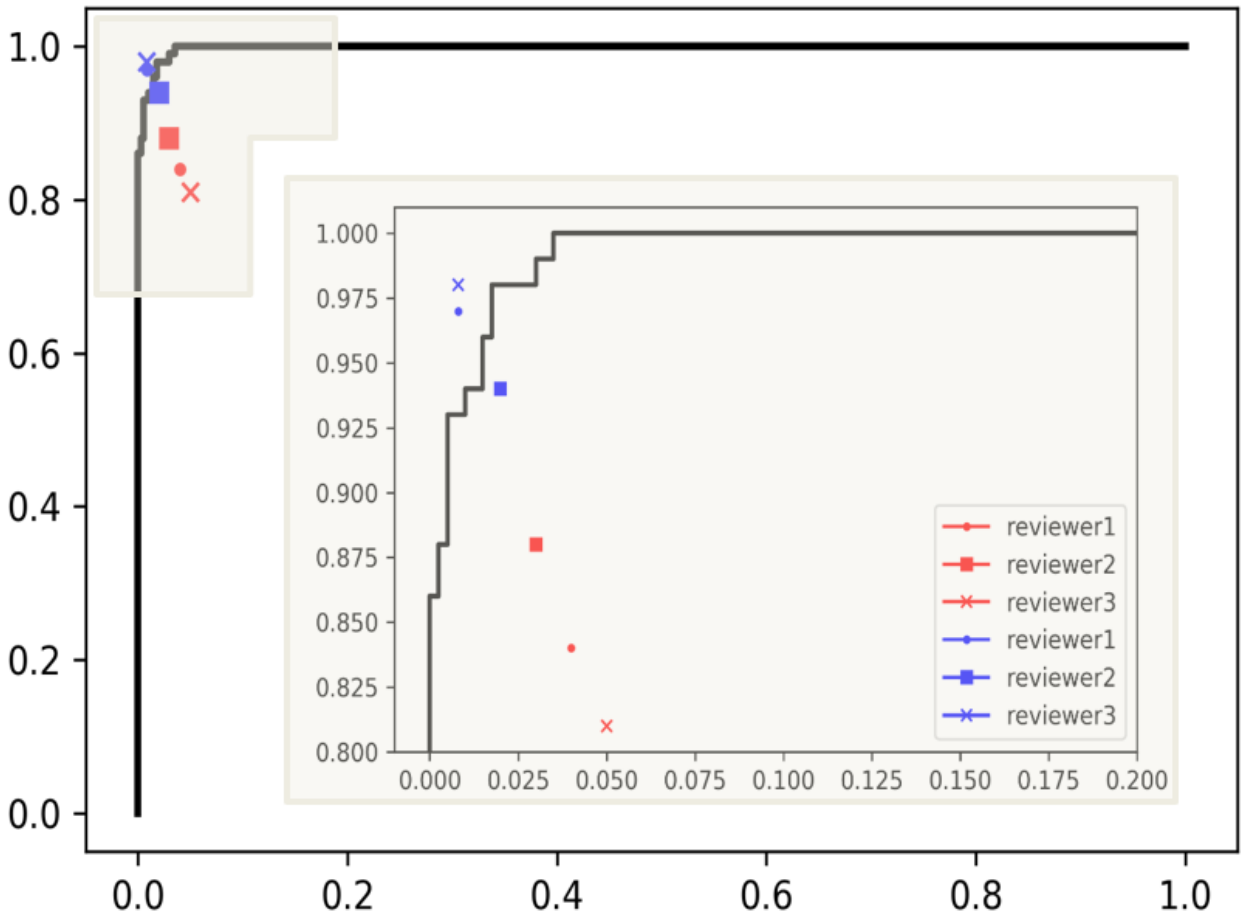

Figure 4

ROC curves for the model and the top-2 accuracy of human reviewers during first (red) and second (blue) sessions. Note the increase of accuracy in all reviewers in second session.