Text Guidance and Distance Gating Improve Detection and Segmentation of Focal Cortical Dysplasia

German Mikhelson[1], Annalena Lange[1,2,3], Theodor Rüber[2,3,4,5], MELD Working Group[7], and Thomas Schultz[1,6]

[1] b-it and Institute of Computer Science, University of Bonn, Bonn, Germany
[2] Department of Neuroradiology, University Hospital Bonn, Bonn, Germany
[3] Department of Epileptology, University Hospital Bonn, Bonn, Germany
[4] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[5] Center for Medical Data Usability and Translation, University of Bonn, Bonn, Germany
[6] Lamarr Institute for Machine Learning and Artificial Intelligence, North Rhine-Westphalia, Germany
[7] MELD Working Group, MELD FCD study website
s17gmikh@uni-bonn.de, {annalena.lange,theodor.rueber}@ukbonn.de, MELD.study@gmail.com, schultz@cs.uni-bonn.de

Abstract. Focal Cortical Dysplasia (FCD) is a leading cause of drug-resistant focal epilepsy, and its identification on Magnetic Resonance Imaging (MRI) can enable potentially curative surgical treatment. FCD features are often subtle and heterogeneous and current deep learning models for automated detection exhibit limited sensitivity. In clinical workflows, prior knowledge about the suspected localization of seizure onset can guide lesion search. We hypothesize that this prior knowledge can similarly guide a deep learning model in the presence of ambiguous imaging features. We propose a text-guided graph neural network that incorporates prior knowledge to guide FCD detection on cortical features derived from MRI. We further introduce a cluster-wise distance-based gating mechanism that leverages the per-vertex distance regression head at inference to suppress false-positive predictions. Experimental results demonstrate that hemisphere- and lobe-level localization information improves sensitivity and segmentation quality, while the gating mechanism substantially increases specificity. Code is available in a GitHub repository.

Keywords: Focal Cortical Dysplasia · Lesion Detection · Surface-Based Features · Multimodal Learning

1 Introduction

Automatic analysis of medical images using deep learning has achieved strong performance across a range of diagnostic tasks. However, the detection of epileptogenic lesions, in particular FCD, remains substantially more challenging. FCD

is a localized abnormality of cortical development and a common cause of drug-resistant focal epilepsy. These lesions are typically small and often lack clear visual boundaries, making them difficult to identify even for expert neuroradiologists.

Earlier approaches for automated FCD detection have predominantly relied on volumetric convolutional neural networks applied to 3D MRI data [1,4]. However, these methods did not explicitly model the cortical surface, where FCD-related abnormalities manifest as subtle changes in cortical morphology and signal. Many studies were evaluated on limited cohort sizes, which can constrain robustness and generalization across sites and scanners. More recent work, such as the Multicentre Epilepsy Lesion Detection (MELD) framework [8], addresses these limitations by representing the cortex as a surface-based graph and applying graph neural networks trained on the largest multi-center dataset of FCD currently available. Although this represents one of the most advanced approaches to date, detection performance is still not on par with clinical expectations, highlighting the intrinsic difficulty of the task. None of these approaches leverages prior clinical knowledge of lesion localization, which is typically available during preoperative examination and could help narrow the model's search area.

Beyond purely visual approaches, recent advances in multimodal and text-guided learning suggest that incorporating non-visual information can help guide models toward clinically relevant regions, particularly in settings with limited annotated data or weak supervision [14,11,5,6].

Contributions. We propose a novel multimodal segmentation framework that integrates surface-based visual features with text descriptions that encode prior knowledge on anatomical localization of the seizure onset (e.g., hemisphere- or lobe-level regions) to guide FCD detection. Our architecture (Fig. 1) uses a pretrained MELD graph surface representation as the vision encoder and a pretrained language model as the text encoder, with fusion via a text-guided attention mechanism [14] adapted to cortical surface data. We introduce a cluster-wise distance-based gating mechanism that repurposes the per-vertex distance regression head from the MELD pipeline at inference time to suppress false-positive cluster predictions, substantially improving specificity. We further investigate how different forms of textual information influence segmentation performance.

2 Methods

2.1 Encoder

The proposed framework (Fig. 1) formulates FCD detection as a surface-based segmentation task, following the MELD framework. It consists of a Vision Encoder, a Text Encoder, and GuideDecoders that fuse vision and text features across decoder stages.

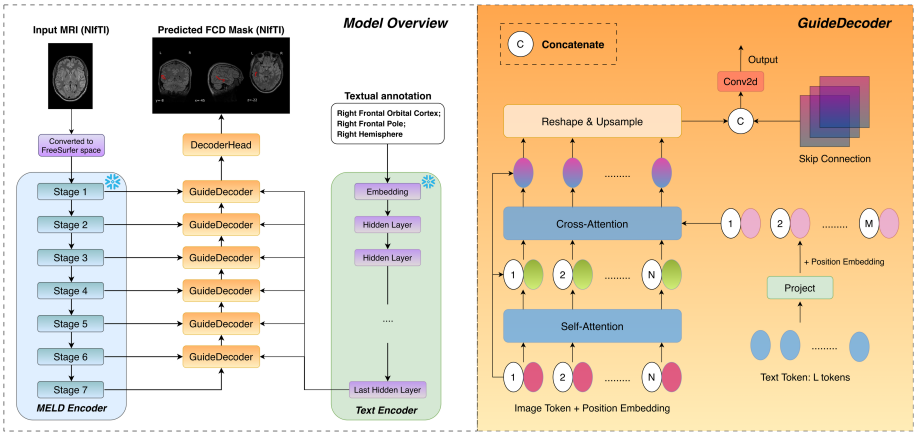

Fig. 1. Overview of the proposed multimodal segmentation framework. Surface-based visual features extracted in FreeSurfer space from MRI are processed by a pretrained MELD encoder and fused with atlas-derived textual embeddings through GuideDecoder blocks. The fused representations are decoded into a vertex-wise FCD prediction and mapped back to the individual subject (scanner) space in NIfTI format. The weights of both pretrained encoders are frozen.

Vision encoder Structural MRI scans are first mapped into FreeSurfer [2] space and processed by the MELD pipeline, yielding multi-resolution surface features at successive decoder stages.

Text encoder We encode textual anatomical descriptions (see Dataset section) using a pretrained RadBERT model [12], chosen for its specialization in radiology reports. The final hidden-layer representation is used as the text embedding and is provided to all decoder stages. Both encoders are kept frozen throughout training.

2.2 GuideDecoder

Multimodal fusion is performed using the GuideDecoder architecture proposed by Zhong et al. [14]. At each decoder stage, the visual features are first refined through a stack of multi-head self-attention layers. The projected textual features then serve as keys and values in a cross-attention module, allowing semantically relevant information from the textual description to guide the visual representations.

Textual embeddings are projected to match the visual feature dimension and used in cross-attention to guide visual representations, following the GuideDecoder design [14]. Positional encodings are added, and the fused features are upsampled and combined with visual skip connections before the final prediction.

In our implementation, the overall GuideDecoder design is preserved, with two modifications to adapt it to surface-based representations. First, we replace

standard 2D upsampling with the HexUnpool operator [10], which performs mean unpooling on the icosphere mesh. Unlike generic interpolation, HexUnpool preserves vertex correspondence across icosphere resolutions and therefore integrates naturally with SpiralConv on the cortical surface. Second, the standard convolution is replaced by a SpiralConv operator [3]. This enables convolutional feature aggregation directly on the cortical surface mesh after upsampling and fusion with skip connections.

2.3 Distance-Based Gating

Text-guided predictions can include false-positive clusters, especially for healthy subjects where the *"unclear"* input provides no spatial localization cue. To reduce the number of false positives, we introduce a cluster-wise distance-based gating mechanism applied only at inference time. The MELD framework includes a per-vertex distance regression head trained to predict the signed distance of each vertex from the lesion boundary [8]: vertices inside the lesion receive negative values, while vertices outside receive positive values proportional to their distance from the boundary. We repurpose this head as a cluster-level gate: for each predicted cluster, the within-cluster distance scores are aggregated by median, and clusters whose aggregated score exceeds a threshold are suppressed as likely false positives. For healthy subjects, predicted clusters typically have high positive distance scores and are therefore filtered out, improving specificity. The threshold is selected only on the validation split using Youden's J statistic [13]. Because this threshold is positive, true-positive clusters that contain near-boundary or negative-distance vertices are retained regardless of lesion size, so the gate does not explicitly favor small or narrow lesions. As shown in Table 4, median, mean, and min aggregation worked similarly well.

3 Experimental Settings

3.1 Dataset

All experiments use the preprocessed MELD dataset [8], a large multi-center collection of presurgical MRI scans from patients with FCD and healthy controls. The dataset was collected from 23 international epilepsy surgery centers using routine 1.5T and 3T MRI protocols. After excluding subjects with missing or corrupted data, 1128 participants were included. Cortical surface features and lesion masks were derived using the MELD preprocessing pipeline and labeling protocol [8].

For model development and evaluation, the dataset was split according to the predefined MELD evaluation protocol. A main cohort of 963 subjects was used for training and testing, with 511 subjects assigned to training and 452 subjects reserved for testing (259 patients with FCD and 193 healthy controls). An independent test cohort of 165 subjects (82 patients and 83 controls) from the Bonn center [9] was used for evaluation.

Table 1. Text conditioning variants used as input to the text encoder.

Name	Description	Example
hemi	Hemisphere name	Left hemisphere / Right hemisphere
lobe	Cortical atlas regions within each lobe	Frontal lobe; Temporal lobe; Limbic lobe ...
hemi+lobe	Hemisphere combined with lobe names	Left hemisphere; Frontal lobe; Temporal lobe; Limbic lobe ...

In addition to the surface-based imaging features, we derive textual anatomical descriptions from cortical atlas information. For patient subjects, the ground-truth lesion ROI is registered to template space and overlaid with anatomical atlases, following the lesion localisation procedure in MELD. Using the AtlasReader library [7], we compute overlap statistics between the ROI and atlas regions and generate textual descriptions of the anatomical context. In this work, we use the Harvard–Oxford probabilistic atlas for regional and lobe-level labels. Because these descriptions are derived from ground-truth lesion masks, they represent oracle localization cues rather than free-form clinical text.

We evaluate three informative text conditioning variants, *hemi*, *lobe*, and *hemi+lobe* (Table 1). The fixed token "unclear" is used whenever no localization cue is provided for healthy controls.

3.2 Loss Function and Evaluation Metrics

To ensure a fair comparison with the MELD framework, we employ the same composite loss function as proposed in [8]. The loss combines binary cross-entropy, Dice loss, distance regression, and weakly supervised subject-level classification, together with deep supervision at intermediate decoder stages. All loss weights follow the original MELD configuration.

Model performance is assessed using IoU, cluster-level PPV (mean across subjects), sensitivity, specificity, and the mean number of false-positive (FP) clusters per FCD patient, following the MELD evaluation protocol [8]. Confidence intervals are obtained via non-parametric bootstrap resampling (10,000 samples).

3.3 Implementation Details

To ensure a fair comparison, we followed the MELD training protocol [8]: unless stated otherwise, all optimization settings, deep supervision, class balancing, and data augmentation were identical to MELD. Training used a batch size of 8 (validation: 4) and a OneCycleLR schedule with an initial learning rate of 3×10^{-4}, a peak learning rate of 3×10^{-3}, 10% warm-up, and cosine annealing thereafter. Models were trained for up to 100 epochs with early stopping after 20 epochs without validation improvement. To enable early stopping and distance-gate calibration, the training cohort was split into 80% training and 20% validation data within each of the five cross-validation folds. The validation-selected

distance-gating thresholds were 0.48 for mean, 0.48 for median, and 0.16 for min aggregation. At test time, predictions were produced by an ensemble of five independently trained models with different random seeds (42–46); per-vertex scores were averaged across ensemble members.

The network jointly integrates surface-based graph features and contextual text embeddings. The graph encoder uses feature channel dimensions $[32, 32, 64, 64, 128, 128, 256]$ across successive encoder stages, while the corresponding text sequence lengths are $[16, 16, 32, 32, 64, 64, 128]$, with a maximum input length of 256 tokens. The text encoder was initialized from RadBERT with a projection dimension of 768.

To improve robustness to missing or imprecise localization information, two text conditioning strategies were applied during training. First, for each patient all three variants (hemi, lobe, hemi+lobe) were pre-generated and one was sampled randomly at each training step. Second, with probability 10% the text input was replaced by *"unclear"*, encouraging robustness when localization is unavailable.

Experiments ran on a single NVIDIA A100 (80 GiB) using PyTorch 2.1.0 (CUDA 12.1), PyTorch Geometric 2.5.3, and Python 3.9.

4 Results

Across both cohorts, Vision+LM improves lesion detection and segmentation compared to the MELD baseline, yielding higher sensitivity and IoU.

Table 2. Median performance on the main cohort (95% CI in parentheses). Gate: – = no gating; med = median distance gating. **Bold** indicates best result per column.

Model	Text	Gate	Cl. PPV ↑	IoU ↑	Spec. ↑	Sens. ↑	FP/FCD ↓
MELD	–	–	**0.515**	0.130 (0.057, 0.190)	58.0%	65.6%	**0.28**
Vision+LM	hemi	–	0.335	0.135 (0.070, 0.191)	27.5%	69.1%	0.88
		med	0.487	0.135 (0.070, 0.191)	**93.3%**	69.1%	0.80
Vision+LM	lobe	–	0.347	0.130 (0.075, 0.204)	27.5%	69.1%	0.80
		med	0.497	0.130 (0.075, 0.204)	**93.3%**	69.1%	0.77
Vision+LM	hemi+lobe	–	0.343	**0.140** (0.079, 0.206)	27.5%	**69.5%**	0.81
		med	0.494	**0.140** (0.079, 0.206)	**93.3%**	**69.5%**	0.76

Table 3. Median performance on the independent (Bonn) cohort (95% CI in parentheses). Gate: – = no gating; med = median distance gating. **Bold** indicates best result per column.

Model	Text	Gate	Cl. PPV ↑	IoU ↑	Spec. ↑	Sens. ↑	FP/FCD ↓
MELD	–	–	0.513	0.218 (0.117, 0.303)	59.0%	69.5%	**0.20**
Vision+LM	hemi	–	0.275	0.306 (0.186, 0.412)	37.3%	**78.0%**	1.41
		med	0.504	**0.318** (0.182, 0.412)	**94.0%**	73.2%	0.70
Vision+LM	lobe	–	0.341	0.251 (0.165, 0.388)	37.3%	72.0%	0.65
		med	**0.547**	0.251 (0.149, 0.388)	**94.0%**	69.5%	0.52
Vision+LM	hemi+lobe	–	0.337	0.292 (0.185, 0.406)	37.3%	74.4%	0.77
		med	0.545	0.292 (0.185, 0.406)	**94.0%**	70.7%	0.56

On the main cohort (Table 2), sensitivity increases from 65.6% (MELD) to 69.5% (Vision+LM, hemi+lobe). On the independent cohort (Table 3), the gain is larger, reaching 78.0% with hemisphere-level prompts. However, without distance gating Vision+LM exhibits an increased number of false positives, reflected in reduced specificity (27.5% / 37.3% on the main / independent cohorts) and higher FP per FCD.

Table 4. Ablation on the main cohort (95% CI in parentheses). *Gate*: cluster-wise distance gating using different aggregation rules (–/mean/med/min). **Bold** indicates best result per column.

Model	Text	Gate	Cl. PPV ↑	IoU ↑	Spec. ↑	Sens. ↑	FP/FCD ↓
MELD	–	–	0.515	0.130 (0.057, 0.190)	58.0%	65.6%	0.28
		mean	0.900	0.131 (0.057, 0.195)	**100.0%**	65.6%	0.07
		med	0.900	0.131 (0.057, 0.195)	**100.0%**	65.6%	0.07
		min	**0.945**	0.131 (0.057, 0.195)	**100.0%**	65.6%	**0.04**
Vision+LM	hemi+lobe	–	0.343	**0.140** (0.079, 0.206)	27.5%	**69.5%**	0.81
		mean	0.517	0.135 (0.078, 0.204)	95.3%	68.7%	0.71
		med	0.494	**0.140** (0.079, 0.206)	93.3%	**69.5%**	0.76
		min	0.533	**0.140** (0.079, 0.206)	93.3%	69.1%	0.65

Applying cluster-wise median distance gating suppresses most false positives and restores specificity to 93.3% on the main cohort and 94.0% on the independent cohort, while largely preserving sensitivity and IoU. As Table 4 shows, MELD with gating can reach even higher specificity than Vision+LM (100.0% vs. 93.3% on the main cohort), but with lower sensitivity. Since missed lesions remain the main clinical bottleneck in FCD detection, Vision+LM prioritizes sensitivity, detecting more lesions (up to 69.5% vs. 65.6% for MELD on the main cohort) while keeping specificity high after gating. This trade-off, together with text guidance, also yields more spatially coherent lesion masks (Fig. 2).

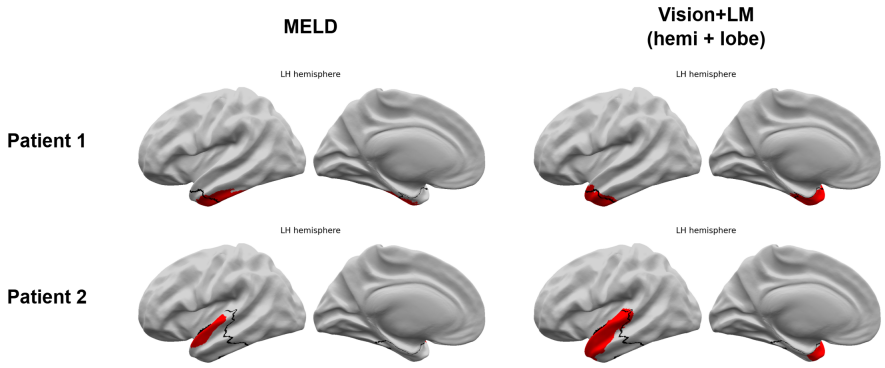

Fig. 2. Qualitative comparison between the MELD baseline and our Vision+LM model trained with hemisphere- and lobe-level textual descriptions. The red overlay shows predicted FCD vertices; the black contour marks the ground-truth (GT) lesion boundary, so red vertices inside the contour correspond to correct lesion predictions.

5 Conclusions

In this work, we proposed a text-guided graph neural network for FCD detection that incorporates atlas-derived anatomical localization cues into a surface-based segmentation framework. Across both cohorts, text-guided models consistently detect more lesions and produce higher-quality segmentations than the MELD baseline, with the most pronounced gains on the independent cohort. Even coarse prompts, such as hemisphere- or lobe-level descriptions are sufficient to improve detection, confirming that the benefit does not require detailed anatomical specificity.

Importantly, the textual descriptions used in this study are derived from ground-truth lesion masks via AtlasReader and therefore represent an oracle, upper-bound evaluation of text guidance. This design isolates the potential value of accurate localization information, but real clinical text from semiology, EEG, or radiology reports is likely to be noisier and less precise; performance may therefore degrade when such free-form clinical narratives are used at test time.

At the same time, text-guided models without gating generate substantially more false-positive clusters than the MELD baseline. The proposed distance gating mechanism effectively reduces false positives and raises specificity, largely closing this gap; nevertheless, false-positive rates remain somewhat elevated compared to MELD, suggesting that further targeted suppression strategies could yield additional gains. In future work, we aim to identify the root cause of these residual false positives — in the text-conditioning pathway, the decoder architecture, or the training objective.

Additional future directions include selectively fine-tuning the text encoder to improve cross-modal alignment, extending the framework to noisy free-form clinical narratives that better reflect clinical variability, and evaluating the model on larger and more heterogeneous cohorts to assess generalizability.

Acknowledgments. We thank Aditya Rastogi for providing access to GPU compute resources used in the experiments and for helpful discussions on the manuscript. We also thank the MELD Working Group for providing the data used in the experiments and for supporting this study. Liaison: Konrad Wagstyl, members: Sophie Adler-Wagstyl, Mathilde Ripart, Carmen Pérez, Ignacio Delgado, Sofia Gonzalez Ortiz, Fernando Cendes, Clarissa Yasuda, Laeticia Ribeiro, Graeme Jackson, Mira Semmelroch, Pasquale Striano, Domenico Tortora, Mariasavina Severino, Ilaria Lagorio, Nuria Bargalló, Saul Pascual-Diaz, José Carlos Pariente, Estefania Conde-Blanco, Maria Eugenia Caligiuri, Angelo Labate, Antonio Gambardella, Lucy Vivash, Ben Sinclair, Patricia Desmond, Elaine Lui, Terry O'Brien, Anna Willard, Wenhan Hu, Kai Zhang, Jia-Jie Mo, Matteo Lenge, Renzo Guerrini, Carmen Barba, Gavin P Winston, John Duncan, Xiaozhen You, William D Gaillard, Nathan T Cohen, Irene Wang, Ryuzaburo Kochi, Yingying Tang, Marcus Likeman, Khalid Hamandi, Shirin Davies, Reetta Kalviainen, Yawu Liu, Stephen Foldes, Zachary Humphreys, Jonathan O'Muircheartaigh, Katy Vecchiato, Nandini Mullatti, Eugenio Abela, Lars Pinborg, Giske Opheim, Ane G Kloster, Jay Shetty, Ailsa McLellan, Jothy Kandasamy, Drahoslav Sokol, Antonio Napolitano, Luca De Palma, Alessandro De Benedictis, Camilla Rossi-Espagnet, Christopher Güttler, Anna Tietze, Terence O'Brien, Rafael Toledano Delgado, Lennart Walger, and Theodor Rüber.

Disclosure of Interests. T.R. receives funding from the German Ministry for Research, Technology and Space (epi-center.ai). The remaining authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Dev, K.M.B., et al.: Automatic detection and localization of focal cortical dysplasia lesions in mri using fully convolutional neural network. Biomedical Signal Processing and Control **52**, 218–225 (2019). https://doi.org/10.1016/j.bspc.2019.04.020
2. Fischl, B.: Freesurfer. NeuroImage **62**(2), 774–781 (2012). https://doi.org/10.1016/j.neuroimage.2012.01.021
3. Gong, S., Zhao, Y., Liu, M., Peng, J.: Spiralnet: A fast and robust mesh convolution operator. arXiv preprint arXiv:1905.00160 (2019)
4. House, P.M., et al.: Automated detection and segmentation of focal cortical dysplasias (fcds) with artificial intelligence: Presentation of a novel convolutional neural network and its prospective clinical validation. Epilepsy Research **172**, 106594 (2021). https://doi.org/10.1016/j.eplepsyres.2021.106594
5. Lee, G.E., Kim, S.H., Cho, J., Choi, S.T., Choi, S.I.: Text-guided cross-position attention for segmentation: Case of medical image. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14224, pp. 537–546. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43904-9_52
6. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: Language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging **43**(1), 96–107 (2023). https://doi.org/10.1109/TMI.2023.3261533
7. Notter, M.P., Gale, D., Herholz, P., Markello, R., Notter-Bielser, M.L., Whitaker, K.: Atlasreader: A python package to generate coordinate tables, region labels, and informative figures from statistical mri images. Journal of Open Source Software **4**(34), 1257 (2019). https://doi.org/10.21105/joss.01257

8. Ripart, M., Spitzer, H., Williams, L.Z., Walger, L., Chen, A., Napolitano, A., et al.: Detection of epileptogenic focal cortical dysplasia using graph neural networks: a MELD study. JAMA Neurology **82**(4), 397–406 (2025). https://doi.org/10.1001/jamaneurol.2024.5406
9. Rüber, T., et al.: An open presurgery mri dataset of people with epilepsy and focal cortical dysplasia type ii. Scientific Data **10**(1), 475 (2023). https://doi.org/10.1038/s41597-023-02386-7
10. Spitzer, H., Ripart, M., Fawaz, A., Williams, L.Z.J., MELD Project, Robinson, E.C., Iglesias, J.E., Adler, S., Wagstyl, K.: Robust and generalisable segmentation of subtle epilepsy-causing lesions: A graph convolutional approach. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14227, pp. 420–428. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43993-3_41
11. Xie, Y., Zhou, T., Zhou, Y., Chen, G.: SimTxtSeg: Weakly-Supervised Medical Image Segmentation with Simple Text Cues. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15008, pp. 634–644. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72111-3_60
12. Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A., Hsu, C.N.: RadBERT: Adapting transformer-based language models to radiology. Radiology: Artificial Intelligence **4**(4), e210258 (2022). https://doi.org/10.1148/ryai.210258
13. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950). https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
14. Zhong, Y., et al.: Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14223, pp. 724–733. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_69