Classification of Sagittal Lumbar Spine MRI for Lumbar Spinal Stenosis Detection using Transfer Learning of a Deep Convolutional Neural Network

Friska Natalia[1] [0000-0002-3857-2405] and Sud Sudirman[2] [0000-0003-4083-0810]

[1] Department of Information System, Universitas Multimedia Nusantara, Tangerang, Indonesia
[2] Department of Computer Science, Liverpool John Moores University, Liverpool, UK
{friska.natalia@umn.ac.id;s.sudirman@ljmu.ac.uk}

Abstract. Analysis of sagittal lumbar spine MRI images remains an important step in automated detection and diagnosis of lumbar spinal stenosis. There are numerous algorithms proposed in the literature that can measure the condition of lumbar intervertebral discs through analysis of the lumbar spine in the sagittal view. However, these algorithms rely on using suitable sagittal images as their inputs. Since an MRI data repository contains more than just these specific images, it is, therefore, necessary to employ an algorithm that can automatically select such images from the entire repository. In this paper, we demonstrate the application of an image classification method using deep convolutional neural networks for this purpose. Specifically, we use a pre-trained Inception-ResNet-v2 model and retrain it using two sets of T1-weighted and T2-weighted images. Through our experiment, we can conclude that this method can reach a performance level of 0.91 and 0.93 on the T1 and T2 datasets, respectively when measured using the accuracy, precision, recall, and f1-score metrics. We also show that the difference in performance between using the two modalities is statistically significant and using T2-weighted images is preferred over using T1-weighted images.

Keywords: Medical Image Classification, Magnetic Resonance Imaging, Lumbar Spinal Stenosis, Transfer Learning, Deep Convolutional Neural Networks.

1 Introduction

Millions of people around the world suffer from chronic lower back pain. It is a chronic disease that is detrimental to the health, social life, and employment of its sufferers. Lumbar spinal stenosis (LSS), a narrowing of the lumbar spinal canal that is resulted from bone or soft tissue inflammation, is one of the most common causes of chronic lower back pain. The pressure on the spinal nerve roots caused by this inflammation is responsible for the pain that is experienced by patients with LSS. A diagnosis of LSS in these patients is often carried out through an inspection of Magnetic Resonance Imaging (MRI) of the patients' lumbar spine by an expert radiologist. Recent advances in medical image processing allow the application of computer algorithms to help radiologists carried out this procedure. Some of these algorithms work only on mid-sagittal MRI images [1–3] whereas some others work only on traverse images that cut through the mid-height of an intervertebral disc (IVD) [4–8]. Since a patient's data repository contains more than just these specific images, the process to select suitable images as inputs to these algorithms is often done manually. To make the whole process more automated it is, therefore, essential that a reliable algorithm to select such images is applied as well.

The objective of this study is to design a suitable solution for selecting all suitable sagittal images from a database of sagittal images, that can be used as inputs to other algorithms that diagnose LSS. Our solution is using a pre-trained Deep Convolutional Neural Network (DCNN) model that has been developed for general image classification and retraining it to make the model suitable for medical image classification.

2 Literature Review

The task of selecting medical images that possess certain characteristics from a collection of medical images falls into the category of image classification, which is a fundamental task in computer vision that categorizes images into one of several predefined classes. The traditional approach in image classification involves two stages, with the first being the extraction of relevant information from the images via the calculation of low-level handcrafted features [9–11]. This is then followed by a classification of the calculated features using trainable classifiers. Despite the success of this approach, it has a significant drawback when used in a wider image classification problem since the features are often task-dependent. In other words, the handcrafted image features that are optimized for a particular task often perform poorly when used in a different task, and the accuracy of the classification is very dependent on the design of these features.

The first DCNN model was originally proposed to overcome the problems associated with the traditional approach of image classification by allowing learning of such features through forward and backpropagation of information in a series of convolutional and non-linear neural network layers [12, 13]. This approach however has a significant practical problem due to its high computational cost and the amount of data it needs to create a general set of features applicable for typical images. But only recently, in the advent of huge computational power from using Graphics Processing Units and the large-scale acquisition and availability of image data resulting from the proliferation of the internet and social media, that this approach has gained renewed attention from the research community which generated faster and better algorithms [14].

The popularity of DCNN compared to the more traditional approach of image classification is that the features are no longer manually handcrafted but instead are automatically learnable. These features are sufficiently general that they can be used in many different types of image classification tasks through Transfer Learning. Transfer Learning is a widely accepted method in Machine Learning where a model developed for a task, often by training using a very large dataset, is used as a starting point for developing another model to solve a different task. This approach is less data-and-label-dependent than other more traditional machine learning approaches and gains popularity recently, especially when a deep learning model is concerned because developing one from scratch requires a vast amount of computational and time resources [15]. One example application of a bespoke DCNN for medical image classification is CheXNet [16]. It is a 121-layer DCNN trained on a dataset with more than 100,000 frontal-view chest X-rays and is claimed to achieve a better performance than the average performance of four radiologists.

3 Material and Method

The material used in this research is taken from our Lumbar Spine MRI Dataset which is available publicly [6, 17]. This dataset contains anonymized clinical MRI studies of 515 patients with symptomatic back pains. The dataset consists of 48,345 T1-weighted and T2-weighted traverse and sagittal images of the patients' lumbar spine. The images were taken using a 1.5 Tesla Siemens Magnetom Essenza MRI scanner mostly when the patients were in Head-First-Supine position. From the entire dataset, we took 19,176 sagittal images for this study. This consists of 9,903 T1-weighted and 9,273 T2-weighted images. The summary of the technical information of the scanning parameters carried out when recording these images is provided in **Table 1**.

Table 1. Sagittal MRI Scanning Parameters

Sequence Types	**T1-weighted**	**T2-weighted**
Number of Echoes (ETL)	3	15 to 18
Repetition Time (ms)	330 to 926	3190 to 4000
Echo Time (ms)	9.2 to 12.0	67.0 to 96.0
Slice Thickness (mm)	3.0 to 4.0	3.0 to 5.0
Spacing Between Slices (mm)	3.3 to 4.8	3.3 to 6.5
Field of View (mm)	280	280
Matrix (Freq. x Phase)	100%	100%
Imaging Frequency (MHz)	63.7	63.7
Number of Phase Encoding Steps	288 to 540	408 to 544
Scanning Sequence	SE	SE
Sequence Variant	SK\SP\OSP	SK\SP\OSP
Scan Options	SAT1\FS	SAT1
Number of Averages	1, 2, 3 or 4	2
Echo Train Length	3	15 or 17
Percent Sampling	50 to 70	66 to 90
Percent Phase Field of View	100	100
Pixel Bandwidth	150 or 235	150, 160 or 195
Flip Angle	150	150

Based on the advice from an expert radiologist, we categorize the sagittal images into two groups, namely *Mid* and *Lateral* groups, corresponding to their suitability for the analysis of the IVD. The first group consists of midsagittal slices that divide the left and right sides of an IVD into two symmetrical parts. It also contains neighboring sagittal slices to the midsagittal slice which still show clear cross-sectional views of the lumbar vertebrae and discs. The second group consists of other sagittal slices that do not meet this requirement. An example of this grouping is shown in **Fig. 1** and **Fig. 2**. The first figure shows the intersection lines of fifteen sagittal images with the shown traverse image whereas the second figure shows a collage of the fifteen sagittal images. The midsagittal slice is identified as slice number 8 in the figures. That slice and four nearest adjacent slices, highlighted in yellow in the figures, are then put in the Mid group. The other ten sagittal slices, highlighted in red, are then put in the Lateral group. The population distribution of these classes for both T1-weighted and T2-weighted datasets is shown in **Table 2**.

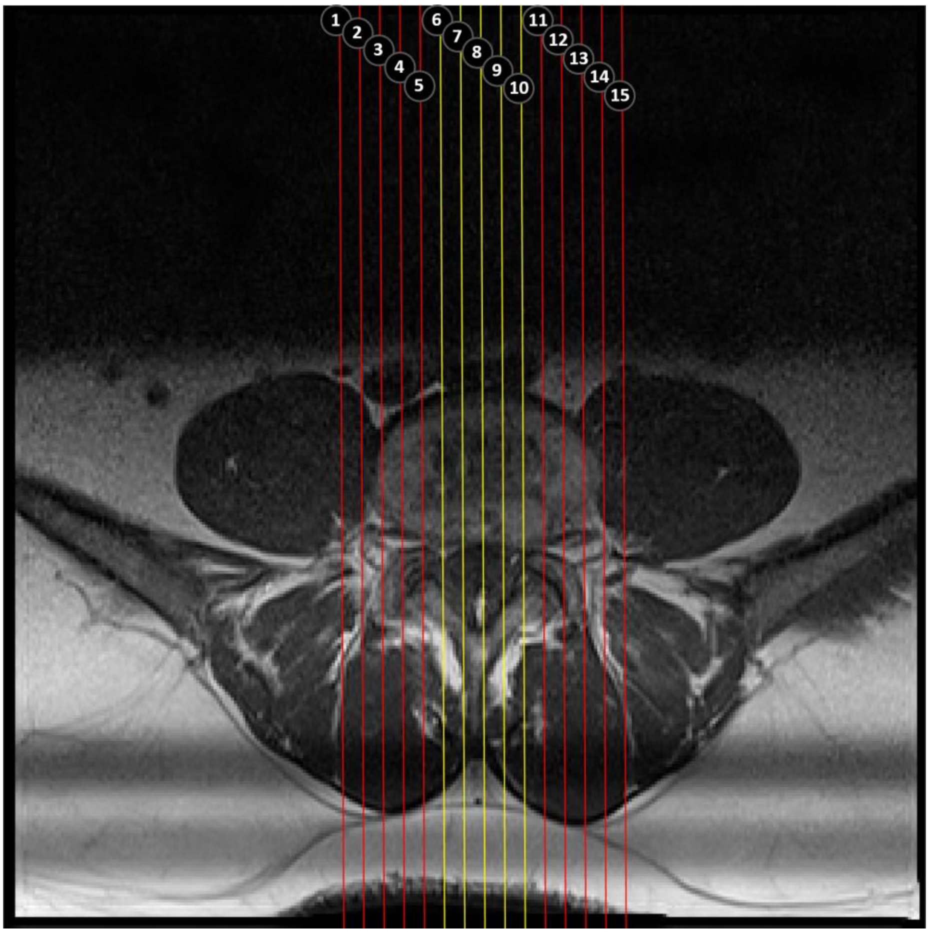


Fig. 1. A traverse image of a lumbar spine showing its intersection lines with fifteen sagittal images that are shown in **Fig. 2**. The yellow lines mark the sagittal images that are categorized in the *Mid* group and the red lines mark those categorized in the *Lateral* group.

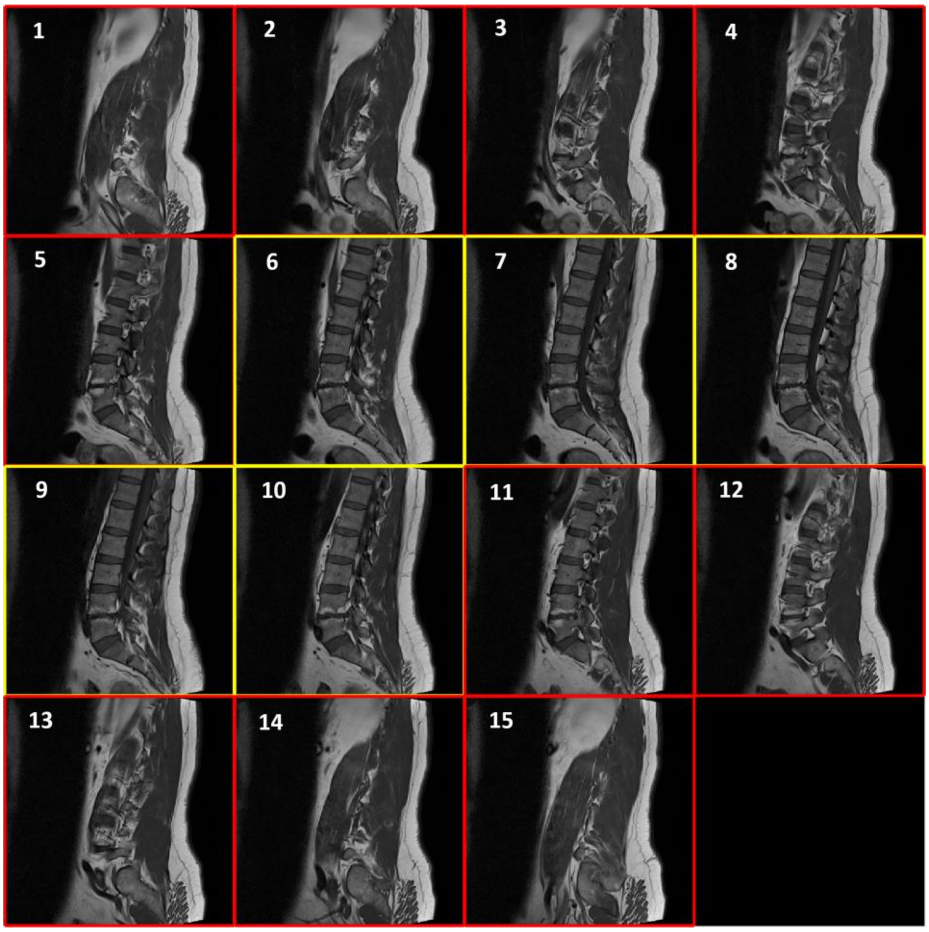


Fig. 2. An example of fifteen sagittal images of a patient's lumbar spine which intersections with a traverse image are shown in **Fig. 1**. The yellow rectangles highlight the sagittal images that are put in the *Mid* group and the red rectangles highlight those put in the *Lateral* group.

Table 2. Class Distribution in the Dataset

Class \ Sequence Types	T1-weighted	T2-weighted
Mid	4,667	5,236
Lateral	4,316	4,957
Total	9,903	9,273

We performed validation of this dataset by checking the *Slice Location* information that is stored as part of the DICOM metadata. The value of the slice location attribute of a DICOM image metadata is the relative position, expressed in mm, of the image plane in the patient 3D axis system. In the absence of the slice location attribute in the DICOM image metadata, we calculated its value from the *Image Position* and *Image Orientation* attributes of the DICOM image metadata using the following technique.

Let us denote p as the 3D coordinate of the top-left point on the plane specified in the Image Position attribute. Also, let u and v be two orthogonal unit vectors that lie on the sagittal plane as specified in the Image Orientation attribute. The slice location s of the plane is defined as the shortest distance from the origin point of the patient coordinate axes to the plane and can be calculated as the dot product $n \cdot p$, where n is the unit vector perpendicular to the plane which can be calculated as the cross product $u \times v$. The directions of the patient coordinate axes themselves are defined by the patient's orientation. The positive direction of the x-axis goes from the right-hand side of the patient to the left-hand side. The positive direction of the y-axis goes from the anterior side of the patient to the posterior side, whereas the z-axis is increasing from the feet toward the head of the patient. The position and orientation of the elements used in this calculation with respect to the patient's coordinate axes are illustrated in **Fig. 3**.

Using either the stored or the calculated slice location information, we could confirm that the sagittal planes in the Mid groups of the dataset are within the acceptable range of normal IVD width [18].

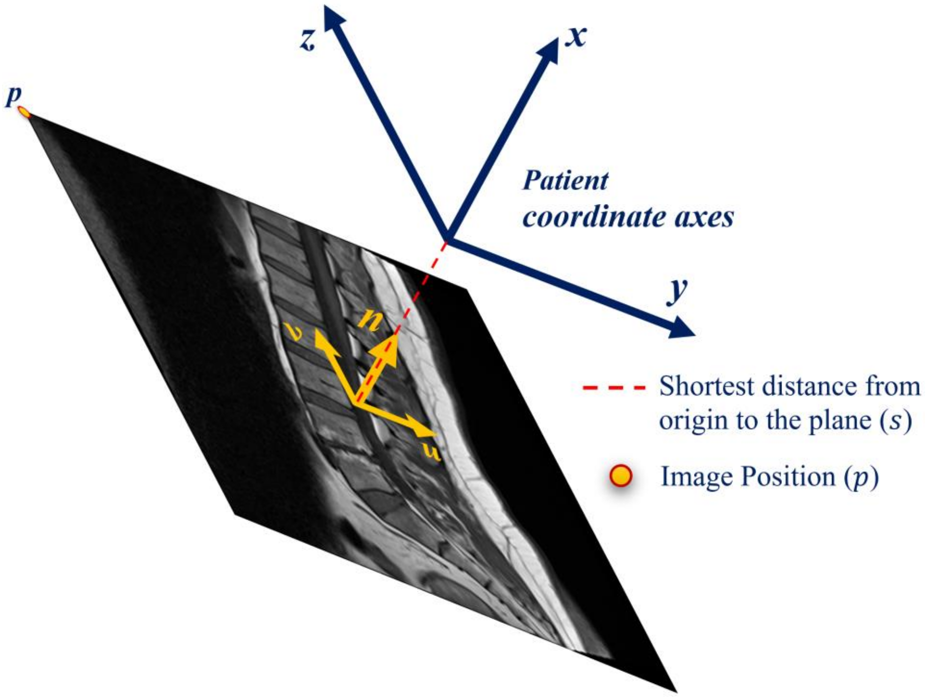
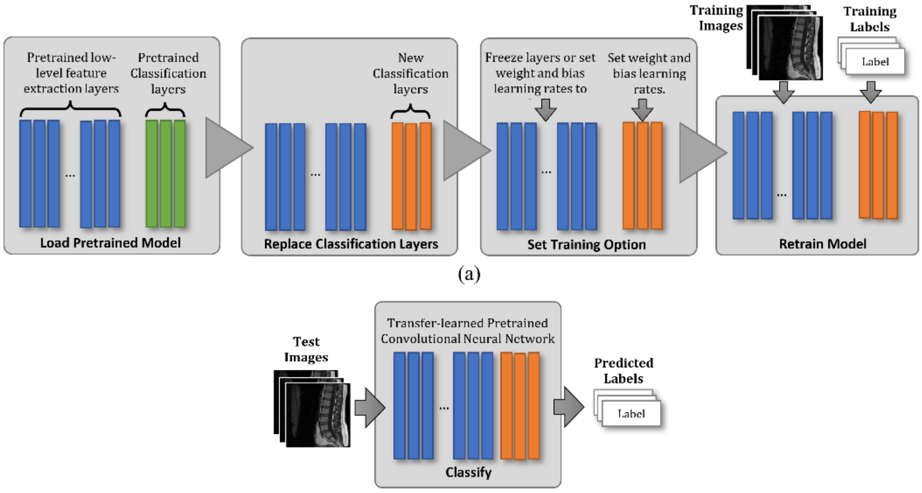


Fig. 3. An illustration of the position and orientation of the elements used to determine the Slice Location s with respect to the patient's coordinate axes.

Once the dataset has been developed, we then use it to train an Inception-ResNet-v2 model. Inception-ResNet-v2 [19] is a convolutional neural architecture that improves on the Inception family of architectures by incorporating the ResNet approach of using residual connections to replace the filter concatenation stage of the Inception architecture. Inception-Resnet-v2 is one of the newer generations of deep convolutional neural networks that is gaining popularity in the classification and annotation of medical images [20]. We adopted a methodology called Transfer Learning that transfers the learned network parameters of the model that was pre-trained using the ImageNet database [21] and then retrain the whole model after replacing its classification layers using the new dataset. The flowchart of the process is illustrated in **Fig. 4**.


Pretrained low-level feature extraction layers
Pretrained Classification layers
...
Load Pretrained Model
New Classification layers
Replace Classification Layers
Freeze layers or set weight and bias learning rates to
Set weight and bias learning rates.
Set Training Option
Training Images
Training Labels
Label
Retrain Model
Test Images
Transfer-learned Pretrained Convolutional Neural Network
Classify
Predicted Labels
Label

(a)

(b)

Fig. 4. A flowchart illustrating a) the retraining process of a pre-trained deep convolutional neural network and b) the classification of the images in the test dataset using the retrained model.

To evaluate the suitability of the method, we use four performance metrics namely overall accuracy (A), precision (P), recall (R), and f1-score (F). Using the standard notations of true positive (tp), true negative (tn), false positive (fp), and false negative (fn), the metrics are calculated as:

$$A = \frac{tp+tn}{tp+tn+fp+fn} \tag{1}$$

$$P = \frac{1}{C}\sum_{i}^{C} P_i \tag{2}$$

$$R = \frac{1}{C}\sum_{i}^{C} R_i \tag{3}$$

$$F = \frac{2}{C}\sum_{i}^{C} \frac{P_i \cdot R_i}{P_i + R_i} \tag{4}$$

where $i \in \{1,2\}$ is the index of the i^{th} class and P_i, R_i and F_i are the class precision, class recall, and class f1-score of the i^{th} class, respectively and are defined as:

$$P_i = \frac{tp_i}{tp_i + fp_i} \tag{5}$$

$$R_i = \frac{tp_i}{tp_i + fn_i} \tag{6}$$

$$F_i = 2 \times \frac{P_i \cdot R_i}{P_i + R_i} \tag{7}$$

We applied the aforementioned transfer learning approach on the T1-weighted dataset and T2-weighted dataset separately and compare the two results. To allow statistical analysis of the results, we carried out the process twenty times, each with different subsets of training and test data.

4 Experimental results, discussion, and analysis

For each experiment repeat, each dataset was randomly split into two sub-groups namely the training and test dataset by an 80:20 ratio. When training the Inception-ResNet-v2 model, the training set was further split into a smaller training set and a validation set. The validation set was used solely to provide an unbiased estimate of the trained model's accuracy during training and was not used to adjust the model's weights or biases. They were instead adjusted by backpropagating the errors calculated using the training set which process was optimized using the stochastic gradient descent with momentum technique. The training is done in small batches of ten samples for up to four epochs. The order of the samples in the training set was shuffled in every epoch to prevent the model from learning the order of the samples. The initial learning rate of the transferred (feature extraction) layers was set to 10^{-4} and that of the new classification layer was set 20 times higher. The training duration for the twenty repeats ranges from 750 to 765 minutes when run on a Windows 10 PC with an Intel i9-7900X CPU @ 3.30GHz with 128 GB RAM and four NVIDIA Titan XP GPUs. A plot showing the validation accuracy during the training process is shown in **Fig. 5**. As can be seen from this figure, the accuracy values for both datasets plateau roughly before the final epoch. The final recorded validation accuracy for the T1 and T2 datasets is 0.92 and 0.91, respectively.

Once the training process is completed, it is used to predict the classes of each sample in the test set. These predicted classes are then used to calculate the four performance metrics defined in Eq. (1) to (4). The results are shown as box plots in **Fig. 6** and **Fig. 7**. The figures show the minimum, the maximum, the sample median, and the first and third quartiles of each metric for ease of visual inspection of the classification performance.

We performed a statistical test to show that the difference in the classification results produced by using the T1 and T2 datasets is statistically significant. We use the Kolmogorov-Smirnov (KS) test [22] and the Bartlett's test for Homogeneity of Variance [23] to check if the populations follow a Normal distribution and if they have identical variances, respectively. Satisfying both tests allows us to use the standard t-test hypothesis testing to prove or disprove the null hypothesis. The null hypothesis here being there is no difference between the two results. If any one of the two populations does not satisfy the KS and Bartlett's tests then we use the Welch's t-test [24] instead.

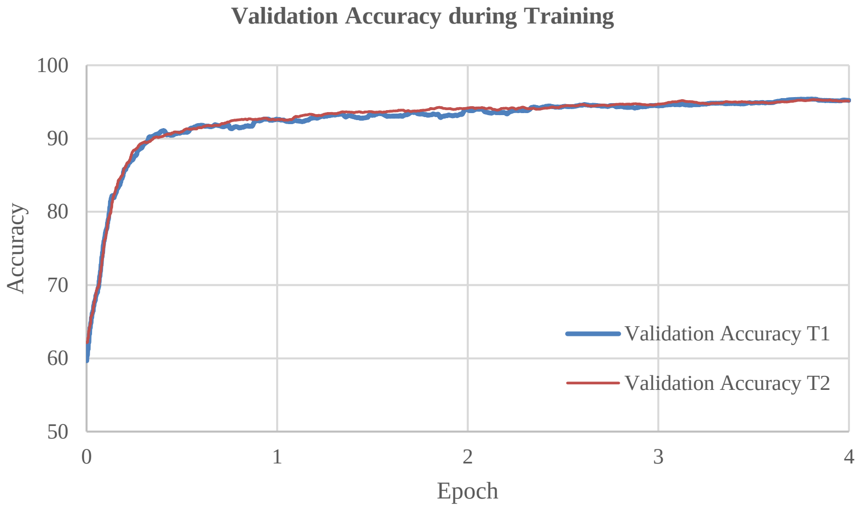


Fig. 5. A history of the validation accuracy during the training process.

The results of the statistical test are provided in **Table 3**. From the table, we can conclude that the classification performance using the T2-weighted dataset as measured using four performance metrics is statistically better than that using the T1-weighted dataset.

Table 3. Summary of the statistical tests

Metrics	**T1 Mean**	**T2 Mean**	**Difference**	**t-test Type**	***p*-value**
Accuracy	0.90	0.93	0.03	t-test	< 0.05
Precision	0.91	0.93	0.02	Welch's t-test	< 0.05
Recall	0.90	0.93	0.03	t-test	< 0.05
F1-Score	0.90	0.93	0.03	t-test	< 0.05

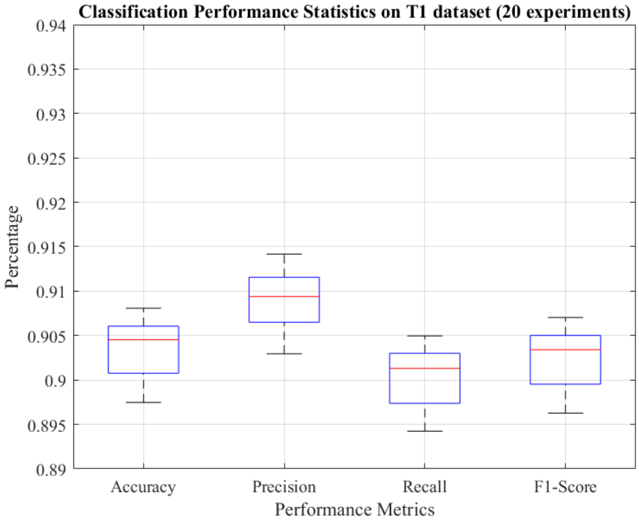


Fig. 6. A box plot showing the statistics of classification performance using the T1 dataset.

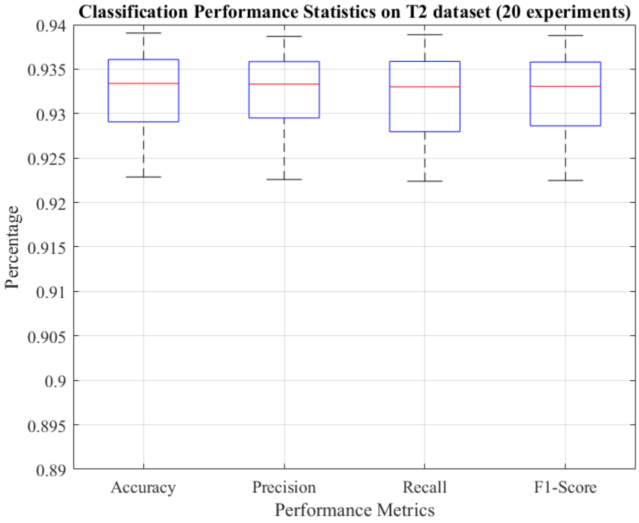


Fig. 7. A box plot showing the statistics of classification performance using the T2 dataset.

5 Conclusion

We have detailed a method to automatically select suitable sagittal images from a collection of lumbar spine sagittal MRI images that can be used as inputs to other algorithms that detect abnormalities in lumbar intervertebral discs. The method is based on transfer learning of a pre-trained Inception-Resnetv2 deep convolutional neural network model. We experimented with this method on a subset of a publicly available lumbar spine MRI dataset that consists of 9,903 T1-weighted and 9,273 T2-weighted MRI images. Our experimental results show that the average classification performance is 0.90 and 0.93 on the T1 dataset and the T2 dataset, respectively. We have also shown through statistical analysis of our experimental results that the classification performance using the T2 dataset is statistically better than that using the T1 dataset. We plan in the future to expand the method to include a wider range of deep convolutional neural network models as well as investigate the effect of applying the dimensionality reduction method on the generalization performance of this approach. We will also explore the possibility of applying the method to classify lumbar spine traverse MRI images.

References

1. Davarpanah, S.H., Liew, A.W.C.: Spatial Possibilistic Fuzzy C-Mean Segmentation Algorithm Integrated with Brain Mid-sagittal Surface Information. Int. J. Fuzzy Syst. 19, 591–605 (2017). https://doi.org/10.1007/s40815-016-0247-0.
2. Alomari, R.S., Ghosh, S., Koh, J., Chaudhary, V.: Vertebral Column Localization, Labeling, and Segmentation. In: Li, S. and Yao, J. (eds.) Spinal Imaging and Image Analysis. pp. 193–229. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-12508-4_7.
3. Ghosh, S., Chaudhary, V.: Supervised methods for detection and segmentation of tissues in clinical lumbar MRI. Comput. Med. Imaging Graph. 38, 639–649 (2014). https://doi.org/https://doi.org/10.1016/j.compmedimag.2014.03.005.
4. Natalia, F., Meidia, H., Afriliana, N., Young, J.C., Yunus, R.E., Al-Jumaily, M., Al-Kafri, A., Sudirman, S.: Automated measurement of anteroposterior diameter and foraminal widths in MRI images for lumbar spinal stenosis diagnosis. PLoS One. 15, 1–27 (2020). https://doi.org/10.1371/journal.pone.0241309.
5. Paul, C.P.L., Smit, T.H., de Graaf, M., Holewijn, R.M., Bisschop, A., van de Ven, P.M., Mullender, M.G., Helder, M.N., Strijkers, G.J.: Quantitative MRI in early intervertebral disc degeneration: T1rho correlates better than T2 and ADC with biomechanics, histology and matrix content. PLoS One. 13, e0191442 (2018).
6. Al-Kafri, A.S., Sudirman, S., Hussain, A., Al-Jumeily, D., Natalia, F., Meidia, H., Afriliana, N., Al-Rashdan, W., Bashtawi, M., Al-Jumaily, M.: Boundary Delineation of MRI Images for Lumbar Spinal Stenosis Detection Through Semantic Segmentation Using Deep Neural Networks. IEEE Access. 7, 43487–43501 (2019). https://doi.org/10.1109/ACCESS.2019.2908002.
7. Zhang, Q., Bhalerao, A., Hutchinson, C.: Weakly-supervised evidence pinpointing and description. In: International Conference on Information Processing in Medical Imaging. pp. 210–222 (2017).
8. Al Kafri, A.S., Sudirman, S., Hussain, A.J., Al-Jumeily, D., Fergus, P., Natalia, F., Meidia, H., Afriliana, N., Sophian, A., Al-Jumaily, M., others, Al-Kafri, A.S., Sudirman, S., Hussain, A.J., Al-Jumeily, D., Fergus, P., Natalia, F., Meidia, H., Afriliana, N., Sophian, A., Al-Jumaily, M., Bashtawi, M., Al-Rashdan, W.: Segmentation of lumbar spine MRI images for stenosis detection using patch-based pixel classification neural network. In: 2018 IEEE

Congress on Evolutionary Computation (CEC). pp. 1–8. , Rio de Janeiro (2018).

9. Baloch, S.H., Krim, H.: Flexible skew-symmetric shape model for shape representation, classification, and sampling. IEEE Trans. image Process. 16, 317–328 (2007).
10. Song, Y., Cai, W., Zhou, Y., Feng, D.D.: Feature-based image patch approximation for lung tissue classification. IEEE Trans. Med. Imaging. 32, 797–808 (2013).
11. Koitka, S., Friedrich, C.M.: Traditional Feature Engineering and Deep Learning Approaches at Medical Classification Task of ImageCLEF 2016. In: CLEF (Working Notes). pp. 304–317 (2016).
12. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems. pp. 396–404 (1990).
13. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Comput. 1, 541–551 (1989).
14. Rawat, W., Wang, Z.: Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput. 29, 2352–2449 (2017).
15. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A Comprehensive Survey on Transfer Learning. Proc. IEEE. (2020). https://doi.org/10.1109/JPROC.2020.3004555.
16. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., others: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv Prepr. arXiv1711.05225. (2017).
17. Sudirman, S., Kafri, A. Al, Natalia, F., Meidia, H., Afriliana, N., Al-Rashdan, W., Bashtawi, M., Al-Jumaily, M.: Lumbar Spine MRI Dataset, https://data.mendeley.com/datasets/k57fr854j2/2, last accessed 2019/05/13. https://doi.org/10.17632/k57fr854j2.2.
18. Zhou, S.H., McCarthy, I.D., McGregor, A.H., Coombs, R.R.H., Hughes, S.P.F.: Geometrical dimensions ,of the lumbar vertebrae - Analysis of data from digitised CT images. Eur. Spine J. 9, 242–248 (2000). https://doi.org/10.1007/s005860000140.
19. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv Prepr. arXiv1602.07261. (2016).
20. Pelka, O., Nensa, F., Friedrich, C.M.: Annotation of enhanced radiographs for medical image retrieval with deep convolutional neural networks. PLoS One. 13, 1–18 (2018). https://doi.org/10.1371/journal.pone.0206229.
21. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255 (2009).
22. Massey Jr, F.J.: The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46, 68–78 (1951).
23. Snedecor, G.W., Cochran, W.G.: Statistical Methods. Wiley India (2014).
24. Mendenhall, W.M., Sincich, T.L.: Statistics for Engineering and the Sciences. CRC Press (2016).