MRI-Based Surgical Planning for Lumbar Spinal Stenosis

Gabriele Abbati[1], Stefan Bauer[2], Sebastian Winklhofer[3], Peter J. Schüffler[4], Ulrike Held[5], Jakob M. Burgstaller[5], Johann Steurer[5], and Joachim M. Buhmann[2(✉)]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
gabb@robots.ox.ac.uk
[2] Department of Computer Science, ETH Zürich, Zürich, Switzerland
{stefan.bauer,jbuhmann}@inf.ethz.ch
[3] Neuroradiology, University Hospital Zürich, Zürich, Switzerland
sebastian.winklhofer@usz.ch
[4] Computational Pathology, Memorial Sloan Kettering Cancer Center, New York, USA
schueffp@mskcc.org
[5] Horten Centre for Patient Oriented Research and Knowledge Transfer, University of Zürich, Zürich, Switzerland
{ulrike.held,jakob.burgstaller,johann.steurer}@usz.ch

Abstract. The most common reason for spinal surgery in elderly patients is lumbar spinal stenosis (LSS). For LSS, treatment decisions based on clinical and radiological information as well as personal experience of the surgeon show large variance. Thus a standardized support system is of high value for a more objective and reproducible decision. In this work, we develop an automated algorithm to localize the stenosis causing the symptoms of the patient in magnetic resonance imaging (MRI). With 22 MRI features of each of five spinal levels of 321 patients, we show it is possible to predict the location of lesion triggering the symptoms. To support this hypothesis, we conduct an automated analysis of labeled and unlabeled MRI scans extracted from 788 patients. We confirm quantitatively the importance of radiological information and provide an algorithmic pipeline for working with raw MRI scans. Both code and data are provided for further research at www.spinalstenosis.ethz.ch.

Keywords: Machine learning · Deep learning · Lumbar spinal stenosis

1 Introduction

The lumbar spine consists of the five vertebrae (*levels* or *segments*) L1–L5. The vertebral discs connect adjacent levels and are denoted as L1/L2, L2/L3, L3/L4,

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-66179-7_14) contains supplementary material, which is available to authorized users.

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 116–124, 2017.
DOI: 10.1007/978-3-319-66179-7_14

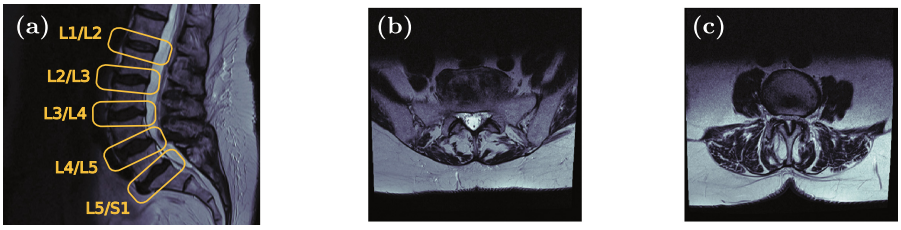

Fig. 1. Examples of T2-weighted MRI. **(a)** The five segments are highlighted yellow in a sagittal scan. **(b)** Axial scan of a patient without symptoms and without narrowing of the spinal channel (white spot in the center). **(c)** Example with extreme narrowing.

L4/L5, L5/S1, where S1 is the first vertebra of the underlying *sacral region* (see Fig. 1). Lumbar Spinal Stenosis (LSS) is the most common indicator for spine surgery in patients older than 65 years [1]. The North American Spine Society defines LSS as "[...] diminished space available for the neural and vascular elements in the lumbar spine secondary to degenerative changes in the spinal canal [...]" [2]. Symptoms such as gluteal and/or lower extremity pain and/or fatigue might occur, possibly associated with back pain. Magnetic resonance imaging (MRI, illustrated in Figs. 1(b) and (c)) and the patient's clinical course contribute to diagnosis and treatment formulation. When conservative treatments such as physiotherapy or steroid injections fail, decompression surgery is frequently indicated [1]. Depending on the clinical presentation of the patient and corresponding imaging findings, surgeons decide which segments to operate.

This decision process exhibits wide variability [3,4], while associations between imaging and symptoms are still not entirely clear [5]. These issues motivate the search for objective methods to help in surgery planning. Since the definition of LSS implies anatomic abnormalities, MRI plays a fundamental role in diagnosis [6]. Andreisek *et al.* [7] identified 27 radiological criteria and parameters for LSS. However, correlations between imaging procedures, clinical findings and symptoms is still unclear, and research efforts show contradictory results [8,9].

This paper comprehensively determines the important role of radiological parameters in LSS surgery planning, in particular by modeling surgical decision-making: to the best of our knowledge, no machine learning approach has been applied in this direction before. In Sect. 2, we automatically predict surgery locations with 22 manual radiological features comparing five different classifiers. We obtain accuracies of 85.4% using random forests and show features associated with stenosis are commonly chosen by all classifiers. In Sect. 3, the highly heterogeneous MRI dataset is preprocessed and a convolutional neural network and convolutional autoencoder are trained to accomplish the same task as before, without any knowledge of the underlying structure of LSS. The automatic preprocessing of raw MRI scans is a key contribution of this work and code with examples will be released in the final version. Both algorithms achieve accuracies of 69.8% and 70.6%, respectively, in mimicking surgeons' decisions, showing the

high relevance of radiological features in LSS treatment. Finally, we conclude with a discussion in Sect. 4.

2 Surgical Prediction from Numerical Dataset

The Numerical Dataset. Radiological T1-weighted and T2-weighted scans from 788 LSS patients have been collected in a multi-center study by Horten Zentrum (Zürich, CH). For every segment and patient, radiologists manually scored 6 quantitative features (e.g. area of spinal canal in mm^2) and 16 qualitative features (e.g. severity grade of compromise of a given vertebral region) known to be most relevant for assessing stenosis (a subset of the ones identified in [7]), forming the "numerical" dataset. Notice only one reading per image is available. A description of the features can be found in the Supplement (A.1).
431 of 788 patients underwent surgery. The Numeric Rating Scale (NRS) [10] for pain assessment was employed to understand whether the intervention improved a certain patient's condition or not. NRS differences larger than 2 points before and six months after surgery were considered as improvement, as failure otherwise. In total, 321 of 431 patients exhibited improvement of NRS after surgery. As there is no information gain from unsuccessful operations, the following analysis addresses the subset of the 321 improved patients, yielding a total of 1385 segments as data points.

Methods. We consider every segment independently as a data vector $\mathbf{x}$ consisting of its 22 feature values. The target is represented by a binary variable y (to operate/not to operate). This binary classification framework is tackled with the following algorithms: K-nearest neighbors (KNN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), support vector machine (SVM), and random forest (RF). Implementations from the *scikit-learn* [11] library are employed. The area under the receiver operating characteristic (ROC) curve is a natural choice for evaluating binary classifiers' performances, and it is combined with 20-fold cross validation.
To evaluate the influence of individual features, forward selection and backward selection are employed to choose the best 3, 5, 8, 10, 12, 15 and 18 features: with 5 different classifiers, a single feature can be chosen for a total of 70 times (7 sets $\times$ 5 classifiers $\times$ 2 algorithms). Thus we can evaluate how often a feature is considered to be among the most relevant ones for surgery prediction. This procedure is again validated through 20-fold cross validation.

Results. For parameter-optimized binary classifiers, box plots describing the area under the ROC curve (AUC) obtained with 20-fold cross validation are shown in Fig. 2(a). The best results are achieved with an optimized random forest classifier: the mean over the AUC returned by the cross validation is 85.4%, with a standard deviation of 3.26%. The precision obtained here is particularly significant if we consider the relatively low agreement rates between doctors in determining treatments for LSS [12,13]. Feature selection indicates that

`SegCentralZone` (assesses the compromise of the central zone of the vertebra), `SegCSarea` (area of the section of the spinal cord in mm^2) and `SegFluidSign` (relation from fluid to cauda equina) as the most important features for assessing stenosis: these are chosen in 88.57%, 87.14% and 70.00%, respectively, of the total trials with feature selection algorithms (total ranking in Fig. 2(b)). All three features are known to be strongly related to spinal stenosis [7]. The results show that radiological data actually helps in assessing LSS and planning surgical treatments.

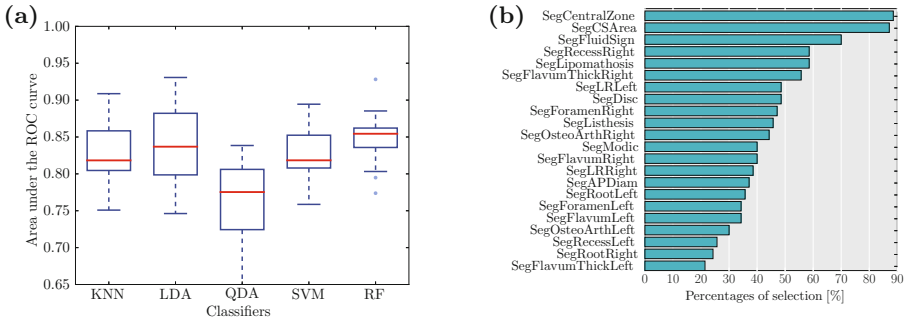

Fig. 2. Summary of the classifiers for segmental surgery prediction. **(a)** The box plots of the 20-fold cross validation. All classifiers show a strong signal between radiological data and surgical treatments. **(b)** Feature ranking as described in the text. The three most important features are `SegCentralZone`, `SegCSarea` and `SegFluidSign`.

3 Surgical Prediction from Radiological Images

Fully automated MRI-based surgery planning would be a helpful tool, as it can substantially speed up the process by skipping manual scoring while reducing the variability of human assessment. Therefore, we aim to directly learn features from raw MRI scans.

The Image Dataset. The above described dataset of 788 LSS patients contains a great variety of T1-weighted and T2-weighted sagittal, coronal and axial series scans (see Fig. 3 for four typical examples). Since the images come from seven different institutions, the dataset is heterogeneous: not all types of MRI scans listed above are always available, and often only a small subset of the segments is accessible. Further, different machines vary in resolution (from 320×320 to 1024×1024 pixels) and scanning frequency (0.2 to 1 scan/mm).

To keep the same segment-wise approach as before, we decide to employ only the T2-weighted axial scans (e.g. Fig. 3(c)), as they picture the whole lumbar spine and can be easily chopped into single segment sub-series. T2-weighted imaging pictures the spinal canal white in contrast to T1-weighted images, in

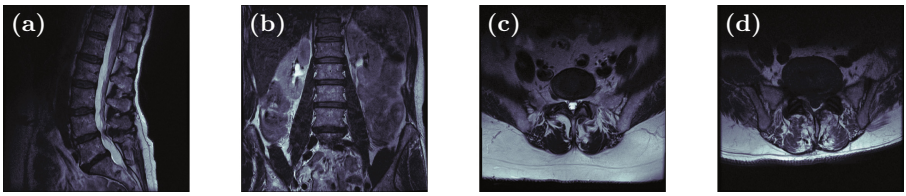

Fig. 3. Typical examples of the different MRI scans: (a)–(c) T2-weighted (a, sagittal; b, coronal; c, axial), (d) T1-weighted axial.

which the canal is dark and hardly visible (Fig. 3(d)). Further, T2-weighted axial scans are the most common series in the dataset. The image dataset includes the same 321 operated patients with improved NRS.

Image Preprocessing & Data Augmentation. All images are cropped and resized to 128×128 pixels, in order to keep the central section. Because of the various scanning frequencies, we then linearly interpolate to a desired number of equally spaced slices: to sufficiently describe the vertebral disc, yet keep the data structure simple, we use four subimages for each segment. **We employed following data augmentation**: rotation by a random angle $\alpha \in [-10°; 10°]$; sagittal mirroring; inversion of the order of the slides (since the MRI machine can scan upwards or downwards); application of random Gaussian noise (zero-mean and 5% standard deviation); random brightness alteration (maximum alteration at 5%). Each image is augmented 20 times by this pipeline, each time every augmentation technique is randomly applied or not applied.

Methods. Deep learning algorithms have already shown great success in a variety of image recognition problems. Convolutional Neural Networks (CNN, implementation details can be found in [14]) are image processing algorithms that are able to extract image features regardless of their position, which is especially useful in our case since scans are not always optimally centered on the spine. Due to the small sample size, a simple architecture is needed to prevent overfitting. **Our CNN has the following structure:** first convolutional layer (filters size 5×5, 128 masks), followed by a max-pooling layer; second convolutional layer (filters size 5×5, 64 masks), followed by a max-pooling layer; a fully connected layer, 2048 nodes; a further fully connected layer, 1024 nodes. Rectifier Linear Units (ReLU) are a common choice for this kind of network. The network structure is illustrated in Fig. 4, step 3. The cost function minimized during training is the mean of the softmax cross-entropy function between the output $\mathbf{x}$ and the actual label vector $\mathbf{z}$, $\mathcal{L} = -\mathbf{z} \log \sigma(\mathbf{x}) - (1 - \mathbf{z}) \log [1 - \sigma(\mathbf{x})]$, where $\sigma(\mathbf{x})$ is the softmax function. The optimizer used for the minimization is AdaGrad [15]. Implementation is done in Python using TensorFlow [16].

The major inherent vice in this approach is the need of labeled examples. We learn from 1576 labeled scanned segments from 321 successfully operated

patients. On the other hand, if we were able to include unlabeled segments in the analysis, we could take advantage of all 4031 segments from the 788 patients. Unsupervised learning methods do not need labeled examples. The autoencoder algorithm [18] is used to reduce the dimensionality of the problem: it consists of an encoder function $\mathbf{h} = f(\mathbf{x})$ and a decoder function $\mathbf{r} = g(\mathbf{h})$. The autoencoder is trained to copy the input to the output, but it is not given the resources to do so exactly (undercompleteness property). In this way an approximation of the input is returned and the model is forced to prioritize the most relevant aspects of the input. As the autoencoder does not need labels for the surgery, all 4031 segments can be used. **An autoencoder sufficient for our needs** can be built by mirroring the CNN and learning how to "invert" the convolutional and the max pooling layers [17] into *deconvolutional layers*: first convolutional layer (filters size 5×5, 128 masks), followed by a max-pooling layer; second convolutional layer (filters size 5×5, 64 masks), followed by a max-pooling layer; a fully connected layer, 1024 nodes; a fully connected layer, 128 nodes (*bottleneck*); a fully connected layer, 1024 nodes; first unpooling and deconvolutional layer (filters size 5×5, 64 masks); second deconvolutional layer (filters size 5×5, 128 masks). This autoencoder reconstructs the original 3D image, and in the middle layer (the bottleneck), we find a 128-number code that identifies each image sufficiently for its reconstruction. We train the autoencoder on all unlabeled images to minimize the difference tensor $\mathbf{J} = (\mathbf{X}_{\text{orig}} - \mathbf{X}_{\text{reconstr}})^2$, where $\mathbf{X}_{\text{orig}}$ is the original image and $\mathbf{X}_{\text{reconstr}}$ is its reconstruction. After training, the autoencoder is used to encode all labeled images and their 128-number codes are used as features in the same classification experiments as in Sect. 2.

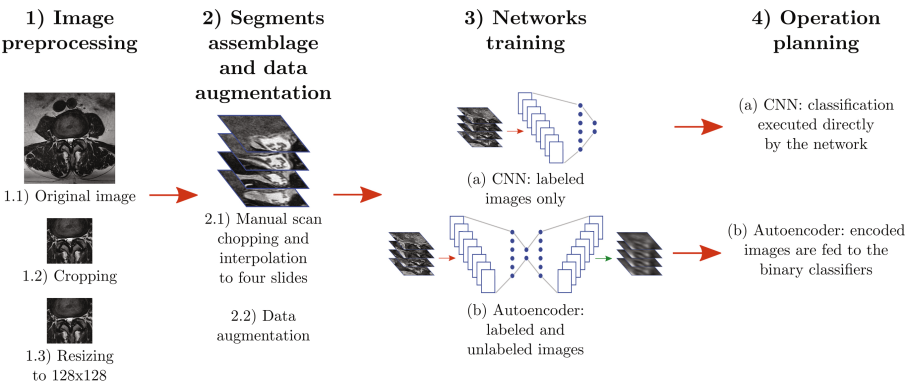

Fig. 4. Proposed computing pipeline from preprocessing of raw MRI pictures to learning of surgical planning

Results. The complete pipeline from the MRI preprocessing to the surgery classification is depicted in Fig. 4. For both CNN and autoencoder, the available image datasets are split into training and test set with a 80/20 ratio. The training sets are augmented as previously described and the networks are trained for 100

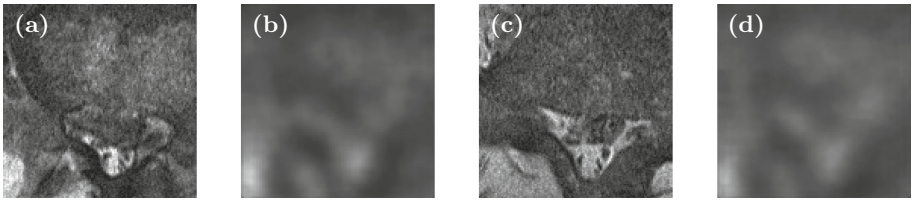

Fig. 5. Image reconstruction examples by the autoencoder. (a), (c): 2 out of 4 slices of the original 3D image. (b), (d): Reconstructed image slices.

epochs. Learning curves are available in the Supplement (A.2). On the test set, the CNN reaches an AUC of 69.8%. This is significantly lower than the AUC obtained with the numerical dataset, but it is still confirming the existence of a signal in the MRI images, and enforces the idea that radiological data are linked to stenosis diagnosis and treatment. Considering the small size of the training data, we are confident that higher precisions can be obtained if the present dataset is improved and expanded. The autoencoder learns successfully to reconstruct the images (Fig. 5). While some details are missed, it is noticeable that the dimension of a picture is now extremely reduced from $128 \times 128 \times 4 = 65536$ numbers to 128. When training and testing the binary classifiers from Sect. 3 with the codes from the labeled segments, the highest mean AUC for a 20-fold cross validation test is given by a optimized LDA classifier, at 70.6%, with a corresponding standard deviation of 6.69%. The mild improvement can be explained by the extension of the dataset to the non-labeled segments.

4 Discussion

While the influence of MRI scans on surgical decisions for LSS was previously unclear, our results quantitatively confirm the importance of medical imaging in LSS diagnosis and treatment planning. We started by effectively modeling surgical decision-making for lumbar spine stenosis through binary classifiers, on the sole basis of manually-assessed radiological features. To reduce human bias and errors in the selection and calculation of features, we developed an automatic pipeline (Fig. 4) to work on raw MRI scans. To the best of our knowledge these are the first and initial steps towards benchmarking LSS. Supervised (CNN) and semi-supervised (convolutional autoencoders) deep learning algorithms were trained on the transformed images and accuracies around 70% on surgical planning were achieved. Compared to the results with the numerical dataset, the differences in accuracy (of about 15%) can be justified by the modest number of MRI scans. We are confident that further systematic efforts aimed at enlarging the image catalog could significantly improve the classification results and thus patient outcome.

Acknowledgments. This research was partially supported by the Max Planck ETH Center for Learning Systems, the SystemsX.ch project SignalX, the Baugarten Foundation, the Helmut Horten Foundation, the Pfizer-Foundation for geriatrics & research in geriatrics, the Symphasis Charitable Foundation, the OPO Foundation, NIH/NCI Cancer Center Support Grant P30 CA008748 and an Oxford - Google DeepMind scholarship.

References

1. Deyo, R.A.: Treatment of lumbar spinal stenosis: a balancing act. Spine J. **10**, 625–627 (2010)
2. Kreiner, S., Shaffer, W.O., Baisden, J., Gilbert, T., et al.: Evidence-based clinical guidelines for multidisciplinary spine care diagnosis and treatment of degenerative lumbar spinal stenosis. North Am. Spine Soc. (2014)
3. Weinstein, J.N., Lurie, J.D., Olson, P.R., Bronner, K.K., Fisher, E.S., United States' trends, regional variations in lumbar spine surgery: 1992–2003. Spine **31**, 2707–2714 (2006)
4. Irwin, Z.N., Hilibrand, A., Gustavel, M., McLain, R., et al.: Variation in surgical decision making for degenerative spinal disorders. Part I: lumbar spine. Spine **30**, 2208–2213 (2005)
5. Jensen, M.C., Brant-Zawadzki, M.N., Obuchowski, N., Modic, M.T., et al.: Magnetic resonance imaging of the lumbar spine in people without back pain. N. Engl. J. Med. **331**, 69–73 (1994)
6. Steurer, J., Roner, S., Gnannt, R., Hodler, J.: Quantitative radiologic criteria for the diagnosis of lumbar spinal stenosis: a systematic literature review. BMC Musculoskelet Disord. **12**, 175 (2011)
7. Andreisek, G., Deyo, R.A., Jarvik, J.G., Porchet, F., et al.: LSOS working group and others, Consensus conference on core radiological parameters to describe lumbar stenosis - an initiative for structured reporting. Eur. Radiol. **24**, 3224–3232 (2014)
8. Haig, A.J., Tong, H.C., Yamakawa, K.S., et al.: Spinal stenosis, back pain, or no symptoms at all? a masked study comparing radiologic and electrodiagnostic diagnoses to the clinical impression. Arch. Phys. Med. Rehabil. **87**, 897–903 (2006)
9. Ishimoto, Y., Yoshimura, N., Muraki, S., et al.: Associations between radiographic lumbar spinal stenosis and clinical symptoms in the general population: the Wakayama Spine Study. Osteoarthr. Cartil. **21**, 783–788 (2013)
10. Downie, W.W., Leatham, P.A., Rhind, V.M., Wright, V., et al.: Studies with pain rating scales. Ann. Rheum. Dis. **37**, 378–381 (1978)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
12. Lurie, J.D., Tosteson, A.N., Tosteson, T.D., Carragee, E., et al.: Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. Spine **33**, 1605–1610 (2008)
13. Fu, M.C., Buerba, R.A., Long, W.D., Blizzard, D.J., et al.: Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. Spine J. **14**, 2442–2448 (2014)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**, 2278–2324 (1998)
15. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)

16. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., et al.: Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint (2016)
17. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: IEEE Conference on CVPR, pp. 2528–2535 (2010)
18. Zemel, R.S.: Autoencoders, minimum description length and Helmholtz free energy. NIPS (1994)