ABSTRACT

Objective. To test the hypothesis that a multicenter-validated computer deep learning algorithm detects MRI-negative focal cortical dysplasia (FCD).

Methods. We used clinically-acquired 3D T1-weighted and 3D FLAIR MRI of 148 patients (median age, 23 years [range, 2-55]; 47% female) with histologically-verified FCD at nine centers to train a deep convolutional neural network (CNN) classifier. Images were initially deemed as MRI-negative in 51% of cases, in whom intracranial EEG determined the focus. For risk stratification, the CNN incorporated Bayesian uncertainty estimation as a measure of confidence. To evaluate performance, detection maps were compared to expert FCD manual labels. Sensitivity was tested in an independent cohort of 23 FCD cases (13±10 years). Applying the algorithm to 42 healthy and 89 temporal lobe epilepsy disease controls tested specificity.

Results. Overall sensitivity was 93% (137/148 FCD detected) using a leave-one-site-out cross-validation, with an average of six false positives per patient. Sensitivity in MRI-negative FCD was 85%. In 73% of patients, the FCD was among the clusters with the highest confidence; in half it ranked the highest. Sensitivity in the independent cohort was 83% (19/23; average of five false positives per patient). Specificity was 89% in healthy and disease controls.

Conclusions. This first multicenter-validated deep learning detection algorithm yields the highest sensitivity to date in MRI-negative FCD. By pairing predictions with risk stratification this classifier may assist clinicians to adjust hypotheses relative to other tests, increasing diagnostic confidence. Moreover, generalizability across age and MRI hardware makes this approach ideal for pre-surgical evaluation of MRI-negative epilepsy.

Classification of evidence. This study provides Class III evidence that deep learning on multimodal MRI accurately identifies FCD in epilepsy patients initially diagnosed as MRI-negative.

INTRODUCTION

Focal cortical dysplasia (FCD), a surgically-amenable developmental epileptogenic brain malformation, presents with cortical thickening on T1-weighted MRI, as well as hyperintensity and blurring of the gray-white matter interface on FLAIR images. While these features are often visible to the naked eye, FCD may be overlooked and only found at surgery [1]. MRI-negative patients represents a major diagnostic challenge [2].

Currently, benchmark automated detection methods fail in 20–40% of patients [3-6], particularly those with subtle FCD, and suffer from high false positive rates [7]. Conversely, deep neural networks outperform state-of-the-art methods at disease detection (see [8,9] for review). Specifically, convolutional neural networks (CNNs) learn abstract concepts from high-dimensional data alleviating the challenging task of hand-crafting features [10]. The integration of convolutional operators that implicitly encode spatial covariance of neighboring voxels (rather than treating each voxel independently) with nonlinearity capturing complex patterns and variability is expected to optimize the detection of the full FCD spectrum. Notably, with regards to diagnostic performance, the deterministic nature of conventional algorithms does not permit risk assessment of the automated decisions, a requirement to be integrated into clinical diagnostic systems. Alternatively, Bayesian CNNs provide a distribution of predictions from which the mean and variance can be computed, the latter being interpreted as a measure of uncertainty [11].

Here, we tested the hypothesis that a multicenter-validated computer deep learning algorithm operating directly on T1-weighted and FLAIR MRI voxel detects MRI-negative focal cortical dysplasia (FCD).

METHODS

The primary question of this study was to assess whether a deep learning algorithm operating on multimodal MRI has significant diagnostic value, including in MRI-negative patients. Our

automated algorithm was trained and validated on a multicenter dataset of patients with histologically confirmed FCD. We ruled out sources of spectrum bias [12] by evaluating specificity against healthy individuals as well as a disease control cohort of patients with temporal lobe epilepsy (TLE) and histologically confirmed hippocampal sclerosis (HS). To minimize incorporation bias [12], the classifier was iteratively trained and tested using a leave-one-site-out scheme; i.e., the classifier was trained iteratively on all sites minus the one held-out for testing; this guaranteed the out-of-fold validation in which tested cohorts were never part of the training. Moreover, the classifier trained on the full dataset was tested on an independent cohort of patients that were never part of training. According to the Classification of evidence schemes of the American Academy of Neurology (https://www.neurology.org/sites/default/files/ifa/loe.pdf) [13], this study satisfies the rating for Class III evidence for diagnostic accuracy, demonstrating that deep learning operating on multimodal MRI has significant diagnostic value, including in MRI-negative patients, with 85% sensitivity.

Subjects

We studied consecutive retrospective cohorts from nine tertiary epilepsy centers worldwide with histologically validated FCD lesions collected from October 2012 to January 2018 and in whom both 3D T1-weighted MRI and 3D FLAIR were acquired as part of the clinical presurgical investigation [14]. The TLE cohort included both patients with MRI-visible HS (n=49; comparable to MRI-positive FCD) and those in whom the MRI was unremarkable, but the histological examination of the surgical specimen revealed the presence of subtle HS (n=40; comparable to MRI-negative, histology-positive FCD). Patients had been investigated for drug-resistant epilepsy with a standard presurgical workup including neurological examination, assessment of seizure history, neuroimaging, and video-EEG recordings.

On histological examination of the surgical specimen [15], FCD Type-II was defined as disrupted cortical lamination with dysmorphic neurons in isolation (IIA, n=70) or together with balloon cells (IIB, n=78). At a mean±SD postoperative follow-up of 31.2±14.4 months (range: 12-78 months), 103 patients (70%) became seizure-free (Engel-I), 33 (22%) had rare disabling seizures (Engel-II), nine (6%) had worthwhile improvement (Engel-III) and three (2%) had no improvement (Engel-IV); in patients with Engel-III and IV, the resection was incomplete as the FCD encroached eloquent areas in primary cortices (7 in sensorimotor, 2 in primary visual and 3 in language areas); the residual lesion and extent of resection were evaluated on post-operative MRI.

Standard Protocol Approvals, Registrations, and Patient Consents

The Ethics Committees and institutional review boards at all participating sites (S1-S9) approved the study, and written informed consent was obtained from all participants.

MRI acquisition and image processing

High-resolution 3D T1-weighted and 3D FLAIR MRI images were acquired in all individuals [14]. All images were obtained on 3T scanners; one site provided additional cases with 1.5T MRI. Imaging parameters are listed on **Table e-1** (available from Dryad: doi.org/10.5061/dryad.h70rxwdgm). MRI data was de-identified; files were converted from DICOM to NIfTI with header anonymization. T1-weighted images were linearly registered to the MNI152 symmetric template. FLAIR images were linearly mapped to T1-weighted MRI in MNI space. T1-weighted and FLAIR underwent intensity non-uniformity correction [16] followed by intensity standardization with scaling of values between 0 and 100. Finally, images were skull-stripped using an in-house deep learning method (v1.0.0: https://github.com/NOEL-MNI/deepMask) trained on manually corrected brain masks from patients with cortical malformations. Two experts manually segmented lesions on co-registered T1-weighted and

FLAIR images; inter-rater Dice agreement was 0.92±0.10 [calculated as $2|M_1 \cap M_2|/(|M_1|+|M_2|)$, where M_1 = label 1, M_2 = label 2, $M_1 \cap M_2$ = intersection of M_1 and M_2].

Classifier design

Figure 1 and **Figure e-1 (**available from Dryad: doi.org/10.5061/dryad.h70rxwdgm) illustrate the design. The full methodology is described in **Additional Methods** (available from Dryad: doi.org/10.5061/dryad.h70rxwdgm).

Data sampling and network architecture. In each individual, we thresholded FLAIR images by z-normalizing intensities and discarding the bottom 10 percentile intensities; this internal thresholding resulted in a mask containing voxels within the grey matter (GM) and its interface with the white matter (WM). This mask was then used to extract 3D patches (*i.e.*, regions of interest centered around a given voxel) from co-registered 3D T1-weighted and FLAIR images, which served as input to the network. Notably, 3D patches seamlessly sampled the FCD across orthogonal planes and tissue types. We designed a cascaded system in which the output of the first CNN (CNN-1) served as input to the second (CNN-2). CNN-1 aimed at maximizing the detection of lesional voxels; CNN-2 reduced the number of misclassified voxels, removing false positives (FPs) while maintaining optimal sensitivity. In brief, for each test subject, 3D T1-weighted and FLAIR patches were fed to CNN-1. To discard improbable lesional candidates, the mean of 20 forward passes (or predictions) was thresholded at >0.1 (equivalent to rejecting bottom 10 percentile probabilities); voxels surviving this threshold served as the input to sample patches for CNN-2.

Estimation of prediction uncertainty. Bayesian inference in deep CNNs with large number of parameters is computationally intensive [17]. By probabilistically excluding neurons (or units) after every convolutional layer during training, the Monte Carlo dropout method [18] simulates an ensemble of neural networks with diverse architectures, thus preventing overfitting without compromising on accuracy. This procedure provides a distribution of posterior probabilities at

each voxel resulting from multiple stochastic forward passes through the classifier; their variance provides a measure of uncertainty. Here, we used the mean and variance of 50 voxel-wise forward passes to generate probability and uncertainty maps. The mean probability map was binarized by thresholding at >0.7 (empirically determined by setting the cluster-level FP rate to <6) and underwent a post-processing routine entailing morphological erosion, dilation and extraction of connected components (>75 voxels) to remove flat blobs and noise, a procedure that resulted in non-overlapping clusters. To evaluate performance, this detection map was compared to the manual expert annotation.

Transforming uncertainty into confidence and ranking. For each cluster of the detection map, we estimated confidence by computing the median uncertainty across its voxels; we then aggregated uncertainties across all clusters and normalized values between 0 and 1 to obtain a measure of confidence. All clusters were then ranked based on their confidence estimates with the highest confidence cluster as rank 1, second highest confidence cluster rank 2, and so on until all clusters surviving the threshold (probability >0.7 and spatial extent >75 voxels) had been ranked. Confidence maps were evaluated together with a diagram plotting lesion probability against lesion ranking, with rank 1 signifying highest confidence to be lesional, regardless of cluster size.

Performance evaluation

To assess performance, we employed a *leave-one-site-out* cross-validation by which the classifier trained on eight sites was tested iteratively on the held-out site until all sites had served as testing set. A minimum of one voxel co-localizing with the manually segmented FCD (ground truth label) was deemed as TP, any detection not co-localizing as FP. Consistent with previous FCD detection literature [3,19-21], we deemed partial overlap to be sufficient without requiring the detection to be completely within the expert label. Demographics and dataset stratification are shown in **Table 1**. In addition, we evaluated the algorithm trained on the

complete dataset of the 148 FCD patients on an independent cohort of 23 FCD cases (11 females; 13±10 years; 70% MRI-negative) from S1 and S2.

Patient-level (*i.e.*, lesion-level) evaluation metrics included sensitivity (P, L) = $|P_1 \cap L_1|/|L_1|$ and specificity (P, L) = $|P_0 \cap L_0|/|L_0|$, where P is the model prediction and L the ground truth label; L_1 and L_0 signify voxels predicted as positive (lesional) and negative (not lesional), while P_1 and P_0 represent the same for model predictions. We evaluated specificity as the absence of any findings by applying the algorithm trained on the complete dataset of FCD patients to healthy controls and TLE disease controls; in other words, specificity was calculated as the proportion of healthy or disease controls in whom no FCD lesion cluster was falsely identified. Site-wise area under the receiver operating characteristic curve (AUC) evaluated voxel-wise classification performance (*i.e.*, the true positive (TP) *vs.* FP rate) stratified by sites.

We evaluated the spatial relation between lesional clusters and FPs in patients as well as healthy and disease controls. To this end, we generated a lesional probability map by overlaying all manually segmented FCD labels; the Dice coefficient quantified the overlap between the FCD probability map and both the group-wise probability and uncertainty maps of FPs.

Pearson's correlation quantified associations between probability and uncertainty, and between age and the number of FPs. Biserial correlation evaluated association between MRI-negative status and the number of FPs. Spearman's correlation quantified association between lesion rank and probability. Nonparametric permutations (10,000 iterations with replacement) tested group differences at $p<.05$ (two-tailed), with Bonferroni correction for multiple comparisons.

Data availability statement

These datasets are not publicly available as they contain information that could compromise the privacy of research participants. The source code and pre-trained model weights are available for download online (v1.0.0 at GitHub: https://github.com/NOEL-MNI/deepFCD,).

In addition, a derivative dataset composed of lesional and non-lesional patches from 148 FCD patients is available as a Hierarchical Data Format dataset (available from Zenodo: doi.org/10.5281/zenodo.3239446).

RESULTS

Demographics. The primary site (S1) comprised 62 FCD patients (33 females; mean±SD age=25±10 years) and control groups consisting of age- and sex- matched healthy individuals (n=42; 22 females; 30±7 years), and patients with TLE and histologically verified HS (n=89; 47 females; age: 31±8). Across the remaining eight sites (S2-S9), the cohort comprised 86 FCD patients (36 females; age: 20±14). In 75 patients (51%) in whom routine MRI evaluation was initially reported as unremarkable in the initial readings of the neuroradiologists at each participating center, the location of the seizure focus was established using intracranial EEG.

Patient-level performance. The classifier's overall sensitivity based on leave-one-site-out cross-validation was 93% (137/148 FCD lesions detected), with 6±5 FP clusters per patient. Stratifying children and adults, sensitivity was 98% for the former (52/53; 7±5 FP clusters) and 89% (85/95; 5±5 FP) for the latter. Notably, 85% of MRI-negative and 100% of MRI-positive lesions were detected. When testing the classifier on the independent cohort (using the model trained on the complete dataset of the 148 FCD patients), overall sensitivity was 83% (19/23 FCD lesions detected; 5±3 FP clusters per patient) with 100% of MRI-positive and 75% of MRI-negative lesions detected. Specificity was 90% in healthy (4/42 with 2±1 FP clusters) and 89% in TLE disease controls (10/89, 1±0 FP cluster). With respect to the latter, specificity was similar between MRI-positive HS (92%; 5/49, 1±0 FP cluster) and MRI-negative HS (88%; 5/40, 1±0 FP cluster). Per-site sensitivity and FP rates are shown in **Table 2**.

Voxel-wise performance. The median AUC was 0.83 (range, 0.72–0.87) indicative of high sensitivity (high TP rates) and specificity (low FP rates), with comparable performance across sites (**Figure 2A**).

Analysis of confidence. In 73% of patients, the FCD lesion was among the five clusters with the highest confidence; in half of them, it ranked the highest, with a mean probability of 72% (95% confidence interval, 69%–76%; **Figure 2B**). Lesion rank negatively correlated with probability, *i.e.*, the lower the rank, the higher the probability of being lesional ($r = -0.69$, $p = 0.005$; **Figure 2C**). Moreover, confidence for a cluster to be lesional centered around 1 (*i.e.*, 100% confidence), while for FPs it centered around zero (**Figure 2D**). Representative MRI-negative cases are shown in **Figure 3** and **4**.

Spatial distribution of FCD and FPs. The majority of FCD lesions were located within the frontal lobes (**Figure 5A**). Overall, FPs in patients, healthy and disease controls (**Figure 5B**) were found in the insula and the parahippocampus (Dice overlap with FCD: 21%, 22% and 34%, respectively). Notably, FPs in healthy and disease controls overlapped to a greater extent (Dice: 52%) and exhibited low confidence to be lesional (i.e., high uncertainty); conversely, FPs in FCD patients tended to display high confidence to be lesional ($p = 0.013$). Coordinates for the lesion and FPs are listed on **Tables e-2** and **e-3 (**available from Dryad: doi.org/10.5061/dryad.h70rxwdgm). The incidence of FP clusters was negatively correlated with age ($r = -0.23$, $p = 0.004$), namely the younger the patients the higher the number of FPs. Number of FPs was not significantly different between MRI-positive and MRI-negative patients.

DISCUSSION

MRI-negative FCD represents a major diagnostic challenge. To define the epileptogenic area patients undergo long and costly hospitalizations for EEG monitoring with intracerebral electrodes, a procedure that carries risks similar to surgery itself [22,23]. Moreover, patients without MRI evidence for FCD are less likely to undergo surgery and consistently show worse seizure control compared to those with visible lesions [24,25]. Here, we present the first deep learning method for automated FCD detection trained and validated on histologically verified

data from multiple centers worldwide. The classifier uses T1- and T2-weighted FLAIR, contrasts available on most recent MR scanners [14], operates in 3D voxel space without laborious pre-processing and feature extraction, and pairs predictions with confidence. It yields the highest performance to date with a sensitivity of 93% using a leave-one-site-out cross-validation and 83% when tested on an independent cohort, while maintaining a high specificity of 89% both in healthy and disease controls. Importantly, deep learning detected MRI-negative FCD with 85% sensitivity, thus offering a considerable gain over standard radiological assessment. Results were generalizable across cohorts with variable age, hardware and sequence parameters. Taken together, such characteristics and performance promise potential for broad clinical translation. Notwithstanding these advantages, good quality scans are essential to guarantee valid results, as motion can mimic lesions [14]; we thus advise against analysing low-quality motion-corrupted scans. Notably, while a classification III for the level of evidence was assigned to our study, the current AAN scheme for diagnostic accuracy does not specify criteria for designs based on machine learning algorithms. A revision of these guidelines, ideally disease-specific, would likely better reflect the level of evidence for studies relying on artificial intelligence.

Deep learning: moving beyond conventional automated FCD detection

Over the last decade, several automated FCD detection algorithms have been developed, the most recent relying on surface-based representations [3,6,19,20]. While the majority operate on T1-weighted MRI, recent methods have combined T1-weighted and T2-weighted MRI for improved performance [6,21]. A few have used shallow (single layer) artificial neural networks [4,21]. Notably, all require arduous pre-processing, including manual corrections of tissue segmentation and surface extraction, thus precluding integration into clinical workflow. Importantly, they rely on domain knowledge to engineer features. These procedures generally fail to detect subtle lesions [7]. In comparison, our approach offers several advantages. Firstly, to

optimize lesion detection across the FCD spectrum, we leveraged the power of CNNs that recursively learn complex properties from the data itself. Secondly, contrary to previous medical imaging applications relying on 2D orthogonal sampling, we extracted 3D patches to model the spatial extent of FCD across multiple slices and tissue types. Operating in true volumetric domain allowed assessing the spatial neighborhood of the lesion, whereas prior surface-based methods have considered each vertex location independently. Thirdly, restricting training to the GM reduced nearly infinite dataset to a manageable finite set. Finally, by relying on subject-wise feature normalization, rather than group-wise, our implementation obviates the need for a matched normative dataset, an expensive and time-intensive undertaking. Compared to previous deep learning methods [26-28] in which clinical description was scarce to absent, and information on the FCD expert labels and histological validation of lesions was not provided, our study relied on best-practice multimodal MRI, histologically-validated lesions, and a large dataset. Moreover, in previous work FLAIR images in presumably MRI-positive patients were acquired with inter-slice gap ranging from 0.5 to 1.0 mm [26,27], and the acquisition parameters for the 3D T1-weighted images were different from those in healthy and disease controls [28].

Notwithstanding practical advantages of our method, a general limitation of deep learning is the reduced transparency of the process leading to the predictions, a consequence of the high dimensionality of learned features. The trade-off is a richer encoding and learning of complex spatial covariances of intensity and morphology that is beyond the ability of human eye. To maximize transparency and validity, we trained our algorithm on manual expert labels of histologically-validated FCD lesions. In addition to a rigorous cross-validation design, including applying the classifier to a totally independent cohort of FCD patients, our predictions were stratified according to confidence to be lesional. Notwithstanding these precautions, as for many diagnostic tests, the convergence of findings with independent tests is essential to increase confidence even further.

Estimation of generalizability is key to any diagnostic method. To guarantee unbiased evaluation, training and testing datasets should remain distinct. We thus devised a strategy in which the model was iteratively trained on patient data from all sites, except the one held-out. this guaranteed out-of-distribution validation in which tested cohorts were never part of the training. This leave-one-site-out cross-validation simulated a real-world scenario with optimal bias-variance trade-off compared to conventional train-test split of *k*-folds; it also exploited the full richness of data during training and the out-of-distribution samples from a single site during testing. Moreover, the classifier trained on the full dataset was tested on a totally independent cohort of patients that were never part of training. Consistent high performance across cohorts, as well as modest FPs in healthy and disease controls, demonstrate that our cascaded CNN classifier learns and optimizes parameters specific to FCD, a fact validated by histological confirmation.

Human-in-the-loop machine learning: key to clinical translation

In machine learning, *human-in-the-loop* refers to the need for human interaction with the learner to improve human performance, machine performance, or both. Human involvement expedites the efficient labeling of difficult or novel cases that the machine has previously not encountered, reducing the potential for errors, a requirement of utmost importance in healthcare. In FCD, the outcome of surgery depends heavily on the identification of the lesion; it is thus crucial to decide which putative lesional clusters are significant. In this context, thresholding the final probabilistic mean map is essential to evaluate the balance between true positive and false positives. Notably, to guarantee an objective assessment of sensitivity and specificity across cohorts, in this study we defined an empirical threshold. However, in clinical practice, a judicious approach would imply adaptive thresholding of the maps at single-patient level, taking into account independent tests. Indeed, in 5/11 of undetected MRI-negative cases, the lesion could be resolved when modulating the threshold in light of seizure semiology and

electrophysiology. Besides thresholding, confidence is pivotal in any diagnostic assessment, an aspect so far neglected. To fill this gap, we incorporated a Bayesian uncertainty estimation that enables risk stratification. In practical terms, we ranked putative lesional clusters in a given patient based on confidence, thus assisting the examiner to gauge the significance of all findings. In 73% of cases the FCD was among the top five clusters with the highest confidence to be lesional; in half of them it ranked the highest. In the remaining 27%, lesions manifested with low confidence; in a real-world scenario, when location is unknown (*i.e.*, no FCD label is available), a concerted evaluation including electro-clinical and other imaging tests is likely to increase diagnostic certainty [29]. While the good performance of our classifier is also attributable to the richness of the training set including a large spectrum of anatomical locations, eleven MRI-negative FCD remained unresolved, with six located in the orbitofrontal cortex, an area for which limited data was available for training. The prospective use of our classifier trained on the entire cohort would likely reclaim these lesions.

The analysis of the spatial distribution of FPs was moderately comparable across FCD patients, healthy and disease controls, mainly involving the insula and parahippocampal region bilaterally. A possible explanation may lie in the similarity of the cytoarchitectonic signature of these cortices with FCD histopathological traits. Notably, the three-layered cortex of the hippocampal formation, the transitional mesocortex of the parahippocampus and the mesocortex-like insula present with indistinct boundaries between laminas compared to the typical six-layered neocortex [30,31]; these cortices may thus mimic dyslamination and blurring. Notably, however, our algorithm detected 3/3 FCD lesions in the insula with high degree of confidence. Since these lesions were provided by different sites, the leave-one-site-out strategy guaranteed that each training set had at least one lesion. Nonetheless, adding more lesions to the training set would increase the classifier's ability to learn better discriminative features in the insular region. Alternatively, an impact of developmental trajectory [32] on FPs is suggested

by high prevalence in younger patients, possibly in relation to age-varying tissue contrast, cortical myelination and maturation, which may also manifest as lesion-like on MRI. Conversely, registration errors are less likely in our voxel-based method as compared to surface-based algorithms. For the latter, to align a subject's brain into a standardized stereotaxic space registration strongly depends on the accuracy of GM/WM segmentation, while our method does not require tissue segmentation. Notably, some FPs were only seen in FCD cases, particularly in fronto-central regions and tended to gather around the lesion, suggesting subthreshold peri-lesional anomalies not included in the manually-segmented label [3,33]. Given the favorable surgical outcome, a biological explanation for FPs in our FCD cohort may thus include a combination of normal cytoarchitectural nuances and non-epileptogenic peri-lesional developmental anomalies. In a previous study [3], we found FPs to manifest as abnormal sulcal depth, while the FCD lesions had higher cortical thickness relative to controls. Sulcal abnormalities in cortical malformations have been described in the proximity and at a distance of MRI-visible lesions and are thought to result from disruptions of neuronal connectivity and WM organization [3,34]. Finally, it is also plausible that some FP clusters my represent dysplastic tissue, an entity so far reported only in five cases [35].

While our algorithm was trained on histologically verified FCD-II lesions and is mainly aimed at identifying MRI-negative FCD, it is possible that it could identify difficult-to-detect low grade tumors that may resemble dysplastic lesions, a rare instance occurrence since these tumors are generally easy to see on routine MRI. Regardless, the dilemma of differentiation of FCD from low grade tumors uniquely based on MRI features may arise; the differential diagnosis is then evaluated using additional tests, including MR spectroscopy. On the other hand, our algorithm may be useful in identifying associated often-occult dysplastic lesions in the peritumoral area [36].

Federated machine learning: a path to the future

Traditional machine learning adopts a centralized approach that requires training datasets to be aggregated in a single center. A significant obstacle to clinical adoption of such strategy is privacy and ethical concerns. Federated learning [37], on the other hand, is a distributed approach that enables multi-institutional collaboration without sharing patient data. Our proposed approach of patch-based data augmentation is privacy-preserving since only a portion of each patient's data is collated and randomized before exposure to the neural network, an implementation that can be flexibly re-configured to support federated learning. As the data corpus diversifies and expands to include more edge cases, performance and confidence of future classifiers will inevitably improve.

REFERENCES

1. Bernasconi A, Bernasconi N, Bernhardt BC, Schrader D. Advances in MRI for "cryptogenic" epilepsies. Nat Rev Neurol. Nature Publishing Group; 2011;7:99–108.

2. So EL, Lee RW. Epilepsy surgery in MRI-negative epilepsies. Current opinion in neurology. 2014;27:206–212.

3. 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;83:48–55.

4. Zhao Y, Ahmed B, Thesen T, et al. A Non-parametric Approach to Detect Epileptogenic Lesions using Restricted Boltzmann Machines. KDD '16. New York, New York, USA: ACM Press; 2016. p. 373–382.

5. Snyder K, Whitehead EP, Theodore WH, Zaghloul KA, Inati SJ, Inati SK. Distinguishing Type II Focal Cortical Dysplasias from Normal Cortex: A Novel Normative Modeling Approach. NeuroImage: Clin. 2021;15:102565.

6. Gill RS, Hong S-J, Fadaie F, et al. Automated Detection of Epileptogenic Cortical Malformations Using Multimodal MRI. In: Cardoso MJ, Arbel T, Carneiro G, et al., editors. Med Image Comput Comput Assist Interv. Cham: Springer International Publishing; 2017. p. 349–356.

7. Kini LG, Gee JC, Litt B. Computational analysis in epilepsy neuroimaging: A survey of features and methods. NeuroImage: Clin. 2016;11:515–529.

8. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017;42:60–88.

9. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. Nature Publishing Group; 2019;25:44–56.

10. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–444.

11. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep. Nature Publishing Group; 2017;7:17816.

12. Sica GT. Bias in research studies. Radiology. 2006;238:780–789.

13. Ashman EJ, Gronseth GS. Level of evidence reviews: three years of progress. Neurology. 2012;79:13–14.

14. Bernasconi A, Cendes F, Theodore WH, et al. Recommendations for the use of structural magnetic resonance imaging in the care of patients with epilepsy: A consensus report from the International League Against Epilepsy Neuroimaging Task Force. Epilepsia. John Wiley & Sons, Ltd (10.1111); 2019;73:464.

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

16. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging. 1998;17:87–97.

17. Gal Y, Ghahramani Z. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. arXiv 2015.

18. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014;15:1929–1958.

19. Thesen T, Quinn BT, Carlson C, et al. Detection of Epileptogenic Cortical Malformations with Surface-Based MRI Morphometry. Feany M, editor. PLoS ONE. Public Library of Science; 2011;6:e16430.

20. Ahmed B, Brodley CE, Blackmon KE, et al. Cortical feature analysis and machine learning improves detection of "MRI-negative" focal cortical dysplasia. Epilepsy Behav. 2015;48:21–28.

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

22. Hader WJ, Tellez-Zenteno J, Metcalfe A, et al. Complications of epilepsy surgery: a systematic review of focal surgical resections and invasive EEG monitoring. Epilepsia. 2013;54:840–847.

23. Hedegärd E, Bjellvi J, Edelvik A, Rydenhag B, Flink R, Malmgren K. Complications to invasive epilepsy surgery workup with subdural and depth electrodes: a prospective population-based observational study. J Neurol Neurosurg Psychiatry. 2014;85:716–720.

24. Téllez-Zenteno JF, Hernandez-Ronquillo L, Moien-Afshari F, Wiebe S. Surgical outcomes in lesional and non-lesional epilepsy: a systematic review and meta-analysis. Epilepsy Res. 2010;89:310–318.

25. Noe K, Sulc V, Wong-Kisiel L, et al. Long-term outcomes after nonlesional extratemporal lobe epilepsy surgery. JAMA Neurology. 2013;70:1003–1008.

26. Bijay Dev KM, Jogi PS, Niyas S, Vinayagamani S, Kesavadas C, Rajan J. Automatic detection and localization of Focal Cortical Dysplasia lesions in MRI using fully convolutional neural network. Biomedical Signal Processing and Control. 2019;52:218–225.

27. Thomas E, Pawan SJ, Kumar 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. 2020;PP:1–1.

28. 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.

29. Guerrini R, Duchowny M, Jayakar P, et al. Diagnostic methods and treatment options for focal cortical dysplasia. Epilepsia. 2015;56:1669–1686.

30. Nieuwenhuys R. The insular cortex: a review. Prog Brain Res. Elsevier; 2012;195:123–163.

31. Gogolla N. The insular cortex. Current Biology. 2017;27:R580–R586.

32. Gilmore JH, Knickmeyer RC, Gao W. Imaging structural and functional brain development in early childhood. Nature reviews Neuroscience. Nature Publishing Group; 2018;19:123–137.

33. Hong S-J, Bernhardt BC, Caldairou B, et al. Multimodal MRI profiling of focal cortical dysplasia type II. Neurology. 2017;88:734–742.

34. Besson P, Andermann F, Dubeau F, Bernasconi A. Small focal cortical dysplasia lesions are located at the bottom of a deep sulcus. Brain. Oxford University Press; 2008;131:3246–3255.

35. Fauser S, Sisodiya SM, Martinian L, et al. Multi-focal occurrence of cortical dysplasia in epilepsy patients. Brain. 2009;132:2079–2090.

36. Slegers RJ, Blümcke I. Low-grade developmental and epilepsy associated brain tumors: a critical update 2020. Acta Neuropathol Commun. 2020;8:27.

37. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2:305–311.

Table 1. Demographics and dataset stratification for cross-site validation.

TESTING				TRAINING			
Site	**N**	**Age (mean±SD yrs)**	**% Female**	**Sites**	**N**	**Age (mean±SD yrs)**	**% Female**
S1-I	45	27±9	49%	S1-II, S2-S9	103	20±13	46%
S1-II	17	18±9	65%	S1-I, S2-S9	131	23±13	44%
S2	08	11±6	25%	S1, S3-S9	140	23±12	48%
S3	05	22±17	80%	S1-S2, S4-S9	143	23±12	45%
S4	11	8±7	36%	S1-S3, S5-S9	137	24±12	44%
S5-I	10	23±14	30%	S1-S4, S5-II, S6-S9	138	23±13	48%
S5-II	12	13±12	42%	S1-S4, S5-I, S6-S9	136	22±12	47%
S6	11	31±15	64%	S1-S5, S7-S9	137	22±12	45%
S7	09	33±13	33%	S1-S6, S8-S9	139	22±13	47%
S8	07	24±13	43%	S1-S7, S9	141	22±13	47%
S9	13	26±8	38%	S1-S8	135	22±13	47%

Abbreviations. S = site; N: sample size**;** yrs = years; I and II refer to different MRI scanners for the same site

Table 2. Site-specific demographics and performance metrics.

Site	N	Age (mean±SD yrs)	% Female	MRI+/ MRI-	Sensitivity		FPs
					All patients	MRI-	
S1-I	45	27±9	49%	13/32	39/45 (87%)	26/32 (81%)	7±4
S1-II	17	18±9	65%	2/15	15/17 (88%)	13/15 (87%)	7±4
S2	08	11±6	25%	5/3	8/8 (100%)	3/3 (100%)	6±5
S3	05	22±17	80%	2/3	5/5 (100%)	3/3 (100%)	1±1
S4	11	8±7	36%	11/0	11/11 (100%)	n/a	8±6
S-I	10	23±14	30%	8/2	9/10 (90%)	1/2 (50%)	10±6
S5-II	12	13±12	42%	11/1	12/12 (100%)	1/1 (100%)	6±7
S6	11	31±15	64%	6/5	11/11 (100%)	5/5 (100%)	3±3
S7	09	33±13	33%	2/7	8 /9 (89%)	6/7 (86%)	8±6
S8	07	24±13	43%	6/1	6/7 (86%)	0/1 (0%)	6±5
S9	13	26±8	38%	7/6	13/13 (100%)	6/6 (100%)	1±2
Total	148	23±13	47%	49/51%	137/148 (93%)	64/75 (85%)	6±5
Indep	23	13±10	48%	30/70%	19/23 (83%)	12/16 (75%)	5±3

Abbreviations. N: sample size**;** FPs: false positive rate per cohort; MRI+/-: MRI positive/negative; SD: standard deviation; yrs: years; I and II refer to different MRI scanners for the same site; n/a: not applicable; Indep: independent validation cohort from S1 and S2.

Figure 1. Classifier design. Training and inference (or testing) workflow. In the cascaded system the output of CNN-1 serves as an input for CNN-2. CNN-1 maximizes the detection of lesional voxels; CNN-2 reduces the number of misclassified voxels, removing false positives (FPs) while maintaining optimal sensitivity. The training procedure (indicated by dashed arrows) operating on T1-weighted and FLAIR MRI, extracts 3D patches from lesional and non-lesional tissue to yield tCNN-1 (trained model 1) and tCNN-2 (trained model 2) models with optimized weights (indicated by vertical dashed-dotted arrows). These models are then used for subject-level inference. For each unseen subject, the inference pipeline (solid arrows) uses tCNN-1 and generates a mean ($\mu_{droupout}$) of 20 predictions (forward passes); the mean map is then thresholded voxel-wise to discard improbable lesion candidates ($\mu_{droupout} > 0.1$). The resulting binary mask serves to sample the input patches for the tCNN-2. A mean probability and uncertainty maps are obtained by collating 50 predictions; uncertainty is transformed into confidence. The sampling strategy (identical for training and inference) is only illustrated for testing.

Figure 2. Performance evaluation. A. Site-wise area under the receiver operating characteristic curve (AUC) using the leave-one-site-out cross-validation (solid colored lines with values; black dotted line represents a naïve classifier). **B.** Frequency of lesions according to their rank. Rank 1 signifies highest confidence to be lesional. 73% of lesions were distributed across ranks 1 to 5. **C.** Lesion rank plotted against probability of being lesional shows a significant correlation with FCD voxels having low rank values (high confidence) and high probability. **D.** Distribution (kernel density estimation) of confidence for lesional and false positive (FP) clusters; lesions exhibit high confidence values, while FP clusters show low confidence.

Figure 3. Automated detection of MRI-negative FCD. The left panels show the T1-weighted MRI and the prediction probability maps with the lesion circled. The middle plots show the

probability of the lesion and false positive (FP) clusters sorted by their rank; the superimposed line indicates the degree of confidence for each cluster. The right panels illustrate the location of the FCD lesion (rank 1, highest confidence; purple) and FP clusters (ranks 2-5; blue). In these cases, the lesion has both highest confidence (rank 1) and high probability (>0.8).

Figure 4. Representative FCD detection examples. Seven representative MRI-negative FCD lesions across sites are shown (top row: prediction overlaid on the FLAIR; the flame scale indicates the probability strength). The bottom labels are interpreted as site-patient-ID/age/gender. The arrows indicate the ground-truth lesion location.

Figure 5. Probability distributions of FCD and false positives. A. Lesional probability maps of manually labelled FCD lesions superimposed on glass brains. **B.** Probability maps of confidence of FP clusters across cohorts. Colors indicate proportions (in %) of lesional (A) and FPs (B) voxels.