A Deep Learning Model for Automatic Detection and Classification of Disc Herniation in Magnetic Resonance Images

Tijana Šušteršič, Vesna Ranković, Vladimir Milovanović, *Senior Member, IEEE*, Vojin Kovačević, Lukas Rasulić, Nenad Filipović, *Member, IEEE*

***Abstract*— Localization of lumbar discs in magnetic resonance imaging (MRI) is a challenging task, due to a vast range of diversity in shape, size, number, and appearance of discs and vertebrae. Based on a review of the cutting-edge methods, the majority of applied techniques are either semi-automatic, extremely sensitive to change in parameters, or involve further modification of the results. All of the above represents a motivation for implementing deep learning-based approaches for automatic segmentation and classification of disc herniation in MRI images. This paper proposes a complete automated process based on deep learning to diagnose disc herniation. The methodology includes several steps starting from segmentation of region of interest (ROI), in this case disc area, bounding box cropping and enhancement of ROI, after which the image is classified based on convolutional neural network (CNNs) into adequate classes (healthy, bulge, central, right or left herniation for axial view and healthy, L4/L5, L5/S1 level of herniation on sagittal view). The results show high accuracy of segmentation for both axial view (dice=0.961, IOU=0.925) and sagittal view (dice=0.897, IOU=0.813) images. After cropping and enhancing the region of interest, accuracy of classification was 0.87 for axial view images and 0.91 for sagittal view images. Comparison with the literature shows that proposed methodology outperforms state-of-the-art results when it comes to multiclassification problems. A fully automated decision support system for disc hernia diagnosis can assist in generating diagnostic findings in a timely manner, while human mistakes caused by cognitive overload and procedure-related errors can be reduced.**

***Index Terms* — disc herniation, deep learning, segmentation, convolutional neural network, decision support system.**

This research is funded by Serbian Ministry of Education, Science, and Technological Development [451-03-9/2021-14/200107 (Faculty of Engineering, University of Kragujevac)]. This research is also supported by the project that has received funding from the European Union's Horizon 2020 research and innovation programmes under grant agreement No 952603 (SGABU project). This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains. T. Šušteršič also acknowledges the support from L'OREAL-UNESCO "For Women in Science" National Fellowship program Serbia (2021 Fellows).

T. Šušteršič, and N. Flipović are with Faculty of Engineering, University of Kragujevac, Kragujevac, Serbia and Bioengineering Research and Development Center (BioIRC), Kragujevac, Serbia (e-mail: tijanas@kg.ac.rs, fica@kg.ac.rs).

V. Ranković, V. Milovanović are with Faculty of Engineering, University of Kragujevac, Kragujevac, Serbia (e-mail: vesnar@kg.ac.rs, vlada@kg.ac.rs).

V. Kovačević is with Center for Neurosurgery, Clinical Center Kragujevac, Kragujevac, Serbia and Department of Surgery, Faculty of Medical Sciences, University of Kragujevac, Kragujevac, Serbia (e-mail: vojincnh@gmail.com)

L. Raulić is with School of Medicine, University of Belgrade, Belgrade, Serbia and Department of Peripheral Nerve Surgery, Functional Neurosurgery and Pain Management Surgery, Clinic for Neurosurgery, Clinical Center of Serbia, Belgrade, Serbia (e-mail: lukas.rasulic@kcs.ac.rs)

I. Introduction

Theoretical overview. Lumbar disc herniation is one of the most common intervertebral disc diseases (IDD), resulting in limited movement and unbearable pain levels. IDD is the cause of more than 90% of surgical spine operations [1]. The process by which the gel-like core in disc ruptures through a tear in the fibrous annulus is referred to as a "herniated disc". This gel substance activates the spinal neurons, creating mechanical and chemical irritation, which results in spinal nerve inflammation and edema. Herniated discs most often occur where the spinal nerves exit in between the lumbar vertebrae and recombine to create the sciatic and femoral nerves, which run down the anterior/posterior part of the thigh and leg [2]. Herniated discs are more common in adults between the ages of 30 and 40, with middle-aged and older people being somewhat more at risk if they engage in intensive physical activity. Lumbar disc herniation is a prevalent cause of lower back pain and leg discomfort, occurring 15 times more frequently than cervical herniation. Disc hernia is reported to occur 8% of the time in the neck (cervical) area and just 1–2% of the time in the thoracic area [3].

Herniation can progress in two ways: partial or total pain reduction or it can cause significant damage to nerves, which in result creates a necessity for surgery. As a result, doctors differentiate between different stages of herniation [2, 4]:

- protrusion (disc bulge) - the annulus fibrosis remains intact, but it causes outpouching, which may impinge on the nerves; the annulus fibrosis of the disc bulges without rupturing,
- prolapse - the nucleus pulposus relocates to the most outside fibers of the annulus fibrosis,
- extrusion - genuine herniated disc (also called a slipped disc) happens when the disc ruptures, enabling the gel-

like core to be squeezed out,

- sequestration - if severe enough, there may be a free fragment, which indicates that there is a piece that has broken totally free from the disc material and entered the spinal canal. [2, 4].

Lumbar disc herniation is diagnosed using imaging techniques such as MRI, myelography, X-rays and CT. Magnetic Resonance Imaging (MRI), represents a non-invasive imaging technique that employs radiofrequency radiation and magnetic field to offer a clear image of the spinal soft tissues. MRI images, unlike X-rays, reveal nerves and discs, making them the gold standard for disc herniation detection [5].

Related work. There exist many proposed approaches in intervertebral disc diagnosis of lumbar spine diseases. Peng et al. [6] recommended a visual and quantitative approach for image segmentation in order to extract six features in MRI images in patients. A similar approach based on features extraction (planar shape features, intensity and texture features derived from the Gray Level Co-Occurrence Matrix) was proposed by Ghosh et al. [7] using a majority voting system for the diagnosis of lumbar hernia. Neubert et al. [8] used signal intensity and shape features to diagnose herniation and degeneration in MRI images, towards classification of pathologies. Unal et al. [9] had the aim to diagnose disc abnormalities in the MRI axial view with a hybrid model that includes the use of the features manually extracted by technician. Additionally, Bhole et al. [10] developed a methodology for segmenting lumbar vertebrae and discs in MRI images using geometric data from T2 sagittal, T2 axial and T1 sagittal views. The achieved accuracy on the testing subset of 67 sagittal view images was 98.8%. Other approaches have also been proposed such as pyramidal histogram of directed gradients (HOG) by Oktay et al. [11], active contour model for segmentation of region of interest on spinal images by Koh et al. [12] or probabilistic models by Schmidt et al. [13]. All of these systems, however, are generally semi-automatic because they require user-expert interaction to set up initial conditions, manually extract characteristics, etc. This is visible, for example, in the research of Hoad et al. [14], who proposed a semi-automatic approach for MRI image segmentation of vertebrae and spinal cord. User intervention was required during the setup process to manually determine the midpoint of the spinal cord at each level of the spine or to manually choose four points for each vertebral body. Following that, an automated active contour algorithm was used to perform spinal canal segmentation.

As a result, some fully automated methods have been investigated to overcome the drawbacks of traditional methods. Some authors [15] investigated a hybrid of traditional and novel strategies for detecting lumbar disc herniation in MRI images. Otsu thresholding was paired with feature extraction by measuring the form function, and then classification was performed using the multi-layer perceptron (MLP), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). The classification accuracies of the MLP being 91.90% and KNN being 92.38% proved to be the highest. Similarly, Chevrefils et al. [16] merged two approaches, morphological and watershed methods, in order to detect discs in MR images. This automatic method was proposed in order to control the initial values of the cluster centroid using intuitive fuzzy clustering. Although numerous articles have produced potentially useful data, the findings have been limited due to a lack of standardized magnetic resonance imaging of the phenotype of spinal degeneration [17].

As a result, many researchers have abandoned traditional approaches in favor of deep learning techniques [18]. Jackson et al. [19] created deep learning methods for detecting and segmenting intervertebral discs and vertebrae in 2D images or volumetric data. Authors such as Simonyan et al. [20], He et al. [21], and Dou et al. [22] created several deep learning-based methods to locate intervertebral discs (IVDs) and vertebrae in 2D images. Cai et al. [23] detected IVDs using a 3D hierarchical model and segmented them using characteristics extracted from deep neural networks. In terms of segmentation, Chen et al. [24] and Suzani et al. [25] used deep learning algorithms, where Chen et al. focused on Convolutional Neural Networks (CNN) and Suzani et al. concentrated on feed forward neural networks. Harun et al. [26] developed a convolutional neural network approach that uses radiological ratings to distinguish intervertebral discs and vertebrae. Another study, such as that undertaken by Al-Kafri et al. [27], proposed using deep learning neural networks (specifically the SegNet architecture) and combined with delineating lumbar disc MRI information to aid clinicians in identifying lumbar spinal stenosis. Zhou et al. [28] created a novel detection method that employs a convolutional neural network to distinguish the spinal vertebrae, exceeding previous findings. Wang et al. [29] created a deep learning method for segmenting and labeling axial MRI slices. Mbarki et al. [30] proposed utilizing VGG16 CNNs to detect not only L1 to L5, and L5-S1 discs, but also the apophysis, and to differentiate between ruptured and normal discs. Although the authors reported 93.3% accuracy, still there were still some manual modifications to the image, i.e. cropping, in addition to the fact that they solely employ axial view to diagnose herniation (bulging, protrusion, extrusion or exclusion). One of the best proposed methodologies for automated lumbar vertebral segmentation and classification is the one by Lu et al. which uses deep learning for spinal stenosis analysis [31]. They analyzed both sagittal and axial MRI slices and used CNN for stenosis grading, achieving dice coefficient of 0.93 and class accuracy of 80.4±1.6%. Another paper by Pan et al. [32] use 3555 MRI images to segment ROI, after which middle points of vertebral bodies were calculated, and classification was performed with respect to three classes - disc bulge, disc herniation, and normal discs. Average classification accuracy was 88.76%. Although the results seem promising, and the advantage is a three-class classification, the used network was ResNet101, without the exploration of other possible deep CNNs and optimization of network hyperparameters.

Because of image resolution, presence of noise, but also disc differences in size, form, and appearance of vertebrae and discs, localization of lumbar discs is a difficult task for traditional image processing methods. This is especially true when the condition is coupled with a problem of diagnosing level and side of herniation. Additionally, traditional image processing methods are mostly dependent on selection of

manually set parameters [30], and their success relies on trial-and-error process to decide which features best describe different classes, while CNNs have proven to be an efficient method in solving various image problems, have achieved high accuracy in medical field and require less fine-tuning and expert analysis [32]. Therefore, deep learning approaches should be prioritized over traditional segmentation and thresholding algorithms. Since, simple features do not exist [30], it is difficult and time-consuming to automate the segmentation and classification process in medical images.

Technical significance. With the exception of Lu et al. [31], previous research on computer-aided diagnosis of spinal problems used either axial or sagittal MRI slices as inputs [33, 34, 35]. However, due to the fact that distinct anatomical planes usually produce complementary information and non-isotropic scan slice resolution is prevalent in clinical application, the interpretation of spinal MRI frequently entails a parallel analysis of both views. Only by examining both views (sagittal and axial), it is possible to accurately estimate the level of herniation, which is a direction that has received little attention. In addition, motivated by machine and deep learning-methodologies presented in the section *Related work*, we propose an automatic decision support system to diagnose disc bulging and herniation in MR images using convolutional neural networks (CNNs). For disc-level segmentation, we employ a U-Net architecture after which bounding boxes are created around the segmented region of interest. We then create a CNN for disc hernia diagnosis that is *multi-input* (sagittal and axial view) and *multi-class* (healthy, bulge, central, right or left herniation for axial view and healthy, L4/L5, L5/S1 level of herniation on sagittal view). By combining the input view and classification, it is possible to make a definitive diagnosis of disc herniation (i.e. L5/S1 disc herniation on the left side), which has not been done before. Additionally, we use a with limited dataset, and contrast it with a big data paradigm in order to achieve high quality results even with smaller amount of data that are usually required for deep learning models.

Clinical relevance. The high frequency of disc herniation in the working and senior population necessitates the use of spinal MRI in medical treatment. However, reporting can be time-inefficient and subject to substantial inter-reader variability. By reducing the radiologists' workload while providing the consistency needed to produce standardized diagnosis reports, automated disc herniation analysis based on images will very certainly play an important role in interpretation of spinal MR imaging. Furthermore, the segmented results onto original images will enable more meaningful visual representations of the reported anatomy and disease, as well as big data for longitudinal analysis of patient data to improve the healthcare resource allocation.

II. Materials and Methods

This section describes the dataset used for machine learning and proposed methodology, as well as the description of the evaluation metrics.

A. Dataset for machine learning

The dataset utilized in this study was a combination of publicly available database Lumbar Spine MRI Dataset obtained from Mendeley Data [36] and images obtained from 23 patients from the Clinical Centre of Kragujevac, Serbia [37]. The use of imaging data was approved by the Ethics Committee of the Clinical Centre Kragujevac, decision number 01-11484 from September 8, 2016. Publicly available dataset consisted of 575 patients, however, only patients with L4/L5, L5/S1 or healthy spine were extracted (220 patients). Table I provides an overview of the number of images per view and diagnosis.

TABLE I
DESCRIPTION OF THE USED DATASET WITH NUMBER OF IMAGES PER VIEW AND HERNIATION DESCRIPTION

	Number of patients	Sagittal view			Total
		healthy	L4/L5	L5/S1	
Mendeley Data	220	329	363	318	1010
Local database	23	31	49	79	159
Total	**243**	**360**	**412**	**397**	**1169**

		Axial view					Total
		healthy	bulge	central	left	right	
Mendeley Data	220	205	56	67	85	49	**500**
Local database	23	38	8	8	25	16	**57**
Total	**243**	**243**	**64**	**75**	**110**	**65**	**557**

B. Proposed system

Similar to the general concept of preprocessing and main task performed in machine learning, the proposed workflow includes:

1) Segmentation of region of interest (ROI) – in this step, we utilize U-net convolutional neural network to extract the disc area in both axial and sagittal view MRI images;

2) Bounding box cropping – in this step, we perform contour recognition and create bounding boxes around the segmented area to reduce the search area for the CNN in the classification process;

3) ROI Enhancement - in this step, we apply a variety of image processing techniques to improve the cropped image so that the contrast is improved and characteristic sections of the ROI are enhanced for the use in the diagnostic process;

4) Data augmentation – in this step we perform data augmentation to enlarge the number of images for training;

5) Classification – in this step, the image is classified into adequate classes (healthy, bulge, central, right or left herniation for axial view and healthy, L4/L5, L5/S1 level of herniation on sagittal view).

The described concept of the proposed system is presented in Fig. 1.

The processing hardware included NVIDIA Quadro RTX 6000 GPU, 64GB of RAM and an Intel(R) Xeon(R) Gold, 6240R, CPU running at 2.40GHz. The network was implemented using Tensorflow and Keras [38] in Python programming language.

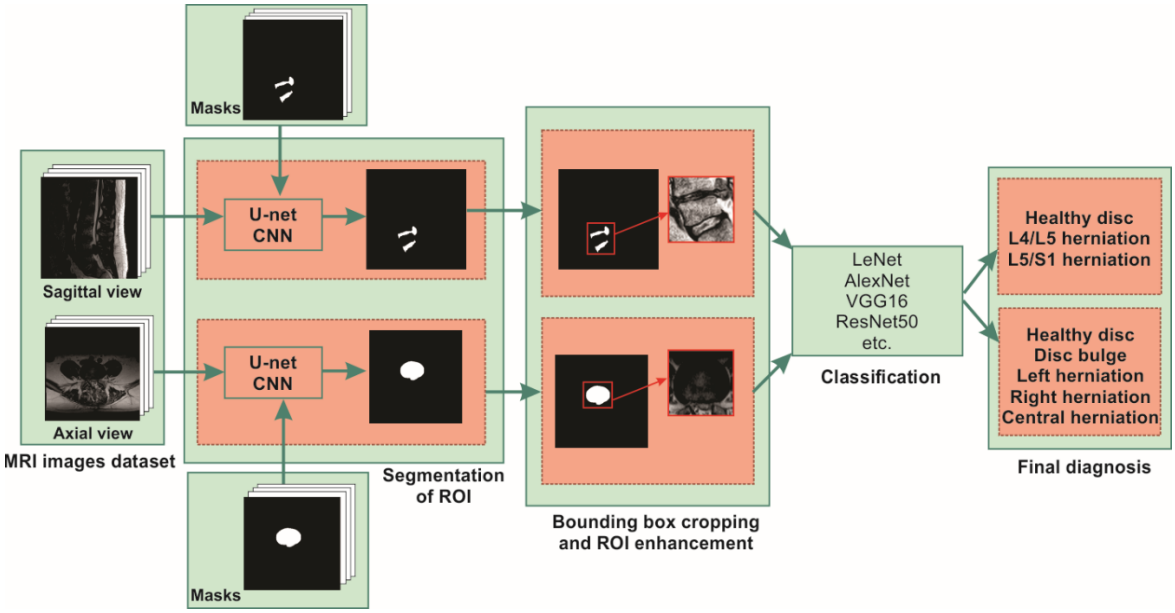


Fig. 1. Proposed workflow of disc hernia diagnosis

Segmentation of region of interest (ROI). The U-net neural network architecture has been demonstrated to be adequate for a wide range of medical image segmentation challenges [39, 40]. U-net architecture, adapted for the purposes of this study, is presented in Fig. 2. While it is explained using axial view images, the same workflow was applied on sagittal view images.

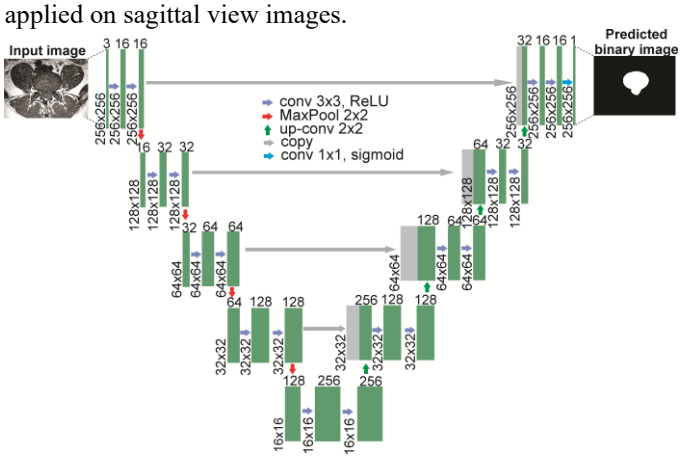


Fig. 2. Architecture of the U-net. Number of channels is written horizontally above each green rectangle denoting multi-channel feature map (i.e. 3->16->16), while dimensions of an image are written vertically next to multi-channel feature map (i.e. 256x256)

As shown in the figure, U-net consists of encoder (contraction path) and decoder (expansion path). First, the encoder consists of two 3x3 convolutional layers followed by two 2x2 max pooling layers at each level. This means that as the depth increases and more advanced features are retrieved, the size of the input image reduces. The red arrow depicts the max pooling, which decreases the picture size by half, where padding = "same" is utilized. Each level of the expansion route (decoder) employs two sequential 2x2 up-conv followed by two 3x3 convolutional layers. In the expansion path, retrieval of information from the original picture is performed. Transposed convolution is a sampling technique used to increase the image size. The image is then merged with the corresponding image from the encoder path after the transposed convolution. The goal of combining data from previous layers is to make a more accurate forecast. Following that, two convolutional layers are added. The decoder path restores the segmentation map's size but discards the localization information. As a result, cross-over connections are employed to provide localization information from the encoder to decoder by concatenating equally sized feature maps. The final step is to reconstruct the image so that it fits the prediction requirements. This shows that the last layer is a convolutional layer with one 1x1 filter.

The training procedure was initially set to 100 epochs, and stochastic gradient descent was utilized with a predetermined learning rate and batch sizes that were varied in this study to determine their influence on the segmentation accuracy. The activation ReLU function was employed. The identical procedure was used for sagittal and axial view photographs separately. Hyperparameter optimization was performed in the following ranges:

- batch size (BS) – examined range of BS=6 to 12, with an increment of 2;
- learning rates (LR) - examined values were LR=0.1, LR=0.01, and LR=0.001, and multiplication with numbers 2 to 9;
- number of layers kernels – i.e. 8, 16, 32, 64 and 16, 32, 64, 128;
- dropout – examined range from 0.1 to 0.5 with an increment of 0.1;
- number of epochs – up to 100.

Bounding box cropping. Normalization of the ROI size was conducted by resizing the cropped images to 64x64. Classification was at first performed without the cropping of the bounding box area. However, due to a limited number of images, but also the diversity of cross sections and lots of details, CNNs were not able to learn the adequate features and

classify the images with high accuracy. Therefore, in order to reduce the search area, bounding boxes were automatically created using contour detection around the segmented region of interest (in both axial and sagittal images) and cropped out of the original image. Further analysis was performed on the cropped regions, and this pre-processing step proved efficient in the overall process of classification. It should be emphasized that 10% of increase in the size of the bounding box was performed in order to not to omit the relevant characteristics of the ROI.

ROI Enhancement and Data augmentation. As machine learning tasks require a large number of images, in fields such as bio-medicine, where a large amount of bio-medical data is often not available [41], to improve segmentation performance, a procedure known as augmentation was used [42]. The augmentation method was employed to artificially increase the number of images in the training dataset [43], while the validation and test dataset contained only original and previously unseen images. In this specific study, several operations were utilized in the augmentation process, specifically:

- horizontal flip (equivalent to the mirroring along the y axis)
- brightness (HSV colorspace is utilized for this task, which means that the greater the values of saturation and value matrices are, the bigger the brightness. Therefore, in order to increase the brightness, matrices were multiplied by a value greater than 1 and to reduce brightness, matrices were multiplied by a value less than 1). Normalization of the intensity was conducted by contrast limited adaptive histogram equalization (CLAHE) algorithm with clipLimit=2.0, tileGridSize=(4,4).

Only geometrical transformations that allowed for retaining the whole image data and avoiding to lose any information were employed. Other transformations, which might have resulted in the removal of crucial areas of the image, were unacceptable owing to the nature of the problem. Other techniques, such as the vertical flip, were also deemed improper since they had no physical significance. The entire dataset was divided into subsets of training, validation and test in the ratio of 80:10:10, respectively, for both views.

Classification. After several preprocessing steps, we have used several CNNs to perform final automatic classification and diagnosis. It should be emphasized that 3-10 images for each patient were available at characteristic cross sections, meaning that in the cases of disc hernia diagnosis only several images are taken to diagnose hernia, not as standard MR imaging with i.e. 100 images per patient. The data are split based on patient identifier rather than individual images, because the images from the same patient are correlated. This implies that convergence is also achieved on a patient-based level. Additionally, due to the fact that segmentation was first performed and classification was introduced on segmented images, correlation between the images of one patient is reduced. The division of the number of patients into training, validation, test subsets was 195, 24, 24, respectively (division 80:10:10).

It should be noted that we also tested the system without the preprocessing steps, and the accuracy was not satisfying due to differences in images/cross sections, level of details, etc. In this research, in order to conduct multiclass classification - healthy, bulge, central, right or left herniation for axial view and healthy, L4/L5, L5/S1 level of herniation in sagittal view, eight different CNN architectures were used:

- LeNet [44],
- AlexNet [45],
- VGG16 [20],
- NiN [46],
- GoogleNet [47],
- DenseNet121 [48],
- ResNet50 [21],
- DiscNet (proposed architecture in this paper, specifically designed to suit the disc herniation classification, therefore named DiscNet)

CNNs indicated above were chosen because of their record of high accuracy classification performance in similar scenarios [30, 31, 32]. CNN architectures such as LeNet, AlexNet, and VGG are all similar because they all involve extracting features via a series of convolution and pooling layers, followed by post-processing of the representations via fully-connected layers [49]. The advances made by AlexNet and VGG over LeNet are mostly viewed in how these subsequent networks broaden and deepen these two mentioned modules. AlexNet consists of nine layers, out of which the first five are convolutional, followed by pooling layers, and the final four layers are the fully connected layers [50]. The previously outlined trend of deeper CNN configuration has led to enhancements of the original AlexNet architecture. VGG16, which was introduced the year after, is one of these architectures. VGG16 is a deeper variant of AlexNet in which the nine-layer architecture is substituted with a 16-layer design, thus the name [20]. In comparison to AlexNet, one of the key advantages of VGG16 is the use of smaller kernels in convolutional layers. It is worth noting that the use of deep CNNs can lead to the reduction of classification accuracy as the depth of the CNN increases. He et al. [21] developed a deep residual network architecture to overcome this problem utilizing the residual block technique, which demonstrated a considerable accuracy impact on the ImageNet dataset. Therefore, we also included ResNet50 as one of the CNNs for classification. GoogLeNet won the ImageNet Challenge in 2014, presenting a framework that represents a combination of the strengths of NiN and repeating block paradigm [47]. One of the major aims of the work was to determine the best size convolution kernels. After all, past successful networks have employed numbers ranging from 1x1 to 11x11. One finding from this study was that it may be beneficial to utilize a mixture of different-sized kernels. When CNNs become more complex and deeper, difficulties occur. This is due to the fact that the information route from the input layer to the output layer (and the gradient in the opposite way) gets so large that it might vanish before reaching the other side [48]. DenseNets simplify the connection pattern established in other architectures [48]. Additionally, and in contrast to deep

architectures, DiscNet was proposed as a result of a large number of different networks that were tested in order to find the most suitable combination of layers, kernels, dropout layers, etc.

No research before has investigated several CNNs in the process of multi-input, multi-class classification for the problem of discus hernia. All of the CNN architectures described above contain predetermined architecture and activation functions, but other hyperparameters such as solver, batch size and number of epochs can be optimized. Grid-search process refers to an extensive search for the best answer in the hyperparameter space. In addition, we have also tested the use of transfer learning in classification of disc hernia. Transfer learning was proposed to tackle the problem of a lack of training data. The procedure included applying transfer learning on images from ImageNet [51]. We applied four Keras deep learning models (VGG16, VGG19, DenseNet121, Xception) to classify disc hernia. The procedure was as follows:

- adopting layers from a previously trained model, except the last classification layer (VGG16, VGG19, DenseNet121, Xception),
- freeze layers, in order to avoid destroying information obtained during future training,
- adding new and trainable layers on top of the frozen layers. These layers will serve to learn to turn the old features into predictions using disc hernia dataset,
- training new layers using disc hernia dataset.

As a last step, fine-tuning was performed, which included unfreezing the whole model acquired above and re-training it on the new data with a very low learning rate (LR=0.00001). This has shown to improve the accuracy by incrementally adapting the pretrained features to the new dataset. It should be emphasized that fine-tuning was done after the model with frozen layers has been trained to convergence.

C. Evaluation metrics

In the process of segmentation, the proposed automated technique was compared with the manual segmentation performed by the neurologist. Evaluation measures such as dice, intersection over union (Jaccard coefficient), accuracy, precision and recall were determined throughout the training, validation, and testing in the segmentation stage. We calculated the overlapping areas between the automated segmentation area indicated as S and the ground truth area marked as G using the dice similarity coefficient D [47]:

$$D = \frac{2|S \cap G|}{|S| + |G|} \tag{1}$$

The Jaccard coefficient (JC), which represents the intersection over union (IOU), was calculated in the same manner as D and is commonly used to compare the similarity between original and automatically segmented regions. It is determined by dividing the total number of pixels in the intersection area by the total pixels number in the union area.

$$JC = \frac{|S \cap G|}{|S \cup G|} \tag{2}$$

In the classification stage, standard evaluation measures such as confusion matrix, accuracy, F1 score, sensitivity (recall), specificity and precision were used. In addition, receiver operating characteristic curve was used, as well as precision recall (PR) curve.

III. Results and Discussion

This section presents results in each of the methodological steps using a machine learning approach.

A. Segmentation results

In order to determine the influence of training parameters, it was investigated how the number of epochs, batch size, learning rate, as well as network depth influence the accuracy of training and validation subsets. It was found that LR=0.01, batch size 8, with the network depth up to 256 kernel features (starting with 8 kernel features) showed the highest accuracy for both sagittal and axial view images. Optimal number of epochs was 35, after which the model has converged. Fig. 3, 4 and 5 show loss function, dice coefficient and IOU during training and validation, respectively.

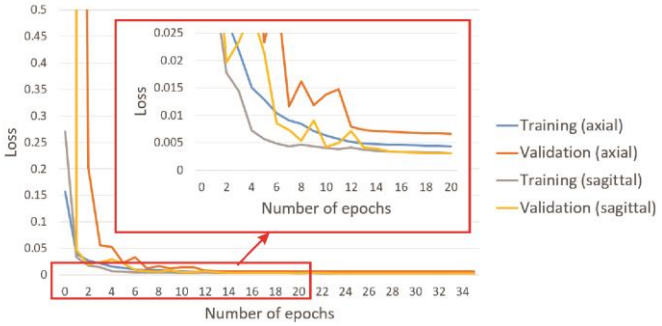


Fig. 3. Loss function during training and validation

The results for dice coefficient showed 96.6% during training and 95.9% during validation for axial view images and 90.5% during training and 90.3% during validation for sagittal view images. The results for IOU (Jaccard coefficient) showed 93.5% during training and 92.1% during validation for axial view images and 82.7% during training and 82.4% during validation for sagittal view images.

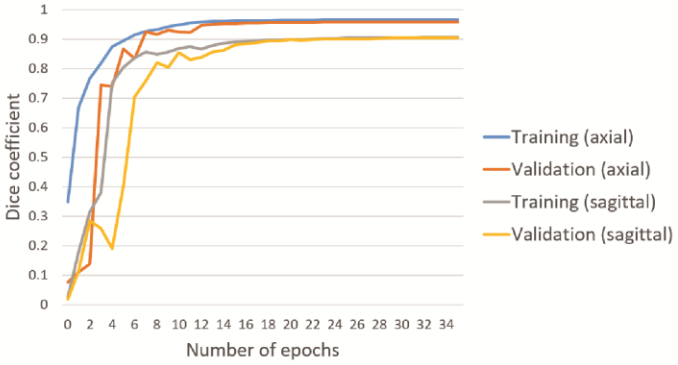


Fig. 4. Dice coefficient during training and validation

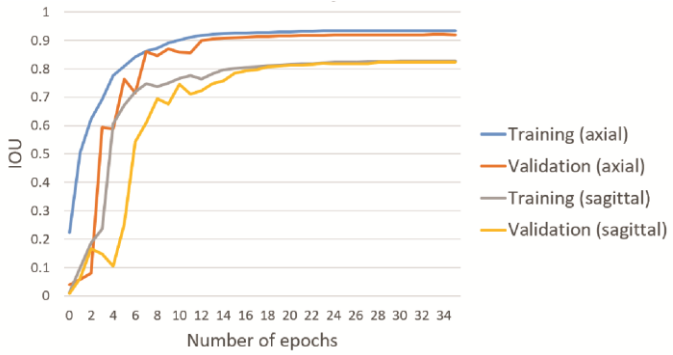


Fig. 5. IOU during training and validation

After finding optimal parameters, the results were evaluated on the unseen images – test subset and the results were reported in Table II. In our study, GPU memory occupation was 4% - 6% at all times and did not have larger deviation. The training time for axial images was roughly 5 minutes, whereas the test segmentation time was less than 56s (1s per image). The training time for sagittal pictures was roughly 9 minutes, whereas the test segmentation time was less than 120s (1s per image).

TABLE II
STATISTICAL MEASURES FOR TRAINING, VALIDATION AND TEST SETS FOR SEGMENTATION USING OPTIMIZED VALUES EP = 35, BS = 8, AND LR = 0.01

		Statistical measure				
		accuracy	**dice**	**IOU**	**precision**	**recall**
Axial view	*Training*	0.997	0.966	0.935	0.992	0.960
	Validation	0.996	0.959	0.921	0.989	0.946
	Test	**0.996**	**0.961**	**0.925**	**0.986**	**0.954**
Sagittal view	*Training*	0.997	0.905	0.827	0.980	0.871
	Validation	0.997	0.903	0.824	0.977	0.871
	Test	**0.997**	**0.897**	**0.813**	**0.976**	**0.860**

Fig. 6 shows the visual result of the segmentation for the patient with a herniated disc in axial view.

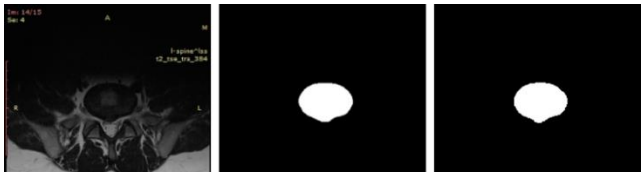


Fig. 6. Original image (left), manually segmented disc (middle) and automatic segmented disc (right) for optimized hyperparameters EP = 35, BS = 8, and LR = 0.01 on axial view images

Fig. 7 shows the visual result of the segmentation for the patient with a herniated disc in sagittal view.

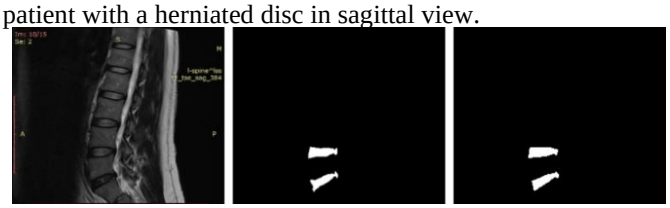


Fig. 7. Original image (left), manually segmented disc (middle) and automatic segmented disc (right) for optimized hyperparameters EP = 35, BS = 8, and LR = 0.01 on sagittal view images

The utilization of a neural network makes the segmentation process fast and robust. It should be also noted that in the dataset for segmentation, both healthy and herniated discs were present in images, meaning that the system was able to segment both herniated and healthy discs with high accuracy. In conclusion of this step in the methodology, based on a machine learning, we provided a fast, accurate, and automatic segmentation approach for spinal (both sagittal and axial) images. This approach covers the initial phase in automated disc herniation diagnosis and analysis. The second step was to crop the segmented area to a size of 64x64 in pixels. This means that all the images were cropped and resized to the same size. It should be emphasized that 10% of increase in the size of the bounding box was performed in order to make sure that possible segmentation errors did not influence further classification. In addition, ROI enhancement using CLAHE histogram equalization enabled an improved visibility level of disc area from the surrounding tissue.

Comparing obtained results with those found in the literature results, it was shown that the proposed methodology overcomes the state of the art (Table III). Taking into account that in most of the literature, only one view is used (either axial or sagittal), as well as the fact that the number of images in our dataset is smaller than in the most of the literature datasets, the proposed approach shows advanced results.

B. Classification results

After successful segmentation of the region of interest and cropping of the relevant bounding box, including ROI enhancement, we have focused on the classification of disc herniation (healthy, bulge, central, right or left herniation for axial view and healthy, L4/L5, L5/S1 level of herniation on sagittal view). In comparison to all trained networks (LeNet, AlexNet, VGG16, NiN, GoogleNet, DenseNet121, ResNet50, and proposed DiscNet), it was found that the best results were achieved with network that is not so deep as the state-of-the-art networks. Large number of trainable parameters proved not be the best solution when having the proposed methodology of cropping the region of interest, therefore reducing the size of image and level of details. Therefore, proposed DenseNet architecture with total number of parameters being 342,691 (trainable params: 256,163 and non-trainable params: 86,528) showed the best results. After using GridSearch to tune hyperparameters, it was found that DiscNet with the following architecture gave the best results:

- Convolutional layer: filters=16, activation function=ReLU and kernel size=7x7,
- Max pooling layer: kernel=3x3, stride=1, padding='same'
- Convolutional layer: filters=32, activation function=ReLU and kernel size=3x3,
- Convolutional layer: filters=64, activation function=ReLU and kernel size=4x4,
- Average pooling layer,
- Flatten layer (return a copy of the array collapsed into one dimension),
- Dropout (0.4),
- Batch normalization,
- Dense layer with softmax activation function.

The most optimal hyperparameters were shown to be:

- sagittal view: Adam optimizer, learning rate 1e-5, $\beta1$=0.9, $\beta2$=0.9, ε=1e-08;
- axial: Adam optimizer, learning rate 1e-5, $\beta1$=0.9,

β2=0.9, ε=1e-08.

In our study, GPU memory occupation was 4-6% at all times and did not have a larger deviation. Since the final network was not as deep as the other proposed networks, training time was in order of seconds for DiscNet.

TABLE III
COMPARISON OF THE ACHIEVED RESULTS WITH THE LITERATURE RESULTS IN SIMILAR SEGMENTATION PROBLEMS.

	Kafri et al. [27]	Mbarki et al. [30]	Lu et al. [31]	Ghosh et al. [7]	Our proposed			
Investigated problem	spinal stenosis	herniated lumber disc	spinal stenosis	disc abnormalities (herniated, bulging, degenerated discs, etc.)	herniated lumber disc			
Type of scan	axial	axial	axial+ sagittal	sagittal	axial		sagittal	
Number of images	48 345	382	22 796	251	557		1169	
Performance metric	IOU	IOU	Dice	Dice	Dice	IOU	Dice	IOU
Accuracy	0.92	0.93	0.93	0.82	0.961	0.925	0.897	0.813

Loss function during training and validation are shown in Fig. 8a for axial view and Fig. 8b for sagittal view. Early stopping was used as a method to avoid overfitting, as well as introduction of dropout layers and regularization coefficients.

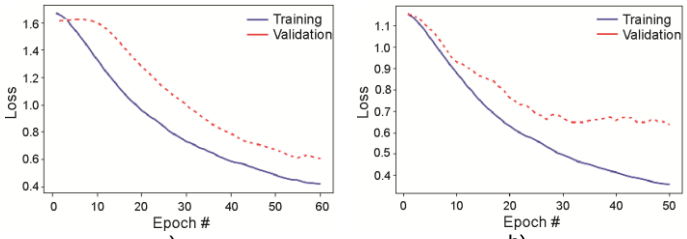


Fig. 8. Loss function during training and validation a) axial view, b) sagittal view

Accuracy curves during training and validation are shown in Fig. 9a for axial view and Fig. 9b for sagittal view. For axial view, average accuracy during validation was 0.87, macro average was 0.82, and weighted average was 0.87. All three types of accuracies are reported due to class imbalance. For sagittal view, average accuracy during validation was 0.91, macro average was 0.91, and weighted average was 0.91.

Detailed results of classification accuracy on the test subset are given in Table IV based on 10-fold cross validation results.

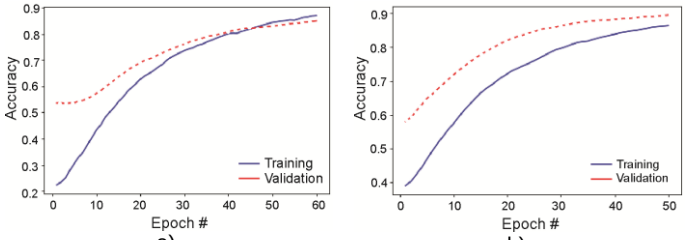


Fig. 9. Accuracy function during training and validation a) axial view, b) sagittal view

Average accuracy for axial view was 0.87 when five classes were target classes. However, an additional classification task was performed, since a disc bulge is often categorized as the first stage where annulus fibrosis of the disc stays intact and therefore can be challenging in distinguishing in comparison to other classes. This was also visible based on the confusion matrix (Fig. 10a). After eliminating one class (disc bulge), average accuracy increased by 3%. It should be noted that hyperparameter optimization was performed again for four-class classification and the most optimal parameters were Adam optimizer, learning rate of 1e-4, β1=0.9, β=2=0.9, and ε =1e-08. Regarding the sagittal view, the results with three class classification showed average accuracy of 0.91.

TABLE IV
CLASS ACCURACY FOR DISC HERNIA DIAGNOSIS (AXIAL AND SAGITTAL VIEW)

Disc hernia classification (axial view) (mean)					
healthy	bulge	central	left	right	class average
0.95	0.74	0.67	0.87	0.86	0.87
healthy	-	central	left	right	class average
1.0	-	0.73	0.82	0.90	0.90
Disc hernia classification (sagittal view) (mean)					
healthy		L4/L5 level		L5/S1 level	class average
0.90		0.91		0.92	0.91

To account for other metrics except accuracy, confusion matrices are given in Fig. 10a for axial view and Fig. 10b for sagittal view. For axial view, it is visible that classes bulge and central are often misclassified meaning that disc bulge and central herniation were the hardest to distinguish from other classes.

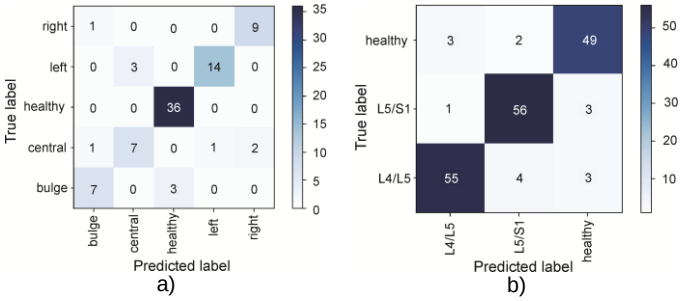
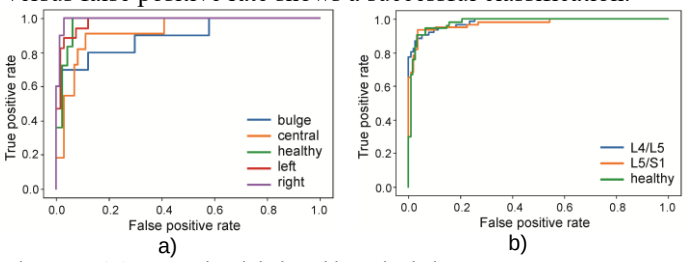


Fig. 10. Confusion matrix a) axial view, b) sagittal view

Disc bulge is in three cases classified as heathy, which is understandable, as there is no rupturing of the annulus. An additional potential problem in the classification was the central classification, which achieved the lowest classification accuracy. This is mainly due to the dataset imbalance with a small number of images for central herniation, as it is also the case in real clinical practice where left and right herniation are

more likely to occur. Also, central protrusion often leans toward left/right paracentral protrusion, so it is understandable why this type of herniation caused some misclassifications (most often as left or right herniation as seen from Fig. 10a). ROC curves are given in Fig. 11a for axial view and Fig. 11b for sagittal view. It can be seen that plotting true positive rate versus false positive rate shows a successful classification.



a) b)

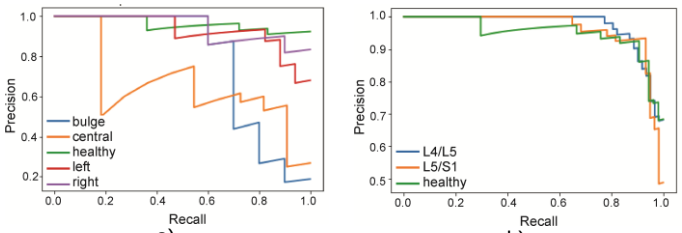
Fig. 11. ROC curve a) axial view, b) sagittal view

Precision recall curves are shown in Fig. 12a for axial view and Fig. 12b for sagittal view. Detailed report on sensitivity (recall), specificity and precision for the test subset are given in Table V.

TABLE V
OTHER STATISTICAL MEASURES FOR DISC HERNIA DIAGNOSIS (AXIAL AND SAGITTAL VIEW)

Disc hernia classification (axial view)					
	healthy	bulge	central	left	right
sensitivity (recall)	1.00	0.70	0.64	0.82	0.90
specificity	0.94	0.97	0.96	0.99	0.97
precision	0.92	0.78	0.70	0.93	0.82
	healthy	-	central	left	right
sensitivity (recall)	1.00	-	0.73	0.82	0.90
specificity	1.00	-	0.95	0.95	0.98
precision	1.00	-	0.73	0.82	0.90
Disc hernia classification (sagittal view)					
	healthy		L4/L5 level		L5/S1 level
sensitivity (recall)	0.91		0.89		0.93
specificity	0.95		0.96		0.95
precision	0.89		0.93		0.90

These results indicate that there are images that represent a borderline between the stages of herniation (i.e. healthy and bulge), and they will be the topic of further investigation in future research.



a) b)

Fig. 12. PR curve
a) axial view, b) sagittal view

After successful classification on each of the views separately, it was tested how successful the proposed methodology was when the model was tested on coupled axial and sagittal images. Assuming that both axial and sagittal images of one patient are available, we tested the model on 100 images from patients that had both axial and sagittal images and were not used in the training and validation phases. The results show accuracy of 0.8 (Table VI).

TABLE VI
CLASS AVERAGE (MEAN) CLASSIFICATION RESULTS FOR ONLY AXIAL INPUT, ONLY SAGITTAL INPUT AND BOTH VIEW INPUTS

Axial only (5 classes)	Sagittal only (3 classes)	Axial + Sagittal
0.87	0.91	0.8

The results from the transfer learning have been summarized in Table VII. The metric reported was accuracy for each class.

TABLE VII
RESULTS OF TRANSFER LEARNING (ACCURACY) WITH DIFFERENT BASE MODELS*

Disc hernia classification (axial view)						
	healthy	bulge	central	left	right	class average
VGG16	0.91	0.86	0.42	0.59	0.48	0.71
VGG19	0.95	0.59	0.56	0.64	0.57	0.75
DenseNet121	0.93	0.57	0.46	0.53	0.40	0.68
Xception	0.93	0.54	0.20	0.44	0.15	0.62
Disc hernia classification (sagittal view)						
	healthy		L4/L5 level		L5/S1 level	class average
VGG16	0.86		0.90		0.89	0.88
VGG19	0.84		0.89		0.86	0.86
DenseNet121	0.81		0.85		0.87	0.85
Xception	0.74		0.83		0.75	0.77

*no pretrained AlexNet model is available in Keras

It can be seen that transfer learning achieves competitive results in comparison to the newly designed CNN. It can be debated that optimal hyperparameters do have a wide range of search space (i.e. number of last layers that will be (un)frozen, the number of last layers retrained, number and type of new layers added). Therefore, obtained results in this study should serve as a guide for future analysis focused only on transfer learning applied for disc hernia diagnosis. Additionally, pretrained weights could be potentially loaded not only from ImageNet dataset, but also a dataset more similar to dataset used in this study (i.e. model trained on MRI spine images), in order to be able to compare the results and explore the possibility of negative transfer learning.

Comparison of the obtained results from this paper with leading-edge results in similar problems are listed in Table VII. It can be seen that most of the available methods investigate only binary classification (healthy versus diseased) such as [33], [34]. Those that investigate multi-classification use only one view, such as [32] and [52] and achieve around 87-88% accuracy on axial view images. However, it should be stated that the number of classes is smaller than in our investigation. With four classes, we achieved beyond state-of-the-art accuracy. It should also be noted that a comparison cannot be made directly, as classes in the literature differ from classes used in this paper. Should the comparison be made, only one paper by Lu et al. [31] uses axial and sagittal view analysis and classification, though in spinal canal stenosis and achieves 80.4±1.6% of accuracy. By comparing Table VI and Table VIII, it can be concluded that the proposed method makes advancement in several directions – conducts analysis of both views (axial and sagittal), performs multiclass classification on relatively small datasets and achieves results comparable and even beyond state of the art.

Nevertheless, there are several limitations of this work.

Dataset used does not contain images of L1-L3 disc level, meaning upper lumbar disc herniations, primarily because it is a less common diagnosis. The plan is to expand the methodology applied on lower lumbar herniations to upper levels in future. Additionally, the dataset used is rather small for the standard use of deep CNNs and also smaller in comparison to the datasets used in the other studies. Also, the dataset included images with both protrusion, extrusion and sequestration levels of herniation, so we were not able to divide the dataset into even smaller subsets, as the main focus was to distinguish the level (L4/L5 or L5/S1) and side (left, right, central, bulge) of herniation. Additionally, due to the fact that number of images in dataset is relatively small for deep learning, implementation of the bootstrap aggregation technique could be investigated. Our future effort will concentrate on two directions areas: (1) refining this system using a more focused technique that analyzes certain elements of MR images, (2) implementation of bootstrapping methods to perform additional validation of the obtained results (3) collecting more MR images in order to produce a completely automatic method and software for disc herniation diagnosis.

IV. Conclusion

An automated approach for diagnosing lumbar disc herniation using MRI axial and sagittal images is proposed in this paper. The goal is to develop a decision-making system that will aid clinicians in terms of diagnostic accuracy and speed. The dataset consisted of combined online available dataset and locally collected images in the Clinical Centre Kragujevac that resulted in 1169 sagittal view images and 557 axial view images. Our methodology was comprised of several steps. The first step was to automatically locate and segment L4/L5 and L5/S1 vertebral discs using a U-net convolutional neural network. This was achieved with high accuracy on both views – axial (dice=0.961, IOU=0.925) and sagittal view (dice=0.897, IOU=0.813) images. CLAHE filter was used to increase the quality and contrast of the segmented region and bounding box was extracted to be forwarded to the final classification. Each cropped region with vertebral disc was classified based on the created convolutional neural network (CNNs) into adequate classes (healthy, bulge, central, right or left herniation for axial view and healthy, L4/L5, L5/S1 level of herniation on sagittal view). The accuracy of classification was 0.87 on axial view images and 0.91 on sagittal view images, while the accuracy was 0.8 when a combined axial and sagittal view was used. Compared to the accuracies in the state-of-the-art literature, our results showed roughly the same or even better accuracy, especially taking into consideration that multi-input and multiclass classification was performed.

TABLE VIII
Overview of the best published results in similar problems. Performance metric is overall accuracy.

	Zhang et al. (2017) [33]	Jamaludin et al. (2017) [34]	Lu et al. (2018) [31]		Salehi et al. (2019) [52]	Pan et al. (2021) [32]
Investigated problem	Spinal canal stenosis	Spinal canal stenosis	Spinal canal stenosis		Disc hernia	Disc hernia
Type of scan	Axial	Sagittal	Axial+Sagittal		Axial	Axial
Number of images	582	12018	22796		2329	3555
Binary/Multi classification	binary (healthy, diseased)	binary (healthy, diseased)	binary (healthy, diseased)	multi (normal, mild, moderate, severe)	multi (normal, bulge, protrusion, extrusion)	multi (healthy, bulge, hernia)
Performance	86.6±3.3	91.4	96.3±0.46	80.4±1.6	87±7	88.76±3.72

References

[1] H. S. An, P. A. Anderson, V. M. Haughton, J. C. Iatridis, J. D. Kang and J. C. e. a. Lotz, "Introduction: disc degeneration: summary," *Spine,* vol. 29, no. 23, pp. 2677-2678, 2004.

[2] H. Winn, Youmans & Winn Neurological Surgery, 7th ed. Elsevier., 2016.

[3] J. Jordan, K. Konstantinou and J. O'Dowd, "Herniated lumbar disc," *BMJ Clinical Evidence Archives,* vol. 2009, p. 1118, 2011.

[4] D. F. Fardon, A. L. Williams, E. J. Dohring, F. R. Murtagh, S. L. G. Rothman and G. K. Sze, "Lumbar disc nomenclature: version 2.0: Recommendations of the combined task forces of the North American Spine Society, the American Society of Spine Radiology and the American Society of Neuroradiology," *The Spine Journal,* vol. 14, no. 11, pp. 2525-2545, 2014.

[5] F. Vitosevic, L. Rasulic and S. M. Medenica, "Morphological Characteristics of the Posterior Cerebral Circulation: An Analysis Based on Non-Invasive Imaging," *Turkish neurosurgery,* vol. 29, no. 5, pp. 625-630, 2019.

[6] B. Peng, J. Hao, S. Hou, W. Wu, D. Jiang, X. Fu and Y. Yang, "Possible pathogenesis of painful intervertebral disc degeneration," *Spine,* vol. 31, no. 5, pp. 56-566, 2006.

[7] S. Ghosh and V. Chaudhary, "Supervised methods for detection and segmentation of tissues in clinical lumbar MRI," *Computerized medical imaging and graphics,* vol. 38, no. 7, pp. 639-649, 2014.

[8] A. Neubert, J. Fripp, C. Engstrom, D. Walker, M. A. Weber, R. Schwarz and S. Crozier, "Three-dimensional morphological and signal intensity features for detection of intervertebral disc degeneration from magnetic resonance images," *Journal of the American Medical Informatics Association,* vol. 20, no. 6, pp. 1082-1090, 2013.

[9] Y. Unal, K. Polat, H. E. Kocer and M. Hariharan, "Detection of abnormalities in lumbar discs from clinical lumbar MRI with hybrid models," *Applied Soft Computing,* vol. 33, pp. 65-76, 2015.

[10] C. Bhole, S. Kompalli and V. Chaudhary, "Context sensitive labeling of spinal structure in MR images," *Medical Imaging 2009: Computer-Aided Diagnosis; International Society for Optics and Photonics,* vol. 7260, p. 72603P, 2009.

[11] A. B. Oktay and Y. S. Akgul, "Localization of the lumbar discs using machine learning and exact probabilistic inference," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Berlin, Heidelberg, 2011.

[12] J. Koh, P. D. Scott, V. Chaudhary and G. Dhillon, "An automatic segmentation method of the spinal canal from clinical MR images based on an attention model and an active contour model," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011.

[13] S. Schmidt, J. Kappes, M. Bergtholdt, V. Pekar, S. Dries, D. Bystrov and C. Schnörr, "Spine detection and labeling using a parts-based graphical model," in *Biennial International Conference on Information Processing in Medical Imaging*, Springer, Berlin, Heidelberg, 2007.

[14] C. L. Hoad and A. L. Martel, "Segmentation of MR images for

computer-assisted surgery of the lumbar spine," *Physics in Medicine & Biology,* vol. 47, no. 19, p. 3503, 2002.

[15] E. Ebrahimzadeh, F. Fayaz, F. Ahmadi and M. Nikravan, "A machine learning-based method in order to diagnose lumbar disc herniation disease by MR image processing," *MedLife Open Access,* vol. 1, no. 1, p. 1, 2018.

[16] C. Chevrefils, F. Chériet, G. Grimard and C. E. Aubin, "Watershed segmentation of intervertebral disk and spinal canal from MRI images. In," in *International Conference Image Analysis and Recognition*, Springer, Berlin, Heidelberg, 2007.

[17] D. Steffens, M. J. Hancock, L. S. Pereira, P. M. Kent, J. Latimer and C. G. Maher, "Do MRI findings identify patients with low back pain or sciatica who respond better to particular interventions? A systematic review," *European Spine Journal,* vol. 25, no. 4, pp. 1170-1187, 2016.

[18] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo and G. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics,* vol. 21, no. 1, pp. 4-21, 2016.

[19] R. P. Jackson, J. E. Cain Jr, R. R. Jacobs, B. R. Cooper and G. E. McMANUS, "The neuroradiographic diagnosis of lumbar herniated nucleus pulposus: II. A comparison of computed tomography (CT), myelography, CT-myelography, and magnetic resonance," *Spine,* vol. 14, no. 2, pp. 1362-1367, 1989.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint,* p. arXiv:1409.1556, 2014.

[21] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[22] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin and P. A. Heng, "3D deeply supervised network for automated segmentation of volumetric medical images," *Medical image analysis,* vol. 41, pp. 10-54, 2017.

[23] Y. Cai, M. Landis, D. T. Laidley, A. Kornecki, A. Lum and S. Li, "Multi-modal vertebrae recognition using transformed deep convolution network," *Computerized medical imaging and graphics,* vol. 51, pp. 11-19, 2016.

[24] H. Chen, Q. Dou, L. Yu, J. Qin and P. A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage,* vol. 170, pp. 446-455, 2018.

[25] A. Suzani, A. Seitel, Y. Liu, S. Fels, R. N. Rohling and P. Abolmaesumi, "Fast automatic vertebrae detection and localization in pathological ct scans-a deep learning approach," in *International conference on medical image computing and computer-assisted intervention*, 2015.

[26] N. F. Harun, K. M. Yusof, M. Z. Jamaludin and S. A. H. S. Hassan, "Motivation in problem-based learning implementation," *Procedia-Social and Behavioral Sciences,* vol. 56, pp. 233-242, 2012.

[27] A. S. Al-Kafri, S. Sudirman, A. Hussain, D. Al-Jumeily, F. Natalia, H. Meidia and M. ... Al-Jumaily, "Boundary delineation of MRI images for lumbar spinal stenosis detection through semantic segmentation using deep neural networks," *IEEE Access,* vol. 7, pp. 43487-43501, 2019.

[28] Y. Zhou, Y. Liu, Q. Chen, G. Gu and X. Sui, "Automatic lumbar MRI detection and identification based on deep learning," *Journal of digital imaging,* vol. 32, no. 3, pp. 513-520, 2019.

[29] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen and T. ... Vercauteren, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE transactions on medical imaging,* vol. 37, no. 7, pp. 1562-1573, 2018.

[30] W. Mbarki, M. Bouchouicha, S. Frizzi, F. Tshibasu, L. B. Farhat and M. Sayadi, "Lumbar spine discs classification based on deep convolutional neural networks using axial view MRI," *Interdisciplinary Neurosurgery,* vol. 22, p. 100837, 2020.

[31] J. T. Lu, S. Pedemonte, B. Bizzo, S. Doyle, K. P. Andriole and M. H. e. a. Michalski, "Deep Spine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning," in *Machine Learning for Healthcare Conference*, 2018.

[32] Q. Pan, K. Zhang, L. He, Z. Dong, L. Zhang and X. e. a. Wu, "Automatically Diagnosing Disk Bulge and Disk Herniation With Lumbar Magnetic Resonance Images by Using Deep Convolutional Neural Networks: Method Development Study," *JMIR medical informatics,* vol. 9, no. 5, p. e14755, 2021.

[33] Q. Zhang, A. Bhalerao and C. Hutchinson, "Weakly-supervised evidence pinpointing and description," in *International Conference on Information Processing in Medical Imaging*, 2017.

[34] A. Jamaludin, T. Kadir and A. Zisserman, "SpineNet: automated classification and evidence visualization in spinal MRIs," *Medical image analysis,* vol. 41, pp. 63-73, 2017.

[35] A. Jamaludin, M. Lootus, T. Kadir, A. Zisserman, J. Urban and M. C. e. a. Battié, " ISSLS PRIZE IN BIOENGINEERING SCIENCE 2017: Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist," *European spine journal,* vol. 26, no. 5, pp. 1374-1383, 2017.

[36] S. Sudirman, A. Al Kafri, F. Natalia, H. Meidia, N. Afriliana and W. e. a. Al-Rashdan, *Lumbar Spine MRI Dataset,* 2019.

[37] "Clinical Centre of Kragujevac," [Online]. Available: https://www.kc-kg.rs/. [Accessed 21 December 2020].

[38] A. Gulli and S. Pal, Deep learning with Keras, Packt Publishing Ltd., 2017.

[39] S. Moradi, A. Alizadehasl and J. e. a. Dhooge, "MFP-Unet: A Novel Deep Learning Based Approach for Left Ventricle Segmentation in Echocardiography," *Physica Medica,* pp. 58-69, 2019.

[40] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.*, 2015, October.

[41] M. D. Bloice, P. M. Roth and A. Holzinger, "Biomedical image augmentation using Augmentor," *Bioinformatics,* vol. 35, no. 21, pp. 4522-4524, 2019.

[42] M. D. Bloice, C. Stocker and A. Holzinger, "Augmentor: an image augmentation library for machine learning," *arXiv ,* vol. arXiv preprint, p. 1708.04680, 2017.

[43] N. A. Farda, J. Y. Lai, J. C. Wang, P. Y. Lee, J. W. Liu and I. H. Hsieh, "Sanders classification of calcaneal fractures in CT images with deep learning and differential data augmentation techniques," *Injury,* 2020.

[44] Y. LeCun, L. Bottou, B. Y and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

[45] A. Krizhevsky, I. Sutskever and H. G. E, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, 2012.

[46] M. Lin, Q. Chen and S. Yan, "Network in network," *arXiv,* vol. preprint arXiv, p. 1312.4400, 2013.

[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed and D. et. al. Anguelov, "Going deeper with convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 1-9, 2015.

[48] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708, 2017.

[49] A. Zhang, Z. C. Lipton, M. Li and A. J. Smola, Dive into deep learning, arXiv preprint arXiv:2106.11342, 2021.

[50] L. Yan, B. Yoshua and H. Geoffrey, "Deep learning," *Nature ,* vol. 521, p. 436–444, 2015.

[51] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, & L. Fei-Fei, "Imagenet: A large-scale hierarchical image database, " *In 2009 IEEE conference on computer vision and pattern recognition,* pp. 248-255, 2009.

[52] E. Salehi, S. Khanbare, H. Yousefi, H. Sharpasand and O. S. & Sheyjani, "Deep Convolutional Neural Networks for Automated Diagnosis of Disc Herniation on Axial MRI," in *IEEE 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019.