







## RESEARCH ARTICLE

# Detection of focal cortical dysplasia: Development and multicentric evaluation of artificial intelligence models

Lennart N. Kersting<sup>1,2</sup>  | Lennart Walger<sup>1,2</sup>  | Tobias Bauer<sup>1,2,3</sup>  |  
 Vadym Gnatkovsky<sup>2</sup> | Fabiane Schuch<sup>2</sup> | Bastian David<sup>2</sup>  | Elisabeth Neuhaus<sup>4,5</sup> |  
 Fee Keil<sup>4</sup> | Anna Tietze<sup>6</sup> | Felix Rosenow<sup>5,7</sup> | Angela M. Kaindl<sup>8,9,10,11</sup>  |  
 Elke Hattingen<sup>4,5</sup> | Hans-Jürgen Huppertz<sup>12</sup> | Alexander Radbruch<sup>1,3,13</sup> |  
 Rainer Surges<sup>2</sup> | Theodor Rüber<sup>1,2,3,13</sup> 

<sup>1</sup>Department of Neuroradiology, University Hospital Bonn, Bonn, Germany

<sup>2</sup>Department of Epileptology, University Hospital Bonn, Bonn, Germany

<sup>3</sup>German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

<sup>4</sup>Department of Neuroradiology, Goethe University Frankfurt, Frankfurt am Main, Germany

<sup>5</sup>LOEWE Center for Personalized Translational Epilepsy Research (CePTER), Goethe-University, Frankfurt am Main, Germany

<sup>6</sup>Charité-Universitätsmedizin Berlin, Institute of Neuroradiology, Berlin, Germany

<sup>7</sup>Epilepsy Center Frankfurt Rhine-Main and Department of Neurology, Goethe-University, Frankfurt am Main, Germany

<sup>8</sup>Charité-Universitätsmedizin Berlin, Department of Pediatric Neurology, Berlin, Germany

<sup>9</sup>Charité-Universitätsmedizin Berlin, Center for Chronically Sick Children, Berlin, Germany

<sup>10</sup>Charité-Universitätsmedizin Berlin, German Epilepsy Center for Children and Adolescents, Berlin, Germany

<sup>11</sup>Charité-Universitätsmedizin Berlin, Institute of Cell and Neurobiology, Berlin, Germany

<sup>12</sup>Swiss Epilepsy Clinic, Klinik Lengg AG, Zurich, Switzerland

<sup>13</sup>Center for Medical Data Usability and Translation, University of Bonn, Bonn, Germany

## Correspondence

Theodor Rüber, Department of  
 Neuroradiology and Epileptology,  
 University Hospital Bonn, Campus  
 Venusberg 1, 53127 Bonn, Germany.  
 Email: [theodor.rueber@ukbonn.de](mailto:theodor.rueber@ukbonn.de)

## Abstract

**Objective:** Focal cortical dysplasia (FCD) is a common cause of drug-resistant focal epilepsy but can be challenging to detect visually on magnetic resonance imaging. Three artificial intelligence models for automated FCD detection are publicly available (MAP18, deepFCD, MELD) but have only been compared on single-center data. Our first objective is to compare them on independent multi-center test data. Additionally, we train and compare three new models and make them publicly available.

**Methods:** We retrospectively collected FCD cases from four epilepsy centers. We chose three novel models that take two-dimensional (2D) slices (2D-nnUNet), 2.5D slices (FastSurferCNN), and large 3D patches (3D-nnUNet) as inputs and trained them on a subset of Bonn data. As core evaluation metrics, we used

Lennart N. Kersting and Lennart Walger share first authorship.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Epilepsia* published by Wiley Periodicals LLC on behalf of International League Against Epilepsy.

voxel-level Dice similarity coefficient (DSC), cluster-level  $F_1$  score, subject-level detection rate, and specificity.

**Results:** We collected 329 subjects, 244 diagnosed with FCD ( $27.7 \pm 14.4$  years old, 54% male) and 85 healthy controls ( $7.1 \pm 2.4$  years old, 51% female). We used 118 subjects for model training and kept the remaining subjects as an independent test set. 3D-nnUNet achieved the highest  $F_1$  score of .58, the highest DSC of .36 (95% confidence interval [CI] = .30–.41), a detection rate of 55%, and a specificity of 86%. deepFCD showed the highest detection rate (82%) but had the lowest specificity (0%) and cluster-level precision (.03, 95% CI = .03–.04,  $F_1$  score = .07). MELD showed the least performance variation across centers, with detection rates between 46% and 54%.

**Significance:** This study shows the variance in performance for FCD detection models in a multicenter dataset. The two models with 3D input data showed the highest sensitivity. The 2D models performed worse than all other models, suggesting that FCD detection requires 3D data. The greatly improved precision of 3D-nnUNet may make it a sensible choice to aid FCD detection.

#### KEYWORDS

computer-aided detection, epilepsy, lesion detection, model comparison, MRI

## 1 | INTRODUCTION

Focal cortical dysplasia (FCD) is the third most common cause of drug-resistant focal epilepsy.<sup>1</sup> Surgical intervention yields seizure freedom in up to 70% of eligible candidates.<sup>2</sup> As part of the preoperative diagnostic workup, magnetic resonance imaging (MRI) is an important modality.<sup>3</sup> Accurate detection of the lesion on MRI is the best clinical predictor for postoperative seizure freedom.<sup>4</sup> FCDs typically exhibit specific features on MRI, including abnormal gyration, transmantle sign, cortical thickening, and gray–white matter blurring.<sup>5</sup> However, FCDs are still difficult to detect, with a high interrater variability and up to 30% of cases missed by conventional visual assessment.<sup>5,6</sup>

To aid the detection of FCDs, various studies have introduced artificial intelligence (AI)-based approaches.<sup>7–9</sup> A prior study introduced specific evaluation criteria for FCD detection based on the Metrics Reloaded Framework<sup>10</sup> and compared three state-of-the-art models (MAP18: Morphometric Analysis Program, version of 2018,<sup>7</sup> MELD: Multi-centre Epilepsy Lesion Detection,<sup>8</sup> deepFCD: deep learning-based model for FCD detection<sup>9</sup>) to human readers with different levels of expertise on single-center data.<sup>6</sup> Detection rates of these models varied from 31% to 73%, with the best model matching the sensitivity of experts, albeit being much less precise. More recently, Zhang and colleagues trained nnUNet, a widely used medical image segmentation framework,<sup>11</sup> for FCD detection, while omitting a detailed evaluation, and not making the model publicly available.<sup>12</sup> All

#### Key points

1. A multicenter cohort of 329 subjects was used to evaluate six AI models, three state-of-the-art and three new models, for the detection of focal cortical dysplasia.
2. The two models with 3D input data performed best, with a detection rate of up to 82%.
3. The newly trained 3D-nnUNet demonstrated superior balance between precision and sensitivity.

of these models employ substantially different approaches. MAP18 predicts FCDs based on single voxels from T1-weighted (T1w) images and morphometric feature maps, whereas MELD uses surface-based features from both T1w and fluid-attenuated inversion recovery (FLAIR) images. deepFCD yields predictions on small three-dimensional (3D) patches ( $16 \times 16 \times 16$  voxels), whereas 3D-nnUNet uses larger 3D patches ( $112 \times 112 \times 192$  voxels). In summary, the advantages of the widely used nnUNet have not been leveraged, and a detailed, multicenter comparison of the various state-of-the-art FCD detection approaches has not been conducted.

Here, we trained three deep learning architectures for FCD detection on an in-house dataset of 118 individuals

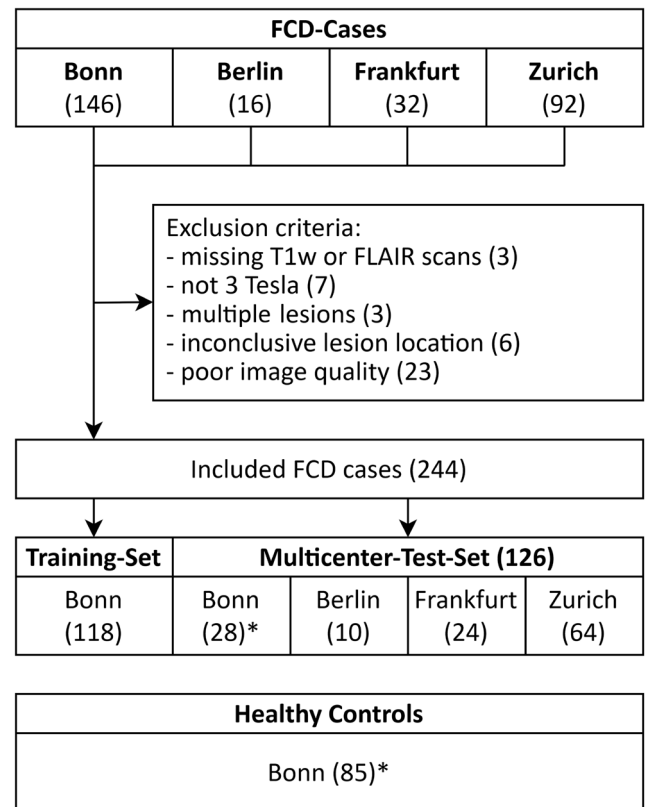
with epilepsy and FCD. For this, we selected the 2D and 3D full-resolution versions of nnUNet,<sup>11</sup> as well as FastSurferCNN, a 2.5D approach.<sup>13</sup> We then compared our newly trained approaches and the three state-of-the-art models (MAP18, MELD, and deepFCD) on 126 FCD cases from four different centers, including 28 publicly available subjects,<sup>14</sup> and an additional 85 published healthy controls (HCs). We hypothesized that there would be significant differences in the overall performance of these models and aimed to determine which approaches are best suited for the automated detection of FCDs.

## 2 | MATERIALS AND METHODS

### 2.1 | Participants

We retrospectively ascertained subjects with epilepsy and FCD who underwent presurgical evaluation at four different centers: Bonn (2006–2021), Berlin (2017–2020), Frankfurt (2007–2020), and Zurich (2008–2019). Exclusion criteria were missing whole brain T1w or FLAIR scans acquired at 3 T, multiple FCDs, inconclusive location, and failed preprocessing due to poor image quality. Lesion masks were created by clinicians experienced in the diagnosis of FCD at each individual center, with access to all other clinical information, such as electroencephalographic recordings, if available. Demographic and clinical information included age at scan, sex, and histopathological diagnosis according to International League Against Epilepsy classification.<sup>15,16</sup> No subjects were excluded due to incomplete demographic or clinical information. In addition, we included 85 HCs from a published dataset from Bonn.<sup>14</sup> To meet the subject requirement for MELD harmonization, non-FCD cases from Berlin and Frankfurt were used. Figure 1 illustrates the inclusion/exclusion process. The study was approved by the internal review boards in Bonn (no. 136/19), Berlin (no. EA2/084/18), and Frankfurt (no. 20–649) and in compliance with the internal review guidelines in Zurich, and written informed consent was obtained from all participants.

FCD cases from Bonn overlap with previous studies.<sup>6,7,14,17–20</sup> Utilizing all cases, Walger and colleagues compared human and AI model performance.<sup>6</sup> In this study, we introduce three new models and compare performances on multicenter data. Eighty-five FCD cases from Bonn have previously been published<sup>14</sup> of which 28 were defined as a representative test set, with an average expert detection rate of 49%.<sup>20</sup> Cases from Frankfurt overlap with previous studies.<sup>21,22</sup> Cases from Zurich were used for the training of MAP18,<sup>7</sup> and had to be excluded from the evaluation of MAP18 in this study due to data leakage, which could lead to overly optimistic performance estimates.



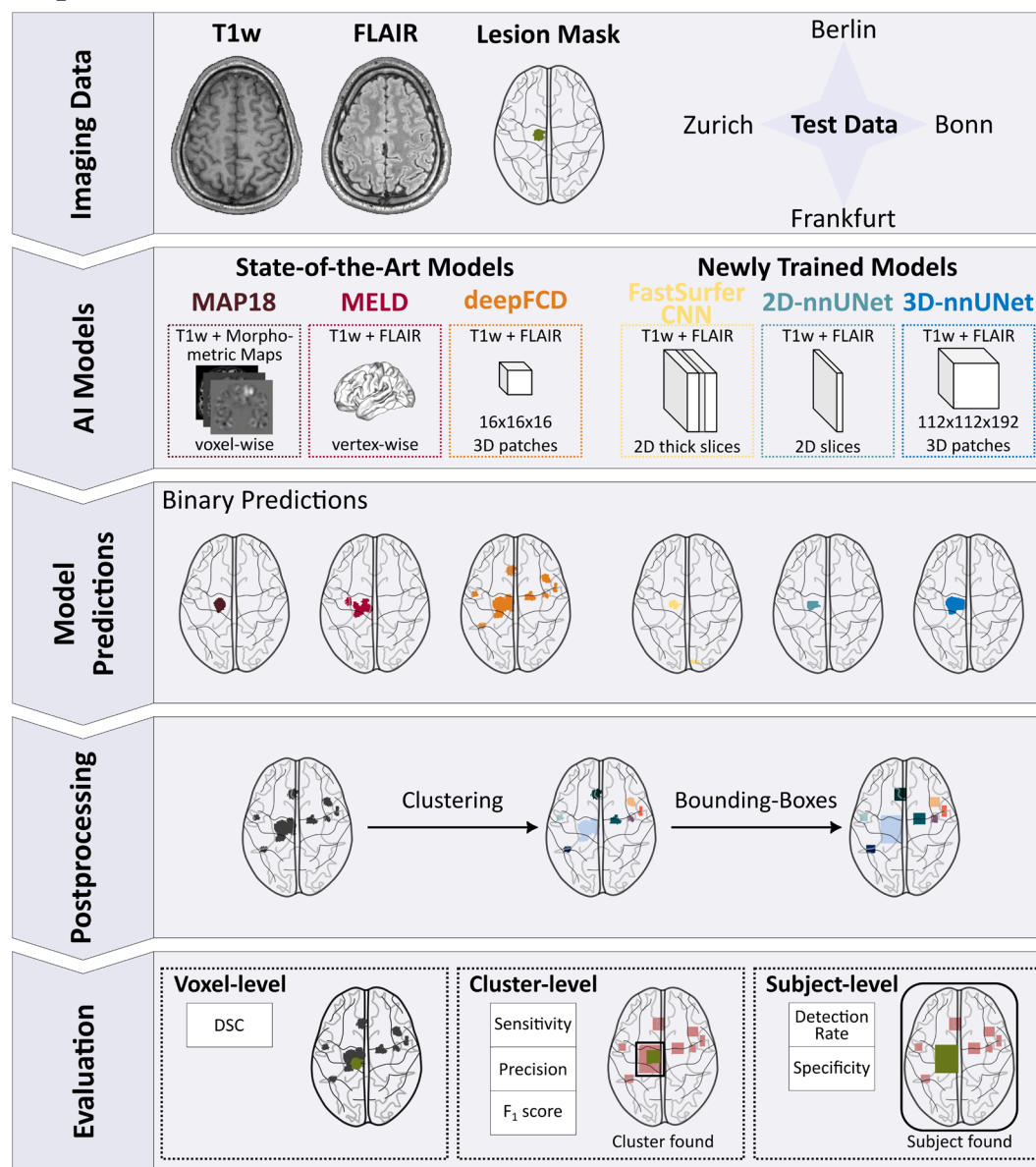
**FIGURE 1** Flowchart illustrating the inclusion and exclusion process. \*These subjects are part of the published focal cortical dysplasia (FCD) dataset from Bonn.<sup>14</sup> FLAIR, fluid-attenuated inversion recovery; T1w, T1-weighted.

### 2.2 | AI models

Inclusion criteria for existing AI models for FCD detection were publicly available code and external validation. They were met by MAP18,<sup>7</sup> MELD,<sup>8</sup> and deepFCD.<sup>9</sup> We chose three models representing a 2.5D, a 2D, and a 3D approach with open-source implementations, namely FastSurferCNN and nnUNet to train from scratch. FastSurferCNN has been developed for whole brain segmentation,<sup>13</sup> and nnUNet is widely used for segmentation tasks in medical imaging.<sup>11</sup> Figure 2 provides a schematic overview of the study processing pipeline. All models were trained and/or run on a Linux machine with an Nvidia 3090 Ti GPU and an Intel Core i9 CPU, except for MAP18, which was run on a Windows computer with an Intel Core i5 CPU.

#### 2.2.1 | Model training

Training of nnUNet and FastSurferCNN was performed with coregistered T1w and FLAIR images as inputs. We used synthseg<sup>23,24</sup> and mri\_easyreg<sup>24,25</sup> for coregistration.



**FIGURE 2** Processing pipeline illustrating the comparison between three newly trained models (FastSurferCNN, 2D-nnUNet, 3D-nnUNet) and three state-of-the-art models (MAP18, MELD, deepFCD). Each model's internal data structure is visualized (one-voxel, one-vertex, two-dimensional [2D] slices, thick slices, 3D patches). The ground truth lesion mask is shown in green, and each model is represented by a unique color. Binary predictions are used for voxel-level evaluation, measured by the Dice similarity coefficient (DSC), whereas bounding boxes of clustered predictions are used for cluster- and subject-level evaluation. The  $F_1$  score is a prediction measure describing the harmonic mean of precision and sensitivity. AI, artificial intelligence; FLAIR, fluid-attenuated inversion recovery; T1w, T1-weighted.

For FastSurferCNN, we concatenated T1w and FLAIR slices as combined inputs and sampled only slices containing lesional voxels. We performed fivefold cross-validation, training each fold for 150 epochs using binary cross-entropy loss. The number of epochs was empirically chosen so the model converged for each fold. Separate models were trained for all three anatomical views. We trained nnUNet without any modifications except for reducing the maximum number of epochs to 400 (from 1000), for which the model converged across all folds. We

omitted the postprocessing step included in its pipeline (which potentially removes all but the largest predicted cluster). We did not change any other hyperparameters.

### 2.2.2 | Model postprocessing

To arrive at a binary prediction for each model, some postprocessing had to be conducted. For FastSurferCNN, we averaged the output for each view across all folds,



thresholded it with .5, and combined the three views keeping the maximum value for each voxel. The output of MAP18 was binarized as described in the original publication.<sup>7</sup> The outputs of FastSurferCNN, 2D- and 3D-nnUNet, and MAP18 were clustered using deepFCD's clustering procedure, which groups together voxels that are immediately connected. MELD and deepFCD already produced a clustered output. The clustered output was not further processed, except for FastSurferCNN, where clusters smaller than 100 voxels were removed (the size was determined empirically).

## 2.3 | Metrics

We evaluated all models at voxel, cluster, and subject levels based on the comparison of model predictions and ground truth lesion masks resampled to an isotropic resolution of 1.0 mm. At voxel level, we report the Dice similarity coefficient (DSC). At cluster level, we first calculated a bounding box around each predicted cluster and the ground truth lesion, that is, the smallest rectangular region enclosing the entire cluster. We then applied two different criteria to determine true positive clusters, as defined in prior work.<sup>6</sup> The first criterion is based on single point localization, which we refer to as “pinpointing.” It is met if the clusterwise center of mass falls within the lesion mask. The second criterion is called “detecting,” which is met if the DSC score between two bounding boxes exceeds a threshold of .22. To not confuse this score with the voxel-level DSC, we will refer to it as “boxDSC.” The specific threshold was determined by Walger and colleagues to specifically reflect the detection performance of experts for FCD detection.<sup>6</sup> Note that these criteria could be met by multiple clusters per subject, even though each had only a single lesion and all clusters were considered in the calculation of the cluster-level metrics. We report cluster-level sensitivity, precision, and  $F_1$  score (a prediction measure combining precision and sensitivity into a single performance metric). At subject level, we report detection rates and pinpointing rates. An FCD case was considered “found” if at least one cluster met the criterion. Subject-level specificity was determined by evaluating model predictions for HCs. An HC was considered “true negative” if the prediction was empty. As primary metrics for comparing model performance, we use voxel-level DSC, cluster-level  $F_1$  score, and subject-level detection rate.

## 2.4 | Statistical analysis

Values were reported as mean and 95% confidence interval (CI). We applied bootstrapping to estimate CIs for the

performance variance across centers. All statistical analyses were performed using Stata.<sup>27</sup>

## 2.5 | Code availability

Instructions for setting up the trained nnUNet models to generate predictions for new individuals are available on GitLab ([https://gitlab.com/lab\\_tni/projects/nnunet\\_fcd](https://gitlab.com/lab_tni/projects/nnunet_fcd)). Adapted FastSurferCNN code can be made available upon reasonable request to the corresponding author.

# 3 | RESULTS

## 3.1 | Participants

A total of 244 FCD cases and 85 HCs were included in the study, of which 211 (126 FCD cases, 85 HCs) formed the multicenter test cohort, as shown in Figure 1. Of the included FCD cases, 49% were histologically confirmed, including five cases with FCD type I, 124 cases with FCD type II, and only one case with FCD type III. For Zurich, the pre-exclusion cohort is presented in Table 1, because demographic and clinical information is not available at the individual level for data protection reasons. The Berlin cohort differs from other centers in that it only contains pediatric subjects (mean age at scan =  $7.5 \pm 2.9$  years). The youngest age at scan in the entire dataset was 3 years. Demographic and clinical information is summarized in Table 1. All T1w images and all but 27 FLAIR images from Bonn, as well as all images from Berlin, Frankfurt, and Zurich, had isotropic resolutions between .5 mm and 1.0 mm. Scanner-specific information is shown in Supplementary Table 1.

## 3.2 | Test set performance

All metrics except for subject-level specificity were calculated using the group of 126 FCD cases. The 85 HCs were only used to calculate subject-level specificity. The average number of clusters per subject varied from .5 (95% CI = .3–.6) for 2D-nnUNet to 24.7 (95% CI = 22.4–27.2) for deepFCD. 2D-nnUNet also generated the highest number of zero predictions with 65% and deepFCD the least with 0%. MAP18 predicted the smallest clusters on average with a volume of .44 mL (95% CI = .31–.79) and 3D-nnUNet the largest with 2.44 mL (95% CI = 1.94–3.18). Whereas Table 2 gives an overview over all metrics for all models, in the following we highlight key differences. Including all preprocessing steps, generating predictions

**TABLE 1** Demographic and clinical Information.

Characteristic	FCD cases					HCs, Bonn
	Bonn, training	Bonn, test	Berlin, test	Frankfurt, test	Zurich, test <sup>a</sup>	
Number, <i>n</i>	118	28	10	24	92	85
Age at scan, years, mean $\pm$ SD	29.5 $\pm$ 14.2	28.0 $\pm$ 10.7	7.5 $\pm$ 2.9	28.3 $\pm$ 15.5	27.2 $\pm$ 14.3	7.1 $\pm$ 2.4
Sex, <i>n</i>						
Female	51	12	5	4	53	43
Male	67	16	5	20	39	42
Histopathology, <i>n</i>						
No surgery	45	8	9	13	59	-
I	3	0	0	0	2	-
II (a/b)	69 (19/50)	20 (6/14)	1 (0/1)	9 (1/8)	25 (NA)	-
IIIb	0	0	0	1	0	-
No definite FCD on histopathology	1	0	0	0	0	-
No classification	0	0	0	0	4	-
No information	0	0	0	1	2	-

Abbreviations: FCD, focal cortical dysplasia; HC, healthy control; NA, not available.

<sup>a</sup>Zurich information was only available for the pre-exclusion cohort (64 of these were included in this study).

**TABLE 2** Performance metrics for the multicenter test cohort.

Level	Metric	MAP18 <sup>a</sup>	MELD	deepFCD	FastSurferCNN	2D-nnUNet	3D-nnUNet
Voxel	Empty	21% (13/62)	12% (15/126)	0% (0/126)	26% (33/126)	65% (82/126)	21% (27/126)
	DSC	.15 [.10–.19]	.21 [.18–.24]	.15 [.12–.17]	.06 [.03, .08]	.07 [.04–.10]	.36 [.30–.41]
Cluster	Number/subject	3.5 [2.7–5.2]	2.1 [1.8–2.5]	24.7 [22.4–27.2]	2.3 [1.9–2.6]	.5 [.3–.6]	.9 [.8–1.0]
	Volume, mL	.44 [.31–.79]	.97 [.80–1.44]	.92 [.87–.97]	.86 [.71–1.02]	.75 [.48–1.28]	2.44 [1.94–3.18]
	Precision	.08 (18/219)	.26 (70/266)	.03 (107/3109)	.06 (16/284)	.26 (16/62)	.63 (69/110)
	Sensitivity	.29 (18/62)	.52 (70/134)	.82 (107/130)	.13 (16/126)	.13 (16/126)	.55 (69/126)
	<i>F</i> <sub>1</sub> score	.13	.35	.07	.08	.17	.58
Subject	Pinpointing rate	66% (41/62)	48% (61/126)	80% (101/126)	25% (31/126)	29% (36/126)	64% (81/126)
	Detection rate	29% (18/62)	49% (62/126)	82% (103/126)	13% (16/126)	13% (16/126)	55% (69/126)
	Specificity	51% (43/85)	55% (47/85)	0% (0/85)	22% (19/85)	95% (81/85)	86% (73/85)

Note: Values in square brackets are 95% confidence intervals; values in parentheses are numerators/denominators. The *F*<sub>1</sub> score is a prediction measure describing the harmonic mean of precision and recall. The 85 healthy controls were only used to calculate subject-level specificity.

Abbreviation: DSC, Dice similarity coefficient.

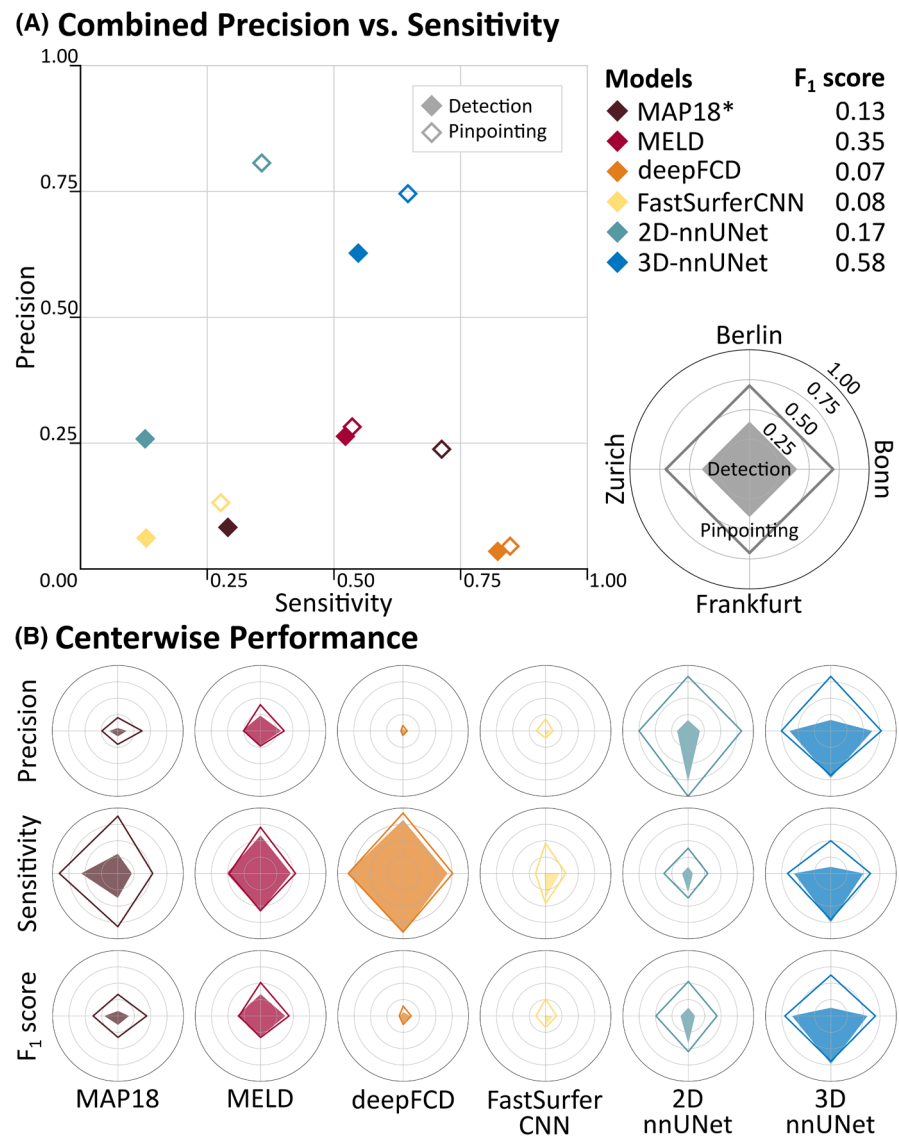
<sup>a</sup>For MAP18, Zurich cases were excluded from evaluation, as they were used in training.

for a single subject took approximately 20 min for MAP18, 7–9 h for MELD (6–8 h for FreeSurfer reconstruction, 45–60 min for feature extraction and harmonization, and 5–10 min for inference), and approximately 2 h for deepFCD (<5 min for preprocessing and 120 min for inference). FastSurferCNN required approximately 8 min and both nnUNets approximately 6 min (for both, preprocessing took approximately 5 min and inference 3 min and <1 min, respectively).

### 3.2.1 | Voxel level

3D-nnUNet achieved the highest DSC score with .36 (95% CI = .30–.41), followed by MELD with .21 (95% CI = .18–.24). MAP18 and deepFCD achieved similar DSC scores of .15 (95% CI = .10–.19) and .15 (95% CI = .12–.17), respectively, for different reasons. Whereas MAP18 showed a low sensitivity of .11 (95% CI = .07–.15) and high precision of .51 (95% CI = .41–.60), the opposite was true

**FIGURE 3** Cluster-level performance of all models for the entire test cohort (A) and centerwise (B). (A) Cluster-level precision and sensitivity for the entire test cohort. \*MAP18 excludes Zurich data. (B) Radar plots showing cluster-level precision, sensitivity, and  $F_1$  score for individual centers for detection criterion (shaded areas) and pinpointing (unfilled lines). The  $F_1$  score is a prediction measure describing the harmonic mean of precision and sensitivity.



for deepFCD (sensitivity = .43, 95% CI = .38–.48; precision = .11, 95% CI = .09–.14). FastSurferCNN and 2D-nnUNet were lowest, with a DSC score of .06 (95% CI = .03–.08) and .07 (95% CI = .04–.10), respectively.

### 3.2.2 | Cluster level

deepFCD achieved the highest average DSC score, choosing the best-overlapping cluster per subject with .38 (95% CI = .34–.42), followed by 3D-nnUNet with .36 (95% CI = .30–.41). However, given that a cluster was detected, 3D-nnUNet showed the highest DSC of .62 (95% CI = .58–.66). Whereas 3D-nnUNet achieved the highest precision of .63, deepFCD showed the lowest with .03. deepFCD was most sensitive (.82), followed by 3D-nnUNet (.55) and MELD (.52). Combining precision and sensitivity, 3D-nnUNet showed the highest  $F_1$  score with .58, followed by

MELD with .35. deepFCD had the lowest  $F_1$  score with .07. Figure 3A shows cluster-level precision and sensitivity.

### 3.2.3 | Subject level

deepFCD achieved the highest pinpointing rate (80%) and detection rate (82%). MAP18 achieved the second highest pinpointing rate (66%), followed by 3D-nnUNet (64%), which had the second highest detection rate (55%). 2D-nnUNet and FastSurferCNN showed both the lowest pinpointing (29% and 25%, respectively) and detection rates (both 13%). 2D-nnUNet produced the fewest nonzero predictions for HCs with 5% (4/85), followed by 3D-nnUNet with 14% (12/85). MELD, MAP18, and FastSurferCNN produced predictions in 45% (38/85), 49% (42/85), and 78% (66/85), respectively. deepFCD had the lowest specificity (0%), with nonzero predictions for 100% (85/85) of the HCs.

### 3.3 | Centerwise performance

The detection rate of the state-of-the-art models did not differ substantially between the test and validation set from Bonn (MAP18: 11%, MELD: 4%, deepFCD: 6%). Detection rates on the validation set were 32% higher for FastSurferCNN (test: 21%, validation: 53%) and 27% higher for 3D-nnUNet (test: 50%, validation: 77%); however, 2D-nnUNet differed by only 7% (test: 7%, validation: 14%). All performance metrics for the validation data are shown in [Supplementary Table 2](#).

The Bonn test set was selected to have an expert performance of 49% by Walger, Schmitz, and colleagues.<sup>20</sup> The highest detection rate was achieved by deepFCD with 68%, followed by 3D-nnUNet with 50% and MELD with 46%. MAP18, FastSurferCNN, and 2D-nnUNet detected 21%, 21%, and 7% of the subjects, respectively. These test cases are also part of the previously published FCD dataset from Bonn. [Supplementary Table 3](#) shows the model performance metrics for all published FCD cases.

Comparing performance per center, MELD showed the least variance in performance across all centers, with detection rates between 46% for Bonn and 54% for Frankfurt. deepFCD was the most sensitive for each center, with a minimum detection rate of 68% for Bonn and a maximum of 92% in Frankfurt. Overall, the Frankfurt cohort represented the “easiest” dataset, with the highest overall detection rate of 57% (Zurich: 42%, Bonn: 35%, Berlin: 33%) and all models achieving their highest detection rate. For Berlin, both 2D-nnUNet and 3D-nnUNet detected only a single subject, but 3D-nnUNet still pinpointed five of 10 (2D-nnUNet pinpointing 2/10). FastSurferCNN detected zero subjects from the Berlin cohort but pinpointed three of 10. The detection rates for all other models did not substantially differ for the Berlin cohort. Similar to the increased detection rates of FastSurferCNN and 3D-nnUNet on our validation data, we observed a 27% higher detection rate for MAP18 on the Zurich cohort. Values for all centerwise metrics are shown in [Table 3](#). [Figure 3B](#) shows the centerwise cluster-level metrics.

## 4 | DISCUSSION

We trained three new models for the detection of FCDs using FastSurferCNN, 2D-nnUNet, and 3D-nnUNet and compared them to the three state-of-the-art FCD detection AI models MAP18, MELD, and deepFCD on an independent multicenter dataset. The new models were trained on 118 FCD cases from Bonn. All models were tested on 85 HCs and 126 FCD cases from four centers, including a pediatric cohort. deepFCD showed the highest detection rate of 82%, followed by 3D-nnUNet with 55%. However, deepFCD also

produced nonzero predictions in all HCs, that is, showed a specificity of 0%, whereas 3D-nnUNet still had a specificity of 86%. 2D-nnUNet and 2.5D FastSurferCNN showed the lowest detection rates, both at 13%. Cluster-level precision was highest for 3D-nnUNet with .63 and lowest for deepFCD with .03. In combination, 3D-nnUNet showed the highest cluster-level  $F_1$  score with .58, followed by MELD with .35. deepFCD had the lowest  $F_1$  score with .07.

Evaluation of AI models for FCD detection is highly dependent on the choice of the criterion for “finding” a lesion.<sup>6</sup> MAP18 was particularly poorly represented by the chosen criterion with 29% detected versus 66% pinpointed. In Berlin, 3D-nnUNet showed its lowest detection rate of 10%, but still pinpointed 50%. Criteria such as one-voxel overlap, as used in the original works of MAP18,<sup>7</sup> MELD,<sup>8</sup> and deepFCD,<sup>9</sup> but also pinpointing, can be exploited, for example, by predicting many clusters. Thus, comparing models based solely on any sort of detection rate does not provide a comprehensive evaluation.<sup>28</sup> We chose the cluster-level  $F_1$  score to highlight the tradeoff between the number of lesions detected and the number of clusters predicted. However, the  $F_1$  score may be less informative compared to the detection rate, especially from a clinical perspective. The acceptable number of false-positive clusters in clinical practice remains an open research question. Although voxel-level metrics are most commonly used to evaluate model performance, it remains unclear how important such metrics are for the care of people with epilepsy; for example, to capture the precise extent of a lesion could be critical for surgical planning. Performing evaluations at each of these levels is necessary to provide a comprehensive analysis and comparison.

Model performance varied between centers. The newly trained models showed better performance on validation data, compared to test data, and also MAP18 performed substantially better on the Zurich cohort, which was used in its training. MELD's performance showed the least variability across all centers, which may be due to the MRI data harmonization scheme used in MELD's pipeline. Regardless of a model's ability to generalize, a given cohort may be “easier” overall, as we observed for Frankfurt data, or a model may be specifically affected by some characteristics within a dataset, as we observed with nnUNet on the Berlin pediatric cohort. Future studies may want to specifically focus on such characteristics, for example, training a model on pediatric cases. Such variation across model performance and cohorts highlights the need for multicenter evaluation to get a better understanding of all the factors that may influence any specific model.

The models evaluated in our study use different methods to process MRI data and generate predictions. MAP18 operates at the single-voxel level,<sup>7</sup> MELD at the single-vertex level,<sup>8</sup> and deepFCD uses small  $16 \times 16 \times 16$  voxel



**TABLE 3** Centerwise focal cortical dysplasia detection performance.

Level	Metric	MAPI8	MELD	deepFCD	FastSurferCNN	2D-nnUNet	3D-nnUNet
Bonn, <i>n</i> = 28							
Voxel	Empty	29% (8/28)	29% (8/28)	0% (0/28)	11% (3/28)	68% (19/28)	29% (8/28)
	DSC	.12 [.08–.18]	.17 [.11–.24]	.15 [.10–.20]	.07 [.03, .13]	.03 [.01–.06]	.30 [.20–.40]
Cluster	Number/subject	1.5 [.9–2.6]	1.5 [1.0–1.9]	11.9 [10.6–13.2]	2.8 [2.1–3.5]	.4 [.2–.6]	.8 [.6–1.0]
	Volume, mL	.45 [.29–.67]	.70 [.52–.91]	.85 [.73–1.01]	.61 [.47–.87]	.35 [.14–.66]	1.75 [1.12–2.75]
	Precision	.14 (6/43)	.32 (13/41)	.06 (19/333)	.08 (6/78)	.18 (2/11)	.64 (14/22)
	Sensitivity	.21 (6/28)	.46 (13/28)	.68 (19/28)	.21 (6/28)	.07 (2/28)	.50 (14/28)
	<i>F</i> <sub>1</sub> score	.17	.38	.11	.11	.10	.56
Subject	Pinpointing rate	50% (14/28)	54% (15/28)	75% (21/28)	29% (8/28)	25% (7/28)	61% (17/28)
	Detection rate	21% (6/28)	46% (13/28)	68% (19/28)	21% (6/28)	7% (2/28)	50% (14/28)
Berlin, <i>n</i> = 10							
Voxel	Empty	10% (1/10)	0% (0/10)	0% (0/10)	10% (1/10)	70% (7/10)	40% (4/10)
	DSC	.13 [.05–.26]	.16 [.07–.28]	.16 [.09–.25]	.02 [.00–.08]	.04 [.00–.16]	.09 [.01–.34]
Cluster	Number/subject	6.9 [4.3–8.8]	3.0 [2.1–3.7]	29.4 [25.8–33.0]	3.3 [2.1–4.1]	.6 [.1–1.8]	.6 [.2–.8]
	Volume, mL	.53 [.21–1.72]	1.90 [.70–6.36]	.97 [.80–1.34]	.63 [.43–1.23]	.50 [.14–1.10]	1.52 [.41–3.63]
	Precision	.04 (3/69)	.23 (7/30)	.03 (9/294)	.00 (0/33)	.17 (1/6)	.17 (1/6)
	Sensitivity	.30 (3/10)	.58 (7/12)	.82 (9/11)	.00 (0/10)	.10 (1/10)	.10 (1/10)
	<i>F</i> <sub>1</sub> score	.08	.33	.06	.00	.12	.12
Subject	Pinpointing rate	80% (8/10)	50% (5/10)	80% (8/10)	30% (3/10)	20% (2/10)	50% (5/10)
	Detection rate	30% (3/10)	50% (5/10)	80% (8/10)	0% (0/10)	10% (1/10)	10% (1/10)
Frankfurt, <i>n</i> = 24							
Voxel	Empty	17% (4/24)	4% (1/24)	0% (0/24)	0% (0/24)	62% (15/24)	8% (2/24)
	DSC	.18 [.11–.27]	.25 [.18–.32]	.28 [.21–.36]	.13 [.07–.21]	.15 [.07–.27]	.51 [.37–.63]
Cluster	Number/subject	4.5 [3.0–9.0]	2.9 [2.0–4.1]	15.4 [12.2–18.9]	4.2 [3.3–5.1]	.4 [.2–.5]	1.0 [.8–1.2]
	Volume, mL	.37 [.24–.60]	1.13 [.83–1.56]	.85 [.73–1.00]	1.05 [.81–1.45]	1.51 [.74–2.93]	2.95 [2.01–4.23]
	Precision	.08 (9/107)	.22 (15/69)	.06 (22/370)	.09 (9/101)	.78 (7/9)	.72 (18/25)
	Sensitivity	.38 (9/24)	.58 (15/26)	.92 (22/24)	.38 (9/24)	.29 (7/24)	.75 (18/24)
	<i>F</i> <sub>1</sub> score	.14	.32	.11	.14	.42	.73
Subject	Pinpointing rate	79% (19/24)	50% (12/24)	88% (21/24)	46% (11/24)	38% (9/24)	71% (17/24)
	Detection rate	38% (9/24)	54% (13/24)	92% (22/24)	38% (9/24)	29% (7/24)	75% (18/24)
Zurich, <i>n</i> = 64							
Voxel	Empty	8% (5/64)	9% (6/64)	0% (0/64)	45% (29/64)	64% (41/64)	20% (13/64)
	DSC	.31 [.26–.36]	.22 [.18–.27]	.09 [.07–.11]	.03 [.01–.05]	.07 [.04–.12]	.36 [.29–.44]
Cluster	Number/subject	4.4 [3.1–7.7]	2.0 [1.6–2.4]	33.0 [30.2–36.2]	1.1 [.8–1.6]	.6 [.4–.9]	.9 [.7–1.0]
	Volume, mL	.39 [293–541]	.76 [.62–.99]	.94 [.88–1.00]	.99 [.71–1.43]	.73 [.33–1.58]	2.58 [1.82–3.94]
	Precision	.13 (36/281)	.28 (35/126)	.03 (57/2112)	.01 (1/72)	.17 (6/36)	.63 (36/57)
	Sensitivity	.56 (36/64)	.51 (35/68)	.85 (57/67)	.02 (1/64)	.09 (6/64)	.56 (36/64)
	<i>F</i> <sub>1</sub> score	.21	.36	.05	.01	.12	.60
Subject	Pinpointing rate	88% (56/64)	45% (29/64)	80% (51/64)	14% (9/64)	28% (18/64)	66% (42/64)
	Detection rate	56% (36/64)	48% (31/64)	84% (54/64)	2% (1/64)	9% (6/64)	56% (36/64)

*Note:* Values in square brackets are 95% confidence intervals; values in parentheses are numerators/denominators. The *F*<sub>1</sub> score is a prediction measure describing the harmonic mean of precision and recall.

Abbreviation: DSC, Dice similarity coefficient.

3D patches.<sup>9</sup> FastSurferCNN uses a 2.5D approach,<sup>13</sup> whereas nnUNet uses either a 2D or 3D approach,<sup>11</sup> using a large 3D patch size (112×112×192 voxel) compared to deepFCD. Our results suggest that models using a 3D framework, especially large 3D patches, are the most promising approach for FCD segmentation. deepFCD was the most sensitive, but least precise model, whereas 3D-nnUNet had the highest  $F_1$  score and second highest sensitivity. A study by Avesta and colleagues compared the performance of 3D models and their 2D or 2.5D counterparts in segmenting three anatomical regions in brain MRI and showed that 3D approaches were superior, even with limited training data.<sup>29</sup> However, differences in input data dimensions or shape may be only one of many factors that affect model performance. The heavy preprocessing employed in MAP18 and MELD may further aid detection ability. Combining preprocessing, for example, generating feature maps of MAP18, with 3D patch-based models may be worth exploring in future approaches.

Several limitations should be acknowledged. First, the localization of FCDs remains challenging even for experts,<sup>6</sup> leading to uncertainty in the ground truth annotations. In this study, lesion masks were drawn by clinicians from different centers, which may introduce additional variations in the annotation style. The employed “pinpointing” and “detecting” criteria aim to counteract such voxel-level variance. Second, only approximately half of all FCD cases were histologically confirmed. Although in a previous study we found no significant differences in detection performance between confirmed and unconfirmed cases in the Bonn cohort,<sup>6</sup> this remains a limitation of our study. Third, because only five of the histologically confirmed cases were FCD type I, our study provides limited insight into the performance of automated lesion detection algorithms for FCD type I. Lastly, we have only included 3-T MRI data, and it remains unclear whether and how model performance is affected by different scanner field strengths.

In conclusion, we presented the first multicenter comparison of publicly available AI-based approaches for FCD detection. Our newly trained 3D-nnUNet offered the best tradeoff between cluster-level sensitivity and precision. Its comparatively fast runtime of only a few minutes per case may aid its integration into clinical practice. Future model development may focus on 3D models for high sensitivity while trying to maintain high precision. Furthermore, it has to be determined what the tradeoff between sensitivity and precision means with respect to how helpful a model is in the diagnostic workup.

## AUTHOR CONTRIBUTIONS

Lennart Kersting and Lennart Walger contributed equally to all aspects of this work. Tobias Bauer, Alexander

Radbruch, Rainer Surges, and Theodor Rüber contributed to the conception and design. Tobias Bauer, Vadym Gnatkovsky, Fabiane Schuch, Bastian David, Elisabeth Neuhaus, Fee Keil, Anna Tietze, Felix Rosenow, Angela M. Kaindl, Elke Hattingen, Hans-Jürgen Huppertz, and Theodor Rüber acquired the data and contributed to its analysis and interpretation. Theodor Rüber helped draft the manuscript, and Elisabeth Neuhaus, Anna Tietze, Felix Rosenow, Angela M. Kaindl, Elke Hattingen, Hans-Jürgen Huppertz, Alexander Radbruch, Rainer Surges, and Theodor Rüber contributed to its revision.

## ACKNOWLEDGMENTS

This work was partially supported by a grant of the federal state of Hesse for the LOEWE Center for Personalized Translational Epilepsy Research, as well as by the Einstein Stiftung Fellowship through the Günter Endres Fond and the Sonnenfeld-Stiftung. Open Access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST STATEMENT

F.R. has received honoraria for lecturing and consultation from Angelini Pharma, Eisai, Jazz Pharma, Roche Pharma, Takeda, and UCB Pharma, and has received financial research support from Dr. Schär Deutschland, Vitaflor Deutschland, Nutricia Milupa, Desitin Pharma, Hamburg, Federal State of Hesse (via the LOEWE program), Chaja Foundation Frankfurt, Reiss Foundation Frankfurt, Dr. Senckenbergische Foundation Frankfurt, Ernst Max von Grunelius Foundation Frankfurt, and Detlev-Wrobel-Fonds for Epilepsy Research Frankfurt outside the submitted work. A.M.K. has served on the advisory boards of Angelini, Desitin, Jazz Pharmaceuticals, Novartis, and UCB. H.-J.H. is the author of the Morphometric Analysis Program v2018 (MAP18). A.R. has served on scientific advisory boards for GE Healthcare, Bracco, Bayer, Guerbet, and AbbVie; has received speaker honoraria from Bayer, Guerbet, Siemens, and Medscape; and has been a consultant for, and has received institutional study support from, Guerbet and Bayer. R.S. has received fees as speaker or for serving on advisory boards from Angelini, Arvelle, Bial, Desitin, Eisai, Janssen-Cilag, LivaNova, Novartis, Precisis, UCB Pharma, UNEEG, and Zogenix. These activities were not related to the content of this article. T.R. has received fees as a speaker from Eisai. The remaining authors have no conflicts of interest. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

## DATA AVAILABILITY STATEMENT

A subset of the 146 FCD cases from Bonn, including the 28 test cases used in this study and the 85 HCs, is available to the public ([doi: 10.18112/openneuro.ds004199.v1.0.5](https://doi.org/10.18112/openneuro.ds004199.v1.0.5)).

Instructions for installing the newly trained nnUNet models are available on GitLab ([https://gitlab.com/lab\\_tni/projects/nnunet\\_fcd](https://gitlab.com/lab_tni/projects/nnunet_fcd)). Additional data can be made available upon reasonable request to the corresponding author.

## ORCID

Lennart N. Kersting  <https://orcid.org/0009-0002-1983-4892>

org/0009-0002-1983-4892

Lennart Walger  <https://orcid.org/0000-0002-3300-6877>

Tobias Bauer  <https://orcid.org/0000-0002-0555-6214>

Bastian David  <https://orcid.org/0000-0002-0146-0629>

Angela M. Kaindl  <https://orcid.org/0000-0001-9454-206X>

org/0000-0001-9454-206X

Theodor Rüber  <https://orcid.org/0000-0002-6180-7671>

## REFERENCES

- Blümcke I, Kobow K, Holthausen H. Die ILAE-Klassifikation fokaler kortikaler Dysplasien im klinischen Gebrauch. *Z Für Epileptol.* 2017;30(3):200–7.
- Lamberink HJ, Otte WM, Blümcke I, Braun KPJ, Aichholzer M, Amorim 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):748–57.
- Guerrini R, Duchowny M, Jayakar P, Krsek P, Kahane P, Tassi L, et al. Diagnostic methods and treatment options for focal cortical dysplasia. *Epilepsia.* 2015;56(11):1669–86.
- Wagstyl K, Whitaker K, Raznahan A, Seidlitz J, Vértes PE, Foldes S, et al. Atlas of lesion locations and postsurgical seizure freedom in focal cortical dysplasia: a MELD study. *Epilepsia.* 2022;63(1):61–74.
- Urbach H, Kellner E, Kremers N, Blümcke I, Demerath T. MRI of focal cortical dysplasia. *Neuroradiology.* 2022;64(3):443–52.
- Walger L, Bauer T, Kügler D, Schmitz MH, Schuch F, Arendt C, et al. A quantitative comparison between human and artificial intelligence in the detection of focal cortical dysplasia. *Investig Radiol.* 2024;1125. <https://doi.org/10.1097/RLI.0000000000001125>
- David B, Kröll-Seger J, Schuch F, Wagner J, Wellmer J, Woermann F, et al. External validation of automated focal cortical dysplasia detection using morphometric analysis. *Epilepsia.* 2021;62(4):1005–21.
- Spitzer H, Ripart M, Whitaker K, D'Arco F, Mankad K, Chen AA, et al. Interpretable surface-based detection of focal cortical dysplasias: a multi-centre epilepsy lesion detection study. *Brain.* 2022;145(11):3859–71.
- Gill RS, Lee H-M, Caldaïrou B, Hong SJ, Barba C, Deleo F, et al. Multicenter validation of a deep learning detection algorithm for focal cortical dysplasia. *Neurology.* 2021;97(16):e1571–e1582.
- Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. *Nat Methods.* 2024;21(2):195–212.
- 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):203–11.
- Zhang S, Zhuang Y, Luo Y, Zhu F, Zhao W, Zeng H. Deep learning-based automated lesion segmentation on pediatric focal cortical dysplasia II preoperative MRI: a reliable approach. *Insights Imaging.* 2024;15(1):71.
- Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage.* 2020;219:117012.
- Schuch F, Walger L, Schmitz M, David B, Bauer T, Harms A, et al. An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II. *Sci Data.* 2023;10(1):475.
- Blümcke I, Thom M, Aronica E, Armstrong DD, Vinters HV, Palmini A, 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):158–74.
- Najm I, Lal D, Alonso Vanegas M, Cendes F, Lopes-Cendes I, Palmini A, et al. The ILAE consensus classification of focal cortical dysplasia: an update proposed by an ad hoc task force of the ILAE diagnostic methods commission. *Epilepsia.* 2022;63(8):1899–919.
- Rácz A, Becker AJ, Quesada CM, Borger V, Vatter H, Surges R, et al. Post-surgical outcome and its determining factors in patients operated on with focal cortical dysplasia type II—A retrospective Monocenter study. *Front Neurol.* 2021;12:666056. <https://doi.org/10.3389/fneur.2021.666056/full>
- Salemdawod A, Wach J, Banat M, Borger V, Hamed M, Haberl H, et al. Predictors of postoperative long-term seizure outcome in pediatric patients with focal cortical dysplasia type II at a German tertiary epilepsy center. *J Neurosurg Pediatr.* 2022;29(1):83–91.
- Wagner J, Weber B, Urbach H, Elger CE, Huppertz HJ. Morphometric MRI analysis improves detection of focal cortical dysplasia type II. *Brain.* 2011;134(10):2844–54.
- Walger L, Schmitz MH, Bauer T, Kügler D, Schuch F, Arendt C, et al. A Public Benchmark for Human Performance in FCD Detection. 2024. <https://doi.org/10.21203/rs.3.rs-4528693/v1>.
- Ahmad R, Maiworm M, Nöth U, Seiler A, Hattingen E, Steinmetz H, et al. Cortical changes in epilepsy patients with focal cortical dysplasia: new insights with T2 mapping. *J Magn Reson Imaging JMRI.* 2020;52(6):1783–9.
- Maiworm M, Nöth U, Hattingen E, Steinmetz H, Knake S, Rosenow F, et al. Improved visualization of focal cortical dysplasia with surface-based multiparametric quantitative MRI. *Front Neurosci.* 2020;14:622.
- Billot B, Greve DN, Puonti O, Thielscher A, van Leemput K, Fischl B, et al. SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. *Med Image Anal.* 2023;86:102789.
- Billot B, Magdamo C, Cheng Y, Arnold SE, das S, Iglesias JE. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. *Proc Natl Acad Sci.* 2023;120(9):e2216399120.
- Iglesias JE. A ready-to-use machine learning tool for symmetric multi-modality registration of brain MRI. *Sci Rep.* 2023;13(1):6657.
- Hoffmann M, Billot B, Greve DN, Iglesias JE, Fischl B, Dalca AV. SynthMorph: learning contrast-invariant registration without acquired images. *IEEE Trans Med Imaging.* 2022;41(3):543–58.
- Anon. Stata Statistical Software. 2023.
- Walger L, Adler S, Wagstyl K, Henschel L, David B, Borger V, et al. Artificial intelligence for the detection of focal cortical

dysplasia: challenges in translating algorithms into clinical practice. *Epilepsia*. 2023;64(5):1093–112.

29. Avesta A, Hossain S, Lin M, Aboian M, Krumholz HM, Aneja S. Comparing 3D, 2.5D, and 2D approaches to brain image auto-segmentation. *Bioengineering*. 2023;10(2):181.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Kersting LN, Walger L, Bauer T, Gnatkovsky V, Schuch F, David B, et al. Detection of focal cortical dysplasia: Development and multicentric evaluation of artificial intelligence models. *Epilepsia*. 2025;66:1165–1176. <https://doi.org/10.1111/epi.18240>