Explainable Artificial Intelligence for Doctors Decision Support in Diagnosing Spinal Pathologies

Aleksandra Vatian[1[0000-0002-5483-716X]], Alexey Zubanenko[1[0000-0001-6953-5239]], Pavel Ulyanov[[0009-0008-0641-901X]], Alexander Golubev[1[0000-0001-7417-6947]], Artem Beresnev[1[0000-0002-4646-6856]] and Natalia Gusarova[1[0000-0002-1361-6037]]

[1]1ITMO University, 49 Kronverksky av., St. Petersburg, Russia
alexvatyan@gmail.com

Abstract. This paper presents an AI-based decision support system for spinal pathology diagnostics using MRI. The system incorporates an ensemble of neural networks and explainable AI (XAI) tools based on Grad-CAM. Our approach is aimed not only at enhancing the transparency of AI predictions, but also at improving clinical decisions in diagnostically complex cases. We experimentally show that (1) XAI can be used to restructure the training dataset to improve model performance, and (2) radiologists make more accurate diagnoses when provided with XAI maps alongside standard images. Our system shows promising results in detecting borderline cases of intervertebral disc protrusions, and lays the foundation for integrating XAI into clinical practice.

Keywords: Explainable AI, Grad-CAM, MRI, Spinal Pathology, Medical Imaging, Neural Networks

1 Introduction and Motivation

High-tech medical imaging, and in particular MRI images, are the first-line information sources in diagnosing spinal disorders [1, 2, 3]. The accuracy of human expert judgement of medical images is far below 100% [4, 5] and is determined not only by the qualifications and experience of radiologists, but also by their subjective differences, as well as by the degree of pathology manifestations.

This fully applies to degenerative spinal diseases [6, 7]. For example, intervertebral protrusions are often visually less pronounced compared to extrusions. Meanwhile, according to the Michigan State University classification [9], they represent significantly more complex diagnostic cases, where the boundaries between normal and pathological are blurred.

Today, artificial intelligence (AI) tools declare equal, if not better, accuracy in classifying spinal lesions than human experts [10, 11, 12]. However, they do not go beyond laboratory conditions to widespread clinical practice. This paradox is explained by a range of reasons, among which until recently priority belonged to the "black box" nature of AI. This challenge was expected to be addressed by the numerous means of explainable AI (XAI) [13, 14], however, their inclusion in medical AI tools did not appear to change the situation much. This is largely because the European Medical Device Regulation (EU MDR) endorsed restrictions regarding transparency and other

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_11

XAI features that have to be met before an AI based tool can be implemented in clinical practice [15].

In this regard, it would make sense to shift the focus of researchers from using XAI in medical domain for its "direct purpose", that is, as a means of explaining decisions made by an AI system, towards using XAI as an additional source of information for professionals, along with the image itself and its AI segmentation. The need for such a shift was clearly stated in [16, 17]. The inclusion of XAI tools in the structure of AI medical imaging systems has been widely reported [18, 19, 20, 21]. However, we could detect only a few works containing methodically verified quantitative estimates of its feasibility.

Several studies [22–24] explored diagnostic support using Grad-CAM or segmentation overlays, but were limited in scope or dataset size.

[23] reported that using similar AI support in MRI diagnostics of brain disorders significantly, by almost 4%, decreased the number of cases erroneously classified as healthy (false negatives), even among experienced radiologists.

In [24], the radiologists of various level of qualification were consecutively presented with two series of images: MR images of vertebral bodies with degenerative changes separately, and then in combination with the results of segmentation according to the Modic scale [25] in a special AI system. The agreement between junior and senior neuroradiologists significantly improved in the latter case (from Cohen's kappa score of 0.52 to 0.58). It is noteworthy that both [23] and [24] cases did not raise any diagnostic doubts among experienced radiologists. In particular, [24] reported senior neuroradiologists having almost identical opinions of all the presented cases regardless of the experimental scenarios.

Of much greater research and practical interest are borderline cases, where the probability of error of both a human professional and an AI diagnostic system could be expected to be higher. For example, when diagnosing intervertebral hernias, the error rate even among experienced radiologists would reach 30% and more [26]. In addition to diagnostic errors themselves, this reduces the quality of the dataset markup used to train the AI system, and, accordingly, fundamentally limits its effectiveness.

Our paper aims to examine the effectiveness of XAI as an additional source of information for professionals to alleviate the above limitations in diagnostically complicated cases, using the example of spinal pathologies. We assume that the demonstration of XAI results will help the professional to discern inconspicuous details of the image and thereby increase the accuracy of diagnosis in complicated cases, including those regarded as borderline. To summarize, our contributions are as follows:

- We propose a novel AI system with ensemble architecture for spinal pathology segmentation, equipped with XAI tools in the form of Grad-CAM activation maps with the ability to select controlled neural network (NN) layers for ensemble elements. This would allow the medical practitioner to choose the most informative level of XAI granularity.
- We experimentally show that the results of XAI can be used as a means of modifying the training dataset, which would lead to an increase in the efficiency of the AI system in segmenting borderline cases.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_11

– We experimentally show that demonstration of XAI results along with the original MRI image leads to increased diagnostic performance of professional radiologists in borderline cases.

The remainder of this paper is structured as follows. Section 2 presents the method, including architecture, preprocessing, and experimental setup. Section 3 discusses the results. Section 4 concludes the paper and outlines future directions.

2 Method and Materials

2.1 Neural Network Configuration

The developed NN for MRI image segmentation is based on the ensemble of models with SegResNet, UNETR, and Swin UNET architectures, respectively. The general structure of the developed NN is presented in Figure 1.

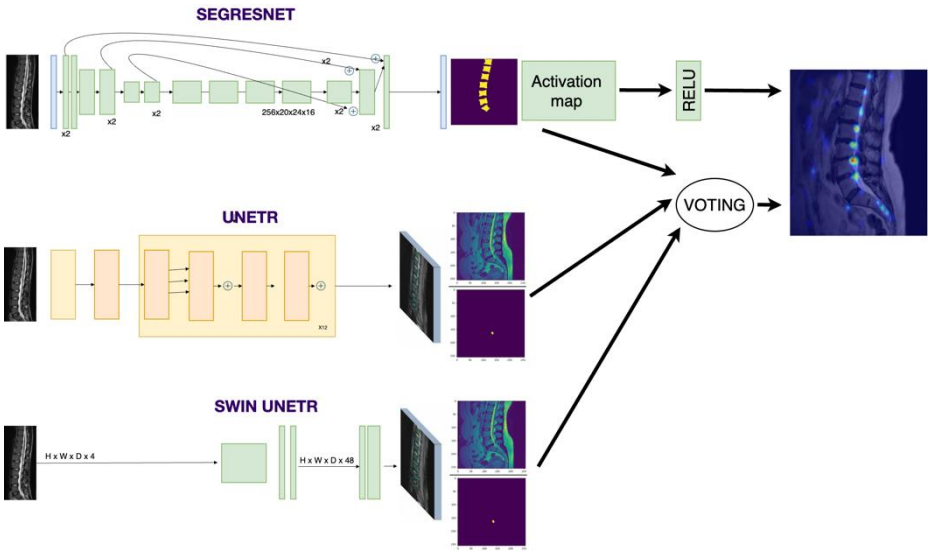

Fig. 1. Developed NN structure

SegResNet [27] from Monai [28] is a deep learning-based segmentation model optimized for medical imaging (e.g. MRI/CT) where edge accuracy and detail preservation are essential [29]. In the proposed NN model SegResNet serves as a backbone. UNETR [30] and Swin UNETR [31] are transformer-based architectures effective for 3D segmentation tasks. Their inclusion ensures a good balance between global context and spatial detail.

The final segmentation of the analyzed image is performed by pixel-by-pixel voting of predictions made by the models participating in the ensemble. The developed NN was trained using the Adam optimizer, an initial learning rate of 0.0001, and the loss

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_11

function Dice + Focal Loss. The 5-fold cross-validation method was used for validation. The plateau criterion was used for early stop of training.

While other segmentation models such as Attention UNet, DeepLabV3+, and V-Net are also popular, they either lack support for volumetric data (e.g., DeepLabV3+) or show limitations in processing long-range dependencies in small datasets. Our selected architectures are complementary in design and offer a well-rounded ensemble suited to the task of detecting subtle spinal abnormalities.

2.2 XAI Implementation

The scheme of XAI implementation in the developed NN is shown in Fig. 2.

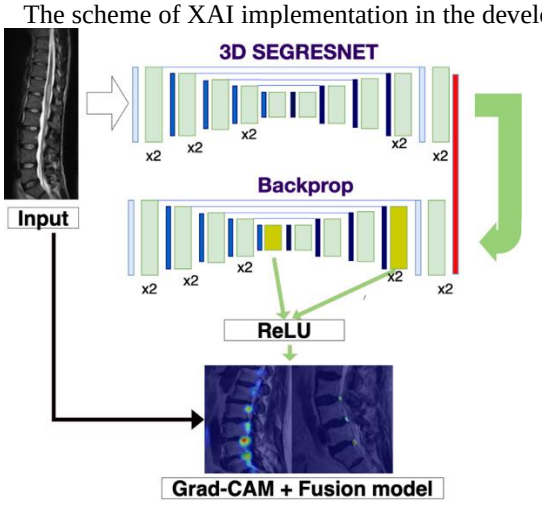

Fig. 2. Scheme of XAI implementation

The method chosen for implementing XAI was Grad-CAM [33], which captures activation maps of the selected NN layer and then overlays them on the input image as a heat map using gradient descent and appropriate spatial transformations. The Grad-CAM method has gained widespread acceptance in medical applications [21, 34, 35] as an intuitive means of visually demonstrating those areas of the input image that appeared to be the most important "from the point of view" of the AI model. Grad-CAM was chosen for its visual clarity and compatibility with CNN architectures, making it especially suitable for radiological interpretation.

Developing the Grad-CAM activation map begins with the selection of the NN target layer, which is shown in yellow in Figure 2. A forward pass is then made through the 3D SEGRESNET architecture, where the input image proceeds through successive convolutional layers (light green blocks) and normalizations. This is followed by backpropagation of the error, calculating gradients in the target layer, which determines the relevance of individual features. The resulting activation map is scaled to the size of the original image and superimposed on the latter as a heat map, which visualizes the areas most highly relevant for the decision made by the model.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_11

The last or penultimate layers of SEGRESNET were used as the target NN layer in our experiments.

2.3 Metrics

The model was evaluated using standard metrics, including Sensitivity, Specificity, Dice Score, and Cohen's Kappa.

2.4 Dataset

A dataset of 1500 axial T2-weighted lumbar spine MRIs (512×512 px, 3 mm slice thickness) was collected from scanners by Siemens, GE, and Philips to ensure imaging variability. All cases were reviewed and labeled by five radiologists in a cross-voting protocol, based on the Lumbar Disc Nomenclature 2.0 and the MSU classification. The dataset includes 200 normal images and 1300 pathological cases, all limited to Grade 1 intervertebral disc protrusions (MSU), deliberately excluding extrusions and sequestrations to focus on diagnostically ambiguous cases. Due to privacy restrictions, the dataset is currently unavailable but is planned for anonymized release.

2.5 Experimental Scenarios

We conducted experiments according to three scenarios.

Scenario 1 We also experimented with restructuring the training dataset based on Grad-CAM sensitivity rankings, which led to improved segmentation performance. Detailed results are omitted due to space constraints.

Scenario 2 was aimed at assessing the impact of Grad-CAM demonstration on the accuracy of diagnostics performed by medical practitioners. Five professional radiologists with different levels of experience (from 2 to 10 years of practical work) were involved in the experiment. A sample of 100 triplets "original medical image + result of segmentation performed by NN + its GradCAM map" was formed. A network trained on a restructured dataset (see Scenario 1) was used as the NN. Twenty triplets from the sample were randomly selected for the demonstration. The radiologists were assigned the task of diagnosing and segmenting the affected area. In the first experiment (**Scenario 2a**), they were demonstrated only the original MRI image, in the second experiment (**Scenario 2b**) - the entire above-described triplet. The randomly selected triplets made repeated demonstration of an image already seen by the doctor highly unlikely.

Scenario 3 Additionally, we evaluated the influence of Grad-CAM granularity levels on diagnostic decisions. These results will be presented in a future extended version.

3 Results and Discussion

Beyond the main diagnostic accuracy study (Scenario 2), we conducted auxiliary experiments involving dataset restructuring (Scenario 1) and varying Grad-CAM

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_11

granularity (Scenario 3). Their results, while promising, are not shown here due to space constraints.

Scenario 2. The results of the experiments for Scenario 2 are presented in Table 2.

Table 2. Performance indicators of medical diagnostics when demonstrating original MRI images and triples of "MRI image + NN segmentation results + GradCAM markup results".

	Scenario 2a				Scenario 2b			
No of participian	Sensitivity	Specificity	Dice	κ	Sensitivity	Specificity	Dice	κ
1	0.87	0.83	0.85	0.81	0.92	0.94	0.87	0.85
2	0.82	0.86	0.80	0.75	0.91	0.92	0.85	0.83
3	0.84	0.82	0.82	0.78	0.90	0.93	0.88	0.86
4	0.86	0.85	0.84	0.80	0.93	0.94	0.86	0.85
5	0.85	0.84	0.84	0.81	0.89	0.92	0.84	0.82
Mean	0.85 ± 0.04	0.84 ± 0.03	0.83 ± 0.03	0.79 ± 0.05	0.91 ± 0.03	0.93 ± 0.03	0.86 ± 0.02	0.84 ± 0.04

Comparison of the results obtained in accordance with scenarios 2a and 2b shows that when demonstrating Grad-CAM maps along with the main MRI images, the efficiency of diagnostics performed by radiologists increases (Sensitivity increased from 0.85 to 0.91, and Specificity – from 0.84 to 0.93). This indicates that medical professional can more accurately determine the presence of pathologies, relying on the visualized attention zones of the model.

4 Conclusion and Future Works

The paper proposes an AI system for segmentation of spinal pathologies with ensemble architecture, equipped with an XAI tool in Grad-CAM form. The efficiency of the proposed system exceeds the SOTA model in diagnosing the most complex, borderline cases of intervertebral hernias.

Our experiments confirm that XAI improves model training and clinical performance in borderline cases. The ability to dynamically adjust XAI granularity enhances diagnostic clarity. Future work includes Active Learning based on XAI focus zones, and adaptive XAI interfaces for clinicians.

Acknowledgments. This work was supported by Russian Science Foundation, Grant № 23-11-00346.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_11

References

1. Merali Z.A., Colak E., Wilson J.R. Applications of Machine Learning to Imaging of Spinal Disorders: Current Status and Future Directions. Global Spine J. 2021 Apr 23;11(1 Suppl):23S–29S. doi: 10.1177/2192568220961353
2. Lee A,, Ong W., Makmur A., , et al. Applications of Artificial Intelligence and Machine Learning in Spine MRI. Bioengineering 2024, 11(9), 894; https://doi.org/10.3390/bioengineering11090894
3. Xuan J., Ke B., Ma W., et al. Spinal disease diagnosis assistant based on MRI images using deep transfer learning methods. Front. Public Health , 24 February 2023. Sec. Digital Public Health
4. Fu M.C., Buerba R.A., Long W.D., et al. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. Spine J. 2014 Oct 1;14(10):2442-8. doi: 10.1016/j.spinee.2014.03.010. Epub 2014 Mar 15.
5. Kim J.-H., van Rijn R.M, van Tulder M.W. Diagnostic accuracy of diagnostic imaging for lumbar disc herniation in adults with low back pain or sciatica is unknown; a systematic review. Chiropr Man Therap. 2018 Aug 21;26:37. doi: 10.1186/s12998-018-0207-x
6. Azimi P., Yazdanian T., Benzel E.C. A Review on the Use of Artificial Intelligence in Spinal Diseases. Asian Spine J 2020; 14(4): 543-571. https://doi.org/10.31616/asj.2020.0147
7. Mbarki W., Bouchouicha M., Frizzi S., et al. Lumbar spine discs classification based on deep convolutional neural networks using axial view MRI, Interdisciplinary Neurosurgery, Volume 22, 2020, 100837, https://doi.org/10.1016/j.inat.2020.100837
8. Fardon D.F., Williams A.L., Dohring E.J., et al. 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. Spine J. 2014 Nov 1;14(11):2525-45. doi: 10.1016/j.spinee.2014.04.022.
9. Mysliwiec L.W., Cholewicki J., Winkelpleck M.D. MSU Classification for herniated lumbar discs on MRI: toward developing objective criteria for surgical selection. Eur Spine J. 2010 Jan 19;19(7):1087–1093. doi: 10.1007/s00586-009-1274-4
10. Liawrungrueang W, Park J-B, Cholamjiak W, et al. Artificial Intelligence-Assisted MRI Diagnosis in Lumbar Degenerative Disc Disease: A Systematic Review. Global Spine J 2024; 21925682241274372LNCS Homepage, http://www.springer.com/lncs, last accessed 2023/10/25
11. Qian J., Su G., Shu X., et al. Lumbar disc herniation diagnosis using deep learning on MRI, Journal of Radiation Research and Applied Sciences, Volume 17, Issue 3, 2024, 100988
12. Amisha P, Malik MP, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care. (2019) 8:2328–31. doi: 10.4103/jfmpc.jfmpc_440_19
13. de Vries B.M., Zwezerijnen G.J.C., Burchell G.L., et al. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front. Med. , 12 May 2023. Sec. Nuclear Medicine. Volume 10 - 2023 | https://doi.org/10.3389/fmed.2023.1180773
14. Hafeez Y., Memon K., AL-Quraishi M.S. Explainable AI in Diagnostic Radiology for Neurological Disorders: A Systematic Review, and What Doctors Think About It. Diagnostics 2025, 15(2), 168; https://doi.org/10.3390/diagnostics15020168
15. Beckers R, Kwade Z, Zanca F. The EU medical device regulation: implications for artificial intelligence-based medical device software in medical physics. Phys Med. (2021) 83:1–8. doi: 10.1016/j.ejmp.2021.02.011

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_11

16. Knapič S., Malhi A., Saluja R. Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain. Mach. Learn. Knowl. Extr. 2021, 3(3), 740-770; https://doi.org/10.3390/make3030037
17. Chen H., Gomez C., Huang C.M. et al. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. npj Digit. Med. 5, 156 (2022). https://doi.org/10.1038/s41746-022-00699-2
18. van der Velden B.H.M., Kuijf H.J., Gilhuijs K.G.A., Max A. et al. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, Volume 79, 2022, 102470, https://doi.org/10.1016/j.media.2022.102470.
19. Borys K., Schmitt Y.A., Nauta M., et al. Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches. European Journal of Radiology, Volume 162, May 2023, 110787.
20. Fontes M., De Almeida J. D. S., Cunha A., Application of Example-Based Explainable Artificial Intelligence (XAI) for Analysis and Interpretation of Medical Imaging: A Systematic Review. IEEE Access, vol. 12, pp. 26419-26427, 2024, doi: 10.1109/ACCESS.2024.3367606
21. M M.M., Manesh T.R., V V.K. et al. Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. BMC Med Imaging 24, 107 (2024). https://doi.org/10.1186/s12880-024-01292-7
22. Chien J.-C., Lee J.-D., Hu C.-S., et al. The Usefulness of Gradient-Weighted CAM in Assisting Medical Diagnoses. Appl. Sci. 2022, 12(15), 7748; https://doi.org/10.3390/app12157748
23. Finck T., Moosbauer J., Probst M., et al. 2022. Faster and Better: How Anomaly Detection Can Accelerate and Improve Reporting of Head Computed Tomography. Diagnostics 12, no. 2: 452. https://doi.org/10.3390/diagnostics12020452
24. Gao K.T., Tibrewala R., Hess M. Automatic detection and voxel-wise mapping of lumbar spine Modic changes with deep learning. JOR Spine. 2022;5:e1204. jorspine.com 1 of 10 https://doi.org/10.1002/jsp2.1204
25. Lange M.B., Petersen L.J., Lausen M.. Influence of Prior Imaging Information on Diagnostic Accuracy for Focal Skeletal Processes—A Retrospective Analysis of the Consistency between Biopsy-Verified Imaging Diagnoses. Diagnostics 2022, 12(7), 1735; https://doi.org/10.3390/diagnostics12071735
26. Modic MT, Steinberg PM, Ross JS, et al. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology. 1988;166(1 Pt 1):193-199.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_11