University of Bonn

Master's thesis for obtaining the academic degree
„Master of Science (M.Sc.)“

Detection of Focal Cortical Dysplasia Type II Using Text Descriptions

Author:
Mikhelson German

First Examiner:
Prof. Dr. Very Smart

Second Examiner:
Prof. Dr. Also Smart

Advisor:
Dr. Lange Annalena

Submitted: December 20, 2025

Declaration of Authorship

I declare that the work presented here is original and the result of my own investigations. Formulations and ideas taken from other sources are cited as such. It has not been submitted, either in part or whole, for a degree at this or any other university. During the preparation of the text, I used AI-based language models as writing assistance tools.

Location, Date

Signature

Abstract

Numerous methods have been developed for the detection of tumors in internal organs such as the lungs, brain, kidneys, and breast. However, detecting epileptogenic lesions remains significantly more challenging. Unlike tumors, these lesions do not typically increase in size over time, and there is a severe shortage of publicly available annotated datasets. As a result, researchers often need to contact hospitals and clinical centers directly to obtain even a minimal number of scans. Although recent studies have demonstrated that deep learning models can detect epileptogenic lesions, their performance remains limited, highlighting the ongoing difficulty of this task.

At the same time, recent work on tumor detection using text-guided approaches has shown promising results, demonstrating that incorporating textual descriptions can significantly improve segmentation accuracy. Inspired by these advances, this study proposes a new method that combines visual and textual features to enhance Focal Cortical Dysplasia (FCD) detection under limited data conditions. We further present a systematic comparison of multiple types of textual annotations and analyze their influence on model performance.

Contents

Declaration of Authorship ii

Abstract iii

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Aims of the Thesis . . . 2
- 1.3 Contributions . . . 2
- 1.4 Thesis structure . . . 3

2 Related Works 4

3 Method 6
- 3.1 Visual Feature Extraction . . . 6
- 3.2 GNN Block . . . 7
- 3.3 Textual Feature Extraction . . . 8
- 3.4 GuideDecoder . . . 8

4 Dataset 10
- 4.1 Bonn Dataset . . . 10
- 4.2 MELD Dataset . . . 11
- 4.3 FreeSurfer Processing . . . 11
- 4.4 Data Augmentation . . . 12
- 4.5 Types of Atlas Descriptions . . . 14

5 Experiments 15
- 5.1 Loss function . . . 15
  - 5.1.1 Cross-Entropy Loss . . . 15
  - 5.1.2 Dice Loss . . . 15
  - 5.1.3 Distance Loss . . . 16
  - 5.1.4 Classification Loss . . . 16
  - 5.1.5 Deep Supervision . . . 16

5.2 Metrics 17
5.2.1 Dice score 17
5.2.2 Positive Predictive Value 17
5.2.3 Intersection over Union 17
5.3 Implementation Details 17
5.3.1 Hardware 18
5.4 Basic Experiments 18
5.5 Linking MELD to GNN 21
5.6 Text Guidance in Decoder 22
5.7 Text–GuideDecoder Connections 23

6 Discussion **25**

7 Conclusion **27**

Appendix **28**

References **31**

1 Introduction

This chapter first highlights the importance of epilepsy detection and discusses the limitations of existing methods. It then outlines the main objectives of the thesis and its research contributions. Finally, the chapter concludes with an overview of the thesis structure.

1.1 Motivation

Automatic detection of tumors using medical imaging is widely studied in many internal organs. Recent advances in deep learning and transformer-based approaches have led to impressive results in the diagnosis of lung cancer [1], brain tumor [2], kidney Computed Tomography (CT) [3], and breast cancer [4]. These studies demonstrate the potential of modern computer vision techniques to achieve clinically significant results in various diagnostic tasks.

On the contrary, the detection of epileptogenic lesions, in particular FCD, remains a much more difficult task. Unlike tumors, such lesions usually do not change in size over time, and their inconspicuous appearance makes it difficult even for experienced neuroradiologists to identify them. In addition, the lack of large publicly available annotated datasets is hindering progress in this area. To address this problem, within the framework of the Multicentre Epilepsy Lesion Detection (MELD) project a large-scale collaborative dataset has recently been created and graph neural network–based approaches have been developed for epileptogenic lesion detection [5]. Although this is the most advanced solution to date, performance remains limited, underscoring the continued complexity of this task.

At the same time, *text-guided* and *multimodal approaches* have gained momentum in medical image analysis. Integrating textual prompts or language-guided embeddings with visual features has been shown to substantially enhance segmentation accuracy in chest X-ray infection detection [6], language-guided multi-level alignments [7], organ-aware segmentation [8], multimodal tumor analysis [9], and pneumothorax segmentation [10]. These advances demonstrate that textual annotations, ranging from atlas-based region names to descriptive clinical reports,

provide valuable complementary information. By projecting textual and visual features into a shared latent space, such information can be effectively aligned and leveraged to enhance model performance.

1.2 Aims of the Thesis

The main objective of this thesis is to investigate how the integration of textual and visual information can improve the accuracy of epileptogenic lesion detection. To this end, several research questions were formulated:

1. Does incorporating textual descriptions from anatomical atlases influence the accuracy of lesion segmentation?
2. Can competitive performance be achieved using only partial atlas information (e.g., hemisphere or lobe labels)?
3. How does the integration of additional Graph Neural Network (GNN) blocks on top of MELD-based representations affect segmentation quality?

We systematically evaluate the impact of different types of textual information, as well as alternative strategies for combining visual features, including the integration of additional GNN layers. We further analyze how these design choices influence segmentation performance, demonstrating that the joint integration of language and visual features can enhance both the accuracy and the sensitivity of epileptogenic lesion detection.

1.3 Contributions

The main contributions of this thesis can be summarized as follows:

1. Proposes a novel multimodal architecture for the detection of epileptogenic lesions, c ombining visual representations with language-guided features.
2. Conducted a comprehensive set of experiments to explore design choices, including the number of connections in the GNN blocks and the integration of different textual branches in the GuideDecoder, in order to optimize segmentation performance.
3. Systematically investigated the effect of different types of textual descriptions (e.g., full atlas annotations, hemispheric information, lobe-level labels) on model accuracy and sensitivity.

4. Provided thorough documentation and usage guidelines for the developed code.

1.4 Thesis structure

Chapter 2 reviews prior research on FCD detection and text-guided medical image segmentation, highlighting key architectural designs, fusion strategies, and limitations that motivate our approach. Chapter 3 introduces the proposed multimodal segmentation framework for epileptogenic lesion detection, describing its surface-based visual feature extraction, GNN block, RadBERT textual embeddings, and their integration within the GuideDecoder. Chapter 4 details the datasets used in this study, namely the Bonn dataset and the large-scale MELD cohort, along with the preprocessing steps, data augmentation strategies, and types of atlas-based textual descriptions. Chapter 5 TODO

2 Related Works

This section reviews existing approaches to FCD detection and text-guided medical image segmentation for tumor detection tasks, focusing on their architecture, fusion strategies, and limitations.

Early work on FCD detection focused exclusively on vision-based methods [1–4]. It is noteworthy that one of the newest models, the MELD Graph [5], represents the surface of the cerebral cortex as a graph with multiple resolutions and applies a GNN-U-Net to segment lesions. It provides high sensitivity and specificity by identifying characteristic peaks (more than 20% in saliency) and determining calibration reliability using the expected calibration error. However, it has not been tested on patients with multiple FCDs, it lacks cross-attentional mechanisms for more complete integration of features, and it does not integrate textual clinical information or evaluate zero-shot generalization across FCD subtypes.

A second line of research explores language-guided segmentation by embedding text semantics into the segmentation pipeline. Early methods [11, 12] simply tokenized text and merged embeddings with image features via attention, but struggled to capture high-level semantics. More recent approaches [6, 13] employ deep pretrained text encoders (e.g., CXR-BERT [14]) fused with ConvNeXt-Tiny [15] image features. In [7], the authors introduce two novel modules: the Target-sensitive Semantic Distance Module (TSDM), which computes contrastive distances between segmentation masks to focus on disease-related regions, and the Language-guided Target Enhancement Module (LTEM), which applies cross-attention to reinforce critical image areas. A bi-directional contrastive loss (averaged cross-entropy in both image $\rightarrow$ text and text $\rightarrow$ image directions) yields more fine-grained guidance and enables models trained on 25% of the data to outperform single-modal baselines trained on full datasets. However, these improvements come at the cost of higher architectural complexity and a strong reliance on high-quality textual annotations, which may limit robustness and generalizability in real-world clinical settings.

To address unpaired multi-modal data, MulModSeg [9] conditions text embeddings on imaging modality using frozen CLIP [16]/BioBERT [17]/Med-CLIP [18] encoders combined with medical prompts, and alternates training between vision and text branches (3D-UNet[19] or SwinUNETR [20]). This scheme improves

generalization without requiring paired CT/MR scans. Experiments show that varying the CT:MR ratio in training data shifts performance, underscoring the importance of balanced modality representation. However, effectiveness depends critically on prompt template design and alternating-training convergence.

Finally, weakly supervised methods such as SimTxtSeg [21] leverage simple text cues (e.g., "lesion in the left hemisphere") to guide segmentation without requiring pixel-level annotations. SimTxtSeg was evaluated using standard segmentation metrics, including Dice similarity coefficient (DSC), Intersection over Union (IoU), Positive Predictive Value (PPV), Normalized Surface Dice (NSD), and the 95% Hausdorff Distance (HD95), alongside cross-entropy loss. These results demonstrate that even coarse textual hints can substantially improve mask quality over vision-only baselines. However, such approaches typically rely on consistent, high-quality text labels and have not yet been validated on rare pathologies such as FCD.

Building on these insights, we draw inspiration from Ariadne's Thread [6], which employs lightweight text prompts and a GuideDecoder to segment infected regions in chest X-rays. Remarkably, this method achieves over a 6% Dice improvement compared to unimodal baselines while using only 10% of the training set, highlighting the potential of multimodal prompting in low-data regimes. For our FCD task, we adopt the MELD Graph model as a pretrained backbone, since it was trained on the largest dataset among the methods surveyed, providing a robust basis for feature extraction and downstream adaptation.

To the best of our knowledge, no prior study has applied text-guided segmentation methods to the FCD detection task, underscoring the novelty of our approach. In summary, while graph-based GNNs excel at modeling cortical geometry and text-guided methods enrich segmentation with semantic context, no existing approach simultaneously addresses surface-space lesion detection, integration of fine-grained clinical narratives, robustness to scanner-induced domain shifts, and zero-shot generalization to unseen FCD subtypes. In this work, we aim to demonstrate that incorporating textual information is also beneficial for the FCD detection task. To this end, we integrate different types of text descriptions derived from anatomical atlases with vision-based features into a multi-resolution GNN framework, using cross-attention fusion and contrastive alignment to enhance robustness and detection accuracy across heterogeneous datasets.

3 Method

As mentioned earlier, the design of our architecture was inspired by the work of [6]. The key differences in our implementation are threefold: (1) we increase the number of feature extraction layers, (2) we introduce a GNN-based block to better aggregate visual features at the top layers, and (3) we disable the text branch in the final decoder layers to reduce overfitting. The overall architecture is illustrated in Figure 3.1, and each component is described in more detail below.

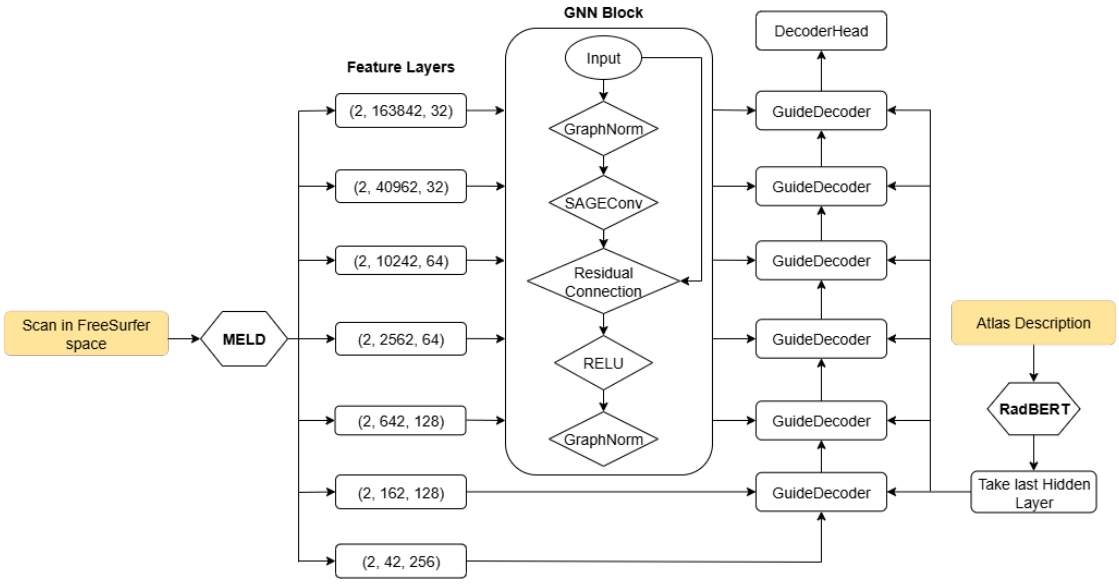

Figure 3.1: Overview of the proposed multimodal segmentation architecture. Visual features extracted from FreeSurfer surface space are processed through hierarchical GNN layers, while textual embeddings from RadBERT guide the decoder via cross-attention.

3.1 Visual Feature Extraction

As input to the vision model, structural MRI scans are first mapped into the FreeSurfer surface space. From the MELD preprocessing pipeline, we obtain a

multi-resolution set of surface-based features across seven hierarchical levels. Each level is represented as a tensor of shape (H, N, C), where:

- H – number of hemispheres (left/right),
- N – number of vertices on the cortical mesh per hemisphere,
- C – number of features per vertex.

To aggregate higher-order geometric information, we process the **6 feature layers** individually through a dedicated **GNN Block**. This choice is motivated by empirical findings (see Chapter 5), where including the lowest two layers led to oversmoothing and diluted discriminative information, while the top 6 layers yielded the best trade-off between expressivity and stability.

3.2 GNN Block

Formally, let

$$\mathbf{X}^{(l)} \in \mathbb{R}^{(B\cdot H\cdot N_l)\times C_l}$$

denote the input feature matrix at layer l, where B is the batch size, H is the number of hemispheres, N_l is the number of vertices per hemisphere at layer l, and C_l is the number of input channels.

Each *GNN Block* then applies the following sequence of operations:

$$\mathbf{H}_0 = \mathrm{GraphNorm}\big(\mathbf{X}^{(l)}, \texttt{batch}\big)\,, \tag{3.1}$$

$$\mathbf{H}_1 = \mathrm{SAGEConv}(\mathbf{H}_0, \texttt{edge\_index}_l)\,, \tag{3.2}$$

$$\mathbf{H}_2 = \mathbf{H}_1 + \mathbf{H}_0 \quad \text{(residual connection)}, \tag{3.3}$$

$$\mathbf{H}_3 = \mathrm{ReLU}(\mathbf{H}_2)\,, \tag{3.4}$$

$$\mathbf{Z}^{(l)} = \mathrm{GraphNorm}(\mathbf{H}_3, \texttt{batch})\,. \tag{3.5}$$

where:

- $\mathbf{X}^{(l)} \in \mathbb{R}^{(B\cdot H\cdot N_l)\times C_l}$ — input feature matrix at level l;
- $\texttt{edge\_index}_l \in \mathbb{N}^{2\times|E_l|}$ — adjacency structure of the cortical surface graph;
- `batch` — batch vector;
- $\mathbf{Z}^{(l)} \in \mathbb{R}^{(B\cdot H\cdot N_l)\times C_l}$ — resulting representation for level l;

- **SAGEConv** implements the GraphSAGE [22] update rule

$$\mathbf{x}_i' = \mathbf{W}_1 \mathbf{x}_i + \mathbf{W}_2 \cdot \underset{j \in \mathcal{N}(i)}{\text{mean}} \mathbf{x}_j,$$

  where $\mathcal{N}(i)$ is the neighborhood of node i;

- **GraphNorm**$(x) = \gamma \cdot \dfrac{x - \mu_g}{\sqrt{\sigma_g^2 + \epsilon}} + \beta$, where μ_g, σ_g^2 are mean and variance per graph, and γ, β are learnable parameters;

- **ReLU**$(x) = \max(0, x)$ — rectified linear activation function;

- **Dropout**(x) randomly sets a subset of elements in x to zero with probability p.

After passing through the GNN Block, the feature dimensionality remains unchanged, ensuring consistency with subsequent blocks. The resulting features are then passed to the GuideDecoder for multimodal fusion.

3.3 Textual Feature Extraction

The text branch of our architecture leverages the pretrained **RadBERT** model [23], which is specifically designed for radiology reports. Prior studies have suggested that selectively unfreezing the last few layers of transformer models can improve downstream performance compared to freezing the entire model [24]. In our implementation, RadBERT is primarily used in a frozen setting, while the potential effect of layer unfreezing is further examined in Chapter 5. The output of the final hidden layer is forwarded to each GuideDecoder for multimodal fusion, except for the topmost three layers.

3.4 GuideDecoder

We adopt the GuideDecoder architecture proposed by Zhong et al. [6], which fuses visual and textual features in a multi–modal fashion. The decoder first projects the textual tokens to align their dimensionality with that of the visual tokens, then applies `multi-head self-attention` and `cross-attention` to exchange semantic information across modalities. Finally, the fused features are upsampled and combined with skip connections from the visual encoder at the same resolution before the final prediction.

In our implementation, the overall design of the GuideDecoder is preserved, but we introduce two modifications to adapt the architecture to surface–based representations. First, we replace the standard 2D upsampling with the `HexUnpool` operator (see Appendix HexUnpool), which performs mean unpooling on the icosphere mesh. Second, instead of the original `DynUNetBlock`[1] from the MONAI framework, which consists of convolution, normalization, and activation layers, we employ a custom `SpiralConv`-based block that directly operates on the cortical surface mesh after upsampling and fusion with skip connections.

[1] https://docs.monai.io/en/0.3.0/_modules/monai/networks/blocks/dynunet_block.html

4 Dataset

We gratefully acknowledge the **MELD Project** for providing access to the dataset used in this work. Without this resource, it would not have been possible to conduct a systematic evaluation of our model.

4.1 Bonn Dataset

The Bonn scientific group has published a publicly available presurgical MRI dataset on **OpenNeuro**, entitled "*An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II*" [25]. This dataset comprises **170 participants**, including **85 individuals with FCD type II** and **85 healthy controls**. For each participant, the following data are available:

- High-resolution **3D T1-weighted** MRI scans (isotropic voxels, $1\,\mathrm{mm}^3$ or $0.8\,\mathrm{mm}^3$ resolution, depending on the subject);
- Corresponding **isotropic 3D FLAIR** imaging, available for most participants;
- **Manually delineated regions of interest (ROIs)** identifying FCD lesions, provided for patients only;
- A set of **clinical and demographic variables** (including age, sex, lesion laterality and location, histopathological subtype IIa/IIb, MRI-negative status, and postoperative outcome according to Engel classification).

Dataset Organization

The dataset follows the BIDS standard. Each subject folder contains an `anat` subfolder with NIfTI files and corresponding JSON metadata. For participants with FCD, lesion ROI masks are also provided. Sequence acquisition parameters are documented in the JSON sidecar files.

4.2 MELD Dataset

The **MELD Project** provides a large-scale neuroimaging dataset comprising Magnetic Resonance Imaging (MRI) scans and clinical data of patients with FCD, as well as healthy controls. In total, the dataset includes **1185 participants** from **23 international epilepsy surgery centers**. Due to missing or corrupted files in the obtained release, we used a subset of **960 participants** in our experiments. The dataset is not publicly available; access must be requested directly from the study authors [5].

For each participant, the dataset includes:

- Structural MRI: **3D T1-weighted** (all participants) and, when available, **3D FLAIR**;
- Lesion annotations: **manually delineated Region of Interest (ROI)** identifying FCD lesions (patients only). For MRI-negative cases, **postsurgical resection cavities** were used to guide ROI definition;
- Demographic and clinical metadata;
- Predefined splits for training, validation, and test cohorts.

4.3 FreeSurfer Processing

Each image was processed with the FreeSurfer framework [26], from which **11 core surface-based features** were extracted:

- **Morphometric features:** cortical thickness, sulcal depth, curvature, and intrinsic curvature;
- **Intensity features:** gray–white matter intensity contrast, and FLAIR intensity sampled at 6 intracortical and subcortical depths.

In addition to the raw measurements, features were further processed into **raw values**, **control-normalized features**, and **asymmetry features** (left vs. right hemisphere). Cortical thickness was additionally adjusted by regressing out curvature. Altogether, this yielded **34 input features per participant**, computed at **163,842 vertices** on a bilaterally symmetrical cortical surface template [5].

Since MRI features vary across scanners and sites, potentially impairing generalization, an **intersite harmonization** procedure was applied using the distributed **ComBat** algorithm, a well-established statistical method for removing scanner/site effects [5].

For conversion from volume space to FreeSurfer surface space, detailed instructions are available in the MELD documentation [27]. In practice, however, it is often more practical to request preprocessed files directly from the authors, since converting a single image to surface space is computationally expensive: approximately 6–7 hours per scan (even with FastSurfer [28] on an NVIDIA A100 GPU, the process required 3–4 hours).

4.4 Data Augmentation

In the original study MELD authors applied augmentations in three stages:

- **Lesion augmentation** — modifying the lesion mask to create new training samples;
- **Mesh–space transforms** — applying geometric transformations to the cortical surface mesh;
- **Intensity transforms** — modifying the intensity features at each vertex.

Each transform is applied independently with probability p defined in the experiment configuration. The order is fixed: lesion augmentation → mesh transforms → intensity transforms.

Lesion–mask augmentation. Given a geodesic distance map D on the cortical surface (negative inside the lesion), we first normalise it by $|\min D|$ and add low–frequency noise, which is generated on a low–resolution icosphere (level 2) and then upsampled to the target resolution using the predefined unpool operators:

$$\tilde{D} \;=\; \texttt{Unpool}\Big(\tfrac{D}{|\min D|} \;+\; \mathcal{N}(0,\sigma^2)\Big)\,, \qquad L' \;=\; \mathbb{1}\{\tilde{D} \le 0\}.$$

In our implementation, we use $\sigma = 0.5$ by default. And as `Unpool` we use the `HexUnpool` operator (see Appendix HexUnpool). After modifying the binary lesion mask L, the geodesic distances and smoothed labels are **recomputed**:

$$D' = \texttt{fast\_geodesics}(L'), \qquad \widetilde{L} = \texttt{smoothing}(L',\, \text{iteration} = 10).$$

This procedure is applied only if the lesion mask is non–empty. In our configurations, however, the recomputed distances and smoothed labels were not used further.

Mesh–space transforms (icosphere re–indexing). We use precomputed index maps (provided by the MELD authors in their GitHub repository) and apply them once per sample to all vertex–wise tensors (features, labels, distances, etc.). The MELD framework defines three types of such maps:

- **Spinning** — index remapping that corresponds to a rigid rotation of the icosphere.
- **Warping** — smooth non-rigid remapping that locally compresses or stretches the mesh.
- **Flipping** — mirror reflection of the mesh, e.g. exchanging left and right hemispheres. *Note:* because SpiralConv is order–sensitive, flipping requires reversing the spiral neighbour order; in our final experiments we set $p(\text{flip})=0$ to avoid this mismatch.

Per–vertex intensity transforms. All intensity transforms act channel–wise on `features`:

- **Gaussian noise**: add $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 \sim \mathcal{U}(0, 0.1)$.
- **Brightness scaling**: multiply each channel by $m_c \sim \mathcal{U}(0.75, 1.25)$.
- **Contrast adjustment**: for each channel c, sample $f \sim \mathcal{U}(0.65, 1.5)$ and apply

$$x_c \leftarrow (x_c - \mu_c)\, f + \mu_c,$$

$$min_c \coloneqq \min(x_c), \quad max_c \coloneqq \max(x_c),$$

$$x_c[x_c < min_c] = min_c, \quad x_c[x_c > max_c] = max_c.$$

- **Gamma**: for $\gamma \sim \mathcal{U}(0.7, 1.5)$,

$$x' \leftarrow \Big(\frac{x - \min(x)}{\max(x) - \min(x) + \varepsilon}\Big)^{\gamma} (\max(x) - \min(x) + \varepsilon) + \min(x);$$

$$x'' \leftarrow \frac{x' - \mu}{\sigma + \varepsilon}, \quad \text{where } \varepsilon = 10^{-7},\ \mu = \text{mean}(x),\ \sigma = \text{std}(x).$$

We also use an *inverted* variant by applying the same operation to $-x$ and negating back.

Disabled/placeholder transforms. The current implementation contains placeholders for `Gaussian blur` and `low-resolution downsampling`. These operations are not implemented in the original authors' code and were not used in our reported results.

4.5 Types of Atlas Descriptions

We generated textual descriptions using `AtlasReader` [29], which registers the input (statistical) MRI map to a template and, for user-specified atlases, produces region names and coordinate tables. In this work we used the Harvard–Oxford (cortical and subcortical) probabilistic atlas [30] and the AAL atlas [31].

However, even strong validation results with full descriptions do not guarantee robustness in real-world settings: in practice, exact overlap percentages and full region names are rarely available. Therefore, in chapter 5 we additionally evaluate reduced-text settings: (i) hemisphere only (Left/Right), (ii) gross lobar labels only (e.g., *Frontal Lobe*, *Temporal Lobe*, *Parietal lobe* ...), and (iii) their combination.

5 Experiments

We investigate three factors: (i) how many MELD feature *stages* are fed into the GNN block; (ii) how many hidden layers from the text encoder are injected into the GuideDecoder; and (iii) the effect of including textual descriptions.

5.1 Loss function

Following best practices for medical image segmentation and to ensure a fair comparison with the MELD model, we employed a composite loss. The loss function is defined as

$$L = L_{ce} + L_{dice} + L_{dist} + L_{class} + \sum_{i \in I_{ds}} w_{ds}^{i} \cdot (L_{ce}^{i} + L_{dice}^{i} + L_{dist}^{i}) \tag{5.1}$$

The individual components are detailed below. Compared to the MELD formulation, we replaced the cross-entropy term with Focal Loss in order to mitigate class imbalance between lesional and non-lesional vertices.

5.1.1 Cross-Entropy Loss

For the segmentation task, we employ the binary cross-entropy loss, which is commonly used in medical image segmentation. Here, y_i denotes the ground-truth label, $\hat{y}_i$ the predicted probability, and n the number of vertices:

$$L_{ce} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right].$$

5.1.2 Dice Loss

The Dice Loss L_{dice} directly optimizes for the overlap between predicted and ground-truth lesion masks. This loss is less sensitive to class imbalance and encourages the network to predict coherent lesion regions. It is defined as

$$L_{dice} = 1 - \frac{2 \sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} \hat{y}_i^2 + \epsilon}$$

It is important to note that we considered two different implementations of this loss. In the MONAI library, the Dice score is calculated for each sample and then averaged over the batch (macro-average), whereas in the MELD implementation the aggregation is performed immediately over the entire batch (micro-average). Under strong class imbalance between background and lesion voxels, the micro-averaged version shifts the loss contribution towards the background, which reduces the gradient signal for rare lesions and leads to training instability. For this reason, we used the MONAI implementation, which provided higher metric values and enabled the model to detect more FCDs.

5.1.3 Distance Loss

To provide the network with additional contextual information and reduce false positives, we include a distance regression loss L_{dist}. The model is trained to predict the normalized geodesic distance d_i of each vertex to the lesion boundary. We use a mean absolute error weighted by $(d_i + 1)^{-1}$, so that errors near the lesion boundary are penalized more strongly than errors in distant non-lesional regions:

$$L_{dist} = \frac{1}{n} \sum_{i=1}^{n} \frac{|d_i - \hat{y}_{i,0}|}{d_i + 1}.$$

5.1.4 Classification Loss

And in the last case, we add a weakly-supervised classification head to mitigate the uncertainty between lesion masks and actual lesions. Each subject is labeled as positive if any vertex belongs to a lesion. The classification head aggregates features across the deepest level (level 1) and predicts a subject-level label $\hat{c}$. The classification loss is then computed as binary cross-entropy:

$$L_{class} = -\sum_{i=1}^{n} c_i \log(\hat{c}_i) + (1 - c_i) \log(1 - \hat{c}_i).$$

5.1.5 Deep Supervision

To encourage gradient flow and stabilize training, we adopt deep supervision at intermediate decoder levels $I_{ds} = \{6, 5, 4, 3, 2, 1\}$. At each level i, auxiliary predictions are generated and the same combination of focal, dice, and distance losses is applied. These auxiliary losses are weighted by w^i_{ds} and added to the total objective:

$$\sum_{i \in I_{ds}} w^i_{ds} (L^i_{foc} + L^i_{dice} + L^i_{dist}).$$

5.2 Metrics

For evaluating the performance of the model we used several commonly applied metrics in medical image analysis: Dice score, Positive Predictive Value (PPV), Intersection over Union (IoU), and Accuracy. Below we briefly describe each of them.

5.2.1 Dice score

The Dice similarity coefficient (DSC) is defined as the harmonic mean between precision and recall. For two sets of predicted positives P and ground truth positives G, it is given by

$$\text{Dice}(P, G) = \frac{2\,|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN},$$

where TP, FP and FN denote true positives, false positives and false negatives, respectively.

5.2.2 Positive Predictive Value

Also known as precision, PPV measures the proportion of correctly identified positive samples among all predicted positives:

$$\text{PPV} = \frac{TP}{TP + FP}.$$

5.2.3 Intersection over Union

The IoU (also called the Jaccard index) quantifies the overlap between predicted and ground truth labels:

$$\text{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|} = \frac{TP}{TP + FP + FN}.$$

5.3 Implementation Details

The proporsed framework was trained with the following hyperparameters. The training batch size was fixed at 8 and the validation batch size at 4. The initial learning rate was set to 3×10^{-4} and optimized using the OneCycleLR scheduler (maximum learning rate 3×10^{-3}). The schedule included a warm-up phase during

the first 10% of training steps, followed by cosine annealing. Training was performed for up to 100 epochs (minimum 20 epochs), with early stopping applied if the validation performance did not improve for 20 epochs. For evaluation, we trained an ensemble of five independently initialized models (seeds 42–46). At test time, we averaged their per-vertex predictions to form the final output.

The model integrated surface-based graph features and contextual text embeddings. Feature channel dimensions across the encoder stages were $[32, 32, 64, 64, 128, 128, 256]$, with corresponding text sequence lengths $[128, 64, 64, 32, 32, 16, 16]$, and a maximum text length of 256 tokens. Deep supervision was employed with levels $I_{ds} = [6, 5, 4, 3, 2, 1]$ and associated weights $w_{ds} = [0.5, 0.25, 0.125, 0.0625, 0.03125, 0.0150765]$. The text encoder was initialized from RadBERT, with a projection dimension of 768.

To address class imbalances, non-lesional hemispheres were undersampled, ensuring that approximately one third of training examples contained a lesion. Data augmentation strategies included random flipping ($p = 0.5$), Gaussian blur ($p = 0.2$), spinning and warping ($p = 0.2$ each), brightness, contrast, gamma correction and Gaussian noise ($p = 0.15$ each), and low-resolution simulation ($p = 0.25$).

5.3.1 Hardware

All experiments were performed on the *Bender* high-performance computing cluster at the University of Bonn, equipped with NVIDIA A100 GPUs (80 GiB each) and AMD EPYC CPUs. A detailed description of the system can be found in the official documentation[1]. Our implementation used PyTorch 2.1.0+cu121, TorchVision 0.16.0+cu121, TorchAudio 2.1.0+cu121, Torch Geometric 2.5.3, Torch Scatter 2.1.2, TorchMetrics 0.11.4, and Python 3.9.18.

5.4 Basic Experiments

Before tuning various hyperparameters such as the number of text connections in the GuideDecoder or the number of unfrozen layers in the LLM, it is essential to first determine the best-performing base model for further experiments. This step saves time and resources by first identifying the most promising model.

We freeze the MELD backbone and unfreeze the last three layers of RadBERT. In all variants, we *train* the same geometry-based upsampling path (HexUnpool + SpiralConv) and the segmentation head. The experiments differ only in whether

[1] https://www.hpc.uni-bonn.de/en/systems/bender

and how the GuideDecoder is used before upsampling, and whether textual features are included.

We consider the following configurations:

- **MELD** serves as the baseline.
- **Exp1 (Unpool+Spiral, no text).** No GuideDecoder is used; features are fed directly into the upsampling path (HexUnpool + SpiralConv) and the segmentation head; no textual input.
- **Exp2 (GuideDecoder, self-attention only).** A GuideDecoder layer is inserted before upsampling, but the text branch is disabled, so only self-attention operates; the same Unpool+Spiral path and segmentation head are trained.
- **Exp3_full (GuideDecoder + Text).** The full GuideDecoder is used (self-attention plus cross-attention to the text encoder); textual descriptions are provided using complete Atlas annotations.
- **Exp3_hemi+lobe (GuideDecoder + Text).** Same as Exp3_full, but Atlas descriptions are reduced to hemisphere and lobe names.
- **Test2 (GuideDecoder + Text).** Same as Exp3_hemi+lobe, but the decoder is initialized from Exp1 and unfrozen from the first epoch.
- **Test3 (GuideDecoder + Text).** Same as Exp3_hemi+lobe, but the decoder is initialized from Exp1 and unfrozen only after 10 epochs.

The last two variants were introduced to investigate the impact of using a pretrained decoder on overall performance. They were applied to the `Exp3_hemi+lobe` model, since later analysis will demonstrate that this variant provides the most consistent balance across metrics.

It is important to note that all experiments in this section were conducted **without** lesion–mask augmentation, flipping, warping, or spinning. Since some augmentations were found to substantially degrade performance, results with augmentation will be presented in the following chapters.

In both tables, the best results are highlighted in green, the second-best in blue, and the third-best in orange.

Table 5.1: Median performance on the main cohort.

Experiment	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
MELD	0.232	0.156	0.707	0.131	162 / 259
Exp1	0.178	0.242	0.685	0.098	157 / 259
Exp2	0.250	0.223	0.645	0.143	166 / 259
Exp3_hemi+lobe	0.260	0.242	0.685	0.149	163 / 259
Test2_hemi+lobe	0.266	0.206	0.558	0.154	172 / 259
Test3_hemi+lobe	0.181	0.271	0.624	0.100	163 / 259
Exp3_full	0.260	0.215	0.457	0.149	175 / 259

Main cohort Among all models, `Exp3_hemi+lobe` exhibits the most stable performance: mid-range Dice, PPV_pixels, and IoU, while its cluster PPV is only about 1.5% lower than MELD. This indicates that hemisphere- and lobe-level textual input provides a balanced trade-off between sensitivity and precision.

The `MELD baseline` takes first place in cluster PPV (0.707) but performs worse on all other metrics. The biggest gap is in PPV_pixels, where MELD is about 12% lower than `Test3_hemi+lobe`, showing weaker pixel-level localization even though cluster detection is fairly consistent.

By contrast, `Exp3_full` achieves the highest lesion detection count (175 vs. 162 for MELD), but at the cost of the lowest cluster PPV (0.457). This shows unstable behavior: the model is very sensitive and segments widely, but also produces many false clusters, reducing precision.

`Test2_hemi+lobe` finds 172 lesions, which is 9 more than `Exp3_hemi+lobe`, but its cluster PPV decreases from 0.685 to 0.558. This shows that immediate fine tuning of the pretrained decoder helps to detect more lesions but lowers precision.

Table 5.2: Median performance on the independent cohort (Bonn Dataset).

Experiment	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
MELD	0.350	0.253	0.709	0.212	54 / 82
Exp1	0.354	0.448	0.802	0.215	49 / 82
Exp2	0.387	0.449	0.704	0.240	51 / 82
Exp3_hemi+lobe	0.396	0.438	0.761	0.247	51 / 82
Test2_hemi+lobe	0.394	0.339	0.590	0.245	53 / 82
Test3_hemi+lobe	0.385	0.452	0.747	0.238	50 / 82
Exp3_full	0.423	0.370	0.415	0.268	57 / 82

Independent cohort (Bonn Dataset) On the Bonn dataset, `Exp3_hemi+lobe` again ranks second in Dice, PPV clusters, and IoU, remaining only 2–3% behind the leading results, which underlines its robustness across datasets.

`Exp3_full` achieves the highest number of detected lesions (57 vs. 54 for MELD),

but as in the main cohort, it yields the weakest cluster PPV (0.415), confirming its unstable, overly sensitive nature.

`Test2_hemi+lobe` detects 2 more lesions than `Exp3_hemi+lobe`, but its performance on Dice and IoU is only marginally lower (by about 0.2%), indicating a small trade-off between sensitivity and segmentation accuracy.

Overall, these results demonstrate that textual guidance consistently improves segmentation quality, even without fine-grained parameter tuning. Importantly, even partial guidance such as hemisphere and lobe information yields meaningful improvements, helping to balance recall and precision.

General conclusions.

- `Exp3_hemi+lobe` is the most balanced and stable model, performing reliably across datasets and metrics.
- `Test2_hemi+lobe` shows slightly less accurate results than `Exp3_hemi+lobe`, but detects more lesions, representing a compromise between accuracy and lesion sensitivity.
- `Exp3_full` achieves the highest sensitivity and lesion coverage but exhibits instability, generating many false positives.
- Decoder tuning strategies (`Test2` vs. `Test3`) reveal a trade-off: early fine-tuning increases recall (higher Dice and lesion counts), while delayed fine-tuning improves precision (higher PPV_pixels).
- The `MELD baseline` consistently underperforms compared to the proposed models, confirming the added value of textual guidance and decoder adaptation.

The full results with confidence intervals are reported in Table 1 and Table 2.

5.5 Linking MELD to GNN

In these experiments we investigated the effect of connecting different numbers of MELD feature stages to the GNN block. The rationale was that higher MELD stages may produce sparse representations, while lower stages provide richer local detail. By progressively adding stages from top to bottom, we aimed to evaluate how multi-stage integration influences model performance. To isolate this effect, the text encoder and GuideDecoder were disabled.

Table 5.3: Different number of MELD stages connected to the GNN block

Experiment	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
Exp1 0 layers	0.148	0.129	0.649	0.080	147 / 259
Exp1 1 layer	0.149	0.141	0.668	0.080	149 / 259
Exp1 2 layers	0.192	0.172	0.553	0.106	156 / 259
Exp1 3 layers	0.142	0.206	0.722	0.076	146 / 259
Exp1 4 layers	0.119	0.174	0.760	0.063	146 / 259
Exp1 5 layers	0.178	0.242	0.685	0.098	157 / 259
Exp1 6 layers	0.234	0.196	0.654	0.133	162 / 259
Exp1 7 layers	0.177	0.174	0.715	0.097	152 / 259

Table 5.4: Different number of MELD stages connected to the GNN block. Independent cohort (Bonn Dataset)

Experiment	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
Exp1 0 layers	0.308	0.372	0.728	0.182	48 / 82
Exp1 1 layer	0.229	0.338	0.718	0.130	48 / 82
Exp1 2 layers	0.379	0.349	0.639	0.234	49 / 82
Exp1 3 layers	0.362	0.475	0.841	0.221	48 / 82
Exp1 4 layers	0.349	0.420	0.868	0.212	48 / 82
Exp1 5 layers	0.354	0.448	0.802	0.215	49 / 82
Exp1 6 layers	0.392	0.352	0.700	0.244	49 / 82
Exp1 7 layers	0.357	0.355	0.879	0.218	47 / 82

Across both cohorts, a clear trade-off emerges. Models with a larger number of connected MELD stages (e.g., with 6 layers) achieved the highest Dice and IoU scores and detected the most lesions, but at the expense of lower precision, as indicated by reduced PPV values. In contrast, configurations with fewer connected stages (e.g., 5 layers) provided more balanced results, yielding higher PPV and thus fewer false positives, but missing more lesions overall.

Given that the primary goal of this work is to maximize lesion detection (i.e., higher sensitivity), we adopt the 6-layer configuration for subsequent experiments. Nevertheless, the 5-layer setting appears preferable in scenarios where precision and false-positive control are prioritized.

5.6 Text Guidance in Decoder

To assess the influence of textual descriptions on segmentation quality, we conducted experiments with different numbers of unfrozen RadBERT layers. We compared settings from a fully frozen text encoder (0 layers) up to complete fine-tuning (12 layers), with intermediate configurations in steps of three layers. To

better assess the effect of unfreezing a small number of layers, we additionally evaluated a configuration with the first three layers unfrozen.

Table 5.5: Different number of unfreezed layers in RadBERT

Experiment	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
0 layers	**0.286**	0.214	0.542	**0.167**	**172 / 259**
3 layers	0.281	0.244	0.630	0.163	164 / 259
6 layers	0.256	0.223	0.587	0.147	166 / 259
9 layers	0.245	**0.263**	**0.634**	0.139	164 / 259
12 layers	0.286	0.202	0.554	0.167	168 / 259

Table 5.6: Different number of unfreezed layers in RadBERT. Independent cohort (Bonn Dataset)

Experiment	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
0 layers	**0.483**	0.427	0.582	**0.319**	**55 / 82**
3 layers	0.456	0.415	**0.803**	0.295	52 / 82
6 layers	0.367	0.398	0.655	0.225	51 / 82
9 layers	0.411	**0.443**	0.721	0.258	52 / 82
12 layers	0.444	0.371	0.663	0.286	53 / 82

The results show a consistent trend across both cohorts. With a fully frozen text encoder (0 layers), the model achieved the highest Dice and IoU, as well as the largest number of detected lesions. This suggests that extensive fine-tuning of RadBERT does not improve overall segmentation performance and may even degrade it.

At the same time, partially unfrozen models (especially with 3 or 9 layers) yielded higher *PPV_clusters* values, indicating fewer false positive clusters and thus a more precise localization of lesions. In particular, the improvement between 0 layers and 3 layers amounts to approximately 9% in the main cohort and about 22% in the independent cohort.

Taken together, these findings reveal a trade-off: freezing RadBERT maximizes sensitivity (number of lesions found), while partial unfreezing improves precision (PPV). Since our primary objective in this study is to detect as many lesions as possible, we adopt the frozen configuration as the default in subsequent experiments. Nevertheless, for applications prioritizing precision, limited fine-tuning of the text encoder may be beneficial.

5.7 Text–GuideDecoder Connections

In these experiments, we examine how the number of LLM connections to the GuideDecoder affects the final performance. We tested configurations ranging

from 0 connections (no text guidance) to 6 connections (text features connected to all GuideDecoder layers).

Table 5.7: Different number of unfreezed layers in RadBERT

Experiment	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
0 connections	0.234	0.196	0.654	0.133	162 / 259
1 connection	0.243	0.184	0.548	0.138	169 / 259
2 connections	0.123	0.116	0.689	0.066	151 / 259
3 connections	0.074	0.076	**0.761**	0.038	139 / 259
4 connections	**0.282**	0.210	0.540	**0.164**	**170 / 259**
5 connections	0.234	0.223	0.614	0.133	161 / 259
6 connections	0.245	**0.227**	0.597	0.140	163 / 259

Table 5.8: Different number of unfreezed layers in RadBERT. Independent cohort (Bonn Dataset)

Experiment	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
0 connections	0.392	0.352	0.700	0.244	49 / 82
1 connection	0.413	0.332	0.545	0.260	**55 / 82**
2 connections	0.339	0.287	0.787	0.204	50 / 82
3 connections	0.319	0.296	**0.875**	0.189	46 / 82
4 connections	**0.433**	0.352	0.555	**0.276**	**55 / 82**
5 connections	0.406	0.397	0.682	0.255	50 / 82
6 connections	0.378	**0.422**	0.760	0.233	50 / 82

In both experiments, using 4 connections yielded the best overall performance, achieving the highest Dice score and IoU, while also detecting the largest number of lesions. This suggests that a moderate level of text guidance provides the optimal balance between leveraging textual information and preserving model stability.

Notably, the difference between using no text guidance (0 connections) and applying it everywhere (6 connections) was relatively small. This implies that excessive text integration may cause overfitting or redundancy, where additional textual connections fail to provide meaningful improvements and can even introduce noise that degrades performance.

Therefore, we adopted 4 connections for all subsequent experiments.

6 Discussion

The experiments conducted in this thesis demonstrate that incorporating textual information into surface-based GNN architectures can substantially improve the detection of FCD type II. At the same time, the results highlight important trade-offs that need to be considered when designing multimodal models.

First, connecting MELD feature stages to the GNN block revealed that deeper integration (6 stages) maximizes sensitivity: Dice, IoU, and the number of detected lesions were highest in this setting. However, precision decreased, as reflected in lower PPV scores. In contrast, fewer connected stages (e.g., 3 or 5) provided better cluster precision, but missed a larger number of lesions. This confirms that multi-stage integration increases lesion coverage, but at the expense of more false positives.

Second, experiments with RadBERT confirmed that full fine-tuning of the language model does not improve segmentation. In fact, the best results were obtained with a frozen text encoder, suggesting that limited training data leads to overfitting when the entire model is updated. Interestingly, partial unfreezing (3–9 layers) improved PPV, i.e. reduced false positives, but consistently lowered Dice and IoU. This indicates that textual guidance can help stabilize localization but does not directly boost lesion sensitivity.

Third, experiments with different forms of textual input demonstrated that even coarse labels (hemisphere or lobe only) meaningfully improved segmentation compared to purely vision-based baselines. Full atlas descriptions increased sensitivity further but produced unstable predictions, generating many false positive clusters. Thus, reduced-text settings may provide the best balance for clinical application.

Several limitations must be acknowledged. The dataset size remains modest compared to other medical imaging benchmarks, and domain shifts across scanners and sites may limit generalizability. Furthermore, the analysis was restricted to type II FCD, and it remains unclear whether the same conclusions extend to other subtypes. Finally, the model relies on atlas-based text generated automatically; the integration of free-form radiology reports may further improve results.

Overall, the discussion of results highlights that textual guidance is a promising direction for FCD detection. However, the balance between sensitivity and precision remains a key design choice that must be adapted depending on whether the

clinical task prioritizes coverage or specificity.

7 Conclusion

This thesis presented a novel multimodal segmentation framework for the detection of focal cortical dysplasia type II, combining surface-based visual features with textual information derived from anatomical atlases. The main findings can be summarized as follows:

- Incorporating textual features consistently improved segmentation performance compared to vision-only baselines, confirming the added value of multimodal integration.
- Connecting six MELD feature stages to the GNN block achieved the highest Dice and IoU scores and maximized the number of detected lesions, although at the cost of reduced precision.
- Frozen RadBERT provided the best overall sensitivity, while partial unfreezing improved PPV and reduced false positives, demonstrating a trade-off between recall and precision.
- Even coarse textual labels such as hemisphere or lobe information yielded substantial improvements, indicating that fine-grained region descriptions are not strictly necessary for performance gains.

Taken together, these results show that multimodal architectures are feasible and beneficial for epileptogenic lesion detection under limited-data conditions. The work establishes a first step toward integrating language and vision for this task. Future research should investigate larger and more heterogeneous datasets, the integration of clinical radiology reports, and alternative fusion strategies to further enhance robustness and generalizability.

Appendix

HexUnpool We implement surface upsampling with a custom `HexUnpool` operator. Given features $X \in \mathbb{R}^{B\times H\times N_{\text{from}}\times C}$ on a coarse icosphere with N_{from} vertices, `HexUnpool` produces $X' \in \mathbb{R}^{B\times H\times N_{\text{to}}\times C}$ on a denser icosphere ($N_{\text{to}} > N_{\text{from}}$). The operation works in two steps:

1. Existing vertices are copied to the corresponding positions in X'.
2. Each new vertex is assigned the mean of its parent vertices from the coarse level, according to predefined upsampling indices $\mathcal{I}$.

Formally, for each new vertex v, we compute

$$X'_v = \frac{1}{|\mathcal{I}(v)|} \sum_{u\in\mathcal{I}(v)} X_u.$$

Experiment results The full results with confidence intervals:

Table 1: Median performance on the main cohort (95% CI in brackets).

Type experiments	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
MELD	0.232 [0.119–0.317]	0.156 [0.089–0.221]	0.707	0.131 [0.063–0.189]	162 / 259
Exp1	0.100 [0.000–0.261]	0.070 [0.000–0.187]	**0.755**	0.053 [0.000–0.150]	139 / 259
Exp2	0.250 [0.150–0.333]	0.223 [0.164–0.352]	0.645	0.143 [0.081–0.200]	166 / 259
Exp3_hemi+lobe	0.260 [0.138–0.336]	0.242 [0.142–0.401]	0.685	0.149 [0.074–0.202]	163 / 259
Exp3_full	0.260 [0.197–0.355]	0.215 [0.129–0.278]	0.457	0.149 [0.109–0.216]	**175 / 259**
Test2_hemi+lobe	**0.266** [0.183–0.334]	0.206 [0.154–0.287]	0.558	**0.154** [0.101–0.201]	172 / 259
Test3_hemi+lobe	0.181 [0.090–0.305]	**0.271** [0.147–0.366]	0.624	0.100 [0.047–0.180]	163 / 259

Table 2: Median performance on the independent cohort (Bonn Dataset, 95% CI in brackets).

Experiment	Dice	PPV pixels	PPV clusters	IoU	Number of FCD found
MELD	0.350 [0.159–0.463]	0.253 [0.093–0.427]	0.709	0.212 [0.086–0.301]	54 / 82
Exp1	0.338 [0.000–0.490]	0.232 [0.000–0.460]	**0.778**	0.203 [0.000–0.325]	46 / 82
Exp2	0.387 [0.092–0.533]	**0.449** [0.142–0.630]	0.704	0.240 [0.048–0.364]	51 / 82
Exp3_hemi+lobe	0.396 [0.048–0.532]	0.438 [0.111–0.686]	0.761	0.247 [0.024–0.363]	51 / 82
Test2_hemi+lobe	0.394 [0.209–0.520]	0.339 [0.155–0.468]	0.590	0.245 [0.117–0.352]	53 / 82
Test3_hemi+lobe	0.385 [0.002–0.503]	0.452 [0.002–0.766]	0.747	0.238 [0.001–0.336]	50 / 82
Exp3_full	**0.423** [0.301–0.537]	0.370 [0.215–0.492]	0.415	**0.268** [0.177–0.367]	**57 / 82**

List of Acronyms

CT Computed Tomography

FCD Focal Cortical Dysplasia

GNN Graph Neural Network

IoU Intersection over Union

MELD Multicentre Epilepsy Lesion Detection

MRI Magnetic Resonance Imaging

PPV Positive Predictive Value

ROI Region of Interest

References

[1] R. Durgam, B. Panduri, V. Balaji, A. O. Khadidos, A. O. Khadidos, and S. Selvarajan. "Enhancing lung cancer detection through integrated deep learning and transformer models". In: *Scientific Reports* 15.1 (2025), p. 15614 (cit. on pp. 1, 4).

[2] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo. "Brain tumor detection based on deep learning approaches and magnetic resonance imaging". In: *Cancers* 15.16 (2023), p. 4172 (cit. on pp. 1, 4).

[3] K. Sharma, Z. Uddin, A. Wadal, and D. Gupta. "Hybrid Deep Learning Framework for Classification of Kidney CT Images: Diagnosis of Stones, Cysts, and Tumors". In: *arXiv preprint arXiv:2502.04367* (2025) (cit. on pp. 1, 4).

[4] A. Mehmood, Y. Hu, and S. H. Khan. "A Novel Channel Boosted Residual CNN-Transformer with Regional-Boundary Learning for Breast Cancer Detection". In: *arXiv preprint arXiv:2503.15008* (2025) (cit. on pp. 1, 4).

[5] M. Ripart, H. Spitzer, L. Z. Williams, L. Walger, A. Chen, A. Napolitano, et al. "Detection of epileptogenic focal cortical dysplasia using graph neural networks: a MELD study". In: *JAMA Neurology* 82.4 (2025), pp. 397–406 (cit. on pp. 1, 4, 11).

[6] Y. Zhong 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*. Cham: Springer Nature, 2023 (cit. on pp. 1, 4–6, 8).

[7] M. Li, M. Meng, S. Ye, M. Fulham, L. Bi, and J. Kim. "Language-guided Medical Image Segmentation with Target-informed Multi-level Contrastive Alignments". In: *arXiv preprint arXiv:2412.13533* (2024) (cit. on pp. 1, 4).

[8] W. Zhang, Z. Zhang, M. He, and J. Ye. "Organ-aware Multi-scale Medical Image Segmentation Using Text Prompt Engineering". In: *arXiv preprint arXiv:2503.13806* (2025) (cit. on p. 1).

[9] C. Li, H. Zhu, R. I. Sultan, H. B. Ebadian, P. Khanduri, C. Indrin, et al. "Mulmodseg: Enhancing unpaired multi-modal medical image segmentation with modality-conditioned text embedding and alternating training". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2025, pp. 3581–3591 (cit. on pp. 1, 4).

[10] Z. Huemann, X. Tie, J. Hu, and T. J. Bradshaw. "ConTEXTual net: a multimodal vision-language model for segmentation of pneumothorax". In: *Journal of Imaging Informatics in Medicine* 37.4 (2024), pp. 1652–1663 (cit. on p. 1).

[11] N. K. Tomar, D. Jha, U. Bagci, and S. Ali. "TGANet: Text-Guided Attention for Improved Polyp Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Cham: Springer Nature Switzerland, 2022, pp. 151–160 (cit. on p. 4).

[12] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong. "LViT: Language meets Vision Transformer in Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* (2023), pp. 1–1 (cit. on p. 4).

[13] G.-E. Lee, S. H. Kim, J. Cho, S. T. Choi, and S.-I. Choi. "Text-Guided Cross-Position Attention for Segmentation: Case of Medical Image". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Cham: Springer Nature Switzerland, 2023, pp. 537–546 (cit. on p. 4).

[14] Microsoft Research. *BiomedVLP-CXR-BERT*. Available at `https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-general`. 2022 (cit. on p. 4).

[15] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. "A ConvNet for the 2020s". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 11976–11986 (cit. on p. 4).

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. PMLR, 2021, pp. 8748–8763 (cit. on p. 4).

[17] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2019), pp. 1234–1240 (cit. on p. 4).

[18] Y. Wang, T. Wang, Y. Li, Y. Gu, M. McDermott, T. Cai, and P. Szolovits. "MedCLIP: Contrastive Learning from Unpaired Medical Images and Texts". In: *arXiv preprint arXiv:2210.10163* (2022) (cit. on p. 4).

[19] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Cham: Springer, 2016, pp. 424–432 (cit. on p. 4).

[20] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, H. R. Roth, and D. Xu. "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022), pp. 2723–2733 (cit. on p. 4).

[21] Y. Xie, T. Zhou, Y. Zhou, and G. Chen. "SimTxtSeg: Weakly-Supervised Medical Image Segmentation with Simple Text Cues". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024*. Vol. 15008. LNCS. Cham: Springer Nature, 2024, pp. 634–644 (cit. on p. 5).

[22] W. L. Hamilton, R. Ying, and J. Leskovec. "Inductive Representation Learning on Large Graphs". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 (cit. on p. 8).

[23] B. H. Park, M. J. Kim, B. S. Kim, H. S. Kim, Y. S. Kim, H. J. Kang, and H. Kim. "Artificial Intelligence for Diagnosis of Focal Cortical Dysplasia: A Narrative Review". In: *Radiology: Artificial Intelligence* 4.6 (2022), e210258 (cit. on p. 8).

[24] M. E. Peters, S. Ruder, and N. A. Smith. "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks". In: *arXiv preprint arXiv:1911.03090* (2019) (cit. on p. 8).

[25] T. Rüber and et al. "An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II". In: *Scientific Data* 10.1 (2023), p. 475 (cit. on p. 10).

[26] B. Fischl. "FreeSurfer". In: *NeuroImage* 62.2 (2012), pp. 774–781 (cit. on p. 11).

[27] *MELD Graph Documentation*. https://meld-graph.readthedocs.io/en/latest/ (cit. on p. 12).

[28] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter. "FastSurfer: Fast and accurate deep learning based neuroimaging pipeline". In: *NeuroImage* 219 (2020), p. 117012 (cit. on p. 12).

[29] M. P. Notter, D. Gale, P. Herholz, R. Markello, M.-L. Notter-Bielser, and K. Whitaker. "AtlasReader: A Python package to generate coordinate tables, region labels, and informative figures from statistical MRI images". In: *Journal of Open Source Software* 4.34 (2019), p. 1257 (cit. on p. 14).

[30] *Harvard–Oxford Cortical and Subcortical Structural Atlases*. RRID:SCR_001476. FMRIB, University of Oxford & Harvard Center for Morphometric Analysis. Available via SciCrunch RRID resolver. 2014 (cit. on p. 14).

[31] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. "Automated Anatomical Labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain". In: *NeuroImage* 15.1 (2002), pp. 273–289 (cit. on p. 14).