Received: 9 October 2022 | Revised: 27 January 2023 | Accepted: 30 January 2023

DOI: 10.1111/epi.17522

Epilepsia®

CRITICAL REVIEW

Artificial intelligence for the detection of focal cortical dysplasia: Challenges in translating algorithms into clinical practice

Lennart Walger[1] | Sophie Adler[2] | Konrad Wagstyl[3] | Leonie Henschel[4] | Bastian David[1] | Valeri Borger[5] | Elke Hattingen[6] | Hartmut Vatter[5] | Christian E. Elger[1] | Torsten Baldeweg[2] | Felix Rosenow[7,8] | Horst Urbach[9] | Albert Becker[10] | Alexander Radbruch[11] | Rainer Surges[1] | Martin Reuter[4,12,13] | Fernando Cendes[14] | Zhong Irene Wang[15] | Hans-Jürgen Huppertz[16] | Theodor Rüber[1]

[1]Department of Epileptology, University of Bonn Medical Center, Bonn, Germany

[2]University College London Great Ormond Street Institute for Child Health, London, UK

[3]Wellcome Centre for Human Neuroimaging, London, UK

[4]German Center for Neurodegenerative Diseases, Bonn, Germany

[5]Department of Neurosurgery, University Hospital Bonn, Bonn, Germany

[6]Department of Neuroradiology, University Hospital and Goethe University Frankfurt, Frankfurt am Main, Germany

[7]Epilepsy Center Frankfurt Rhine-Main and Department of Neurology, Goethe University and University Hospital Frankfurt, Frankfurt am Main, Germany

[8]LOEWE Center for Personalized Translational Epilepsy Research, Goethe University, Frankfurt am Main, Germany

[9]Department of Neuroradiology, Medical Center, University of Freiburg, Freiburg, Germany

[10]Department of Neuropathology, University Hospital Bonn, Bonn, Germany

[11]Department of Neuroradiology, University Hospital Bonn, Bonn, Germany

[12]A. A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, Massachusetts, USA

[13]Department of Radiology, Harvard Medical School, Boston, Massachusetts, USA

[14]Department of Neurology, University of Campinas, Campinas, Brazil

[15]Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, Ohio, USA

[16]Swiss Epilepsy Center, Klinik Lengg, Zürich, Switzerland

Correspondence
Lennart Walger and Theodor Rüber, Department of Epileptology, University of Bonn Medical Center, Bonn, Germany.
Email: lennart.walger@ukbonn.de and theodor.rueber@ukbonn.de

Abstract

Focal cortical dysplasias (FCDs) are malformations of cortical development and one of the most common pathologies causing pharmacoresistant focal epilepsy. Resective neurosurgery yields high success rates, especially if the full extent of the lesion is correctly identified and completely removed. The visual assessment of magnetic resonance imaging does not pinpoint the FCD in 30%–50% of cases, and half of all patients with FCD are not amenable to epilepsy surgery, partly

because the FCD could not be sufficiently localized. Computational approaches to FCD detection are an active area of research, benefitting from advancements in computer vision. Automatic FCD detection is a significant challenge and one of the first clinical grounds where the application of artificial intelligence may translate into an advance for patients' health. The emergence of new methods from the combination of health and computer sciences creates novel challenges. Imaging data need to be organized into structured, well-annotated datasets and combined with other clinical information, such as histopathological subtypes or neuroimaging characteristics. Algorithmic output, that is, model prediction, requires a technically correct evaluation with adequate metrics that are understandable and usable for clinicians. Publication of code and data is necessary to make research accessible and reproducible. This critical review introduces the field of automatic FCD detection, explaining underlying medical and technical concepts, highlighting its challenges and current limitations, and providing a perspective for a novel research environment.

KEYWORDS

digitalization in medicine, focal epilepsy, image processing, neuroimaging, presurgical evaluation

1 | INTRODUCTION

Focal cortical dysplasias (FCDs) are malformations of cortical development and one of the most common pathologies causing pharmacoresistant focal epilepsy.[1,2] Histopathological and genetic analysis may characterize FCDs into different types, potentially exhibiting certain features visible in magnetic resonance imaging (MRI). These are cortical thickening, blurring of the gray–white matter interface, abnormal cortical gyration, hyperintense T2/fluid-attenuated inversion recovery (FLAIR) signal, and transmantle sign, as depicted in Figure 1. Despite common visible abnormalities, almost one third of patients show no abnormal MRI on visual analysis and are deemed "MRI-negative."[3]

The neurosurgical resection of the dysplastic cortex usually yields high success rates, with 70% of patients achieving seizure freedom.[4,5] However, postoperative seizure freedom is highly dependent on presurgical identification of the lesion on MRI,[6] accurate assessment of the lesion extent,[7] lesion location,[5] complete resection of the lesion,[8] and the subtype of FCD.[1,9] Notably, the favorable long-term postoperative seizure outcome rate decreases to approximately 11% for patients with extratemporal epilepsy who are MRI-negative.[10]

Because detecting an FCD on MRI substantially impacts postoperative outcomes, many computer-assisted approaches exist. Conventional examples, such as voxel-based morphometry, use statistical methods to find areas of the brain that differ from normal controls.[11,12] With the increasing capability of artificial intelligence (AI) through machine learning (ML) and especially deep learning (DL), new approaches based on artificial neural networks (ANNs) have emerged. AI has dramatically influenced the field of automatic FCD detection in MRI-negative focal epilepsies, with recent works predominantly involving DL models.[13–18]

In this critical review, we look at current research on computational approaches for FCD detection, highlighting

Key Points

- FCD detection is one of the first clinical grounds where the application of artificial intelligence may translate into an advance for patients' health
- FCD datasets should be openly available and include detailed information about histopathological subtypes and neuroimaging characteristics
- Using appropriate metrics improves model evaluation and increases clinical impact
- Postprocessing strategies for model output aid patient-level diagnosis
- Publishing artificial intelligence models as "ready-to-use" software would support independent validation, ease of use, and adoption of these models by external groups

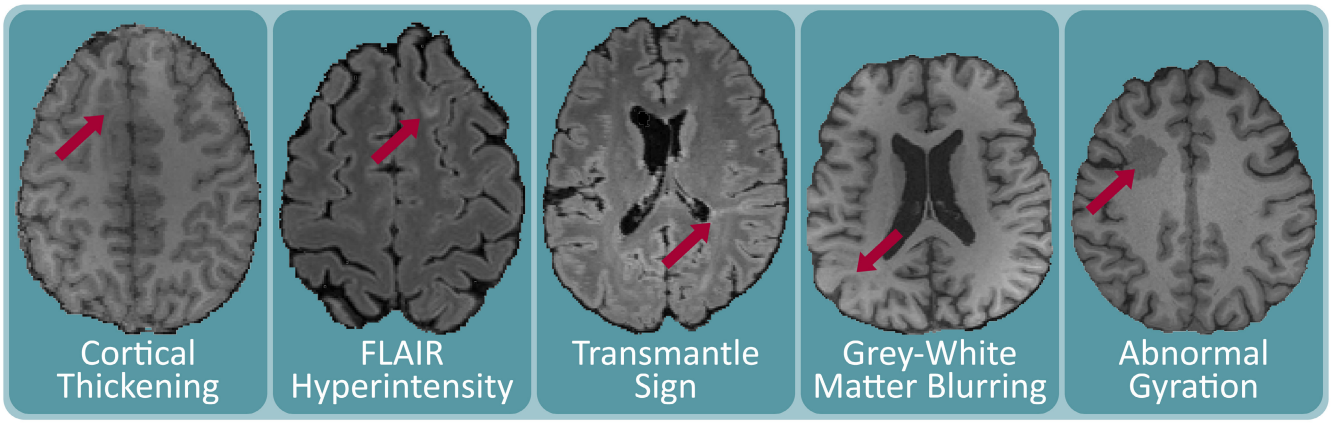

FIGURE 1 Typical features of focal cortical dysplasias (FCDs) visible in magnetic resonance imaging. The prevalence of these features depends on the histopathological type of FCD. FLAIR, fluid-attenuated inversion recovery.

advantages and successful examples and discussing what has prevented their widespread application in routine clinical practice. The latter requires a mutual understanding of the matter by clinicians, translational researchers, and computer scientists. We want to create a common ground by introducing clinical aspects of FCD detection and the basic concepts of AI before diving into specific challenges and solutions for when these different areas of expertise meet. Given the clinical relevance of FCD detection, we think this is one of the first fields where AI algorithms may translate successfully from "bit to bedside"[19] into a measurable advance for health care.

1.1 | AI for computer vision

Computer vision is one of the major applications of AI and laid the foundation for DL model development a decade ago. ML and DL are subfields of AI, and although frequently used interchangeably, these terms have different meanings.[20,21] ML generally refers to methods that *learn* from data, involve statistical models, and depend on human intervention. DL specifically applies models in the form of *deep* ANNs. An ANN is a network of small connected computational units inspired by neurons and axons in the brain. These artificial neurons are commonly arranged in layers with connections from lower to higher layers. For DL, ANN architectures have many (up to hundreds) of these layers, hence the adjective *deep*. ANN-based models for computer vision take data in the form of images as input and produce an output, commonly referred to as a prediction.

In FCD detection, the prediction typically takes a value between 0 and 1. These values can be interpreted as probabilities, for example, a prediction value of .7 means the model predicts the input data to be 70% "lesional." These probabilities can be thresholded to arrive at binary labels: 0 for "nonlesional" and 1 for "lesional." In a later section, we explain different approaches for arriving at such predictions in more detail. The *learning*, or *training*, happens by iteratively letting a model predict input data, comparing the predicted to the *real* value, and changing the computations inside the model (i.e., the model parameters). The *real* value is commonly referred to as *ground truth*, which for the above example, would be whether the image contains a lesion. For FCD detection in general, the ground truth is often derived in the form of *lesion masks*, with clinicians manually demarking the location and extent of the FCD.

The performance of a model is usually only evaluated based on how often the prediction is correct. Analyzing exactly how a model arrives at a particular prediction and what it has learned is nontrivial and is actively researched under the term *explainable AI*. Common terms in the field of FCD detection are summarized in Table 1.

2 | CLINICAL CHALLENGES IN FCD DETECTION

This section summarizes challenges for clinicians when diagnosing FCDs, which center around collecting and assessing patient data. In terms of how AI can help in the diagnostic process, we show how each clinical aspect potentially affects model development and evaluation. We specifically want to draw attention to the two most challenging scenarios for clinicians and for which AI could be most beneficial. First, given a person deemed "MRI-negative" but where clinical evaluation is indicative of an FCD, can an AI still pinpoint the FCD on MRI? Second, given a potentially abnormal region on the MRI, can an AI evaluate whether it is pathologic? We will refer to the first scenario as "hypothesis formation" and the second as "hypothesis refinement."

TABLE 1 Common terminology used in FCD detection research.

Term	Meaning
ANN	Artificial neural network
DL	Deep learning
ML	Machine learning
Prediction	The output of an ML algorithm
Ground truth	True labels for input data (e.g., lesion mask)
Voxel	The smallest part of a three-dimensional image
Vertex	A point on a surface
Patch	Part of a larger image
Lesion	A type of structural abnormality
FCD	An epilepsy-causing malformation of cortical development
T1, FLAIR	Types of imaging sequences resulting in different highlighting of various tissue types (most importantly gray and white matter)
MRI-negative	Lesion could not be found on MRI
FreeSurfer/ FastSurfer	A software tool to reconstruct the cortical surface from brain images[22,23]
MAP18	Software for voxel-based FCD detection using morphometric maps as input for a shallow ANN[24,25]
MELD	Software for vertex-based FCD detection based on using surface features generated with FreeSurfer and an ANN[18]
deepFCD	Software for patch-based FCD detection that takes T1 and FLAIR images as input[16]

Abbreviations: FCD, focal cortical dysplasia; FLAIR, fluid-attenuated inversion recovery; MAP18, Morphometric Analysis Program v2018; MELD, Multi-centre Epilepsy Lesion Detection; MRI, magnetic resonance imaging.

2.1 | Data and ground truth

The data necessary for FCD detection consist of input data and ground truth. Input data, usually MRI data but potentially other imaging or clinical information, can vary significantly across studies. MRI data can stem from different scanners with varying field strengths and may consist of different imaging modalities, for example, T1 and FLAIR sequences. The choice of MRI sequences can be significant, as Demerath et al.[26] show, where specific T1 sequences, for example, affect downstream processing and visual assessment. For human readers, FLAIR sequences are most helpful in diagnosing FCDs.[27]

In most studies, FCDs are labeled manually in all three dimensions on a voxelwise level using a binary mask as a *lesion mask*. The ground truth of these lesion masks has different levels of "certainty." The gold standard for validation of an FCD case is histopathological confirmation, which is only possible after surgery. However, only about half of the patients with FCD will receive surgical treatment. The patients considered MRI-negative, one third overall, present a particular challenge for automated imaging analysis but the most critical group regarding the potential clinical impact of automatic FCD detection. As stated in the introduction, "MRI-negative" loosely means medical experts could not pinpoint the lesion on routine assessment, but the term is underdefined. Some studies define it as the lesion being overlooked at least once in regular radiological assessment.[18] In contrast, in others, it means that even after reevaluation with additional clinical information, a lesion is deemed invisible in MRI.[28] It remains unclear how other modalities, such as electroencephalography (EEG) and positron emission tomography (PET), often known to the clinician, impact the MRI diagnosis.

2.1.1 | On "MRI-negative"

For a finding to be truly MRI-negative has to imply that the MRI contains no information that helps pinpoint the lesion. In this case, the MRI must be deemed entirely unhelpful for FCD detection. As stated at the beginning of this section, current hopes for the help of AI in cases where the visual (human) assessment of MRI is unsuccessful are twofold. On the one hand, there are cases where information for lesion localization is still contained in the image and that algorithms can leverage (hypothesis formation). On the other hand, there are scenarios where MRI diagnosis is only possible in combination with additional nonimaging information, for example, MRI features that are not lesion-specific but that combined with EEG, semiology, and other imaging modalities add to the evidence about lesion location (hypothesis refinement). AI that uses additional imaging modalities and other clinical information must be developed for these cases.

The ratio of histopathologically confirmed seizure-free and MRI-negative patients varies significantly across FCD-related research. For example, the percentage of patients labeled "MRI-negative," regardless of the exact definition, varies between 0% and 100%, as shown in Table 2. This variability makes model comparability difficult. Most datasets are not publicly available, so most works evaluate their model on single-center data. Encouragingly, the number of multicenter studies is increasing.

Many works only use T1 sequences[24,46] or add FLAIR sequences.[16,34] FLAIR sequences can benefit FCD detection because FCDs may exhibit local hyperintensities,

TABLE 2 Ratios of MRI-negative patients across FCD-related studies, if reported and regardless of the exact definition of "MRI-negative."

	Total	MRI-positive	MRI-negative	Ratio of MRI-negative
Adler et al.[13]	**27**	**27**	**0**	**0%**
Ahmed et al.[29]	**31**	**7**	**24**	**77%**
Alaverdyan et al.[30]	**21**	**3**	**18**	**86%**
Chen et al.[31]	39	15	24	61%
Chen et al.[32]	16	8	8	50%
Colombo et al.[33]	118	93	25	21%
David et al.[24]	**173**	**173**	**0**	**0%**
Ganji et al.[34]	**30**	**20**	**10**	**33%**
Gill et al.[16]	**171**	**79**	**92**	**54%**
Hong et al.[35]	**45 3 T|36 1.5 T**	**26|22**	**19|14**	**42%|39%**
Kim et al.[36]	48	30	18	38%
Kral et al.[37]	53	51	2	4%
Lee et al.[38]	21 IIa|25 IIb	5|9	16|16	76%|64%
Martin et al.[12]	22	15	7	32%
Mellerio et al.[39]	25	17	8	32%
Mo et al.[40]	**18 IIa|22 IIb**	**40**	**0**	**0%**
Radhakrishnan et al.[41]	78	60	18	23%
Seong et al.[42]	81	34	47	58%
Spitzer et al.[18]	**538**	**360**	**178**	**33%**
Wagner et al.[43]	17 IIa|74 IIb	11|67	6|7	35%|9%
Wang et al.[44]	43	0	43	100%
Wang et al.[45]	150	80|0[a]	70|150[a]	47%|100%[a]

Note: Automatic FCD detection approaches are highlighted in bold. Some studies differentiate between scanner field strength (1.5 or 3 T) or histopathological types (types IIa and IIb).Abbreviations: FCD, focal cortical dysplasia; MRI, magnetic resonance imaging.

[a]Using two definitions of MRI-negative (after reevaluation with additional clinical information vs. "overlooked once").

which is the case for 50%–70% of FCD type II cases. However, they may be less helpful for the harder-to-detect type I FCDs, where <20% are associated with a FLAIR hyperintensity.[47] For the human eye, hyperintensities are better recognizable than changes in cortical thickness or blurring of the gray–white boundary. Whether the choice of input sequences similarly impacts model performance remains an open question.

2.2 | Histopathological considerations on ground truth

Histopathological assessment categorizes FCDs into several types with several possible features visible in the MRI, as Figure 1 shows. In FCD type I, MRI may show mild hyperintensity of the white matter in T2/FLAIR images or cortical atrophy, albeit most patients have no visible MRI abnormalities.[48] In FCD type IIa, MRI findings include abnormal focal patterns of sulci and gyri, cortical dimple, mild cortical thickening, blurring of the gray–white matter interface, and at times a mild hyperintense T2/FLAIR signal in the subcortical and deep white matter.[49] MRI changes of FCD type IIb, or FCDs with balloon cells, often show deep sulci, abnormal cortical gyration, and hyperintense T2/FLAIR signal in the subcortical white matter, sometimes with a wedge shape that extends to the ventricle ependymal surface defined as transmantle sign.[50] FCD type III lesions share the typical features of the associated lesions.[49] Mild malformations of cortical development (mMCDs) and mMCDs with oligodendroglial hyperplasia (MOGHEs) may exhibit blurring of the gray–white matter interface and FLAIR hyperintensities in children.[51,52] In adults, MRI findings in MOGHE include changes in gyri and sulci morphology with mild or no cortical/subcortical hyperintense FLAIR signal, blurring of gray–white matter interface, and cortical thickening.[53]

Inter- and intraobserver agreement in evaluating the International League Against Epilepsy classification of FCDs strongly depends on the FCD type.[54] Particularly

FCDs of type I represent a considerable challenge for correct diagnosis, which generally requires neuroanatomically excellent brain tissue representation. Although the histopathological diagnosis of type II FCDs is straightforward due to the robust cellular characteristics of dysmorphic neurons and balloon cells, there may be sampling problems and misclassification of FCD type IIb as type IIa. Although resectioning the MRI-documented cortical abnormality in FCD IIb is crucial for a favorable postoperative outcome,[28] these tissue portions often contain mainly dysmorphic neurons. In contrast, balloon cells are primarily present in the white matter adjacent to the abnormal cortex.[55] Because resection of the subcortical hyperintense zone in FCD type IIb is not essential for seizure freedom,[28] neuropathological assessments may miss balloon cells. New molecular–genetic and epigenetic characterizations of FCD biopsies may improve diagnostic accuracy[56] and be integrated into a multilayered diagnostic scheme yielding more fine-grained classifications.[50] Forthcoming advances and the limitations of neuropathological categorization[57] need consideration for image analyses, especially when describing a dataset and defining a ground truth. Like different imaging sequences, the proportion of various FCD types and imaging features within a dataset can impact model evaluation.

2.3 | Neurosurgical considerations on ground truth

Preoperative lesion masks are important for guiding FCD resection. Unlike tumors, FCDs are usually not visible macroscopically in situ, but surgeons may assess lesion extent intraoperatively using intraoperative MRI and, more recently, intraoperative ultrasonography.[58,59] The success of surgery depends on exact intraoperative neuronavigation and complete resection of the FCD.[60,61] There are different definitions of "complete resection." They include pathological proof of "clear margin,"[62] a combination of preoperative evaluation and intraoperative electrocorticography,[63] or the removal of the "abnormal cortex" based on EEG analyses[64] or of all visible abnormalities on MRI.[65] FCD detection approaches yielding accurate lesion segmentation may inform surgeons. Nonetheless, epilepsy surgery has to balance a "complete resection" against the risk of collateral damage. Particularly in the vicinity of eloquent areas—any brain area involved in language, somatic, memory, or sensory processing[66]—epilepsy surgery yields lower seizure freedom rates, likely reflecting conservative resection.[5] To not risk resecting more tissue than necessary,[67] procedures such as MRI-guided stereotactic laser interstitial thermal therapy can be more suitable for patients with discrete lesions.[68] It allows for a nearly "voxel-level" ablation of the lesion, as indicated by the lesion mask. However, the common occurrence of FCDs in the depth of sulci can make them difficult to target.[68]

3 | WHAT IS AUTOMATIC FCD "DETECTION"?

The task of finding FCDs is commonly called "FCD detection." It is noteworthy that the word "detection" originates from the clinical context of finding (i.e., localizing) an FCD and differs from the technical computer vision term "object detection." This section will detail different techniques for "automatic FCD detection." A processing pipeline typically consists of three parts: preprocessing, the AI algorithm, and postprocessing. Pre- and postprocessing steps will be detailed later; first, we want to focus on the different AI models applied.

Figure 2 shows example predictions for two subjects with FCD generated by three recent works.[16,18,24] Although each produces an output for the whole brain, the predictions differ, for example, in the number of predicted voxels, their probability values and distribution, and the results for various evaluation metrics. Generally, models for image processing fall into one of the three major categories of computer vision: classification, object detection, and segmentation. Figure 3 shows an example output for each task.

3.1 | Classification

A classification model predicts what the input data are. Its output is a value between 0 and 1, akin to the probability for the input data belonging to a particular class, as shown in Figure 3A,B. For FCD detection, the possible classes are usually either "lesional" or "nonlesional." Most approaches formulate the problem of FCD detection as a classification task.[13,16,24,34,35,40,46,69,70,71–74] Input data range from raw MRI data to morphometric maps or surface features. They can be one-dimensional, that is, single voxels (or vertices if the input data are surface-based), or two- or three-dimensional images. In the two- and three-dimensional cases, an algorithm often only operates on smaller image parts, so-called "patches." The output is thus generated on the voxel level,[24,46,71–73] vertex level,[13,18,35,40,70] or patch level.[16,34,69,74] Figure 2 shows examples of the outputs from Multi-centre Epilepsy Lesion Detection (MELD),[18] Morphometric Analysis Program v2018 (MAP18),[24] and deepFCD[16] models.

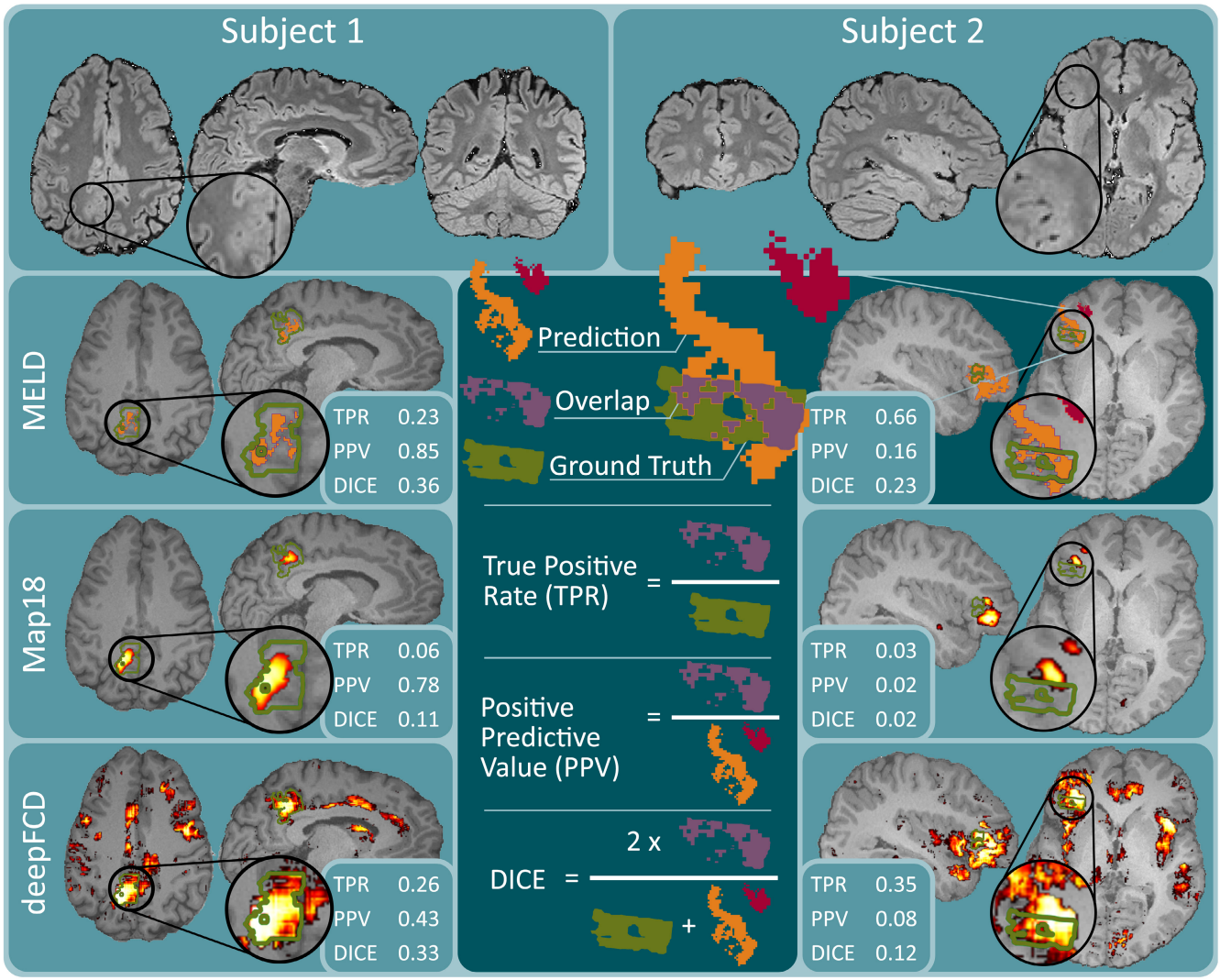

FIGURE 2 Predictions for two example patients with histopathologically confirmed focal cortical dysplasia (FCD) type IIa from three recent works on automatic FCD detection—Morphometric Analysis Program v2018 (MAP18),[24] deepFCD,[16] and Multi-centre Epilepsy Lesion Detection (MELD)[18]—and visualization of suggested evaluation metrics. Two medical experts determined the ground truth lesion mask (green outline) independently using all available imaging sequences and pre- and postoperative information, including histopathological confirmation of FCD. MAP18 and deepFCD pipelines output voxelwise probabilities, and MELD outputs clusters (i.e., groupings of voxels) with an additional report about cluster features (not displayed). Determining whether an algorithm has actually "found" a lesion is not trivial. The current widely used criterion of counting a single voxel overlap as a successful localization leads to all models finding both examples. However, the respective predictions differ significantly, for example, in the number of predicted voxels, resulting in very different values for evaluation metrics such as the true positive rate (TPR), positive predictive value (PPV), and Dice scores.

3.1.1 | Anomaly detection

Anomaly or outlier detection is, despite the name, a case of binary, that is, two-class, classification, which involves comparing test data with an "assumption of what is 'normal.'" This assumption can, for example, stem from a database of healthy subjects.[25] More recent approaches learn from healthy subjects through DL architectures to estimate healthy anatomical variability and identify lesions as outliers from this distribution.[30]

3.2 | Object detection

An object detection model decides what and where an object (i.e., an FCD lesion) is in the input image. The output consists of coarse localization information in the form of a center point and a bounding box in addition to classification information. Figure 3C,D shows two example outputs for single or multiple object detection. DL approaches for object detection, such as the YOLO architecture,[75] have been successfully applied for other areas of medical image processing, such as skin cancer, breast cancer, or brain

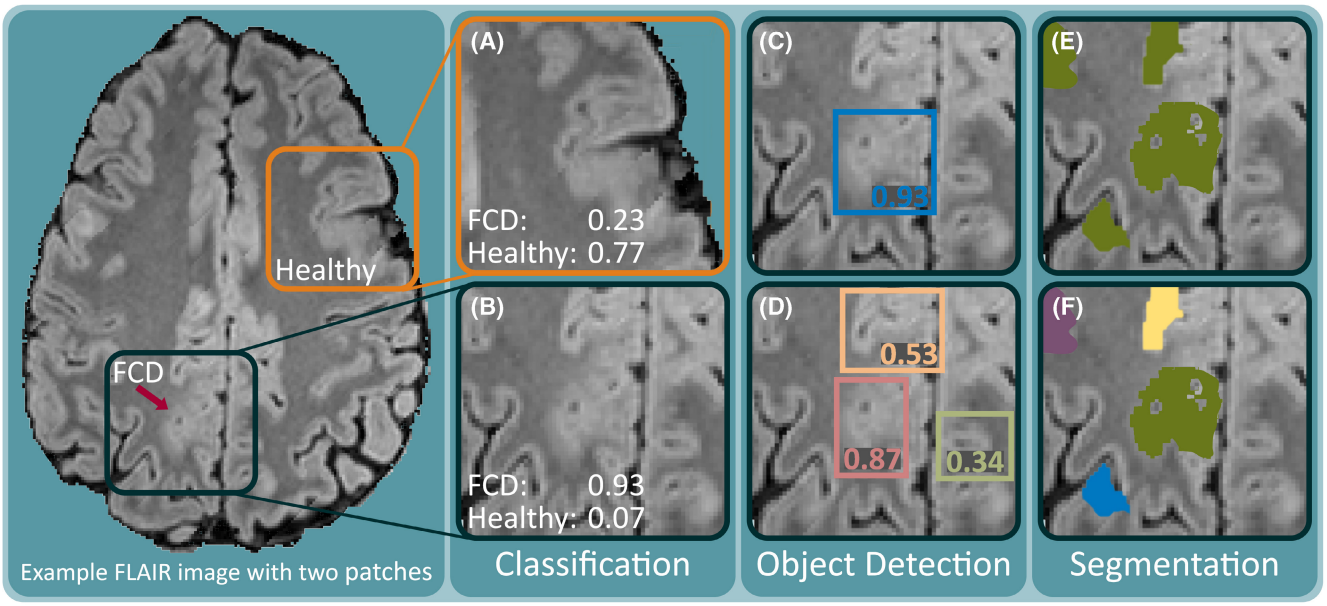

FIGURE 3 Examples of the three theoretical tasks in computer vision: classification, object detection, and segmentation. In medical image analysis, a classification model predicts class memberships for single voxels (or vertices if based on surface data) or for image patches, that is, parts of the whole image (A, B), and outputs a probability score between 0 and 1 (here for the classes "FCD" and "Healthy"). Object detection models define a bounding box around one (C) or multiple (D) objects with accompanying class probabilities. (E) "Semantic" segmentation models predict voxel-level class membership without the notion of objects (i.e., clusters). (F) "Instance" segmentation groups voxelwise class labels into different objects. FCD, focal cortical dysplasia; FLAIR, fluid-attenuated inversion recovery.

tumor segmentation.[76–79] However, such approaches have not been used for FCD detection.

3.3 | Segmentation

A segmentation model also predicts what the input image is, but in contrast to classification, segmentation produces a class label for every input voxel. The output thus has the same dimension as the input data, whereas for classification, the outcome is a single value for the whole input. Additionally to this so-called *semantic* segmentation exists *instance* segmentation, where the voxels are further grouped (clustered) into objects (i.e., clusters). Figure 3E,F gives one example for both. House et al.[80] and Thomas et al.,[81] have published approaches for segmentation based on the successful UNet model[82]—a type of model that is widely applied in medical imaging. Segmentation might be most important for planning surgery, where one goal is to identify and remove as much lesional cortex as accurately as possible.

3.4 | Comparison

Many works do not uniquely belong to one of these three categories and involve a mix of other processing steps. For example, several approaches use morphometric maps and other differences compared to a "normal" cohort as inputs to a classification model.[13,24,71,73] Colliot et al.[83] explore segmentation with coarse localization information as additional input, which could be helpful for hypothesis refinement.

Each task has certain aspects, pros, and cons, which we briefly highlight and put into perspective. Table 3 shows a condensed overview. The ground truth for classification tasks is usually the easiest to obtain in the form of a single label indicating "lesional" or "nonlesional." For voxel-, vertex-, or patch-based approaches, the number of training samples is much larger than the number of lesions. However, these approaches have a limited field of view and can thus only take part of the brain as context. Other approaches can also take whole-brain images as input, primarily for two-dimensional or 2.5-dimensional data. The latter means concatenating slices from coronal, sagittal, and axial slices around a given coordinate. Three-dimensional whole-brain processing is complicated because of its computational cost.

In clinical practice, localization information is usually not communicated by bounding boxes or segmentation maps, so experts must manually label magnetic resonance images to obtain the ground truth. Especially for segmentation (but also for voxel- and vertex-level classification), this is a very time-consuming task. Object detection offers a tradeoff between classification and localization information, requiring less accurate manual labeling. However,

TABLE 3 Description of the three tasks in computer vision, the resulting outputs, and the advantages and disadvantages when developing and applying models.

Task	Result	Pros	Cons
Classification	Single output with probabilities for class membership	Large number of training examples when using voxels or vertices Models generally require fewer parameters than other approaches Anomaly detection Reduces the need for manual labeling	Large amount of work creating manual segmentation labels required for training False positives due to limited fields of view/ lack of whole-brain context Anomaly detection Potentially detects abnormalities not specific to a pathology
Object detection	Probabilities for class membership, bounding box	Spatial information Fast labeling via bounding boxes	Requires annotated data in the form of bounding boxes Detected boxes do not reflect the structure/ shape of the cerebral cortex
Segmentation	Voxel-level probabilities for class membership	Spatial context-aware, lesion-specific	Requires detailed (voxel-level) labeling

lesions might not fit well into rectangular bounding boxes due to the folded morphology of the cerebral cortex. Segmentation approaches provide the most lesion-specific output and naturally incorporate spatial context. Anomaly detection tasks require the least manual labeling, but anomalies alone might not be pathology-specific.

3.5 | "FCD detection" processing pipelines

The whole processing pipeline typically involves additional steps besides the AI model. Preprocessing transforms raw (MRI) data into model input. For voxel-, vertex-, or patch-level processing, input data are typically sampled from the whole brain so that predictions are generated everywhere. However, one can imagine pipelines requiring clinicians to select specific areas they want to be predicted, which has yet to be explored for FCD detection. Postprocessing further alters the raw model prediction to yield the final output. Figures 4 and 5 show examples of components involved in both steps, respectively.

3.6 | Patient-level diagnosis

A patient-level diagnosis is the ultimate goal of clinical FCD detection and, thus, a specific purpose of automated approaches. The question is how low-level predictions, for example, some voxels classified to be lesional, relate to a decision on the patient level, that is, whether someone has FCD. Current works often accept even a single voxel- or vertex-level true positive as a true positive on the patient level, that is, a person with FCD. Equivalently, for control cases, even a single false positive prediction on the voxel level counts as a false positive diagnosis. It remains to be seen how lower level (voxel-, vertex-, or lesion-level) or patient-level predictions impact a clinician's decision for hypothesis formation and refinement.[16] Gill and colleagues, for example, state a patient-level detection rate of 93% "with an average of 6 false positive [cluster]s per patient."[16] However, the most highly ranked cluster coincides with the ground truth in only 36% of cases, leaving open whether a clinician would also correctly diagnose 93% of patients with this approach.

4 | CHALLENGES IN CURRENT FCD DETECTION RESEARCH

We have so far introduced essential parts of FCD detection pipelines and shown examples of how they may look. In the following sections, we will highlight crucial areas for developing algorithms for automatic FCD detection. For the sake of brevity, we will refer to any processing pipeline as a "model." An object detection model might recursively contain a model for classification and clustering postprocessing steps.

4.1 | Model training

4.1.1 | Splitting data

The data are split into three sets for training and testing a model: the training, validation, and test set. The training happens on the training set, and performance is periodically measured on the validation set. Training can be stopped when performance on the validation data saturates. This *early stopping* can protect the model from

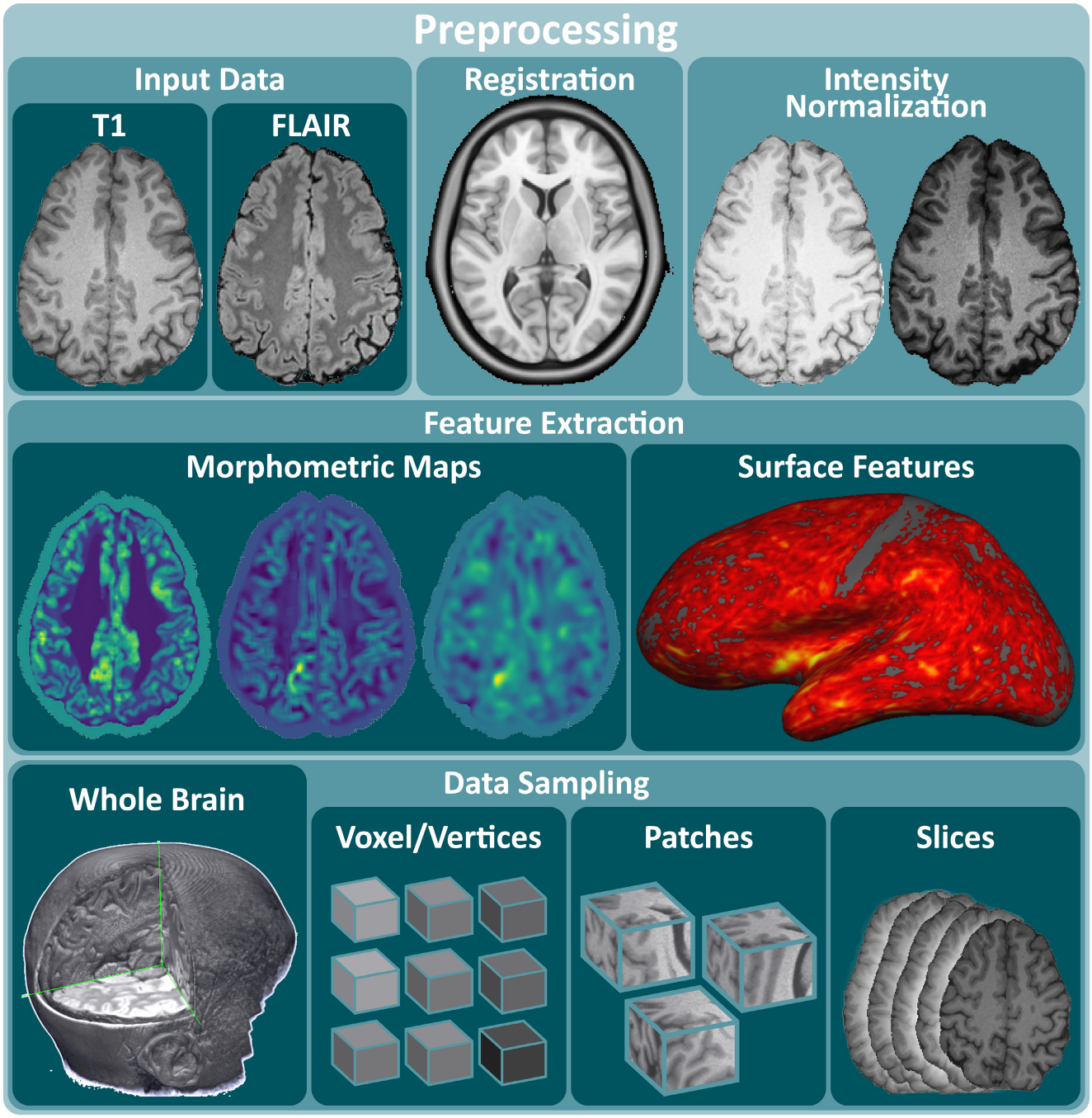

FIGURE 4 Common preprocessing steps. Typically, preprocessing involves standardizing input data through intensity normalization, resampling, or registration to a template. For some algorithms, additional input needs to be calculated, such as volumetric[24] or surface-based[18] feature maps. Finally, voxels, patches, or slices can be extracted from the whole brain volume to generate the artificial intelligence model input. FLAIR, fluid-attenuated inversion recovery.

overfitting the training data and failing to generalize to new data. The performance on the test set is only evaluated after all training has finished; the model never "sees" the test split data during training to avoid learning something from it, an error commonly called *data leakage*.[84] Because of the often-limited number of FCD subjects, approaches might include cross-validation without a separate test set, a practical option, although technically insufficient for model evaluation. Recent works such as Gill et al.[16] and Spitzer et al.,[18] do cross-validation and keep separate test sets aside.

4.1.2 | Hyperparameters

All processing steps involve design choices, that is, so-called hyperparameters. For ANNs, typical hyperparameters relate

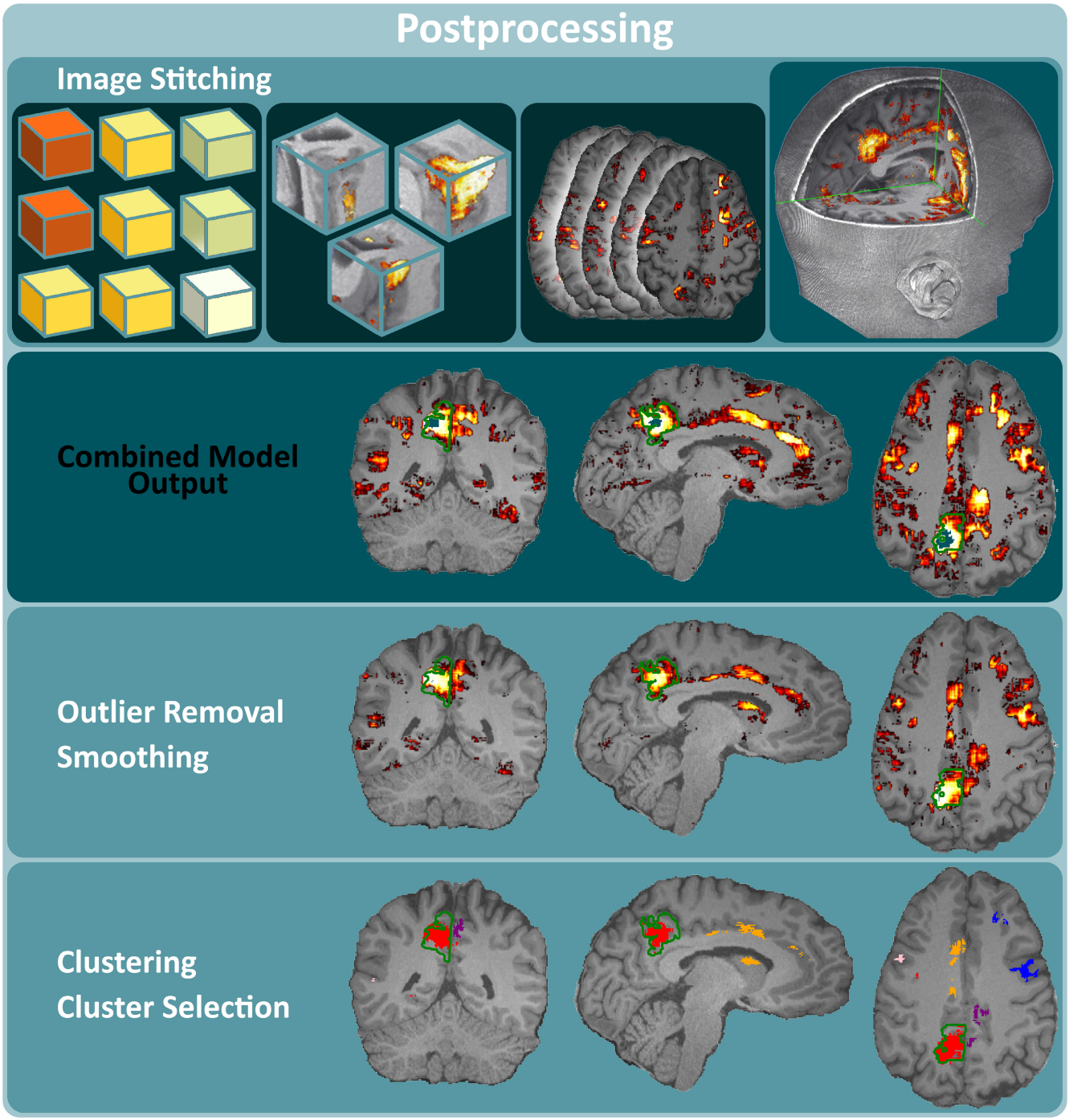

FIGURE 5 Common postprocessing steps. Often, a whole-brain prediction is generated from the model output; for patch-based approaches, this process is called *stitching*. Generally, approaches can smooth the prediction and apply outlier removal by utilizing spatial relationships between classified voxels or patches. Especially for voxel and vertex level classification, pipelines often involve an additional clustering step[13,18,35] to arrive at an object (i.e., lesion)-level diagnosis. Clustering also allows comparing clusters, for example, ranking them and removing them based on their size, mean probability, or rank. Cluster removal can help reduce the number of false positive clusters.

to the network architecture, learning rate, and loss function. Parameter choices for the postprocessing step, such as the probability thresholds, or those for a particular clustering and cluster selection method, are also essential hyperparameters. These are, however, rarely discussed or carefully evaluated and may introduce additional information leakage.[84,85]

4.2 | Model evaluation

4.2.1 | Evaluation metrics

One assesses a model's performance with different evaluation metrics depending on the goal (classification, object

detection, or segmentation). Sensitivity, specificity, and derived scores such as likelihood ratios are typical choices for a classification task. The threshold for translating predictions to class labels is critical and can significantly alter a model's performance on downstream tasks, such as patient-level diagnosis. The same metrics apply to object detection after matching objects and ground-truth clusters. A match is typically accepted if the overlap, measured, for example, by the intersection over union or the Dice coefficient, lies above a threshold. These metrics also help assess segmentation quality. They allow for model evaluation on the level of the model output, that is, voxel, vertex, or patch level. Figure 2 shows examples of three potential evaluation metrics: true positive rate, positive predictive value (PPV), and Dice.

Additionally, the voxel-level output may be postprocessed to allow a lesion-level, that is, cluster-level, or even patient-level evaluation, as is often desired in a clinical context. Some methods involve clustering and ranking clusters according to their size, connectivity, or mean probability and exclude clusters below a given size or rank.[16] As mentioned, cluster- and case-level analyses involve a cluster-matching step. Therefore, a threshold for the overlap to be accepted as a "match" must be chosen. Currently, in most works, the criterion is "one voxel overlap," that is, Dice > 0. Because human performance has yet to be quantified in terms of Dice scores for FCD detection, it is unclear what values would be acceptable. Works that assess the clinical impact do not quantify overlap or the number of false positive clusters.[15,73,74] However, current approaches often mix voxel-, patch-, cluster-, and patient-level evaluation metrics and fall short of showing the actual performance and impact of a claimed methodological novelty.

4.3 | Reproducible science

The confidentiality of medical data presents a significant hurdle in reproducible science. Few works[13,16] have made used data accessible, allowing others to partially reproduce the analysis and compare models with the same input data. Another more straightforward way to facilitate model comparison is to share code and trained models and enable others to apply the model to their data. Two recent works share their code online.[16,18] Developing ready-to-use code, however, is a labor-intensive task. The application of said models also requires high-level technical expertise, for example, programming skills, for structuring data and generating predictions. Improving the ease of use of automatic FCD detection algorithms remains an open challenge.

4.4 | The current best approach is: Humans?

Part of the reason existing approaches have yet to find their way into everyday clinical practice is the difficulty of assessing how they compare to medical experts and how they impact clinical decision-making. A few works exist that examine the impact of MAP18 software[24,25] and qualitatively assess the added benefit of a model's output for medical experts.[73,74] However, their evaluation depends on a human expert and lacks a quantification of the algorithm's performance. Solely comparing model performance, for example, how often the prediction overlaps the ground truth, to human performance on MRI-negative patients has severe shortcomings beyond there being no clear definition of the term "MRI-negative." The pitfalls can best be seen in (but are not exclusive to) the methodologically interesting work of Alaverdyan et al.,[30] where, regarding the MRI-negative patient group, they state that "[...] human performance is at 0%."[30] However, one must be very careful; the output of a human expert's assessment is usually a single or very few areas classified as lesions resulting in one conclusive diagnosis, but models can generate a much higher number of cluster predictions with varying probabilities. For a fair comparison to expert performance, one should thus restrict a model's output to only a few clusters per patient. The highest reported result when outputting a single cluster is a sensitivity of approximately 36%,[16] which lies far below regular performances from medical experts in clinical settings. However, outputting a single cluster may not be the ultimate goal of automated FCD detection, especially for hypothesis formation.

Regarding false positives, experts usually know about evidence from other (nonimaging) modalities such as semiology, EEG, or PET. From our experience, it is rare for medical experts to find false positive lesions in healthy controls. The comparison to healthy control subjects is also biased, as this scenario hardly arises in a clinical setting. Recent studies hint at nonnegligible rates of incident findings of brain abnormalities (not specific to FCDs),[86,87] but no work has quantified expert FCD detection performance on healthy controls or how likely medical experts are to rate multiple brain regions as lesional.

5 | TRANSLATING MODELS INTO CLINICAL PRACTICE

Having highlighted the shortcomings of current FCD detection models and the challenges when comparing approaches, in this section, we want to provide suggestions on how to alleviate these problems and how future works

can best explore methodological advancements and clinical benefits.

5.1 | Collecting and exploring datasets

A simple step to assess the complexity and quality of a dataset and put a model's performance into context is to report results for histopathologically confirmed and MRI-negative patient groups separately from the overall results. Listing the prevalence of MRI features and the different histopathological types across a given dataset would further help estimate its complexity.[24] Because these can still be subjective, other more quantitative metrics could also help. Although it is common to provide clinical information such as age and sex across subjects, metrics such as average lesion size[24,70,80] or the distribution of lesion locations[5,16] should also be reported. These could provide a different categorization than seizure outcome or histopathological type, which might need to be revised considering the previously mentioned challenges for neurosurgery and histopathology. Additionally, a detailed description of MRI features ties into the multilayered classification scheme, specifically Layer 3, proposed by Najm et al.[50]

There are efforts, such as the MELD project (https://meldproject.github.io/), to collect multicenter data into larger datasets and make them available for algorithmic development. Nonetheless, a publicly available volumetric MRI dataset for FCDs has yet to be created. In the future, a public well-annotated benchmark dataset would provide a common ground for evaluating models. Diverse data are vital to ensure algorithms are widely applicable. In children, for example, the developing brain undergoes structural changes, including overall volume increases, gray–white matter contrast changes associated with white matter myelination, and dynamic changes in cortical thickness. These could all impact an algorithm's ability to detect FCD lesions. Therefore, FCD detection algorithms must be trained and evaluated on cohorts containing pediatric patients and controls. Although some approaches have included pediatric cases,[18] specific further development is required to ensure robustness to these development changes.

5.2 | Common evaluation metrics

Table 4 provides an overview of suggested metrics for evaluating an FCD detection algorithm. These metrics

TABLE 4 Overview of metrics and curves for evaluating an FCD detection algorithm.

Metric/curve	Voxel/vertex level	Lesion level	Patient level
PPV (precision)	How many predicted voxels are true positives "20% of predicted voxels lie inside the lesion"	How many predicted clusters overlap with lesions "Half the predicted clusters overlap the real lesions"	How many predicted subjects have FCD "80% of the predicted subjects have FCD"
TPR (sensitivity, recall)	How much of a lesion was found "On average, the model finds 20% of all lesional voxels"	How many lesions were found "The model finds 80% of all lesions"	How many patients were found "The model identified all FCD patients"
Precision–recall curve	Helps determine thresholds for prediction probabilities	Helps determine thresholds for cluster probabilities, sizes, and ranking (same for both levels)	
ROC curve	Problematic because of class imbalance, use precision–recall curve instead	Helps determine thresholds for clustering criteria and cluster-level thresholds (same for both levels)	
AUC	Summarizes information of evaluation curves into a single value (same for all levels)		
FROC curve	Not applicable	Links lesion-level evaluation to patient-level diagnosis "Our model finds 80% of all lesions with an average of three false positive clusters per case"	
Dice similarity coefficient	For evaluating segmentation quality	Not applicable	Not applicable
Hausdorff distance	For evaluating segmentation quality, especially regarding boundaries	Not applicable	Not applicable
Volumetric similarity	Not applicable	For evaluating segmentation quality	Not applicable

Note: These metrics can and should be applied separately to different levels of output: voxel, lesion, or patient level.

Abbreviations: AUC, area under the curve; FCD, focal cortical dysplasia; FROC, free-response ROC; PPV, positive predictive value; ROC, receiver operating characteristic; TPR, true positive rate.

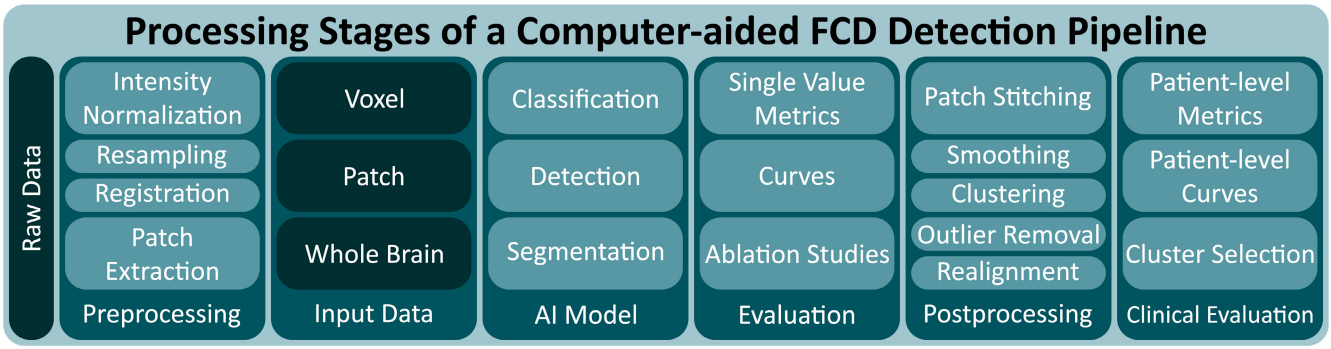

FIGURE 6 Summary of common steps in the processing pipeline of automatic focal cortical dysplasia (FCD) detection approaches. AI, artificial intelligence.

can and should be applied separately to different levels of output: voxel, lesion, or patient level. It remains an open research question how the intrinsic model performance, that is, the evaluation of the raw model output, relates to clinical usefulness and which model outputs best aid visual inspection of MRI data. For example, a single predicted "lesional" voxel might be as helpful for hypothesis formation as a better segmented lesion but with many more false positive predictions. The latter might benefit hypothesis refinement when lesion location is certain because of other clinical characteristics. We propose using voxel-level metrics, such as the PPV, as additional measures to estimate clinical usefulness. Figure 2 shows a visual example of PPV.

The postprocessing stage is especially relevant for translation into clinical practice. Typical steps include thresholding, clustering, voxel- and cluster-level outlier removal, and cluster selection. These clustering steps must be consistent across the dataset; for example, they must not differ for patients and healthy controls. Furthermore, approaches should optimize hyperparameters for this stage separately on a subset of the data (i.e., the training dataset) to ensure that the model is not overfitting and achieves similar performance on new data.

Ablation studies are a standard tool to show the benefit of a claimed novelty. For example, Thomas et al.[81] changed critical features of their proposed novel network architecture to showcase its advantage (while testing against baseline models). Other published works, however, lack this type of analysis. Suppose a claimed novelty lies in the input data type; one way to do an ablation study is to leave the specific input from training and testing by masking specific input channels, alleviating the need to change the network architecture. One can thus analyze the impact of novel feature maps and markers such as morphometric maps or different imaging modalities such as FLAIR.

Another way to evaluate performance is to examine metrics over varying conditions, such as probability thresholds or cluster-matching criteria. The receiver operating characteristic (ROC) curve alongside the area under curve score is one example of such an analysis. Because of the high imbalance between lesional and nonlesional tissue for FCDs, the ROC curve only gives meaningful results on the lesion or patient level. Another such curve would be the free-response ROC (FROC) curve.[88,89] It combines case- and lesion-level metrics, namely the average number of false positive clusters per case plotted against a lesion-level sensitivity.[90] For example, apply FROC to mammographic lesions.

Developing and evaluating a pipeline for FCD detection involves many design choices. Figure 6 gives a summary of the previously mentioned steps. Specifying these steps and carefully choosing adequate evaluation metrics are integral for correctly assessing the AI model and comparing it with other works. Evaluation metrics and acceptable thresholds could be applied differently for hypothesis formation or refinement. In summary, there needs to be more to model evaluation than solely evaluating performance on a downstream task, such as patient-level classification.

5.3 | Making pipelines accessible

The most recent large-scale works on FCD detection have set a strong positive example by making their code available and providing instructions about applying it to new data.[16,18] Publishing code marks a great start and can be further extended by following other steps to reproducibility in neuroscience, such as using the Brain Imaging Data Structure specification for a data structure.[91] Processing pipelines should include detailed pre- and postprocessing information to facilitate comparability further. Preferably, published code should incorporate these steps. Ideally, the final output should be in the same space as the input data, preferably a subject's structural space, to circumvent registration

problems and simplify visual comparison for doctors, as exemplified in Wagstyl et al.[92] Specifically for AI models, a step to improve performance on new data is fine-tuning. A model is fine-tuned by partially retraining it with new (center-specific) data. This is not trivial, however, and this functionality in published code could improve model performance when translating approaches into hospital environments.

To encourage accessibility, the scientific community should gratify the development of ready-to-use and open-source code like peer-reviewed publications. However, it also takes combined institutional efforts to meet requirements for technical equipment and expertise and the long-term maintenance of software. We thus want to emphasize that making computational approaches and AI specifically accessible to a wide array of research and clinical institutions cannot rely solely on researchers.

5.4 | Overtaking humans?

Lastly, we want to highlight future research directions to help close the gap between research approaches and clinical practice. Direct comparison to human performance implies the goal of overtaking human performance. However, we believe that algorithms will not be used as "standalone" MRI analyses and will not substitute for imaging experts. Instead, automated approaches parallel to the standard visual MRI analysis may enhance clinicians' confidence in evaluating FCDs, that is, hypothesis refinement, especially because the visual analysis can be time-consuming, and the level of expertise varies considerably among centers, which explains the variable, and sometimes high, percentage of MRI-negative patients. Although no dedicated work quantifies how a model best aids clinical decision-making, existing approaches have converged upon outputting several clusters. For example, El Tahry et al.[15] and Urbach et al.,[73] have shown that this may benefit hypothesis formation. Thus, existing algorithms may already aid neuroradiological assessment, and we strongly encourage integrating available approaches into clinical practice.

Research efforts should focus on difficult-to-detect FCDs and hypothesis formation to further examine the models' benefits. Although it is necessary for successful training to include as many data as possible, evaluation should differentiate these cases. The additional value for detecting FCDs in patients with FLAIR hyperintensity might be small compared to subjects without typical MRI features. The impact of a model is also more significant for cases of FCD type I than for those with MRI-visible type II. Currently, no models accept nonimaging information as input, so the potential impact for truly MRI-negative cases remains unknown. As soon as we have better quantified the benefit of FCD detection approaches on clinical grounds, they may be implemented as "decision support systems" and relieve radiologists from the additional workload caused by increased multimodal and high-resolution scans.

Ultimately, an algorithm should be evaluated prospectively based on its impact on the quality-adjusted life-years of people with epilepsy. Improved surgical outcome rates, more admissions to surgery, a faster diagnosis, or fewer other diagnostic procedures are other signs of a positive effect.

5.5 | Current limitations

A few factors currently hamper the widespread impact of automatic FCD detection. These include a lack of detailed model evaluation due to small dataset sizes, insufficient data descriptions, and the unmet need for expertise and infrastructure for integrating research works into clinical workflows. There is currently no research or guidelines on how algorithmic output can and should best be included in clinical decision-making and how it would impact hypothesis generation and refinement in practice. Advancements in disease classification, for example, the inclusion of mMCD and MOGHE, still need to be reflected in current models. There is also a significant percentage of cases where current algorithms do not detect a lesion. Recent works have already started to overcome these limitations by combining multicenter data and making models accessible. It remains to be seen how a streamlined model evaluation will further change model development and clinical impact. We deem the research into automated approaches for FCD detection too early to estimate a saturation point for how effective such models can become.

6 | CONCLUSIONS

Research in automatic FCD detection has made significant progress, and modern approaches combine advances in medical imaging with state-of-the-art methods from AI and DL. We have highlighted challenges in both fields and their combination, such as accumulating representative data, evaluating a model's impact, and making findings reproducible. We have suggested how researchers may design future studies to improve patient care. One strategy could be to provide a detailed data description, and another to correctly evaluate models for their intrinsic performance and potential impact on decision-making.

We encourage scientists in the field to combine efforts to share data and code, thereby promoting the field of FCD detection as a template model for successful translational clinical research.

AUTHOR CONTRIBUTIONS

Lennart Walger, Sophie Adler, Konrad Wagstyl, Leonie Henschel, Bastian David, Valeri Borger, Hartmut Vatter, Albert Becker, Fernando Cendes, Zhong Irene Wang, Hans-Jürgen Huppertz, and Theodor Rüber contributed to drafting the manuscript. Elke Hattingen, Christian E. Elger, and Martin Reuter contributed to the conception of this work. Torsten Baldeweg, Felix Rosenow, Horst Urbach, Alexander Radbruch, and Rainer Surges contributed to manuscript revision.

ACKNOWLEDGMENTS

S.A. was funded by the Rosetrees Trust (A2665). K.W. was supported by the Wellcome Trust (215901/Z/19/Z). T.B., S.A., and K.W. were supported by the National Institute for Health Research Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London. F.R. was funded by grants from the Detlev-Wrobel-Fonds for Epilepsy Research, the Deutsche Forschungsgemeinschaft, the Federal Ministry of Education and Research, the LOEWE Program of the State of Hessen, and the European Union. F.C. was supported by FAPESP (São Paulo Research Foundation) grant #2013/07559-3. Z.I.W. was funded by the National Institutes of Health (R01 NS109439). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

V.B. has received fees for serving as clinical consultant from Brainlab. These activities were not related to the content of this article. CEE has received fees from UCB, Desitin, Bial. H.U. has received fees from Eisai, Biogen, and MBits and is on the advisory board of Biogen. A.R. lectures for Guerbet and Bayer, and is on the advisory board for GE, Bracco, and Guerbet. F.R. reports personal fees from Angelini Pharma/Arvelle Therapeutics, Eisai, GW Pharmaceuticals/Jazz Pharma, and UCB Pharma. F.C. is an associate editor of *Epilepsia*, and has received honoraria for consulting and speaking from UCB Pharma, Takeda Pharmaceuticals, Abbott, Torrent, Prati-Donaduzzi, Eurofarma, and Zodiac Pharma. R.S. has received fees as a speaker or for serving on advisory boards from Angelini, Arvelle, Bial, Desitin, Eisai, Janssen-Cilag, LivaNova, Novartis, Precisis, UCB Pharma, UnEEG, and Zogenix. H.-J.H. is the author of the Morphometric Analysis Program v2018 (MAP18). The remaining authors have no conflicts of interest.

ORCID

Lennart Walger https://orcid.org/0000-0002-3300-6877
Konrad Wagstyl https://orcid.org/0000-0003-3439-5808
Bastian David https://orcid.org/0000-0002-0146-0629
Valeri Borger https://orcid.org/0000-0002-5905-4121
Torsten Baldeweg https://orcid.org/0000-0002-5724-1679
Horst Urbach https://orcid.org/0000-0001-7264-4807
Albert Becker https://orcid.org/0000-0003-2661-3705
Rainer Surges https://orcid.org/0000-0002-3177-8582
Fernando Cendes https://orcid.org/0000-0001-9336-9568
Zhong Irene Wang https://orcid.org/0000-0002-3829-5217
Theodor Rüber https://orcid.org/0000-0002-6180-7671

REFERENCES

1. 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. https://doi.org/10.1111/j.1528-1167.2010.02777.x
2. Sisodiya SM, Fauser S, Cross JH, Thom M. Focal cortical dysplasia type II: biological features and clinical perspectives. Lancet Neurol. 2009;8(9):830–43. https://doi.org/10.1016/S1474-4422(09)70201-7
3. Tassi L, Colombo N, Garbelli R, Francione S, Lo Russo G, Mai R, et al. Focal cortical dysplasia: neuropathological subtypes, EEG, neuroimaging and surgical outcome. Brain. 2002;125(8):1719–32. https://doi.org/10.1093/brain/awf175
4. 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. https://doi.org/10.1016/S1474-4422(20)30220-9
5. 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. https://doi.org/10.1111/epi.17130
6. Téllez-Zenteno JF, Hernández Ronquillo L, Moien-Afshari F, Wiebe S. Surgical outcomes in Lesional and non-Lesional epilepsy: a systematic review and meta-analysis. Epilepsy Res. 2010;89(2):310–8. https://doi.org/10.1016/j.eplepsyres.2010.02.007
7. Timoney N, Rutka JT. Recent advances in epilepsy surgery and achieving best outcomes using high-frequency oscillations, diffusion tensor imaging, magnetoencephalography, intraoperative Neuromonitoring, focal cortical dysplasia, and bottom of sulcus dysplasia. Neurosurgery. 2017;64(CN_suppl_1):1–10. https://doi.org/10.1093/neuros/nyx239
8. Widjaja E, Jain P, Demoe L, Guttmann A, Tomlinson G, Sander B. Seizure outcome of pediatric epilepsy surgery: systematic review and meta-analyses. Neurology. 2020;94(7):311–21. https://doi.org/10.1212/WNL.0000000000008966

9. Najm IM, Sarnat HB, Blümcke I. Review: the international consensus classification of focal cortical dysplasia - a critical update 2018. Neuropathol Appl Neurobiol. 2018;44(1):18–31. https://doi.org/10.1111/nan.12462
10. Noe K, Sulc V, Wong-Kisiel L, Wirrell E, Van Gompel JJ, Wetjen N, et al. Long-term outcomes after nonlesional extratemporal lobe epilepsy surgery. JAMA Neurol. 2013;70(8):1003–8. https://doi.org/10.1001/jamaneurol.2013.209
11. Duncan JS, Trimmel K. Advanced neuroimaging techniques in epilepsy. Curr Opin Neurol. 2022;35(2):189–95. https://doi.org/10.1097/WCO.0000000000001007
12. Martin P, Winston GP, Bartlett P, de Tisi J, Duncan JS, Focke NK. Voxel-based magnetic resonance image postprocessing in epilepsy. Epilepsia. 2017;58(9):1653–64. https://doi.org/10.1111/epi.13851
13. Adler S, Wagstyl K, Gunny R, Ronan L, Carmichael D, Cross JH, et al. Novel surface features for automated detection of focal cortical dysplasias in paediatric epilepsy. NeuroImage Clin. 2017;14:18–27. https://doi.org/10.1016/j.nicl.2016.12.030
14. Beheshti I, Sone D, Maikusa N, Kimura Y, Shigemoto Y, Sato N, et al. FLAIR-wise machine-learning classification and lateralization of MRI-negative 18F-FDG PET-positive temporal lobe epilepsy. Front Neurol. 2020;11:580713. https://doi.org/10.3389/fneur.2020.580713
15. El Tahry R, Santos SF, Vrielynck P, de Tourtchaninoff M, Duprez T, Vaz GR, et al. Additional clinical value of voxel-based morphometric MRI post-processing for MRI-negative epilepsies: a prospective study. Epileptic Disord. 2020;22(2):156–64. https://doi.org/10.1684/epd.2020.1152
16. Gill RS, Lee HM, Caldairou 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):e1571–82. https://doi.org/10.1212/WNL.0000000000012698
17. Kini LG, Gee JC, Litt B. Computational analysis in epilepsy neuroimaging: a survey of features and methods. NeuroImage Clin. 2016;11:515–29. https://doi.org/10.1016/j.nicl.2016.02.013
18. 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. https://doi.org/10.1093/brain/awac224
19. Higgins D, Madai VI. From bit to bedside: a practical framework for artificial intelligence product development in healthcare. Adv Intell Syst. 2020;2(10):2000052. https://doi.org/10.1002/aisy.202000052
20. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317–8. https://doi.org/10.1001/jama.2017.18391
21. McCarthy, John. What is artificial intelligence? 2004; p. 15.
22. Fischl B. FreeSurfer. Neuroimage. 2012;62(2):774–81. https://doi.org/10.1016/j.neuroimage.2012.01.021
23. 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. https://doi.org/10.1016/j.neuroimage.2020.117012
24. 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. https://doi.org/10.1111/epi.16853
25. Huppertz H-J, Grimm C, Fauser S, Kassubek J, Mader I, Hochmuth A, et al. Enhanced visualization of blurred gray–white matter junctions in focal cortical dysplasia by voxel-based 3D MRI analysis. Epilepsy Res. 2005;67(1):35–50. https://doi.org/10.1016/j.eplepsyres.2005.07.009
26. Demerath T, Rubensdörfer L, Schwarzwald R, Schulze-Bonhage A, Altenmüller DM, Kaller C, et al. Morphometric MRI analysis: improved detection of focal cortical dysplasia using the MP2RAGE sequence. Am J Neuroradiol. 2020;41:1009–14. https://doi.org/10.3174/ajnr.A6579
27. Wellmer J, Quesada CM, Rothe L, Elger CE, Bien CG, Urbach H. Proposal for a magnetic resonance imaging protocol for the detection of epileptogenic lesions at early outpatient stages. Epilepsia. 2013;54(11):1977–87. https://doi.org/10.1111/epi.12375
28. Wagner J, Urbach H, Niehusmann P, von Lehe M, Elger CE, Wellmer J. Focal cortical dysplasia type IIb: completeness of cortical, not subcortical, resection is necessary for seizure freedom. Epilepsia. 2011a;52(8):1418–24. https://doi.org/10.1111/j.1528-1167.2011.03158.x
29. Ahmed B, Brodley CE, Blackmon KE, Kuzniecky R, Barash G, Carlson C, et al. Cortical feature analysis and machine learning improves detection of "MRI-negative" focal cortical dysplasia. Epilepsy Behav. 2015;48:21–8.
30. Alaverdyan Z, Jung J, Bouet R, Lartizien C. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening. Med Image Anal. 2020;60:101618. https://doi.org/10.1016/j.media.2019.101618
31. Chen X, Qian T, Kober T, Zhang G, Ren Z, Yu T, et al. Gray-matter-specific MR imaging improves the detection of epileptogenic zones in focal cortical dysplasia: a new sequence called fluid and white matter suppression (FLAWS). NeuroImage Clin. 2018a;20:388–97. https://doi.org/10.1016/j.nicl.2018.08.010
32. Chen X, Qian T, Maréchal B, Zhang G, Ren Z, Yu T, et al. Quantitative volume-based morphometry in focal cortical dysplasia: a pilot study for lesion localization at the individual level. Eur J Radiol. 2018b;105:240–5. https://doi.org/10.1016/j.ejrad.2018.06.019
33. Colombo N, Tassi L, Deleo F, Citterio A, Bramerio M, Mai R, et al. Focal cortical dysplasia type IIa and IIb: MRI aspects in 118 cases proven by histopathology. Neuroradiology. 2012;54(10):1065–77. https://doi.org/10.1007/s00234-012-1049-1
34. Ganji Z, Hakak MA, Zamanpour SA, Zare H. Automatic detection of focal cortical dysplasia type II in MRI: is the application of surface-based morphometry and machine learning promising? Front Hum Neurosci. 2021;15:608285. https://doi.org/10.3389/fnhum.2021.608285
35. Hong S-J, Kim H, Schrader D, Bernasconi N, Bernhardt BC, Bernasconi A. Automated detection of cortical dysplasia type II in MRI-negative epilepsy. Neurology. 2014;83(1):48–55. https://doi.org/10.1212/WNL.0000000000000543
36. Kim YH, Kang HC, Kim DS, Kim SH, Shim KW, Kim HD, et al. Neuroimaging in identifying focal cortical dysplasia and prognostic factors in pediatric and adolescent epilepsy surgery. Epilepsia. 2011;52(4):722–7. https://doi.org/10.1111/j.1528-1167.2010.02950.x
37. Kral T, Clusmann H, Blumcke I, Fimmers R, Ostertun B, Kurthen M, et al. Outcome of epilepsy surgery in focal cortical

dysplasia. J Neurol Neurosurg Psychiatry. 2003;74(2):183–8. https://doi.org/10.1136/jnnp.74.2.183
38. Lee HM, Gill RS, Fadaie F, Cho KH, Guiot MC, Hong SJ, et al. Unsupervised machine learning reveals Lesional variability in focal cortical dysplasia at mesoscopic scale. NeuroImage Clin. 2020;28:102438. https://doi.org/10.1016/j.nicl.2020.102438
39. Mellerio C, Labeyrie MA, Chassoux F, Roca P, Alami O, Plat M, et al. 3T MRI improves the detection of transmantle sign in type 2 focal cortical dysplasia. Epilepsia. 2014;55(1):117–22. https://doi.org/10.1111/epi.12464
40. Mo JJ, Zhang JG, Li WL, Chen C, Zhou NJ, Hu WH, et al. Clinical value of machine learning in the automated detection of focal cortical dysplasia using quantitative multimodal surface-based features. Front Neurosci. 2018;12:1008. https://doi.org/10.3389/fnins.2018.01008
41. Radhakrishnan R, Leach JL, Mangano FT, Gelfand MJ, Rozhkov L, Miles L, et al. Prospective detection of cortical dysplasia on clinical MRI in pediatric intractable epilepsy. Pediatr Radiol. 2016;46(10):1430–8. https://doi.org/10.1007/s00247-016-3623-x
42. Seong MJ, Choi SJ, Joo EY, Shon YM, Seo DW, Hong SB, et al. Surgical outcome and prognostic factors in epilepsy patients with MR-negative focal cortical dysplasia. PLOS One. 2021;16(4):e0249929. https://doi.org/10.1371/journal.pone.0249929
43. Wagner J, Weber B, Urbach H, Elger CE, Huppertz HJ. Morphometric MRI analysis improves detection of focal cortical dysplasia type II. Brain. 2011b;134(10):2844–54. https://doi.org/10.1093/brain/awr204
44. 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):1051–8.
45. Wang ZI, Jones SE, Jaisani Z, Najm IM, Prayson RA, Burgess RC, et al. Voxel-based morphometric magnetic resonance imaging (MRI) postprocessing in MRI-negative epilepsies. Ann Neurol. 2015;77(6):1060–75. https://doi.org/10.1002/ana.24407
46. Antel SB, Collins DL, Bernasconi N, Andermann F, Shinghal R, Kearney RE, et al. Automated detection of focal cortical dysplasia lesions using computational models of their MRI characteristics and texture analysis. Neuroimage. 2003;19(4):1748–59. https://doi.org/10.1016/S1053-8119(03)00226-X
47. Krsek P, Maton B, Jayakar P, Dean P, Korman B, Rey G, et al. Incomplete resection of focal cortical dysplasia is the Main predictor of poor postsurgical outcome. Neurology. 2009;72(3):217–23.
48. Blümcke I, Coras R, Busch RM, Morita-Sherman M, Lal D, Prayson R, et al. Toward a better definition of focal cortical dysplasia: an iterative histopathological and genetic agreement trial. Epilepsia. 2021;62(6):1416–28. https://doi.org/10.1111/epi.16899
49. Cendes F, Theodore WH, Brinkmann BH, Sulc V, Cascino GD. Neuroimaging of epilepsy. Handb Clin Neurol. 2016;136:985–1014. https://doi.org/10.1016/B978-0-444-53486-6.00051-X
50. 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. https://doi.org/10.1111/epi.17301
51. Hartlieb T, Winkler P, Coras R, Pieper T, Holthausen H, Blümcke I, et al. Age-related MR characteristics in mild malformation of cortical development with Oligodendroglial hyperplasia and epilepsy (MOGHE). Epilepsy Behav. 2019;91:68–74. https://doi.org/10.1016/j.yebeh.2018.07.009
52. Schurr J, Coras R, Rössler K, Pieper T, Kudernatsch M, Holthausen H, et al. Mild malformation of cortical development with oligodendroglial hyperplasia in frontal lobe epilepsy: a new clinico-pathological entity. Brain Pathol. 2017;27(1):26–35. https://doi.org/10.1111/bpa.12347
53. Mendes Coelho VC, Morita-Sherman M, Yasuda CL, Alvim MMK, Amorim BJ, Tedeschi H, et al. Magnetic resonance imaging findings and clinical characteristics in mild malformation of cortical development with Oligodendroglial hyperplasia and epilepsy in a predominantly adult cohort. Epilepsia. 2021;62(6):1429–41. https://doi.org/10.1111/epi.16907
54. Coras R, de Boer OJ, Armstrong D, Becker A, Jacques TS, Miyata H, et al. Good Interobserver and Intraobserver agreement in the evaluation of the new ILAE classification of focal cortical dysplasias. Epilepsia. 2012;53(8):1341–8. https://doi.org/10.1111/j.1528-1167.2012.03508.x
55. Englund C, Folkerth RD, Born D, Lacy JM, Hevner RF. Aberrant neuronal-glial differentiation in Taylor-type focal cortical dysplasia (type IIA/B). Acta Neuropathol. 2005;109(5):519–33. https://doi.org/10.1007/s00401-005-1005-9
56. Baldassari S, Ribierre T, Marsan E, Adle-Biassette H, Ferrand-Sorbets S, Bulteau C, et al. Dissecting the genetic basis of focal cortical dysplasia: a large cohort study. Acta Neuropathol. 2019;138(6):885–900. https://doi.org/10.1007/s00401-019-02061-5
57. Bedrosian TA, Miller KE, Grischow OE, Schieffer KM, LaHaye S, Yoon H, et al. Detection of brain somatic variation in epilepsy-associated developmental lesions. Epilepsia. 2022;63(8):1981–97. https://doi.org/10.1111/epi.17323
58. Akeret K, Bellut D, Huppertz HJ, Ramantani G, König K, Serra C, et al. Ultrasonographic features of focal cortical dysplasia and their relevance for epilepsy surgery. Neurosurg Focus. 2018;45(3):E5. https://doi.org/10.3171/2018.6.FOCUS18221
59. Prada F, Gennari AG, Del Bene M, Bono BC, Quaia E, DIncerti L, et al. Intraoperative ultrasonography (IoUS) characteristics of focal cortical dysplasia (FCD) type II b. Seizure. 2019;69:80–6. https://doi.org/10.1016/j.seizure.2019.02.020
60. Das KK, Basu M, Basu S, Chou DK, Jungalwala FB. Biosynthesis in vitro of GlcA Beta 1-3nLcOse4Cer by a novel Glucuronyltransferase (GlcAT-1) from embryonic chicken brain. J Biol Chem. 1991;266(8):5238–43.
61. Fauser S, Essang C, Altenmüller DM, Staack AM, Steinhoff BJ, Strobl K, et al. Long-term seizure outcome in 211 patients with focal cortical dysplasia. Epilepsia. 2015;56(1):66–76. https://doi.org/10.1111/epi.12876
62. Cohen-Gadol AA, Ozduman K, Bronen RA, Kim JH, Spencer DD. Long-term outcome after epilepsy surgery for focal cortical dysplasia. J Neurosurg. 2004;101(1):55–65. https://doi.org/10.3171/jns.2004.101.1.0055
63. Hader WJ et al. Cortical dysplastic lesions in children with intractable epilepsy: role of complete resection. J Neurosurg. 2004;100(2 Suppl Pediatrics):110–7. https://doi.org/10.3171/ped.2004.100.2.0110
64. Krsek P, Maton B, Korman B, Pacheco-Jacome E, Jayakar P, Dunoyer C, et al. Different features of histopathological

subtypes of pediatric focal cortical dysplasia. Ann Neurol. 2008;63(6):758–69. https://doi.org/10.1002/ana.21398
65. Rowland NC, Englot DJ, Cage TA, Sughrue ME, Barbaro NM, Chang EF. A meta-analysis of predictors of seizure freedom in the surgical management of focal cortical dysplasia. J Neurosurg. 2012;116(5):1035–41. https://doi.org/10.3171/2012.1.JNS111105
66. Yan H, Ibrahim GM. Resective epilepsy surgery involving eloquent cortex in the age of responsive Neurostimulation: a value-based decision-making framework. Epilepsy Behav. 2019;99:106479. https://doi.org/10.1016/j.yebeh.2019.106479
67. Sezer S, van Amerongen MJ, Delye HHK, Ter Laan M. Accuracy of the neurosurgeons estimation of extent of resection in glioblastoma. Acta Neurochir. 2020;162(2):373–8. https://doi.org/10.1007/s00701-019-04089-8
68. Shimamoto S, Wu C, Sperling MR. Laser interstitial thermal therapy in drug-resistant epilepsy. Curr Opin Neurol. 2019;32(2):237–45. https://doi.org/10.1097/WCO.0000000000000662
69. Aliev R, Kondrateva E, Sharaev M, Bronov O, Marinets A, Subbotin S, et al. Convolutional neural networks for automatic detection of Focal Cortical Dysplasia. In: Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics: Proceedings of the 9th International Conference on Cognitive Sciences, Intercognsci-2020, October 10–16, 2020, Moscow, Russia. Cham: Springer International Publishing; 2021. p. 582–8. https://doi.org/10.1007/978-3-030-71637-0_67
70. Besson P, Bernasconi N, Colliot O, Evans A, Bernasconi A. Surface-based texture and morphological analysis detects subtle cortical dysplasia. Med Image Comput Comput Assist Interv. 2008;22:645–52. https://doi.org/10.1007/978-3-540-85988-8_77
71. Demerath T, Kaller CP, Heers M, Staack A, Schwarzwald R, Kober T, et al. Fully automated detection of focal cortical dysplasia: comparison of MPRAGE and MP2RAGE sequences. Epilepsia. 2022;63(1):75–85. https://doi.org/10.1111/epi.17127
72. Feng C, Zhao H, Li Y, Cheng Z, Wen J. Improved detection of focal cortical dysplasia in normal-appearing FLAIR images using a Bayesian classifier. Med Phys. 2021;48(2):912–25. https://doi.org/10.1002/mp.14646
73. Urbach H, Heers M, Altenmueller DM, Schulze-Bonhage A, Staack AM, Bast T, et al. "Within a minute" detection of focal cortical dysplasia. Neuroradiology. 2022;64(4):715–26. https://doi.org/10.1007/s00234-021-02823-7
74. Wang H, Ahmed SN, Mandal M. Automated detection of focal cortical dysplasia using a deep convolutional neural network. Comput Med Imaging Graph. 2020;79:101662. https://doi.org/10.1016/j.compmedimag.2019.101662
75. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016. p. 779–88. https://doi.org/10.1109/CVPR.2016.91
76. Al-Antari MA, Al-Masni MA, Kim TS. Deep learning computer-aided diagnosis for breast lesion in digital mammogram. In: Lee G, Fujita H, editors. Deep learning in medical image analysis: challenges and applications. Cham: Springer International Publishing; 2020. p. 59–72. https://doi.org/10.1007/978-3-030-33128-3_4
77. Albahli S, Nida N, Irtaza A, Yousaf MH, Mahmood MT. Melanoma lesion detection and segmentation using YOLOv4-DarkNet and active contour. IEEE Access. 2020;8:198403–14. https://doi.org/10.1109/ACCESS.2020.3035345
78. Hossain A, Islam MT, Almutairi AF. A deep learning model to classify and detect brain abnormalities in portable microwave based imaging system. Sci Rep. 2022;12(1):6319. https://doi.org/10.1038/s41598-022-10309-6
79. Ünver HM, Ayan E. Skin lesion segmentation in Dermoscopic images with combination of YOLO and GrabCut algorithm. Diagnostics. 2019;9(3):72. https://doi.org/10.3390/diagnostics9030072
80. House PM, Kopelyan M, Braniewska N, Silski B, Chudzinska A, Holst B, et al. Automated detection and segmentation of focal cortical Dysplasias (FCDs) with artificial intelligence: presentation of a novel convolutional neural network and its prospective clinical validation. Epilepsy Res. 2021;172:106594. https://doi.org/10.1016/j.eplepsyres.2021.106594
81. Thomas E, Pawan SJ, Kumar S, Horo A, Niyas S, Vinayagamani S, et al. Multi-res-attention UNet: a CNN model for the segmentation of focal cortical dysplasia lesions from magnetic resonance images. IEEE J Biomed Health Inform. 2021;25(5):1724–34. https://doi.org/10.1109/JBHI.2020.3024188
82. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III. Cham: Springer International Publishing; 2015. p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28
83. Colliot O, Mansi T, Bernasconi N, Naessens V, Klironomos D, Bernasconi A. Segmentation of focal cortical dysplasia lesions using a feature-based level set. Med Image Comput Comput Assist Interv. 2005;8(Pt 1):375–82. https://doi.org/10.1007/11566465_47
84. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in ML-based science. arXiv preprint arXiv:2207.07048. 2022. https://doi.org/10.48550/arXiv.2207.07048
85. Karageorgiou E. The logic of exploratory and confirmatory data analysis. Cogn Critique. 2011;3:35–48.
86. Li Y, Thompson WK, Reuter C, Nillo R, Jernigan T, Dale A, et al. Rates of incidental findings in brain magnetic resonance imaging in children. JAMA Neurol. 2021;78(5):578–87. https://doi.org/10.1001/jamaneurol.2021.0306
87. Vernooij MW, Ikram MA, Tanghe HL, Vincent AJ, Hofman A, Krestin GP, et al. Incidental findings on brain MRI in the general population. N Engl J Med. 2007;357(18):1821–8. https://doi.org/10.1056/NEJMoa070972
88. Egan JP, Greenberg GZ, Schulman AI. Operating characteristics, signal detectability, and the method of free response. J Acoust Soc Am. 1961;33(8):993–1007. https://doi.org/10.1121/1.1908935
89. Miller H. The FROC curve: a representation of the Observers performance for the method of free response. J Acoust Soc Am. 1969;46(6B):1473–6.
90. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer

aided detection of mammographic lesions. Med Image Anal. 2017;35:303–12.
91. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, das S, Duff EP, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data. 2016;3(1):160044. https://doi.org/10.1038/sdata.2016.44
92. Wagstyl K, Adler S, et al. Planning stereoelectroencephalography using automated lesion detection: retrospective feasibility study. Epilepsia. 2020;61(7):1406–16. https://doi.org/10.1111/epi.16574

How to cite this article: 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;64:1093–1112. https://doi.org/10.1111/epi.17522