Detection of epileptogenic focal cortical dysplasia using graph neural networks: a MELD study

Mathilde Ripart, MSc 1;
Hannah Spitzer, PhD 2,3;
Logan Z. J. Williams, MD 4;
Lennart Walger, MSc 5,6;
Andrew Chen, PhD 7;
Antonio Napolitano, PhD 10;
Camilla Rossi-Espagnet, MD, PhD 11;
Stephen T. Foldes, PhD 12;
Wenhan Hu, MD 13;
Jiajie Mo, MD 13;
Marcus Likeman, MD 14;
Theodor Rüber, MD 5,6;
Maria Eugenia Caligiuri, PhD 15;
Antonio Gambardella, MD 15;
Christopher Guttler, MD 17,18,19;
Anna Tietze, MD, PhD 17,18,19;
Matteo Lenge, PhD 20,21;
Renzo Guerrini, MD 20,21;
Nathan T. Cohen, MD 23;
Irene Wang, PhD 24;
Ane Kloster, MD 27;
Lars H. Pinborg, MD 27;
Khalid Hamandi, MD 28;
Graeme Jackson, MD, PhD 29,30;
Domenico Tortora, MD, PhD 31;
Martin Tisdall, MD 33;
Estefania Conde-Blanco, MD, PhD 36;
Jose C. Pariente, MSc 34;
Carmen Perez-Enriquez, PhD 40,41;
Sofia Gonzalez-Ortiz, MD, PhD 42;
Nandini Mullatti, MD 44;
Katy Vecchiato, MD, PhD 33;
Yawu Liu, MD, PhD 46;
Reetta Kalviainen, MD, PhD 47,46,35;
Drahoslav Sokol, PhD 49;
Jay Shetty, MD 49;
Ben Sinclair, PhD 51;
Lucy Vivash, PhD 51;
Anna Willard, MD, PhD 51;
Gavin P. Winston, MD, PhD 53,54;
Clarissa Yasuda, MD, PhD 55,56;
Fernando Cendes, MD, PhD 55,56;
Russell T. Shinohara, PhD 7;
John S Duncan, MD, PhD 53,58;
J. Helen Cross, MD, PhD 1;
Torsten Baldeweg, MD 1;

Emma C. Robinson, PhD 4;
Juan Eugenio Iglesias, PhD 59,60;
Sophie Adler †, MD, PhD 1,35;
Konrad Wagstyl †, MD, PhD 4,1;
MELD FCD writing group

MELD FCD writing group

Abdulah Fawaz, PhD 4;
Alessandro De Benedictis, MD, PhD 8;
Luca De Palma, MD 9;
Kai Zhang, MD 13;
Angelo Labate, MD 16;
Carmen Barba, MD, PhD 20,21;
Xiaozhen You, PhD 23;
William D Gaillard, MD 23;
Yingying Tang, MD 24,25;
Shan Wang, MD 24,26;
Shirin Davies, MSc 28;
Mira Semmelroch, PhD 29;
Mariasavina Severino, MD 31;
Pasquale Striano, MD, PhD 32;
Aswin Chari, MD, PhD 33;
Felice D'Arco, MD 33;
Kshitij Mankad, MD 33;
Nuria Bargallo, MD, PhD 34,35;
Saul Pascual-Diaz, PhD 37;
Ignacio Delgado-Martinez, MD, PhD 38,39;
Jonathan O'Muircheartaigh, PhD 45;
Eugenio Abela, MD 43;
Jothy Kandasamy, MD 48;
Ailsa McLellan, MD 49;
Patricia Desmond, MD 50;
Elaine Lui, MD, PhD 50;
Terence J. O'Brien, MD 51,52;
Kirstie Whitaker, PhD 57;

1 UCL Great Ormond Street Institute of Child Health, UK
2 Institute for Stroke and Dementia Research, University Hospital, LMU Munich, Germany
3 Institute of Computational Biology, Helmholtz Munich, Germany
4 School of Biomedical Engineering & Imaging Sciences, King's College London, UK
5 Department of Neuroradiology, University Hospital Bonn, Bonn, Germany
6 Department of Epileptology, University Hospital Bonn, Bonn, Germany
7 Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, USA
8 Neurosurgery Unit, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy
9 Neurology Epilepsy and Movement Disorders, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy
10 Medical Physics Unit, Bambino Gesù Children's hospital, IRCCS, Rome, Italy
11 Functional and Interventional Neuroimaging Unit, Bambino Gesù Children's hospital, IRCCS, Rome, Italy

12 Barrow Neurological Institute, Phoenix, AZ, USA
13 Beijing Tiantan Hospital, Beijing, China
14 Bristol Royal Hospital for Children, Bristol, UK
15 Department of Medical and Surgical Sciences, University Magna Graecia, Catanzaro, Italy
16 University of Messina, Italy
17 Charité - Universitätsmedizin Berlin, Germany
18 Freie Universität Berlin, Germany
19 Institute of Neuroradiology, Humboldt-Universität zu Berlin, Germany
20 Neuroscience and Human Genetics Department, Meyer Children's Hospital IRCCS, Florence, Italy
21 University of Florence, Florence, Italy
22 Meyer Children's Hospital IRCCS, Florence, Italy
23 Center for Neuroscience, Children's National Hospital, USA
24 Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH, USA
25 Department of Neurology, West China Hospital of Sichuan University, Chengdu, Sichuan, China
26 Epilepsy Center, Department of Neurology, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China
27 Epilepsy Clinic & Neurobiology Research Unit, Department of Neurology, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark
28 University Hospital of Wales, Cardiff, UK
29 The Florey Institute of Neuroscience and Mental Health, University of Melbourne, Australia
30 Comprehensive Epilepsy Program, Austin Health, University of Melbourne, Victoria, Australia
31 Department of Neuroradiology, IRCCS Istituto Giannina Gaslini, Genova, Italy
32 Department of Neurosciences, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health, IRCCS Istituto Giannina Gaslini, Genova, Italy
33 Great Ormond Street Hospital for Children, London, UK
34 Department of Neuroradiology, Hospital Clinic & Fundació de Recerca Clínic Barcelona-Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain
35 Member of EpiCARE ERN
36 Department of Neurology, Hospital Clínic & Fundació de Recerca Clínic Barcelona-Institut d'Investigacions Biomèdiques August Pi i Sunyer , Barcelona, Spain
37 Fundació de Recerca Clínic Barcelona-Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain
38 Human Anatomy and Embryology Unit, Department of Morphological Sciences, Faculty of Medicine, Universitat Autònoma de Barcelona, Bellaterra, Spain
39 Department of Neurosurgery, Hospital del Mar, Barcelona, Spain
40 Epilepsy Monitoring Unit, Department of Neurology, Hospital del Mar, Barcelona, Spain
41 Epilepsy Unit, Department of Neurology, Hospital Vithas Málaga, Spain
42 Department of Neuroradiology, Hospital Clinic, Barcelona, Spain
43 Zentrum für Entwicklungs- und Neuropsychiatrie, Klinik für Konsiliar, Alters- und Neuropsychiatrie, Psychiatrische Dienste Argau AG, Windisch, Switzerland
44 Dept of Clinical Neurophysiology and Epilepsy, Kings College Hospital, London, UK
45 Forensic and Neurodevelopmental Sciences, Institute of Psychiatry, King's College London, UK
46 Department of Neurology, University of Eastern Finland, Finland
47 Kuopio Epilepsy Center, Kuopio University Hospital, Finland
48 Department of Clinical Neurosciences, Royal Hospital for Children and Young People, Edinburgh, UK
49 Paediatric Neurosciences, Royal Hospital for Children and Young People, Edinburgh, UK
50 Department of Radiology, Royal Melbourne Hospital, University of Melbourne, Australia
51 Department of Neuroscience, School of Translational Medicine, Alfred Health and Monash University, Melbourne, Australia
52 Department of Medicine, The Royal Melbourne Hospital, The University of Melbourne, Victoria, Australia
53 UCL Queen Square Institute of Neurology, London, UK
54 Department of Medicine, Queen's University, Kingston, Canada
55 UNICAMP University of Campinas, Campinas, Brasil
56 Brazilian Institute of Neuroscience and Neurotechnology, Brazil
57 The Alan Turing Institute, London, UK
58 National Hospital for Neurology and Neurosurgery, London, UK
59 Massachusetts General Hospital & Harvard Medical School, USA
60 Centre for Medical Image Computing, UCL, UK

† These authors contributed equally.

Key points

Question: Can we improve the diagnosis of epilepsy-causing focal cortical dysplasias using state-of-the-art AI?

Findings: In this diagnostic study of 703 patients with epilepsy due to focal cortical dysplasia (FCD), a context-aware graph neural network (MELD Graph) detected 64% of lesions previously missed by radiologists, with a high positive predictive value. Interpretable reports facilitate clinical integration by characterizing lesion location, size and morphology alongside the algorithm's prediction confidence.

Meaning: This publicly available interpretable algorithm, validated on a large cohort, advances the use of AI-based radiological adjuncts for early detection and neurosurgical planning in patients with focal epilepsy due to FCD.

Abstract

Importance: A leading cause of surgically remediable, drug-resistant focal epilepsy is focal cortical dysplasia (FCD). FCD is challenging to visualize and often considered "MRI-negative". Existing automated methods for FCD detection are limited by high numbers of false positive predictions, hampering their clinical utility.

Objective: To evaluate the efficacy and interpretability of graph neural networks in automatically detecting FCD lesions on MRI scans.

Design: In this diagnostic study, retrospective MRI data were collated from 23 epilepsy centers worldwide between 2018 and 2022 as part of the MELD Project, and analyzed in 2023. Data from 20 centers were split equally into training and testing cohorts, with data from three centers

withheld for site-independent testing. A graph neural network (MELD Graph) was trained to identify FCD on surface-based features. Network performance was compared to an existing algorithm. Feature analysis, saliencies and confidence scores were used to interpret network predictions.

Setting: Multicentre

Participants: 34 surface-based MRI features and manual lesion masks collated from a convenience sample of 1185 participants, 703 patients with FCD-related epilepsy and 482 controls. 57 participants were excluded during MRI quality control.

Main Outcome(s) and Measure(s): Sensitivity, specificity and positive predictive value (PPV) of automatically identified lesions.

Results: In the test dataset, MELD Graph had a sensitivity of 81.6% in histopathologically-confirmed patients seizure-free one year after surgery and 63.7% in MRI-negative patients with FCD. The PPV of putative lesions from the 260 patients in the test dataset (125 [48%] female, 18.0 [IQR 11.0-29.0] years) was 67% (70% sensitivity, 60% specificity), compared to 39% (67% sensitivity, 54% specificity) using an existing, baseline algorithm. In the independent test cohort (116 patients, 62 [53%] female, 22.5 [IQR 13.5-27.5] years), the PPV was 76% (72% sensitivity, 56% specificity), compared to 46% (77% sensitivity, 47% specificity) using the baseline algorithm. Interpretable reports characterize lesion location, size, confidence and salient features.

Conclusions and Relevance: MELD Graph represents a state-of-the-art, openly available and interpretable tool for FCD detection on MRI scans, with significant improvements in positive predictive value. Its clinical implementation holds promise for early diagnosis and improved management of focal epilepsy, potentially leading to better patient outcomes.

Introduction

Focal cortical dysplasia (FCD) is a leading cause of drug-resistant focal epilepsy in children and adults[1]. It is surgically remediable, with postoperative seizure freedom rates of around 65%[2], but lesions are often considered MRI-negative, and post-surgical outcome is highly affected by early and accurate detection of lesions on MRI[3,4].

Machine learning methods to detect FCD on MRI have improved in recent years[5–7]. However, clinical utility of these approaches is hampered by the multiple putative lesions identified, resulting in a low positive predictive value for each being a true FCD[8]. Previous approaches have addressed constraints on sample sizes and computational memory by subdividing MRI scans into smaller independent "patches"[5–7]. This limits their capacity to prioritize across disparate putative abnormalities.

Here, through the Multicentre Epilepsy Lesion Detection (MELD) project, we leveraged advances in deep learning[9–14] to train a graph neural network to segment FCD on a large multicentre MRI cohort of patients (Figure 1). The novel MELD Graph algorithm, which incorporates whole-brain context to improve specificity and accuracy, was directly compared against our previously published patch-based Multi-Layer Perceptron[5], including on a multicentre independent test cohort. We aimed to create a state-of-the-art, clinically translatable, open-access AI algorithm for the automated detection of FCDs.

Methods

Dataset

The MELD project used an MRI dataset of 1185 participants collated from 23 international epilepsy surgery centers (Figure 1, eTable1). Each center received local ethical approval to retrieve and anonymise retrospective, routinely available clinical data, without requiring explicit consent. Standards for Reporting of Diagnostic Accuracy (STARD) guidelines were followed.

Patients were included if they had a radiological or histopathological diagnosis of FCD. Controls were included if scanned for research purposes or for headache, if they had no other neurological conditions and a normal MRI. For full details on demographic and clinical data and MRI data processing see eMethods.

3D T1-weighted (all participants) and T2-weighted Fluid-Attenuated Inversion Recovery (FLAIR) (where available) MR images were processed using FreeSurfer[15] and 11 features were extracted: cortical thickness, gray-white matter intensity contrast, intrinsic curvature, sulcal depth, curvature and FLAIR intensity sampled at six intra- and sub-cortical depths.

3D regions of interest (ROIs) for each FCD were manually delineated on the MRI[16]. If there was no visible lesion ("MRI-negative"), resection cavities on post-surgical scans were used to guide masking. ROIs were intersected with FreeSurfer surfaces and the features and ROIs were registered to a bilaterally symmetrical template[17] (Figure 1A). Quality control involved outlier identification and removal of patients with missing lesion masks[18].

The main cohort, comprising data from 20 centers, was randomly split 50:50 into training and testing cohorts (Table 1), matching the randomisation of a previous study[5] to enable direct

comparison. Each center had participants in both training and testing cohorts. An additional independent test cohort included three additional centers.

Surface-based features were preprocessed into: raw features, control-normalized features and feature asymmetries (eMethods). An additional feature was cortical thickness with curvature regressed out[19]. This resulted in 34 input features per participant calculated at 163,842 cortical surface vertices. MRI features vary depending on the scanner, worsening algorithm performance on untested scanners. To adjust for scanner effects, extracted features in the independent test cohorts underwent inter-site harmonization to the training cohort using distributedComBat[20]. To estimate the number of subjects needed to generate consistent harmonization parameters, a series of subsampling experiments was performed (eMethods). Model performance on the independent test cohorts was evaluated with and without harmonization.

MELD Graph for FCD Lesion Segmentation

Previous multicentre-validated FCD detection tools restricted model inputs to small $1cm^3$ patches of cortex. Deep learning models, which have the capacity to extract increasingly abstract, complex and large-scale features, have revolutionized biomedical image segmentation. nnU-Net[13], a convolutional neural network, achieves state-of-the-art performance across a range of biomedical segmentation tasks, but is typically applied to regular 2D/3D grids with translation-invariant convolutions. To address the irregularity of folded cortical surfaces, we developed MELD Graph, a graph-based nnU-Net implementation (Figure 1B, eFigure 1, eMethods)[11,13,21]. MELD Graph processes hemispheres with 34 features and 163,842 vertices, predicting lesion segmentations, with neighbor-connected lesional vertices grouped into single “clusters”.

Evaluation of the MELD Graph model

The MELD Graph model and a baseline, previously published MELD Multi Layer Perceptron (MELD MLP) model[5], were evaluated on the test and independent test cohorts (Figure 1C). The following metrics were calculated: sensitivity, specificity, positive predictive value (PPV) and intersection over union (IoU) - a metric for segmentation accuracy (eMethods). MELD Graph lesion detection rates, where a predicted lesion cluster overlapped with (or was within 20mm of) the manual lesion mask, in the test dataset were stratified by demographic factors (eMethods). Confidence intervals and significant differences between models and demographic factors ($p<0.05$) were calculated using bootstrapping and permutation-based null models (eMethods). As an additional measure of model specificity, the number of false positive clusters (defined as predictions beyond 2cm of a lesion mask) in patients and controls was calculated.

MELD Graph performances were compared to a similar model but trained solely on a subset of patients with MRI-negative FCD confirmed histopathologically and controls, to establish whether selective training improved performance on these patients compared to the larger, heterogeneous dataset (eMethods).

Interpretable AI

For all predicted lesional vertices, integrated gradients saliency was computed[22], identifying the relative importance of input variables for a given prediction. Saliency can be averaged across input features to identify the most salient vertices within a lesion, or averaged across lesional vertices to identify which input features were most important to the model's prediction (eMethods). To highlight salient regions, the 20% most salient lesional vertices of each predicted lesion were identified. For these salient vertices, we computed the: prediction confidence (max model prediction score), mean lesional feature values and mean feature salience. The calibration of confidence scores was assessed using Expected Calibration Error.

To characterize the true positive and false positive predictions, as well as the missed FCD lesions (false negatives), mean feature values for these regions were calculated. Detected FCDs were characterized according to their histological subtypes. Additionally, we characterized model performance (sensitivity and PPV) according to lesion size and lesion location (eMethods).

Individual reports

The trained model was incorporated into a pipeline that takes input MRI data (Figure 1D, eFigure 2), performs feature extraction and preprocessing, runs inference on the MELD Graph model, and generates an interpretable report (Figure 1E, eFigure 2). The report details the locations of predicted lesions (and the 20% most salient vertices) on the cortical surface and native T1 as well as model confidence in lesional predictions, lesional features and the associated saliencies of features.

Code and data availability

All code is available to download from www.github.com/MELDProject/meld_graph. Lesion masks in template space are available from https://github.com/MELDProject/pool. Access to the MELD surface-based FCD dataset is via request.

Results

Participants

57 participants were removed during quality control (37 patients with missing lesion masks, 13 outliers and 7 participants where FreeSurfer failed). Data from 20 centers were split into training (278 patients, 180 controls) and testing (260 patients, 193 controls) cohorts (eTable 1). Subjects

from three centers were withheld as an independent test cohort (116 patients, 101 controls). Lesion masks for patients previously considered “MRI-negative” were significantly smaller (median 29% smaller) than visible lesions (Mann Whitney U: $Z=-2.07$, $p<0.04$).

Model evaluation

The performance of the MELD Graph model compared to the baseline multi-layer perceptron (MELD MLP) on the test dataset is presented in Table 1. MELD Graph had significantly higher PPV than the baseline MELD MLP model on test (67% vs 39%) and independent (76% vs 46%) test cohorts, primarily due to a reduction in the number of false positive clusters (Figure 2). On the test dataset, MELD Graph had a maximum of three false positive clusters in patients (median,[IQR]; 0,[0-1]) and two in controls (0,[0-1]), in comparison to MELD MLP for which patients had a maximum of 12 clusters (1,[0-2]), and eight in controls (0,[0-1]) (eTable 2). This significant reduction in false positives for MELD Graph was also seen on the independent test cohort (eTable 2). In patients whose lesions were detected by both MELD Graph and MELD MLP models, the segmentation accuracy, computed as Intersection Over Union, was significantly higher using MELD Graph in the test (0.3 vs 0.23, N=160) and independent test (0.36 vs 0.29, N=78) cohorts (Table 1). Figure 2 provides examples of individual predictions using the MELD Graph and MLP models. MELD Graph predictions do not have additional predictions (false positives), and segmentations are smoother and more contiguous.

Performance on the independent test cohort was calculated with and without harmonization (eTable 3). Sensitivity remained stable, 72% with and 70% without harmonization, whereas specificity dropped from 56% with to 39% without harmonization. Subsampling experiments indicate that a minimum of 20 subjects is required to generate reliable harmonization parameters (eFigure 3). Moreover, training MELD Graph solely on MRI-negative, histopathologically confirmed FCDs, showed a significant drop in the PPV, from 72% to 58% (eTable 4).

Table 2 provides a performance breakdown according to demographic factors. There was a 63.7% detection rate (51/80) in patients previously reported MRI-negative. MELD Graph detected 75.4% of 57 Type IIA and 76.3% of the 93 Type IIB lesions. In seizure-free patients with histopathologically-confirmed FCDs, 81.6% of lesions were detected. Of note, MELD Graph was able to detect 84.6% (11/13) of the particularly subtle FCD Type I lesions. Histopathological subgroups had no significant impact on sensitivity. However the histopathologically unconfirmed group, combining unoperated patients and operated patients in whom histopathology was either unavailable or inconclusive, had lower sensitivity. Lower detection rates in this group might reflect uncertainty in their manually-defined masks.

Analysis of detected and missed lesions in the test cohort (eFigure 4) revealed features driving the algorithm's prediction of lesional vertices. Overall, lesions were characterized by abnormally deep sulci, increased intrinsic curvature and cortical thickness, decreased gray-white matter contrast, decreased gray matter FLAIR intensity and increased white matter FLAIR intensity. FCD Types 2A and 2B demonstrated pronounced folding and thickness abnormalities, with FCD Type 2B having the additional distinctive FLAIR hyperintensity in the white matter (transmantle sign). In comparison, FCD Type 1 and 3 had more subtle cortical tissue abnormalities. Lesions not detected by MELD Graph were characterized by having less abnormal features ($p<0.05$, eFigure 4). Model performance was not significantly associated with either lesion size or location (eFigure 4). In detected clusters in the test cohort, IoU correlated significantly with the confidence score ($r=0.55$, $p<0.01$), indicating higher segmentation accuracies in higher confidence predictions (eFigure 5).

Interpretable AI models:

The MELD Graph tool outputs model predictions in native space on the T1w scan, in NIfTI format, and an interpretable report, in PDF format, containing predicted lesion locations and associated model confidence, as well as characterizing the lesional features and their salience for each putative lesion. Figure 3 and eFigure 6 contain example components from individual patient reports. Patient 1 has a lesion in the precuneus. MELD Graph was 93% confident in its prediction. The lesion, as evident by the feature z-scores, is characterized by blurring of the gray-white matter boundary (decreased gray-white contrast), increased cortical thickness and abnormal FLAIR signal intensity in the gray matter. Most of these features have high saliency scores indicating their importance to the classifier's prediction. Patients 2-4, are examples of "MRI-negative" lesions, not identified by five expert raters[23]. MELD Graph identified these lesions with 7-45% confidence. Interpretable reports for these patients are available at www.github.com/MELDProject/meld_graph.

The Expected Calibration Error, which quantifies the discrepancy between MELD Graph's confidence scores and the observed accuracies, was calculated as 0.10 (where 0 represents perfect calibration). This indicates that model confidence scores were well-calibrated (eFigure 5).

Discussion

MELD Graph is a graph convolutional neural network for automated segmentation of FCD that integrates information across whole cortical hemispheres. The model was tested on a large dataset of 453 participants (from 20 centers) as well as an independent test cohort of 217 participants from three additional centers. MELD Graph accurately localized 81.6% of histopathologically confirmed FCDs in patients seizure-free one year after surgery. Over the entire

dataset, MELD Graph accurately localized 70% of the FCDs from the test dataset and 72% from the independent test cohort, with a fourfold reduction in the number of false positive clusters and a significantly improved PPV over previous approaches. MELD Graph is released as an open-access tool, that takes input MRI data (a preoperative T1-weighted scan plus an optional FLAIR scan) from an epilepsy patient with a suspected FCD, performs feature extraction and processing, runs inference using the MELD Graph model, and generates an interpretable report depicting lesion location, model confidence, and characterizing lesional features and their salience.

Automated tools to detect FCDs

FCD is the most common histopathological finding in surgical patients with MRI-negative epilepsy[24,25]. In MRI-negative patients, the absence of a confident lesional hypothesis can prohibit surgical candidacy and additional noninvasive[26] and invasive[27] investigations to identify a lesion introduce significant delays to surgery. Increased duration of epilepsy is associated with poorer outcome and progressive cognitive sequelae[2,4]. Surgery is curative in around 65% and seizure freedom combined with reductions in anti-seizure medications may halt or even reverse neuropsychological decline[28]. The development of tools like MELD Graph which can successfully localize FCDs across a range of scanners, ages and subtypes, including detecting 64% of MRI-negative FCDs, may therefore improve both seizure freedom and developmental outcomes[28] through earlier diagnosis and surgery[29].

Previous models for detecting FCD have been validated on multicentre data[5–7], but have a recognised propensity to produce multiple false positive predictions in patients[8], hampering their utility as radiological adjuncts. These models have analyzed limited portions of the brain in isolation, either as small patches of surface vertices[18,30,31] or voxels[6,7]. In contrast, the MELD Graph model incorporates whole-brain context, leading to multiple improvements: increased specificity (Figure 2), heightened sensitivity, especially to subtle FCD Type I lesions (Table 2),

better accuracy (Table 1), and well-calibrated confidence estimates (eFigure 5). Evaluated on the same dataset, MELD Graph produced a maximum of three clusters in patients, whereas the MELD MLP model identified up to twelve. For clinical translation, this will reduce the time necessary for clinicians to review model outputs and improve confidence in AI radiological adjuncts. Future integration of MELD Graph into clinical practice may lead to earlier diagnosis of some patients, as well as the detection of subtle lesions in some patients currently considered non-lesional. MELD Graph outputs could be considered in conjunction with other investigations such as electroencephalography (EEG), magnetoencephalography (MEG), and positron emission tomography (PET), or used in the planning of stereoelectroencephalography (sEEG)[32,33]. This may expedite patients' journeys to epilepsy surgery, increasing the likelihood of seizure freedom[2] as well as cessation of neurocognitive decline[28].

Interpretable models for clinical translation

Model interpretability is critical for incorporating AI into radiological review. MELD Graph reports (Figure 3 and eFigure 6) highlight lesion location, imaging characteristics, the lesional patches that were most salient to the model[22] and the model's confidence in its prediction[34]. High-confidence predictions are rarely incorrect, while lower-confidence outputs may highlight subtle abnormalities which warrant careful review (eFigure 5).

MRI-identification and delineation of FCD is often carried out at multiple stages during a patient's journey and the development of lesion detection tools must address these use-cases. These include 1) early review of an MRI by a general radiologist at a non-specialist center, 2) expert neuroradiological review of a patient referred to a surgical centre[35], 3) multidisciplinary review at an epilepsy surgery meeting[36], 4) neurosurgical planning of sEEG implantation[32,33], and 5) neurosurgical planning of lesion resection or ablation[37]. By providing confidence estimates for each putative lesion, clinicians and researchers using MELD Graph can individually choose a

confidence threshold, balancing their need for sensitivity versus specificity. For example, a neuroradiologist at a specialist epilepsy surgery center, may choose to review low confidence clusters in a complicated “MRI-negative” patient. The interpretable lesion reports alongside confidence scores are a step towards clear communication of model outputs to facilitate translation.

Validation of neurotechnology

One challenge for many promising AI algorithms is that their performance typically drops on data from new patients in new settings. Previously unseen MRI scanners, with differing sequence parameters resulting in subtly different image contrasts, represent a domain shift from the training data distribution. Using multicentre data during training is one mitigation strategy as the algorithm is exposed to some of this heterogeneity. However, an essential validation of new machine-learning technologies is testing them on data from withheld, independent test sites. Here, we demonstrate that MELD Graph maintains performance on unseen data from three centers. Optimal specificity on independent test site data was dependent on harmonization of new data to the training cohort. However, this harmonization process requires a minimum of 20 scans acquired on the same scanner (eFigure 3). Therefore, to facilitate re-use of the algorithm on individual MRI scans from patients from new centers, for which such harmonization is not possible, a non-harmonization version is provided with documented performance statistics (eTable 3). Furthermore, the release of MELD Graph as an open-source tool serves to enable independent validation and reuse, including prospective studies to assess whether it accelerates diagnosis, alters treatment and ultimately improves outcomes.

Limitations

Surface-based features have a number of advantages including interpretability and the ability to share anonymised data for reanalysis[38,39]. Nevertheless, the use of predetermined

features limits the potential for the network to learn novel features. Future work, applying deep learning to multimodal data, including MRI, PET, MEG and EEG, will enable complex features to be learnt that may not be visible or known to neuroradiologists. Finally, MELD Graph has not been evaluated on patients with multiple FCDs.

Conclusions

MELD Graph, is a state-of-the-art, openly available, graph convolutional network for FCD detection. With improved performance over existing methods, interpretable predictions, model confidence scores and individual patient reports, MELD Graph will support the integration of lesion detection tools into the radiological workflow.

Acknowledgements

The MELD project, M.R. and S.A. are supported by the Rosetrees Trust (A2665), Epilepsy Research Institute (P2208) and the NIHR GOSH BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. A.N. was funded by PCDH19 Alliance. The work of R.G., C.B., and M.L. from the Meyer Children's Hospital IRCCS was supported, in part, by funds from the 'Current Research Annual Funding' of the Italian Ministry of Health. C.G. and A.T. were funded by the German Research Foundation (DFG, SFB295RETUNE). N.T.C. was supported by a CNF/PERF Shields Award, the CNRI Chief Research Officer Award, Hess Foundation, and Children's National IDDRC. I.W. was funded by NIH R01 NS109439. L.H.P was funded by the Lundbeck Foundation BrainDrugs (R279-2018-1145). The Florey Institute of Neuroscience and Mental Health acknowledges the strong support from the Victorian Government and in particular the funding from the Operational Infrastructure Support Grant. The authors from the Florey Institute of Neuroscience and Mental Health acknowledge the facilities and scientific and technical assistance of the National Imaging Facility, a National Collaborative Research Infrastructure Strategy (NCRIS) capability. Y.L and R.K

acknowledge the Saastamoinen Foundation and Academy of Finland. J.S. was funded by NHS Research Scotland. G.P.W was funded by the MRC (G0802012, MR/M00841X/1). C.Y acknowledges support from CNPQ(315953/2021-7), FAPESP (2013/07559-3). J.D. was funded by the NIHR. J.E.H was funded by the NIH (1RF1MH123195, 1R01AG070988, 1R01EB031114, 1UM1MH130981, 1RF1AG080371). K.W was funded by the Wellcome Trust.

References

1. Blumcke I, Spreafico R, Haaker G, et al. Histopathological Findings in Brain Tissue Obtained during Epilepsy Surgery. *N Engl J Med*. 2017;377(17):1648-1656.

2. Lamberink HJ, Otte WM, Blümcke I, et al. Seizure outcome and use of antiepileptic drugs after epilepsy surgery according to histopathological diagnosis: a retrospective multicentre cohort study. *Lancet Neurol*. 2020;19(9):748-757.

3. Téllez-Zenteno JF, Ronquillo LH, Moien-Afshari F, Wiebe S. Surgical outcomes in lesional and non-lesional epilepsy: A systematic review and meta-analysis. *Epilepsy Res*. 2010;89(2):310-318.

4. Wagstyl K, Whitaker K, Raznahan A, et al. Atlas of lesion locations and postsurgical seizure freedom in focal cortical dysplasia: A MELD study. *Epilepsia*. Published online November 29, 2021. doi:10.1111/epi.17130

5. Spitzer H, Ripart M, Whitaker K, et al. Interpretable surface-based detection of focal cortical dysplasias: a Multi-centre Epilepsy Lesion Detection study. *Brain*. 2022;145(11):3859-3871.

6. David B, Kröll-Seger J, Schuch F, et al. External validation of automated focal cortical dysplasia detection using morphometric analysis. *Epilepsia*. Published online February 27, 2021. doi:10.1111/epi.16853

7. Gill RS, Lee HM, Caldairou B, et al. Multicenter Validation of a Deep Learning Detection Algorithm for Focal Cortical Dysplasia. *Neurology*. Published online September 14, 2021. doi:10.1212/WNL.0000000000012698

8. Walger L, Adler S, Wagstyl K, et al. Artificial intelligence for the detection of focal cortical dysplasia: Challenges in translating algorithms into clinical practice. *Epilepsia*. Published online January 31, 2023. doi:10.1111/epi.17522

9. Cucurull G, Wagstyl K, Casanova A, et al. Convolutional neural networks for mesh-based parcellation of the cerebral cortex. Published online April 11, 2018. Accessed June 14, 2018. https://openreview.net/pdf?id=rkKvBAiiz

10. Gong S, Chen L, Bronstein M, Zafeiriou S. SpiralNet++: A Fast and Highly Efficient Mesh Convolution Operator. *arXiv [csCV]*. Published online November 13, 2019. http://arxiv.org/abs/1911.05856

11. Spitzer H, Ripart M, Fawaz A, et al. Robust and Generalisable Segmentation of Subtle Epilepsy-Causing Lesions: A Graph Convolutional Approach. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer Nature Switzerland; 2023:420-428.

12. Fawaz A, Williams LZJ, Alansary A, et al. Benchmarking Geometric Deep Learning for Cortical Segmentation and Neurodevelopmental Phenotype Prediction. *bioRxiv*. Published online December 2, 2021:2021.12.01.470730. doi:10.1101/2021.12.01.470730

13. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*.

2021;18(2):203-211.

14. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Process Mag*. 2017;34(4):18-42.

15. Fischl B. FreeSurfer. *Neuroimage*. 2012;62(2):774-781.

16. Adler S, Whitaker K, Semmelroch M, Wagstyl K. MELD protocol 4 - lesion masking v2. *protocols.io*. Published online April 5, 2018. doi:10.17504/protocols.io.n9udh6w

17. Greve DN, Van der Haegen L, Cai Q, et al. A surface-based analysis of language lateralization and cortical asymmetry. *J Cogn Neurosci*. 2013;25(9):1477-1492.

18. Spitzer H, Ripart M, Whitaker K, et al. Interpretable surface-based detection of focal cortical dysplasias: a MELD study. *bioRxiv*. Published online December 14, 2021. doi:10.1101/2021.12.13.21267721

19. Sigalovsky IS, Fischl B, Melcher JR. Mapping an intrinsic MR property of gray matter in auditory cortex of living humans: a possible marker for primary cortex and hemispheric differences. *Neuroimage*. 2006;32(4):1524-1537.

20. Chen AA, Luo C, Chen Y, Shinohara RT, Shou H, Alzheimer's Disease Neuroimaging Initiative. Privacy-preserving harmonization via distributed ComBat. *Neuroimage*. 2022;248:118822.

21. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing; 2015:234-241.

22. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *arXiv [csLG]*. Published online March 3, 2017. Accessed August 19, 2021. http://proceedings.mlr.press/v70/sundararajan17a/sundararajan17a.pdf

23. Walger, Lennart, Tobias Bauer, David Kügler, Matthias H. Schmitz, Fabiane Schuch, Christophe Arendt, Tobias Baumgartner, et al. 2024. "A Quantitative Comparison between Human and Artificial Intelligence in the Detection of Focal Cortical Dysplasia." *Investigative Radiology*, November. https://doi.org/10.1097/RLI.0000000000001125.

24. Wang ZI, Alexopoulos AV, Jones SE, Jaisani Z, Najm IM, Prayson RA. The pathology of magnetic-resonance-imaging-negative epilepsy. *Mod Pathol*. 2013;26(8):1051-1058.

25. Eriksson MH, Whitaker KJ, Booth J, et al. Pediatric epilepsy surgery from 2000 to 2018: Changes in referral and surgical volumes, patient characteristics, genetic testing, and post-surgical outcomes. *Epilepsia*. Published online June 2, 2023. doi:10.1111/epi.17670

26. Czarnetzki C, Spinelli L, Huppertz HJ, et al. Yield of non-invasive imaging in MRI-negative focal epilepsy. *J Neurol*. 2024;271(2):995-1003.

27. McGonigal A, Bartolomei F, Régis J, et al. Stereoelectroencephalography in presurgical assessment of MRI-negative epilepsy. *Brain*. 2007;130(Pt 12):3169-3183.

28. Eriksson MH, Prentice F, Piper RJ, et al. Long-term neuropsychological trajectories in children with epilepsy: does surgery halt decline? *Brain*. Published online April 20, 2024.

doi:10.1093/brain/awae121

29. Hale AT, Chari A, Scott RC, et al. Expedited epilepsy surgery prior to drug resistance in children: a frontier worth crossing? *Brain*. 2022;145(11):3755-3762.

30. Adler S, Wagstyl K, Gunny R, et al. Novel surface features for automated detection of focal cortical dysplasias in paediatric epilepsy. *Neuroimage Clin*. 2017;14:18-27.

31. Jin B, Krishnan B, Adler S, et al. Automated detection of focal cortical dysplasia type II with surface-based magnetic resonance imaging postprocessing and machine learning. *Epilepsia*. 2018;59(5):982-992.

32. Wagstyl K, Adler S, Pimpel B, et al. Planning stereoelectroencephalography using automated lesion detection: Retrospective feasibility study. *Epilepsia*. Published online June 13, 2020. doi:10.1111/epi.16574

33. Chari A, Adler S, Wagstyl K, et al. Lesion detection in epilepsy surgery: Lessons from a prospective evaluation of a machine learning algorithm. *Dev Med Child Neurol*. Published online August 9, 2023. doi:10.1111/dmcn.15727

34. Naeini MP, Cooper G, Hauskrecht M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *AAAI*. 2015;29(1). doi:10.1609/aaai.v29i1.9602

35. Urbach H, Kellner E, Kremers N, Blümcke I, Demerath T. MRI of focal cortical dysplasia. *Neuroradiology*. 2022;64(3):443-452.

36. Duncan JS. Multidisciplinary team meetings: the epilepsy experience. *Pract Neurol*. Published online May 9, 2022. doi:10.1136/practneurol-2022-003350

37. Baumgartner C, Koren JP, Britto-Arias M, Zoche L, Pirker S. Presurgical epilepsy evaluation and epilepsy surgery. *F1000Res*. 2019;8. doi:10.12688/f1000research.17714.1

38. Cohen NT, You X, Krishnamurthy M, et al. Networks Underlie Temporal Onset of Dysplasia-Related Epilepsy: A MELD Study. *Ann Neurol*. 2022;92(3):503-511.

39. Mellor S, Timms RC, O'Neill GC, et al. Combining OPM and lesion mapping data for epilepsy surgery planning: a simulation study. *Sci Rep*. 2024;14(1):2882.

Figures

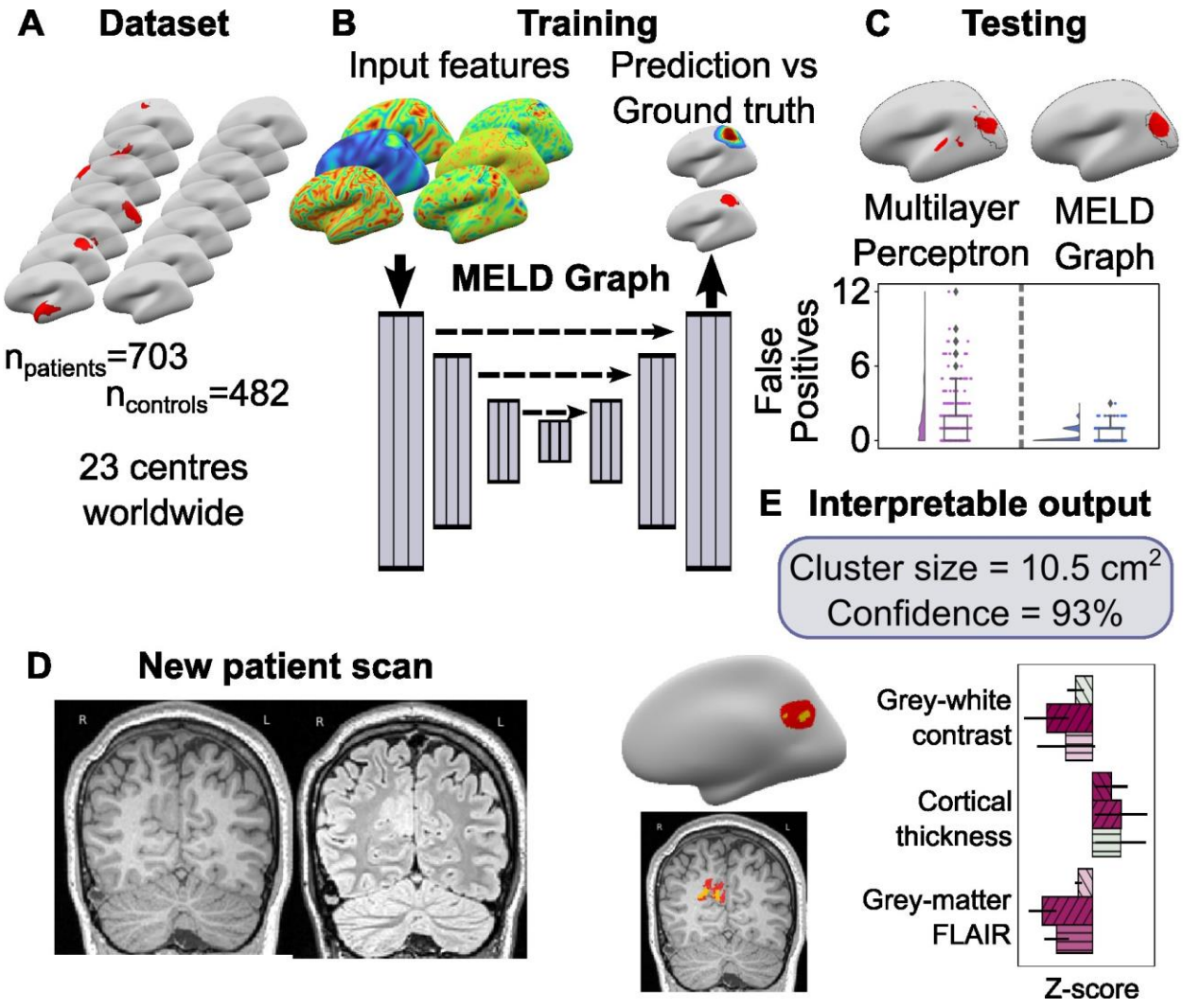


Figure 1. Overview of MELD Graph. (A) *The MELD FCD dataset contains surface-based MRI data and ground truth lesion masks for 703 patients and 482 controls from 23 centers worldwide.* ***(B)*** *The MELD Graph U-Net model was trained to recognise patterns of abnormal features and their broader hemispheric context, to segment FCD lesions.* ***(C)*** *The model was compared against a previously published, and widely reused baseline algorithm - a multilayer perceptron[5]. MELD Graph had a significantly increased positive predictive value driven by a decrease in the number of false positive clusters.* ***(D)*** *Images from a new patient with a suspected FCD can be analyzed*

*with MELD Graph to generate an interpretable report **(E)** that displays predicted lesion location, size and features, alongside confidence and relative feature importance (salience).*

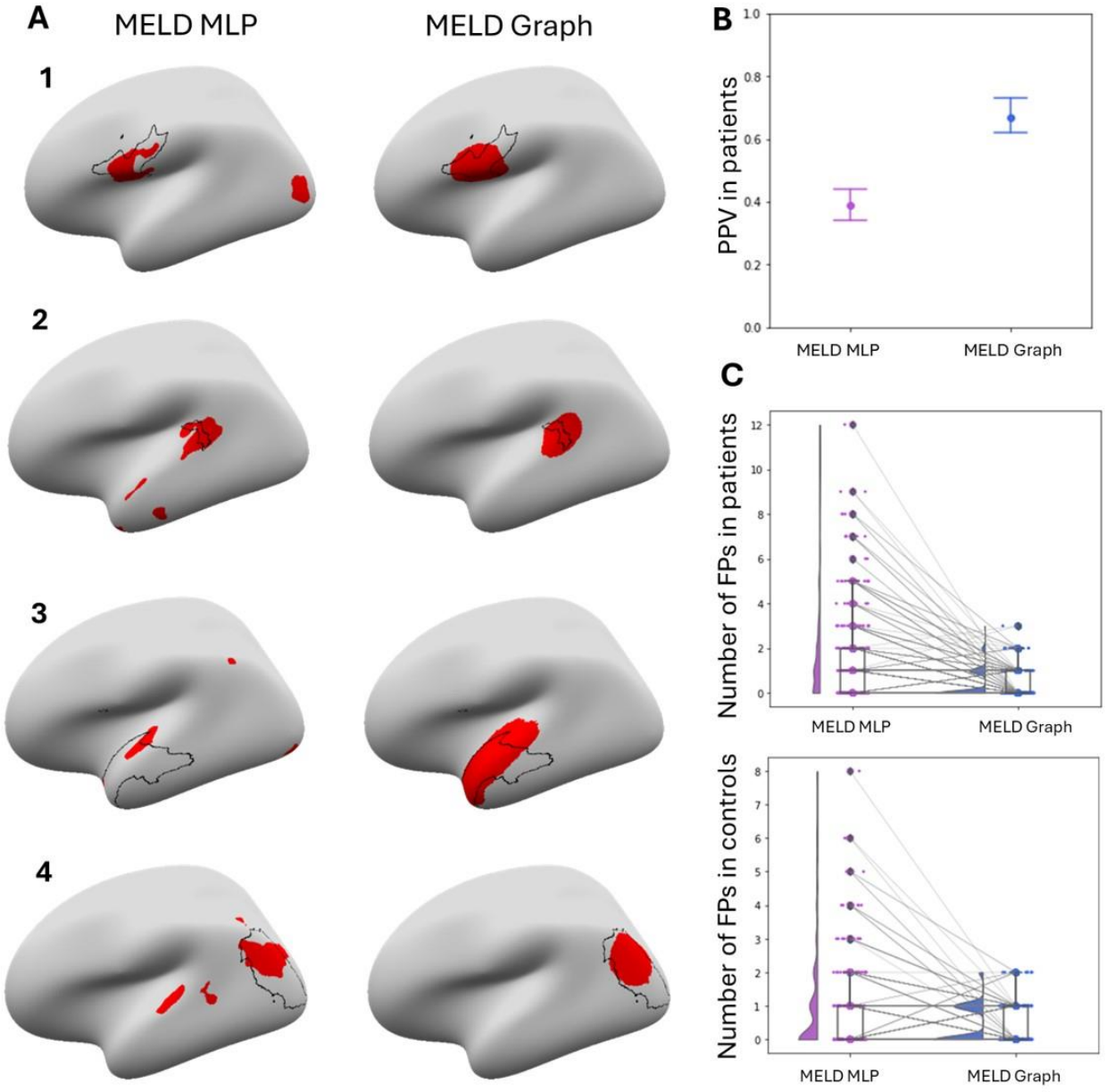


***Figure 2. Reduction in false positive clusters with MELD Graph. (A)** Example classifier predictions for four patients using MELD Graph and the baseline multilayer perceptron (MELD MLP). Black line = manual lesion mask. Red = classifier predictions. **(B)** Box-and-whisker plot showing the mean positive predictive value (PPV) and the confidence interval (CI) in the test patients detected by both MELD MLP and MELD Graph models (N=161)**. (C)** Box-and-whisker*

plot showing median, interquartile range and outlier numbers of false positive (FPs) clusters predicted on patients and controls in the test dataset using MELD Graph compared to MELD MLP. Gray lines connect identical subjects between the models.

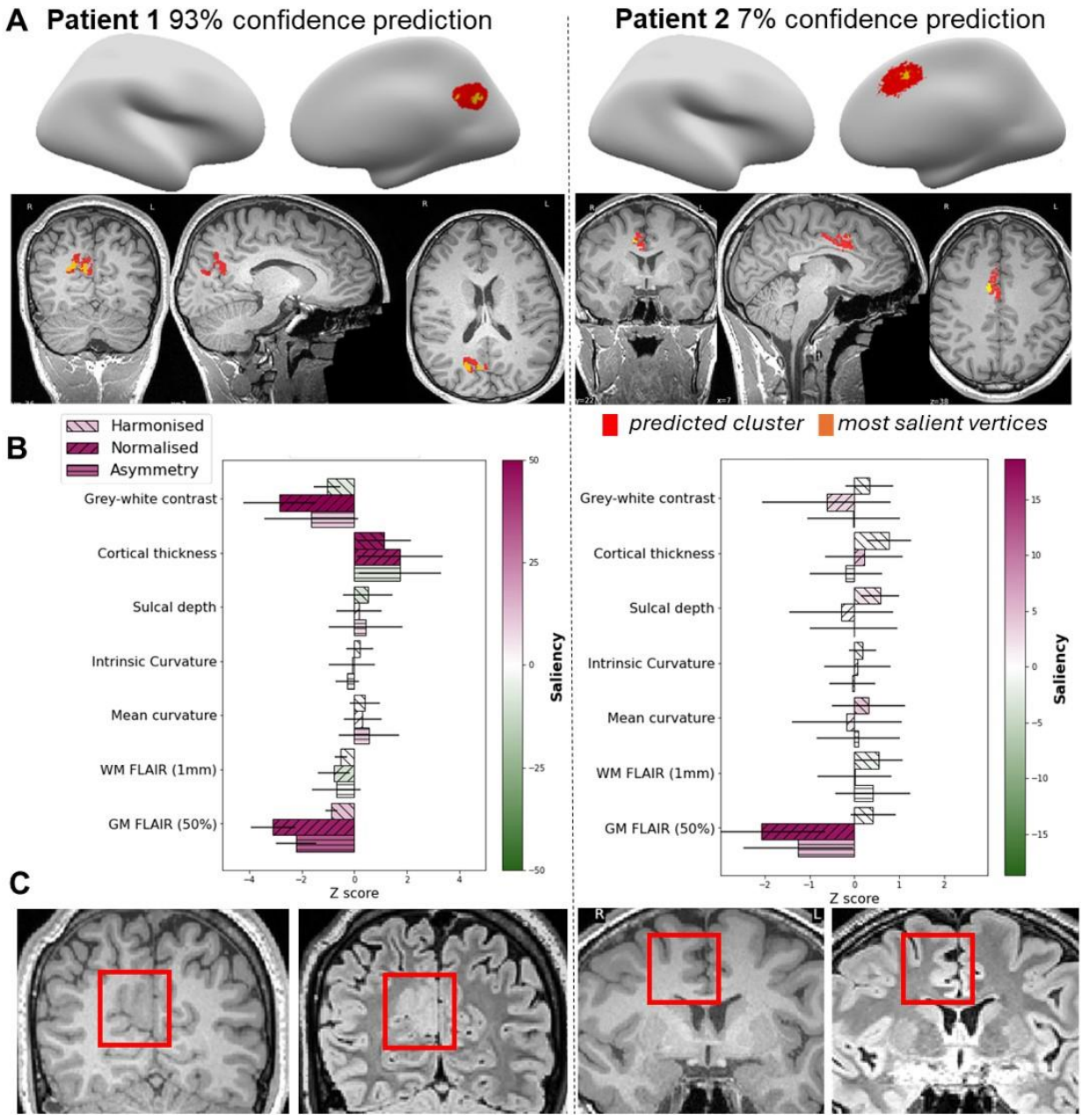


Figure 3. *Examples of interpretable patient reports. MELD Graph outputs for two patients from the independent test cohort. Patient 1 is an example of an MRI-positive FCD detected by MELD Graph with a high confidence prediction (93%). Patient 2 has a FCD that was not identified by 5*

*expert radiologists but detected by MELD Graph with low confidence (7%). (**A**) Classifier predictions (red) and 20% most salient vertices (orange) visualized on brain surfaces of the lesional hemisphere and on the T1 volume. (**B**) Z-scored mean feature values within 20% most salient vertices of predicted lesions. Color represents saliency scores. Features driving the classifier's prediction are positive (pink). Features inconsistent with MELD Graph prediction are negative (green). (**C**) T1 and FLAIR coronal sections with a red box indicating the lesional cortex.*

Tables

		MELD MLP	**MELD Graph**
Test dataset	Sensitivity (n=260 patients)	67% (61-73%)	70% (64-75%)
	Specificity (n=193 controls)	54% (47-61%)	60% (53-67%)
	PPV	39% (35-44%)	**67% (62-73%)**
	IoU (n=160)	0.23 (0.22-0.26)	**0.3 (0.27-0.33)**
Independent test cohort	Sensitivity (n=116 patients)	77% (69-84%)	72% (61-78%)
	Specificity (n=101 controls)	47% (37-56%)	**56% (47-66%)**
	PPV	46% (40-53%)	**76% (61-79%)**
	IoU (n=78)	0.29 (0.25-0.33)	**0.36 (0.31-0.41)**

Table 1. *Comparison of the performance of the MELD Graph model to the original Multilayer Perceptron (MELD MLP) on the same withheld and independent test cohort derived from three centers. Reported are the test results and the bootstrapped 95% confidence intervals. In bold are metrics with significantly changed performance ($p<0.05$) in comparison to MELD MLP.*

		% Detected	Patients (n)
Age group			
	Adult	67.9	131
	Pediatric	71.3	129
Sex			
	Female	62.4	125
	Male	**76.3**	135
Ever-reported MRI-negative			
	Visible	72.2	180
	MRI-negative	63.7	80
Postoperative seizure freedom			
	Seizure-free	79.2	106
	Not seizure-free	**62.7**	51
Histology			
	FCD I	84.6	13

	FCD IIA	75.4	57
	FCD IIB	76.3	93
	FCD III	75.0	8
	Not available	**56.2**	89
Histology confirmed + Seizure-free			
	True	**81.6**	98
	False	62.3	162
Modality			
	T1w only	67.3	150
	T1w + FLAIR	72.7	110

Table 2. *MELD Graph performance on the test dataset grouped according to demographic factors. FCD detection rate is broken down by age group, sex, MRI status, seizure freedom, histology and available MRI modalities. Lesions termed "MRI-negative" were from MRI scans reported as non-lesional at some point during the patient's clinical evaluation.*

Supplemental Online Content

Table of contents:

eMethods

Participants

23 international epilepsy centers participated in this study. The inclusion criteria for patients was as follows: age over 3 years, 3D preoperative T1-weighted MRI scan (1.5T or 3T) available, radiological diagnosis of a single focal cortical dysplasia (FCD) or MRI-negative with histopathological confirmation of a single FCD lesion. Exclusion criteria were as follows: previous neurosurgeries, large structural abnormalities in addition to the FCD and T1 scans with gadolinium enhancement. The control participants inclusion criteria were: age over 3 years, no neurological conditions and availability of a T1-weighted MRI brain scan (1.5 or 3 T). Patients scanned for headache could be included as controls if they had no other neurological conditions and the MRI was normal.

In patients, the following demographic / clinical data were retrieved: age at preoperative scan, sex, age of epilepsy onset, duration of epilepsy (time from age of epilepsy onset to age at preoperative scan), ever reported MRI-negative and histopathological diagnosis (ILAE three-tiered classification system),[1] seizure-freedom (Engel class I or other) and follow-up time in operated patients. In controls, the following demographic variables were collected: age at scan and sex.

The main cohort included data from 20 centers. Using the same randomisation as in a previous study, participants in the main cohort were randomly split 50:50 into training and testing cohorts (Table 1). The independent test cohort included participants from a further three additional centers. Two of these centers were used in our previous work as independent test sites, with the addition of a recent openly available FCD dataset[2]. No participants from these three centers were included in the training cohort.

MRI features

For each participant, 3D T1-weighted (all participants) and FLAIR (where available) MR images were processed using FreeSurfer[3] and the following 11 surface-based features (cortical thickness, gray-white matter intensity contrast, intrinsic curvature, sulcal depth, curvature and FLAIR intensity sampled as 6 intra- and sub-cortical depths) were extracted (Fig. 1) as per previous work[4]. Surface-based features were registered to a bilaterally symmetrical template, fsaverage_sym, using folding-based registration[5].

The following pre-processing steps were performed on the surface-based MRI features:

1. **Smoothing.** Surface-based smoothing was performed using a Gaussian kernel of 3mm for all features, except for intrinsic curvature where a kernel of 20mm was used.
2. **Intra- and inter-subject normalization.** Intra-subject z-scoring of features was performed to account for age and sex-related changes. Inter-subject z-scoring by the mean and standard deviation of features at each vertex from healthy controls was used to account for inter-regional differences.
3. **Asymmetry calculations.** To enhance inter-hemispheric asymmetries, right hemisphere per-vertex z-scored feature values were subtracted from left hemisphere values and vice versa.

Additionally, cortical thickness with curvature regressed out was computed[6].

The final surface-based feature set consisted of the 11 "smoothed" (Steps 1+2), 11 "normalized" (Steps 1+2) and 11 "asymmetry" features (Steps 1-3), as well as curvature-regressed cortical thickness - resulting in 34 input features.

For data from the independent test sites, an additional preprocessing step, termed "harmonization" was included after smoothing (step 1). Data from the independent test sites was harmonized to the cohort used to train the model using distributed-ComBat[7].

Lesion masking

At each center, for each patient, FCDs were manually drawn on T1w or fluid-attenuated inversion recovery (FLAIR) images following a lesion masking protocol[8]. If there was no visible lesion on MRI ("MRI negative"), resection cavities on post-surgical scans were used to assist in the creation of a lesion mask. A non-parametric Mann-Whitney U test was used to compare the size of lesion masks (which do not follow Gaussian distributions) between patients with visible lesions on MRI and patients who had at some point in their preoperative evaluation been considered "MRI-negative". The lesion masks were projected onto individual FreeSurfer surfaces and then registered to fsaverage_sym.

Given the subtlety of FCDs on MRI, they are challenging to mask and the borders of the manually delineated lesion masks are imprecise. We have previously shown that in this dataset, feature abnormalities extend approximately 40 mm beyond the lesion mask and mean fraction mask overlap between expert neuroradiologists was 42%[9]. As a result, regions of uncertainty, termed "border-zones", were created around each lesion mask extending 20 mm across the cortical surface. A 20mm ring of uncertainty has been previously shown to increase the mean fraction mask overlap between rater–rater pairs to 82%[9]. Predicted lesion clusters within 20 mm of the lesion masks are considered detected (see network evaluation section).

MELD Graph - a graph convolutional network for surface-based lesion segmentation

For surface-based lesion segmentation, we created a graph-based implementation of nnU-Net[10,11] (eFigure 1). U-Nets are convolutional neural networks that were designed for biomedical image segmentation[10]. nnU-Net is a deep learning-based segmentation method that automatically configures a U-Net according to the input data, minimizing manual design and optimization of the network. However, unlike typical imaging data represented on rectangular grids, our dataset is surface-based - represented on vertices with connected triangular faces. As such it required customized convolutions, downsampling and upsampling steps.

Convolutions, downsampling and upsampling

We used a spiral convolution[12], which translates standard 2D convolutions to irregular meshes by defining the filter by an outward spiral (eFigure 1) . Similar to how a 2d filter captures a ring of information around the input pixel, this spiral convolution captures a ring of information around the input node. We use a spiral length of 7, representing the central node/vertex and 6 adjacent neighbors on a hexagonal mesh. This is roughly equivalent to a 3x3 2D kernel. The spiral convolution enables the network to learn information from neighbors and across the entire hemisphere in contrast to previous approaches which see the features from each vertex in isolation[9].

For upsampling and downsampling through the U-Net, we created a series of seven successive icospheres, triangulated meshes. The icosphere templates were generated by successively upsampling an icosahedral icosphere, S^1, with 42 vertices and 80 triangular faces. Icosphere S^{i+1}, where i is the resolution of the icosphere, is generated from S^i by adding vertices at every edge. As input to our model we use icosphere S^7 (163842 vertices).

In the decoder, to upsample from S^i to S^{i+1}, the mean of each vertex in S^i is assigned to all neighbors at level S^{i+1}. In the encoder, to downsample from S^{i+1} to S^i, all the neighbors of a vertex at S^{i+1} are aggregated - a similar translation to 2D max pooling. The U-Net has seven levels (mirroring the seven icospheres S7-S1). Each level consists of three convolutional layers using spiral convolutions and leaky Relu as the activation function (eFigure 1).

Loss Functions

Loss functions quantify the difference between predicted and labeled values, and enable training and learning. The MELD graph model uses the following loss functions:

Segmentation loss. Cross-entropy and dice loss functions were used for the segmentation, following best practices for U-Net segmentation models[11].

Distance loss. A distance regression task was added to the U-Net to encourage the network to learn whole-brain context and reduce the number of false positive clusters. The model is trained to predict the geodesic distance to the lesion boundary at every vertex. The distance loss is a mean absolute error loss, weighted by the distance. It is weighted by the distance to avoid overly penalizing small errors in predicting large distances from the lesion.

Classification loss. A weakly-supervised classification loss was added to mitigate uncertainty in the correspondence between lesion masks and lesions. A hemisphere was labeled as lesional (positive) if it contained a lesion mask. To predict whether a hemisphere was lesional, we added a classification head to the deepest level (level 1) of the U-Net (eFigure 1). The classification head contained a fully connected layer aggregating over all filters, followed by a fully-connected layer aggregating over all vertices and was trained using cross-entropy.

Deep supervision. We included deep supervision at levels one to six of the U-Net to encourage the flow of gradients throughout the entire network. The model is trained on the weighted sum of the dice, cross entropy and distance losses at each level. The network must create segmentations at deeper levels of the network (deep supervision). This ensures learning at all levels of the network.

Data augmentation

Data augmentations involve modifying the existing training dataset to increase the size and heterogeneity of the training dataset to prevent models overfitting and improve model performance. Following the recommendations outlined in nnU-NET, we use spatial augmentations and intensity augmentations (eFigure

1). Spatial augmentation techniques, included rotations, inversions, and non-linear deformations of the surface-based data[13]; while intensity-based augmentations included the addition of Gaussian noise to alter feature intensity, contrast adjustments, uniform scaling of brightness, and the application of a gamma intensity transform.

Network implementation details

The graph-based convolutional implementation of nnU-Net (MELD Graph) was trained with a maximum of 1000 epochs, each of them passing through the whole training data and a maximum patience of 1000 epochs. The initial learning rate was set to 10^{-4} with a learning rate decay of 0.9 and a momentum at 0.99. The training batch size was 8. The probabilities of augmentation were set as follows: 0.5 for inversion, 0.2 for rotation and deformation, 0.15 for Gaussian noise, contrast, brightness and gamma.
The weights for the deep supervision levels l_ds=[6,5,4,3,2,1] were w_ds=[0.5,0.25,0.125,0.0625,0.03125,0.0150765]. To address the imbalance of classes during training, the non-lesional hemispheres were undersampled to ensure that 33% of the training examples contained a lesion. The model from the epoch with the best validation loss was stored for evaluation.
Training was carried out through 5-fold cross-validation on the train cohort, with folds being determined by random partition of the subjects in the train cohort. For testing, the 5 train models were ensembled, averaging the predictions of each model to generate a final model output.

The training and evaluation were performed on a High Performance Cluster (HPC) with Single NVIDIA A100 GPU and 1000 GiB RAM. The code was written in Python 3.9.13 and the main python packages employed were PyTorch 1.10.0+cu11.1 and PyTorch Geometric 2.0.4.

Post-processing

For each vertex in an individual hemisphere, the model generates a prediction value indicating the likelihood of that vertex being lesional. These per-vertex predictions were first thresholded according to the following function:

$$threshold = \begin{cases} 0.5, & if\ mp \geq 0.5 \\ \max(mp * 0.2, 0.01) & if\ mp < 0.5 \end{cases}$$

Where mp is the maximum prediction value across all vertices. The minimum threshold below which no lesions are outputted is 0.01. This was identified as the prediction value where the number of false positives in the validation cohort first falls.
The thresholded predictions were then organized into spatially connected clusters of vertices on the surface mesh, to regroup any fragmented clusters. Vertex clusters containing fewer than 100 vertices (approximately 0.5 cm^2) were discarded, as they disproportionately correspond to false positives.

Evaluation of MELD Graph model
To evaluate the MELD Graph model, we compared its performance on the test dataset and independent test sites, to a previously published algorithm for FCD detection, which uses a multi-layer perception (MLP). The following metrics were calculated for the MELD graph and MLP models:

Sensitivity - defined as the proportion of patients where a predicted lesion cluster overlapped with (or was within 20mm of) the manual lesion mask. The 20mm expansion of the lesion masks was chosen for evaluation of the algorithm as previous work has demonstrated the difficulty of manually masking these lesions and that adding a 20mm border zone increases mean fraction mask overlap between radiologists from 42% to 82%[4].

Specificity - defined as the proportion of controls with no predicted clusters.

Positive predictive value - defined as the number of detected lesions divided by the total number of predicted clusters across the whole patient cohort.

$$PPV = TP / (TP + FP)$$

Number of false positive clusters - defined as the total number of predicted clusters per participant minus any predicted clusters that overlap with the manual lesion mask.

Intersection Over Union (IoU) - defined as the number of vertices overlapping between the predicted lesion segmentation and the manual lesion mask divided by the number of vertices in the union of both masks. Calculated on patients in the test cohort and independent test sites detected by both MELD Graph and MLP models.

Bootstrap resampling, sampling the cohort randomly with replacement 10,000 times, was used to calculate confidence intervals on the sensitivity, specificity, positive predictive value and IoU estimates between the MELD Graph and MLP models. Permutation-based null-models were used to test for statistically significant ($p<0.05$) differences between model performance estimates. Per-patient metrics were randomly permuted between models 1000 times and test statistics recalculated to generate a distribution against which the actual metric difference was compared.

Breakdown lesion detection rates in the test dataset were calculated according to demographic factors. Significant differences ($p<0.05$) between each group within a demographic factor (e.g. pediatric vs adults) were assessed by permuting the group's assignments 1000 times and recomputing sensitivity estimates to generate a null distribution.

To understand whether using harmonization of new site data to the cohort used to train the model was necessary, MELD Graph was evaluated on the independent test sites with and without distributed-ComBat harmonization. In this experiment, the MRI features from the independent test sites underwent the same smoothing, and optionally distributed-Combat[7], before intra- and inter-subject normalization and asymmetry calculations.

To estimate the number of subjects needed to generate consistent harmonization parameters, we carried out a number of subsampling experiments, comparing the harmonized cortical thickness maps computed with estimates derived from 3 - 70 subjects from a single site and 100 independent subjects from the same site (eFigure 3).

We sought to assess whether a model trained specifically to detect the most subtle FCDs would outperform the model trained on the full heterogeneous dataset. We trained the same model architecture on a subset of patients from the training dataset who were MRI-negative and histopathologically confirmed. This new model was tested on patients in the test cohort who were also MRI-negative and histopathologically confirmed and compared with MELD Graph evaluated on the same patients. We calculated each model's sensitivity, specificity, PPV, IoU and number of false positive clusters (eTable 4). Bootstrap resampling and permutation-based null-models (see above) were applied to calculate confidence intervals and to test for significant differences ($p<0.05$) in these metrics.

Interpretability of MELD Graph outputs

To understand which specific features and which vertices drove network predictions, integrated gradients saliency was computed[14]. This method computes which features are important to the network by looking at the integral (Riemann approximation) of the gradients computed from a baseline input (0 for each feature) to the actual feature values for each vertex. Within each predicted lesion, for each vertex, the integrated saliencies were averaged across the 34 features and the vertices with the 20% most salient vertices were identified. For small predicted lesions, where the 20% most salient vertices correspond to less than 125 vertices, the 125 most salient vertices were used instead. For these most salient vertices, the following metrics were calculated:

1. Mean Z-score of each feature
2. Mean integrated gradients saliency of each feature
3. Classifier confidence. To determine classifier confidence in its predictions, averaged across the 20% most salient vertices for each predicted cluster, the maximum prediction score for the ensembled MELD graph models was computed.

These data were used to characterize the detected FCD lesions (the true-positive clusters), the missed FCD lesions (the false negative clusters) and the false positive clusters in the test cohort (eFigure 4A). For each feature, the distribution of the mean asymmetries was plotted for each group. A one-way ANOVA was applied to test whether the feature means varied according to the group, followed by a Tukey Honest Significant Difference test to characterize the direction of any significant differences. Additionally, this characterization was broken down by known histology subtypes: FCD 1 and FCD 3 combined together, FCD 2A and FCD 2B (eFigure 4B).

Additionally, we characterized model performance (sensitivity and PPV) according to lesion size and lesion location. To assess the impact of lesion size on model performance, we fit a logistic regression model to predict whether a lesion mask was detected based on its size (sensitivity), or whether the MELD Graph cluster size was predictive of it being a true or false positive (PPV). For visualization, lesions/clusters were grouped into 5 quintiles according to size and sensitivity and PPV was calculated for each group (eFigure 4C). To assess the impact of lesion location on model performance, we fit similar logistic regression models at every vertex, predicting whether lesion masks overlapping this particular vertex were more likely to be detected (sensitivity map) or predicted lesions were more likely to be TP/FP clusters (PPV map). For vertices where the logistic regression failed because the matrix was singular (e.g. vertex with no lesions), p-values were disregarded. P-values were corrected for multiple comparisons using the Holm method at 5% significance. For visualization of any trends, sensitivity at each vertex and PPV for lesions at each vertex were mapped to the cortical surface (eFigure 4D).

A Spearman’s rank correlation was used to assess the (typically nonlinear) relationship between cluster confidence and IoU in true positive clusters from the test cohort that were identified by the MELD Graph model.

eTable 1. **Demographics of train cohort, test cohort and independent test sites.**

	Train cohort Patients (n=278)	Train cohort Controls (n=180)	Test cohort Patients (n=260)	Test cohort Controls (n=193)	Independent test sites Patients (n=116)	Independent test sites Controls (n=101)
Age at preoperative scan (median, IQR)	20.0 (11.0 - 32.8)	29.0 (19.0 - 37.9)	18.0 (11.0 - 29.0)	29.0 (19.5 - 39.2)	22.5 (13.1 - 27.5)	27.5 (22.5 - 37.5)
Sex (f:m)	150 : 127	105 : 75	125 : 135	104 : 88	62 : 54	51: 50
Age of epilepsy onset (median, IQR)	6.0 (2.5 - 12.0)		6.0 (3.0 - 11.0)		2.8 (0.8 - 5.5)	
Duration of epilepsy (median, IQR)	10.0 (4.3 - 18.4)		10.2 (5.0 - 18.2)		2.65 (1.2 - 7.2)	
FLAIR available	132 / 278 (47%)	28 / 180 (16%)	110 / 260 (42%)	28 / 193 (15%)	33 / 116 (28%)	18 / 101 (18%)
Scanner (1.5T:3T)	41 : 237	18 : 162	56 : 204	15 : 178	0 : 116	0 : 101
Surgery	208/278 (75%)		190 / 260 (73.0%)		69 / 116 (59.0%)	
Histology available	193/208 (93%)		171 / 190 (90%)		68 / 69 (99%)	
Seizure free	123/183 (67%)		106/157 (68.0%)		52/64 (81%)	
Follow up time (median, IQR)	2.0 (1.0 - 3.0)		2.0 (1.0 - 3.4)		2.3 (1.5 - 3.3)	

eTable 2. False positive clusters in patients and controls in the test dataset and independent test cohort.

	Test cohort		**Independent test cohort**	
False positive clusters	Controls (median (IQR), max)	Patients (median (IQR), max)	Controls (median (IQR), max)	Patients (median (IQR), max)
MELD MLP	0 (0-1), 8	1 (0-2), 12	1 (0-1), 16	1 (0-2), 7
MELD Graph model	0 (0-1), 2	0 (0-1), 3	0 (0-1), 3	0 (0-0), 2

eTable 2*. Results table comparing the number of false positive clusters using the MELD Graph model to the MELD MLP model in patients and controls in the test dataset and independent test cohort. Reported are the median, interquartile range and maximum number of false positive clusters.*

eTable 3. Comparison of MELD Graph model evaluated on harmonized vs non-harmonized independent test cohort.

		MELD Graph with harmonization	MELD Graph without harmonization
Independent test cohort	Sensitivity (n=116 patients)	72% (63-79%)	70% (61-78%)
	Specificity (n=101 controls)	56% (47-66%)	**39% (29-48%)**
	PPV (per patient)	76% (68-85%)	**70% (61-79%)**

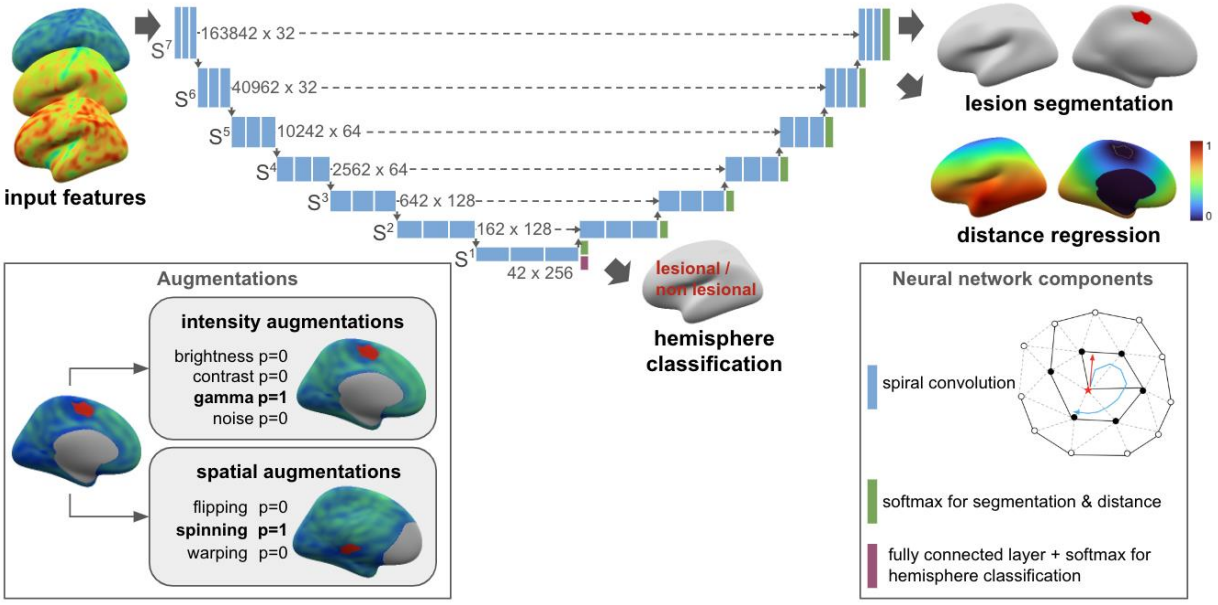
eTable 3*. Results table comparing the performance of the MELD Graph model on the independent test cohort with and without inter-site harmonization of the surface-based features. Reported are the test results and the bootstrapped 95% confidence intervals. In bold are metrics with significantly changed performance (p<0.05) in comparison to the MELD Graph model with harmonized surface-based features.*

eTable 4. Comparison of MELD Graph model vs model trained only with MRI-negative and FCD histology-confirmed patients.

	MELD Graph (all training cohort)	MELD Graph (MRI-negative + histology-confirmed FCD)
Sensitivity (n=58)	72% (60-84%)	67% (55-79%)
Specificity (n=193)	60% (53-67%)	54% (47-61%)
PPV (per patient)	72% (61-84%)	**58% (46-72%)**
IoU (n=58)	0.25 (0.19-0.31)	0.22 (0.17-0.28)
number FPs patients (median [IQR], max)	0 [0,0], 3	0 [0-1], 4

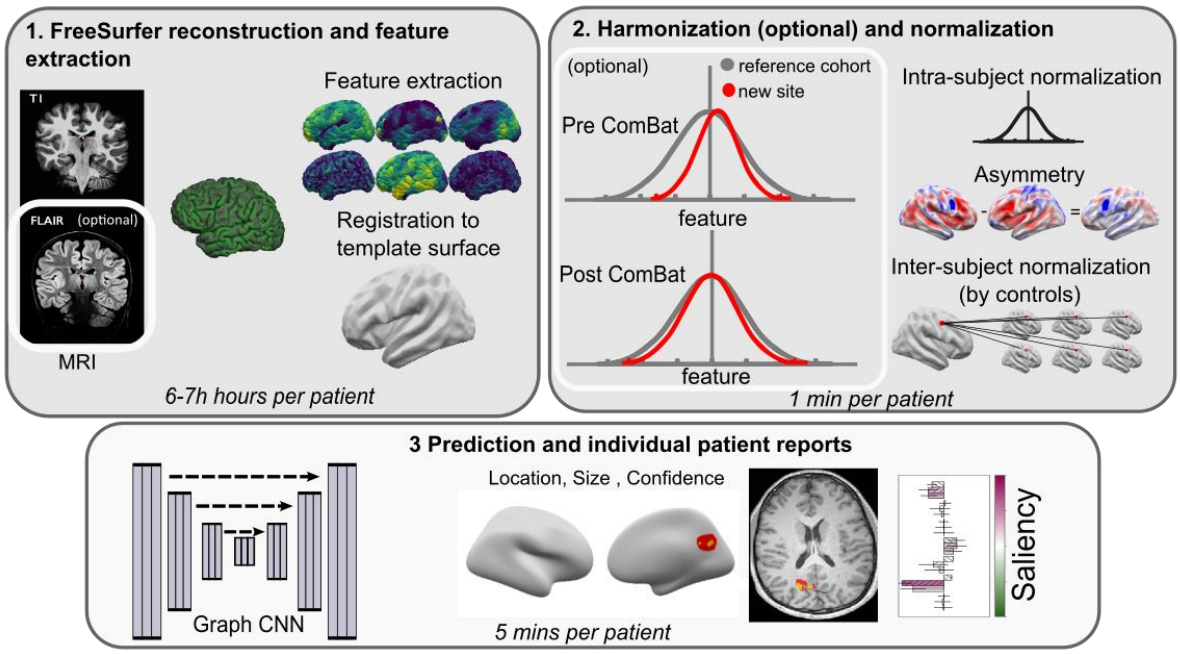
eTable 4. *Results table comparing the same model architecture (MELD Graph) trained on the whole training cohort vs a subset of only MRI-negative and histologically confirmed FCD cases and evaluated on the test cohort of MRI-negative & histologically confirmed cases. Reported are the test results and the bootstrapped 95% confidence intervals. In bold are metrics with significantly changed performance (p<0.05) in comparison to the MELD Graph model trained using the whole training cohort.*

eFigure 1. MELD Graph model architecture.



eFigure 1. *MELD Graph model architecture. Graph-based implementation of nnU-Net for surface-based lesion segmentation, with auxiliary distance regression, hemisphere classification and object detection tasks. Lower left box: Types of data augmentation employed. Examples show the result of gamma intensity augmentation (top) and spinning (bottom). Lower right box: Neural network components. Visualization of spiral convolution used on surface-based mesh.*

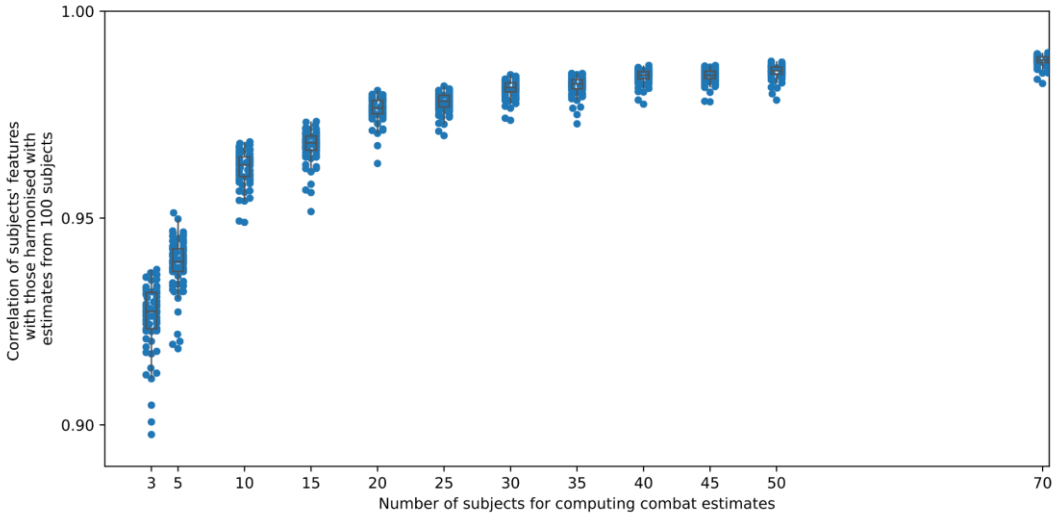
eFigure 2. Pipeline for running a new patient's MRI scan through MELD Graph.



eFigure 2. *Pipeline for running a new patient's MRI scan through MELD Graph. (1) T1w scan and, if provided, FLAIR scans, are processed through FreeSurfer to extract surface-based morphological features. Features are then coregistered to a symmetric template surface. (2) Feature harmonization and*

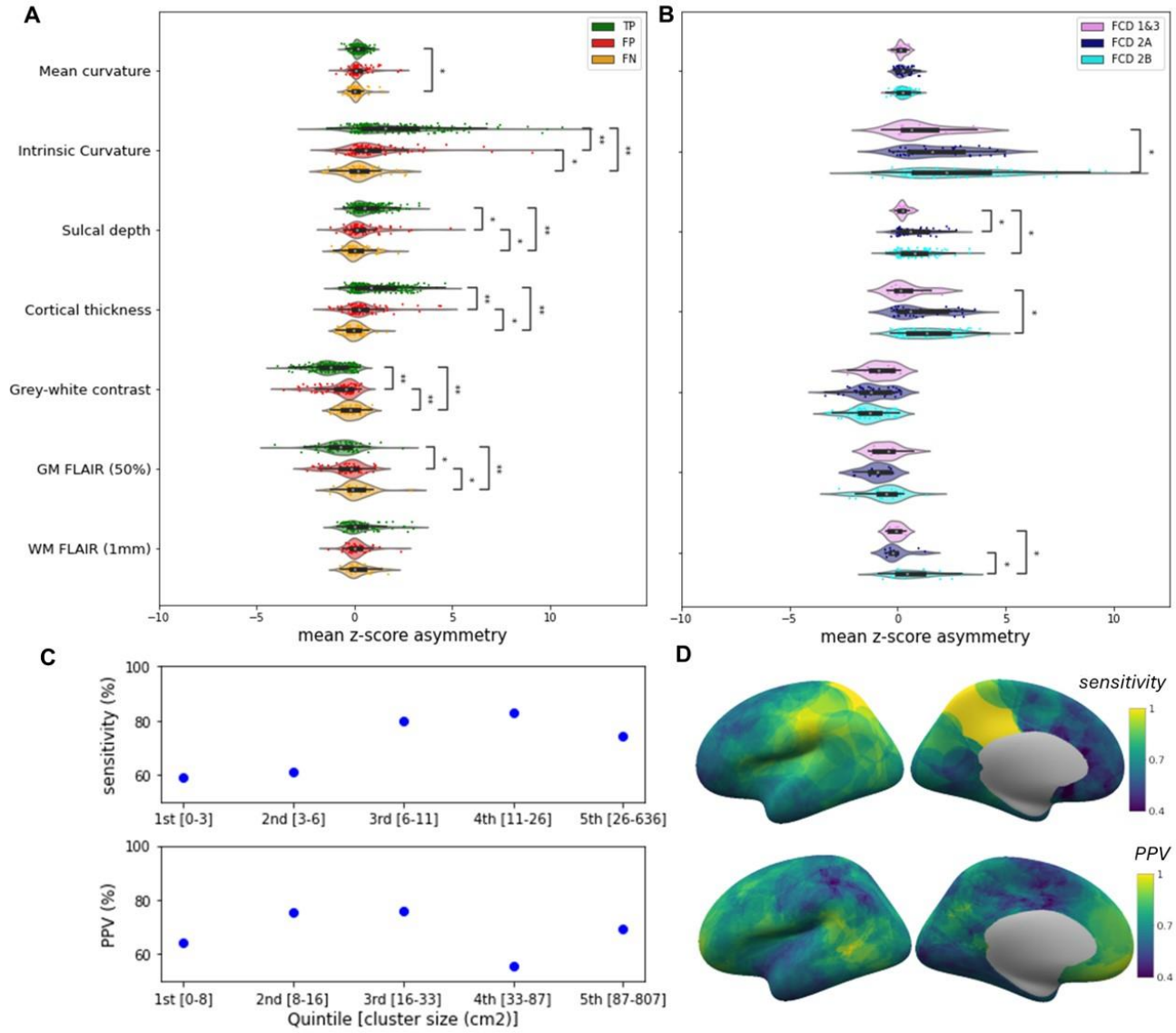
normalization. First, morphological features are smoothed. Then, an optional step of harmonization can be applied to minimize inter-scanner feature differences. Finally, features undergo intra-subject, inter-hemispheric and inter-subject normalisation. (3) The normalized features are used as input in the trained MELD Graph CNN model to predict lesional vertices. Results of the MELD Graph model are outputted on individual,interpretable reports. Each predicted cluster (red) and its most salient vertices (orange) can be visualized on the cortical surfaces and on the original T1w volume, alongside information about the cluster size, location, confidence, feature values and feature saliencies.

eFigure 3. Stability of NeuroCombat (harmonization to adjust for site and scanner differences) as a function of sample size.



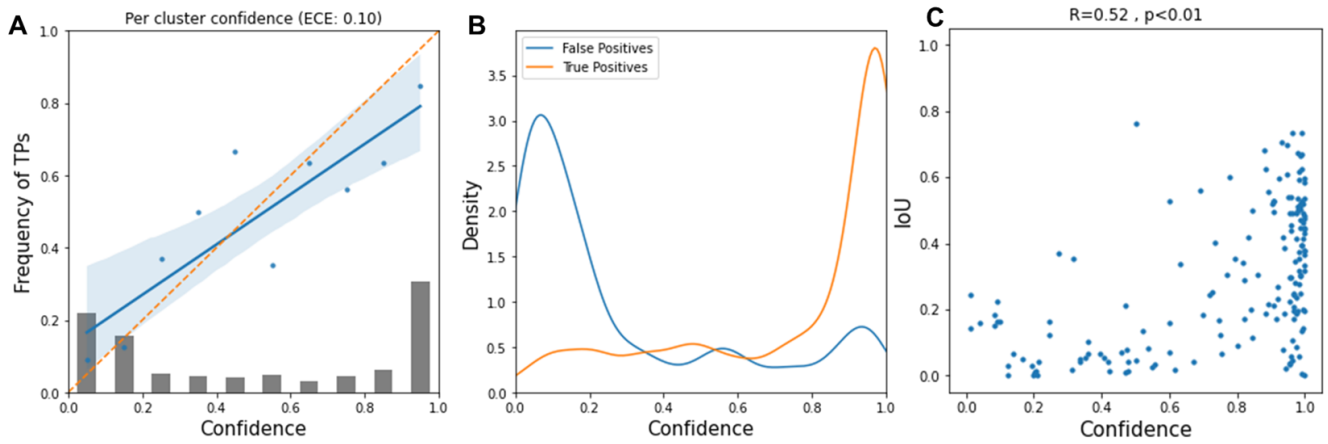
eFigure 3*. Stability of NeuroCombat (harmonization to adjust for site and scanner differences) as a function of sample size. Estimating scanner-specific harmonization parameters from increasing numbers of subjects (3-70) results in harmonized cortical thickness maps that are more highly correlated to maps from the same subjects that were harmonized using parameters calculated using an independent sample of 100 subjects. Based on these data we recommend a minimum of 20 subjects are used to calculate these parameters.*

eFigure 4. Characterization of detected FCD lesions.



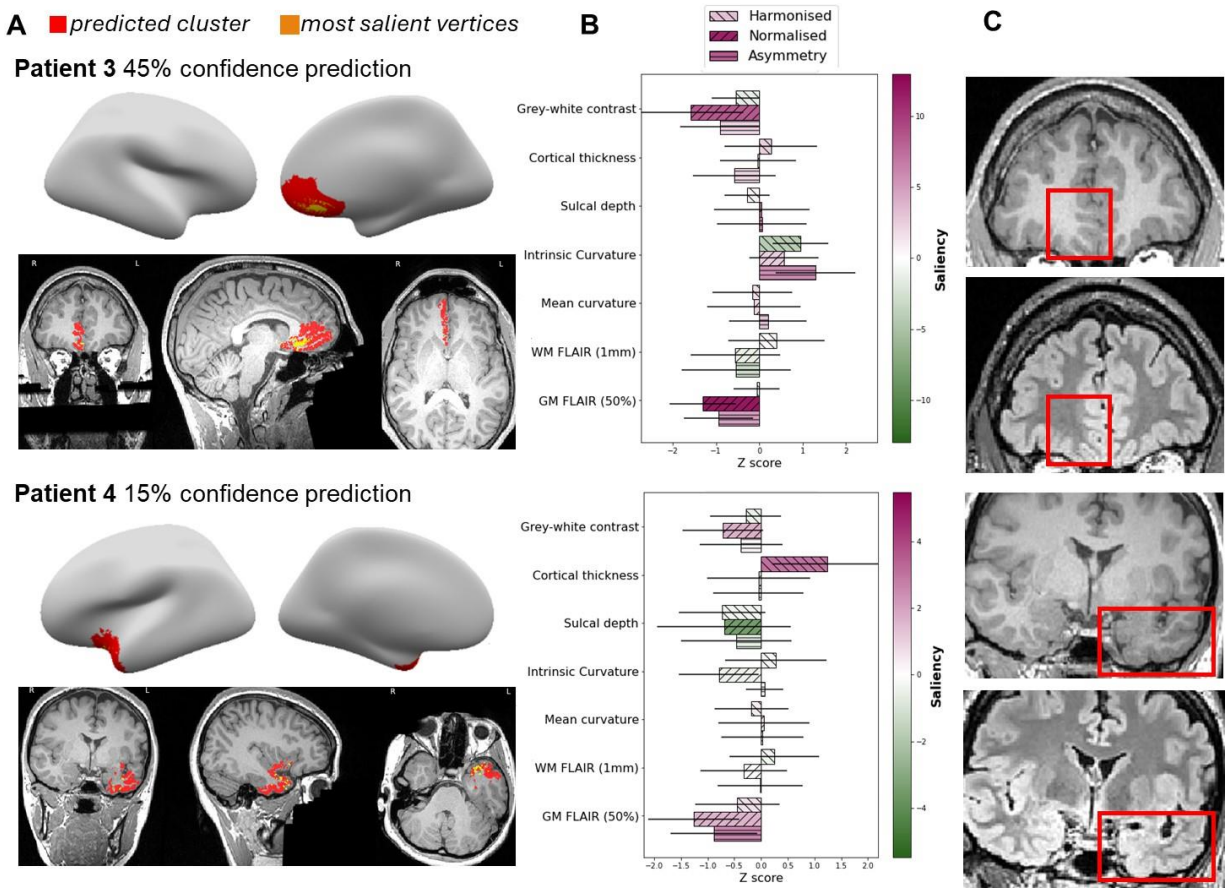
eFigure 4. *Characterization of detected FCD lesions. (**A**) Characterization of MRI features for correctly predicted lesions (True Positives, TP), incorrectly predicted lesions (False Positives, FP) and manual lesion masks not detected (False Negatives, FN). The distribution of mean feature asymmetry is presented for TP, FP and FN lesions in the test dataset. Tukey HSD test shows significant differences between each pair of distribution (* p<0.05, ** p<0.001). **(B)** Characterization of MRI feature differences for FCD histological subtypes. (**C**) Sensitivity and positive predictive value (PPV) grouped by cluster size. Average values are plotted in five different bins corresponding to the quintiles of the cluster size distribution. Cluster size was not predictive of model sensitivity (p=0.09) or PPV (p=0.76). (**D**) Sensitivity and PPV as spatial distributions. Vertex location was not significantly predictive of sensitivity or PPV (all corrected p-values > 0.05), likely due to small sample sizes at each location.*

eFigure 5. Calibration of confidence scores.



eFigure 5*. Calibration of confidence scores.* ***(A)*** *The relative frequency of true positives among all predicted clusters is plotted against MELD Graph model confidence scores grouped by decile. Per cluster confidence scores are well-calibrated, closely matching the relative frequency of true positives among predicted clusters in patients with an Expected Calibration Error of 0.10. Blue line = line of best fit, orange dashed line = perfect calibration. Gray bars represent the total number of clusters in each decile.* ***(B)*** *The distribution of confidence scores is plotted for true positive and false positive clusters in patients. High-confidence predictions are more likely to be true positives, while low-confidence clusters are more likely to be false positives. Providing per cluster confidence scores in the individualized patient reports improves model interpretability.* ***(C)*** *In detected lesions in the test cohort, intersection over union (IoU) scores are plotted against cluster confidence. IoU correlated significantly with confidence score (Spearman r=0.52, p<0.01), indicating that higher confidence predictions are associated with higher segmentation accuracies.*

eFigure 6. Examples of interpretable patient reports from two MRI-negative patients from the independent test cohort.



eFigure 6*. Examples of interpretable patient reports from two MRI-negative patients from the independent test cohort. Patient 3 and Patient 4 have FCDs that were not identified by 5 expert radiologists but detected by MELD Graph with low confidence predictions (<50%).* ***(A)*** *Classifier predictions (red) and 20% most salient vertices (orange) visualized on brain surfaces of the lesional hemisphere and on the T1 volume.* ***(B)*** *Z-scored mean feature values within 20% most salient vertices of predicted lesions. Color represents saliency scores. Features driving the classifier's prediction are positive (pink). Features inconsistent with MELD Graph prediction are negative (green).* ***(C)*** *T1 and FLAIR coronal sections with a red box indicating the lesional cortex.*

eReferences

1. Blümcke I, Thom M, Aronica E, et al. The clinicopathologic spectrum of focal cortical dysplasias: a consensus classification proposed by an ad hoc Task Force of the ILAE Diagnostic Methods Commission. *Epilepsia*. 2011;52(1):158-174.

2. Schuch F, Walger L, Schmitz M, et al. An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II. *Sci Data*. 2023;10(1):475.

3. Fischl B. FreeSurfer. *Neuroimage*. 2012;62(2):774-781.

4. Spitzer H, Ripart M, Whitaker K, et al. Interpretable surface-based detection of focal cortical dysplasias: a Multi-centre Epilepsy Lesion Detection study. *Brain*. 2022;145(11):3859-3871.

5. Greve DN, Van der Haegen L, Cai Q, et al. A surface-based analysis of language lateralization and cortical asymmetry. *J Cogn Neurosci*. 2013;25(9):1477-1492.

6. Sigalovsky IS, Fischl B, Melcher JR. Mapping an intrinsic MR property of gray matter in auditory cortex of living humans: a possible marker for primary cortex and hemispheric differences. *Neuroimage*. 2006;32(4):1524-1537.

7. Chen AA, Luo C, Chen Y, Shinohara RT, Shou H, Alzheimer's Disease Neuroimaging Initiative. Privacy-preserving harmonization via distributed ComBat. *Neuroimage*. 2022;248:118822.

8. Adler S, Whitaker K, Semmelroch M, Wagstyl K. MELD protocol 4 - lesion masking v2. *protocols.io*. Published online April 5, 2018. doi:10.17504/protocols.io.n9udh6w

9. Spitzer H, Ripart M, Whitaker K, et al. Interpretable surface-based detection of focal cortical dysplasias: a MELD study. *bioRxiv*. Published online December 14, 2021. doi:10.1101/2021.12.13.21267721

10. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing; 2015:234-241.

11. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211.

12. Gong S, Chen L, Bronstein M, Zafeiriou S. Spiralnet++: A fast and highly efficient mesh convolution operator. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. ; 2019:0-0.

13. Fawaz A, Williams LZJ, Alansary A, et al. Benchmarking Geometric Deep Learning for Cortical Segmentation and Neurodevelopmental Phenotype Prediction. *bioRxiv*. Published online December 2, 2021:2021.12.01.470730. doi:10.1101/2021.12.01.470730

14. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *arXiv [csLG]*. Published online March 3, 2017. Accessed August 19, 2021. http://proceedings.mlr.press/v70/sundararajan17a/sundararajan17a.pdf