Convolutional Neural Networks for Pediatric Refractory Epilepsy Classification Using Resting-State Functional Magnetic Resonance Imaging

Ryan D. Nguyen[1], Emmett H. Kennady[1], Matthew D. Smyth[5], Liang Zhu[4], Ludovic P. Pao[1], Shannon K. Swisher[1], Alberto Rosas[1], Anish Mitra[6], Rajan P. Patel[2], Jeremy Lankford[3], Gretchen Von Allmen[3], Michael W. Watkins[3], Michael E. Funke[3], Manish N. Shah[1]

■ OBJECTIVE: **This study aims to evaluate the performance of convolutional neural networks (CNNs) trained with resting-state functional magnetic resonance imaging (rfMRI) latency data in the classification of patients with pediatric epilepsy from healthy controls.**

■ METHODS: **Preoperative rfMRI and anatomic magnetic resonance imaging scans were obtained from 63 pediatric patients with refractory epilepsy and 259 pediatric healthy controls. Latency maps of the temporal difference between rfMRI and the global mean signal were calculated using voxel-wise cross-covariance. Healthy control and epilepsy latency *z* score maps were pseudorandomized and partitioned into training data (60%), validation data (20%), and test data (20%). Healthy control individuals and patients with epilepsy were labeled as negative and positive, respectively. CNN models were then trained with the designated training data. Model hyperparameters were evaluated with a grid-search method. The model with the highest sensitivity was evaluated using unseen test data. Accuracy, sensitivity, specificity, F1 score, and area under the receiver operating characteristic curve were used to evaluate the ability of the model to classify epilepsy in the test data set.**

■ RESULTS: **The model with the highest validation sensitivity correctly classified 74% of unseen test patients with 85% sensitivity, 71% specificity, F1 score of 0.56, and an area under the receiver operating characteristic curve of 0.86.**

■ CONCLUSIONS: **Using rfMRI latency data, we trained a CNN model to classify patients with pediatric epilepsy from healthy controls with good performance. CNN could serve as an adjunct in the diagnosis of pediatric epilepsy. Identification of pediatric epilepsy earlier in the disease course could decrease time to referral to specialized epilepsy centers and thus improve prognosis in this population.**

INTRODUCTION

Epilepsy is the most common chronic neurologic condition in children.[1] Approximately 1 of every 150 children younger than 10 years receives a diagnosis of epilepsy.[1] One third of children diagnosed with epilepsy do not respond to treatment with antiepileptic drugs alone.[2,3] Uncontrolled seizures in children negatively affect brain network development and contribute to severe developmental delay.[4,5] Two major epilepsy research forums in the United States and the European Union identified early and accurate diagnosis of epilepsy as a top priority for epilepsy research to address these irreversible developmental

Key words
- Convolutional neural networks
- Machine learning
- Pediatric epilepsy
- Pediatric refractory epilepsy
- Resting-state functional MRI
- Temporal latency

Abbreviations and Acronyms
BOLD: Blood-oxygen-level-dependent
CMHH: Children's Memorial Hermann Hospital
CNN: Convolutional neural network
EEG: Electroencephalography
MRI: Magnetic resonance imaging
PCC: Posterior cingulate cortex
rfMRI: Resting-state functional magnetic resonance imaging
SLCH: St. Louis Children's Hospital

From the Departments of [1]Pediatric Surgery and Neurosurgery, [2]Diagnostic and Interventional Imaging, and [3]Pediatric Neurology and [4]Biostatistics and Epidemiology Research Design Core, Institute for Clinical and Translational Sciences, McGovern Medical School at UTHealth, Houston, Texas; and [5]Department of Neurological Surgery and [6]Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, Missouri, USA

To whom correspondence should be addressed: Ryan D. Nguyen, B.S. [E-mail: ryan90nguyen@tamu.edu]

Citation: World Neurosurg. (2021) 149:e1112-e1122. https://doi.org/10.1016/j.wneu.2020.12.131

*Journal homepage: **www.journals.elsevier.com/world-neurosurgery***

*Available online: **www.sciencedirect.com***

impacts.[6,7] The pediatric epilepsy care paradigm involves comprehensive epilepsy specialty care after recognition that the patient's epilepsy is refractory to antiepileptic drugs. However, the timing of referral and seeking of comprehensive care is highly variable, as shown by the 19–20 years average time from seizure onset to epilepsy surgery referral for eligible patients.[6,8,9] Both research forums proposed changing the epilepsy care paradigm to include earlier referral to comprehensive epilepsy specialty care either at initial diagnosis or after the first antiepileptic drug has failed.[6,7] En masse, this paradigm change could lead to unnecessary costs because most children with an unprovoked seizure achieve remission without comprehensive care.[10] Therefore, early and accurate diagnosis is paramount to prevent overdiagnosis and underdiagnosis of epilepsy.

The diagnostic paradigm of suspected epilepsy includes history, interictal electroencephalography (EEG), and structural neuroimaging. An accurate history is the most inexpensive and useful method for diagnosing epilepsy. However, obtaining an accurate history from patients and parents can sometimes be a challenge. Interictal EEG is an inexpensive test for epilepsy; however, it can be an imperfect predictor of seizure recurrence. In patients with normal EEG results, the 2-year risk of seizure recurrence after first unprovoked seizure is as high as 27%, and in patients with epileptiform activity present on interictal EEG, the risk of recurrence is only 58%.[3] This situation can lead to overdiagnosis or underdiagnosis of epilepsy. In addition, not all patients with epilepsy have detectable structural abnormalities on neuroimaging. As a result, additional diagnostic tools are needed to aid in early and accurate epilepsy diagnosis to correctly identify patients who would benefit from comprehensive epilepsy care.

Machine learning–enhanced neuroimaging could serve as an adjunct in the diagnosis of pediatric epilepsy. Previous studies have shown that machine learning algorithms can successfully analyze large clinical data sets and make conclusive predictions in multiple diseases. Machine learning has recently been applied with magnetic resonance imaging (MRI) to assist in epilepsy classification.[11-13] Deep machine learning algorithms, specifically convolutional neural networks (CNNs), imitate neural processing to find features or patterns in training images to apply to the classification of unseen data. CNN has most notably shown success in image classification of diabetic retinopathy and malignant skin lesions.[14,15]

The usefulness of resting-state functional MRI (rfMRI) is an increasingly popular subject in the epilepsy research community. rfMRI can detect changes in brain network architecture that differentiate epileptic and nonepileptic patients.[16-18] Brain network architecture is commonly quantitatively analyzed by seed-based or component-based analysis of rfMRI blood-oxygen–level-dependent (BOLD) data. Recent studies have shown success using independent component analysis generated resting-state networks and machine learning methods to analyze epilepsy.[19-21] However, these methods require individual inspection of each component map and thus are labor intensive. Recently, voxel-wise temporal latency differences compared with the global mean BOLD signal have been suggested as a new way to characterize brain networks.[22,23] Altered temporal latency patterns on rfMRI has been shown to reflect changes in resting-state network architecture.[24] In previous studies, gross rfMRI latency changes were shown to be related to laterality of temporal and extratemporal epilepsies.[25,26] However, additional work is needed in this area, and seizure foci localization using rfMRI latency changes remains elusive.

In this study, we use CNN trained with rfMRI latency data to classify pediatric patients with epilepsy from healthy control individuals. To our knowledge, this is the first study using this novel application of CNN trained with rfMRI latency data to classify pediatric epilepsy.

METHODS

Data Acquisition

After institutional review board approval, rfMRI and anatomic MRI scans were reviewed from epilepsy and healthy control groups: 80 prospectively registered patients undergoing refractory epilepsy surgery from St. Louis Children's Hospital (SLCH) at Washington University in St. Louis, 5 retrospectively registered patients with refractory epilepsy from Children's Memorial Hermann Hospital (CMHH) in Houston, and 585 healthy control patients who were labeled as "typically developing" individuals by respective institutions in the multi-institution ADHD 200 data set.[27] Signed informed consent was acquired by each respective institution for all participating patients. Inclusion criteria were patients who were reviewed by multidisciplinary epilepsy conference at either SLCH or CMHH with available preoperative and postoperative structural and rfMRI. Patients diagnosed with extratemporal and temporal refractory epilepsy were included. Patients who did not meet these requirements or whose rfMRI data did not pass quality control were excluded from the study. After application of these criteria, 63 patients with epilepsy and 259 healthy control individuals were included. The average age of the epilepsy and healthy cohort was 14.48 $\pm$ 6.07 and 10.66 $\pm$ 2.65, respectively. The percentage of male patients in the epilepsy and healthy cohort was 65% and 48%, respectively.

rfMRI Processing Methods

The MRI collection methods varied between collection sites (http://fcon_1000.projects.nitrc.org/indi/adhd200/; **Supplementary Table 1**). MRI data were processed with Washington University Neuroimaging Laboratory Linux scripts, as described in previously.[26] BOLD sequences of controls and patients with epilepsy were registered to standard atlas volumetric sequences. Images were in addition processed with spatial smoothing using a Gaussian kernel of 6 mm full-width half-maximum temporal low-pass filtering >0.1 Hz, regression of nuisance waveforms, and zero-meaning of each voxel time course. Frames with excessive motion were excluded from analysis.[23] Latency maps for patients in both groups were generated by computing a voxel-wise lagged cross-covariance function. Lag, or latency, is the value at which the absolute cross-covariance function shows extremum between processed rfMRI BOLD signal and the whole-brain mean signal, which is then determined by parabolic interpolation.[23] The normal distributions of 36 seed regions in average latency maps of patients with epilepsy and healthy controls were then analyzed for overlap.[28] Latency z score maps of healthy controls and patients with epilepsy were created by voxel-wise z score

calculation using whole-brain healthy control mean and standard deviation latency maps as well as Fslmaths in the FSL suite (https://fsl.fmrib.ox.ac.uk/fslcourse/lectures/practicals/intro3/index.html).[26,29]

Data Partitioning

Before partitioning, data were stratified to ensure that equal proportions of healthy individuals and patients with epilepsy were represented in each phase of model training and evaluation. The 322 participants were then partitioned into training, validation, and test data sets, representing 60%, 20% and 20% of the data, respectively.

CNN Data Processing

A three-dimensional CNN was created using Python 3.6.3 with packages TensorFlow-CPU 1.3.0, nibabel 2.20, nilearn 0.4.0, numpy 1.13.3, pandas 0.20.3, and matplotlib 2.1.0. A multidimensional array size of number of patients by 3 (48 × 64 × 48 image array, 1 × 2 one-hot label array, and patient identification number). Healthy controls and patients with epilepsy were labeled as negative and positive classes, respectively. Data were visualized with Matplotlib Pyplot function (https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html) for quality control before and after multidimensional array creation.

CNN Architecture

Using TensorFlow, a CNN was created with 1 input layer, 2 convolutional layers with max pooling, a fully connected layer with dropout, and an output layer. The CNN architecture was similar to a standard MNIST (Modified National Institute of Standards and Technology) CNN example given in the TensorFlow tutorial that was adapted for a three-dimensional input.[30] In the next section, we discuss how CNN hyperparameters were tested to determine optimal settings. The architecture of our CNN models are as follows (see Figure 1 for three-dimensional diagram):

1) The first convolutional layer included thirty-two 3 × 3 × 3 feature filters with a 1 × 1 × 1 stride length. Weights were initialized using the Xavier method, and biases were initialized at zero.[31]
2) In convolution, each 3 × 3 × 3 filter scanned the whole input rfMRI latency z score image (48 × 64 × 48). Input rfMRI latency z score images were padded with zeros at the borders to maintain input image dimensions throughout convolution. The area scanned by the 3 × 3 × 3 feature filter was called a receptive field because at that instant, the program was processing only the part of the image in the 3 × 3 × 3 filter.
3) The double dot product was calculated using the rfMRI latency z score image values in the receptive window and weight value array of each filter (n = 32). A bias value was added to the double dot product and the final value was saved into a spatially indexed voxel in a resulting convolutional image (w*x+b; w = weight value array, x = receptive window array, b = bias value). Each voxel in the resulting convolutional image corresponded to a receptive window as the feature filter scanned the image. A rectified linear unit activation function was applied to the resulting convolutional activation image (n = 32, 48 × 64 × 48).[32]
4) The resulting activation maps were then down sampled using max pooling where another 2 × 2 × 2 window scanned the activation images at a 2 × 2 × 2 stride length. As the max pooling window scanned the activation image, the maximum value was scanned into a spatially corresponding voxel in a down sampled activation image with half the dimensions of the original activation image (n =

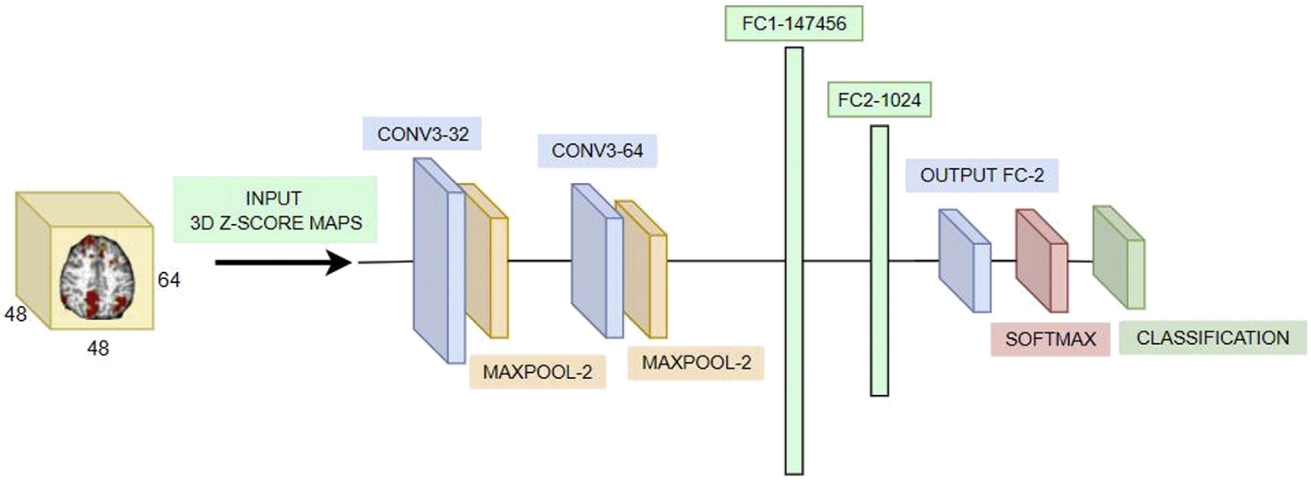

Figure 1. Three-dimensional convolutional neural network architecture. Input image (dimension 48 × 64 × 48) undergoes convolution using thirty-two 3 × 3 × 3 filters (*CONV3-32*) and then max pooling using a 2 × 2 × 2 filter (*MAXPOOL-2*). The resulting layers undergo further convolution using sixty-four 3 × 3 × 3 filters (*CONV3-64*) and additional max pooling using 2 × 2 × 2 filters (*MAXPOOL-2*). The result is flattened into a 1 × 147456 layer (*FC1-147456*), connected to a fully connected layer of 1 × 1024 (*FC2-1024*) and connected to an output layer (*OUTPUT FC-2*). A SoftMax activation function is applied to the 2 output nodes (*SOFTMAX*) to obtain a final classification decision (*CLASSIFICATION*).

32, 24 × 32 × 24). The down sampled activation image represented condensed spatial feature information for the first set of features.

5) This process of convolution and max pooling was repeated with the previous down sampled images to scan for higher-level features (n = 32, 24 × 32 × 24). This strategy created the final set of down sampled images (n = 64, 12 × 16 × 12).

6) The final set of down sampled activation images were then flattened into one dimension (1 × 147456), where each node represented a high-level feature.

7) This flattened image was then used as the input for a fully connected layer of size 1 × 1024, where each node to node connection was modeled linearly as summarized in equation 1, where w is the weight value of node to node connection, x is the fully connected node value, b is the bias value of node to node connection, and Y is the activation function for each node.

$$Y = \sum_{n} (w_n \times x_n) + b \quad (1)$$

Like a neural network, each fully connected node remained connected to all nodes in the flattened layer. The sum of all the contributions from the flattened layer (presynaptic neurons) then determined if the fully connected node (postsynaptic neuron) fired by using rectified linear unit.[32]

8) Each fully connected node was linked to the output layer, where each node to node connection was again modeled linearly. Each fully connected node was also connected to every node in the output layer (n = 2). All contributions from the fully connected node were put into a SoftMax activation function that generated a probability value for the 2 output nodes (1 epilepsy node and 1 healthy node).[31] The probability value was used to classify the participants as patients with epilepsy or healthy controls.

9) Each image was fed into the CNN algorithm in batch size of 1 for training. The image was processed by each layer of the CNN architecture to produce a 1 × 2 matrix of 2 probability values for healthy controls and patients with epilepsy. A SoftMax cross entropy function that outputs a cost or error value was then used to compare the 1 × 2 matrix with the ground truth 1 × 2 label matrix. This cost or error value was used to estimate how incorrect the predicted results/labels of the model were compared with the true label.

10) The algorithm then updated all weights and biases in the model through a process called back propagation. The amount added or subtracted from the weights and biases was determined by the calculated cost value referenced in step 9 and Adam, an adaptive stochastic gradient descent meta-algorithm.[33]

11) This process was repeated for each subject in the training data set.

These steps constitute a full training cycle or epoch, where each patient's rfMRI latency z score image was fed through the CNN architecture in forward propagation, and the CNN model was modified through weight and bias updates in back propagation to better model the training data with each epoch.

During training, 50%–90% of flattened layer to fully connected layer connections in steps 7 and 8 were shut off. This process is known as dropout. When nodes were shut off, weights and biases associated with the shut-off node were not considered in the classification decision and were not updated in back propagation for the corresponding training cycle. Dropout prevented fully connected nodes from overlearning from the training data, which improved the overall generalizability of the model.[34]

CNN Hyperparameter Testing

CNN has many hyperparameters that affect how the algorithm learned from the data. We explored how 3 of these hyperparameters affected the performance of our CNN models.

Learning Rate. Learning rate affected how quickly the model learned from the training examples. During stochastic gradient descent, the meta-algorithm tried to find the point of least error, global minima, in a multidimensional space of weight and bias values. Learning rate was used to determine the step size of the meta-algorithm. The larger the learning rate, the larger the step size toward the global minima. Larger step sizes led to quicker training times and may have allowed the meta-algorithm to escape local minima. However, larger step sizes may have also caused the meta-algorithm to skip the global minima altogether. We explored a common range of learning rates, including 1e-4, 1e-5, 1e-6, and 1e-7.[35]

Dropout Rate/Regularization. Dropout determined what percentage of fully connected nodes was shut off for a training cycle. The nodes that were affected by dropout were randomly chosen for each training cycle. The nodes that remained fully connected were used for classification during forward propagation, and weights and biases for remaining nodes were updated during back propagation. Dropout improved model generalizability by preventing nodes from overlearning the training data. The aggregation of randomly pruned neural networks in each training cycle could theoretically be analogous to ensemble learning.[34] Previous studies have suggested an optimal dropout rate of 50%.[34,36] However, because of the small size of our data, we explored dropout rates of 0.9, 0.8, 0.7, 0.6, and 0.5 to compensate for expected overfitting.

Epoch. Epoch was used to decide the number of training cycles. Each model was trained for 1–200 epochs with step size of 1. The epoch with the best results for each model was selected for further evaluation.

Model Selection

Models were primarily evaluated based on sensitivity. Using grid searching, CNN was trained using hyperparameter combinations represented by a matrix or grid which corresponds with the number of learning rates, dropout rates, and epochs for a total of 4000 different models.[37] After validation, the CNN model that

yielded the highest sensitivity was used to evaluate the final unseen test data. The difference between validation and test data was evaluated for generalizability. A validation-test difference of 10% was determined to be sufficient.

Subgroup Analysis

After final model selection, subgroup analysis was performed based on institution, age, seizure cause, and surgical outcome to determine if there were significant differences in model performance within these subgroups. During subgroup analysis by age, patients younger or at the mean age of the study population were compared with patients older than mean age because of limited sample size in certain ages. Subgroup analysis of surgical outcome was determined by Engel classification. Engel classification qualifies the patient's surgical outcome based on the extent of seizure freedom achieved.[38] Class I indicates complete seizure freedom.[38] Class II indicates rare disabling seizures.[38] Class III indicates the presence of worthwhile improvement (seizure reduction), and class IV indicates no worthwhile improvement.[38] Seizure cause was classified in 5 different categories: cortical malformations, inflammation, stroke, tumor, and no pathology information available. To determine if there were significant differences in model performance, pairwise comparisons within subgroupings were performed using logistic regression using a generalized linear model in R.

Subgroup analysis was only performed on validation and test data sets. Performance on training data set patients should theoretically be high and would not be an accurate indicator of potential bias of machine learning models trained on the same data.

RESULTS

General demographics of the study population are shown in Table 1. Mean and median age of the study population (epilepsy and healthy control) was 11.4 and 11 years, respectively. Additional information on patients with epilepsy is detailed in Table 2, including cause, use of sedation during imaging acquisition, laterality of focal epilepsy determined by postoperative imaging if applicable, and Engel class outcome of surgery.

The overlap of the control and epilepsy normal distributions of latency z score in 36 seed regions was compared, with a range of 0.44–0.58. Figure 2 shows the overlap of distributions in healthy and control groups for the posterior cingulate cortex (PCC) as well as representative axial slices of average latency maps for healthy controls and patients with epilepsy. The CNN model with the best validation sensitivity had the following hyperparameters: dropout rate of 50%, learning rate of 1e-4, and 181 training epochs. This performance of the CNN model classifying both validation and test data sets is shown in Table 3. The receiver operating characteristic curve for both validation and testing data sets is shown in Figure 3.

Subgroup analysis results of model performance by institution, age, cause, and surgical outcome are listed in Tables 4–7, respectively. There was no significant difference in model performance by age and cause. Model performance on data obtained at the Kennedy Krieger Institute and the University of Pittsburg were significantly different from data obtained at other institutions. Model performance on data with surgical outcome Engel classification III, worthwhile improvement, was significantly different from data with other surgical outcomes, but there were only 3 patients with an Engel classification III surgical outcome.

Table 1. Epilepsy and Cohort Demographic Data

Characteristic	Study Cohort: With Epilepsy	Study Cohort: Without Epilepsy	*P* Value
Number of patients (%)	63 (19.6)	259 (80.4)	
Gender, n (%)			
Men	41 (65.1)	125 (48.3)	0.02 (0.0166)
Women	22 (34.9)	134 (51.7)	0.02 (0.0241)
Age (years), mean (standard deviation)	14.5 (6.07)	10.7 (2.65)	<0.0001

P values were obtained by 2-tailed *t* test (age) and χ^2 test (gender).

Table 2. Epilepsy Cohort Demographic Data

Cause	Number of Patients
Cortical malformation	15
Inflammation	11
Stroke	6
Tumor	7
No pathology information available	24
Imaging acquisition sedation use	
Sedated	36
Not sedated	22
No information available	5
Laterality	
Right	28
Left	17
Not applicable*	13
No information	5
Surgery outcome (Engel class)	
I	27
II	3
III	8
IV	4
No information	16

*Not applicable refers to nonfocal epilepsy in which a corpus callosotomy was performed.

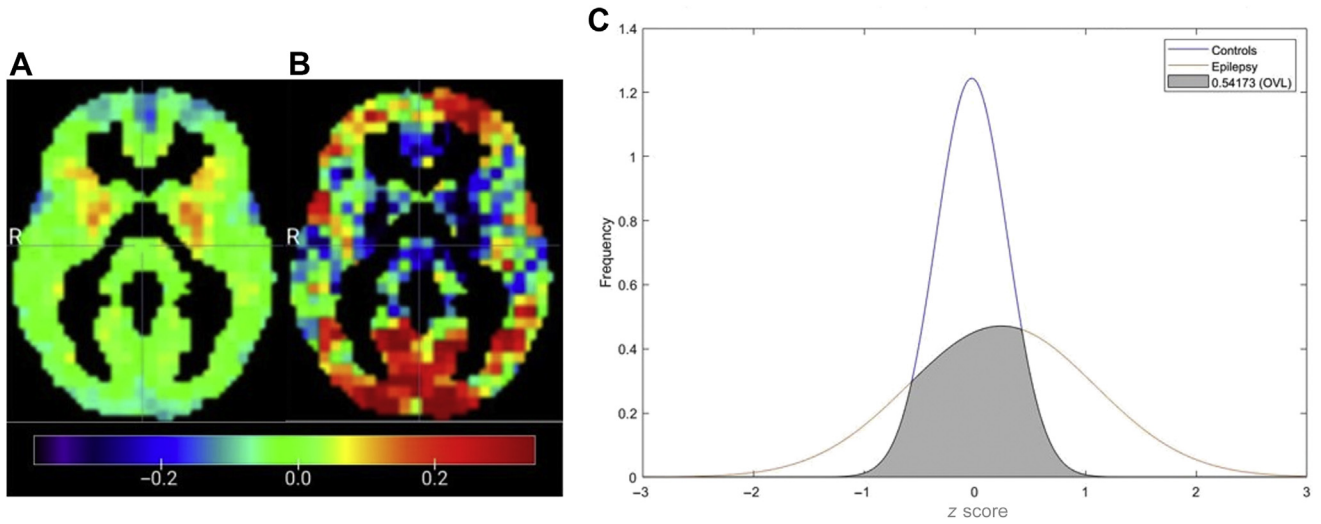

Figure 2. Analysis of healthy control and epilepsy group distribution overlap in the posterior cingulate cortex. (**A**) A representative axial slice of the mean healthy control temporal latency *z* map. (**B**) A representation axial slice of the mean epilepsy temporal latency z map. (**C**) Overlap coefficient (OVL) map of a standard posterior cingulate cortex seed showing adequacy of *t* testing between control and epilepsy groups.

DISCUSSION

There is a highly variable time to referral to comprehensive epilepsy centers in pediatric patients with epilepsy who need specialized care. During this time, these patients are at risk of poor development and sudden death.[39,40] To enable early referral to comprehensive specialized epilepsy care, additional diagnostic tools are essential to improve early diagnosis of epilepsy. CNN, a deep learning algorithm commonly used in imagery analysis, uses receptive fields to scan for features. With each training cycle, CNN learns which features best inform the model on the specific classification task at hand through weight and bias updates in back propagation. CNN has been used to predict epileptic seizures with EEG data.[2,41,42] Altered temporal latency patterns, seen in rfMRI, have been proposed to reflect changes in resting-state network architecture.[22-24] Previous studies[25,26] have shown that latency changes were related to laterality of epileptogenic foci in temporal and extratemporal epilepsy. To

Table 3. Comparison of Validation, Test, and Validation-Test Difference Results for Convolutional Neural Network with Respect to Area Under Curve, Accuracy, Sensitivity, Specificity, and F1 Score

Method	Area Under Curve	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score
Convolutional neural network					
Validation	0.90	0.78	0.92	0.75	0.61
Test	0.86	0.74	0.85	0.71	0.56
Validation-Test Difference	0.04	0.04	0.07	0.04	0.05

Lower validation-test difference may suggest better model generalizability.

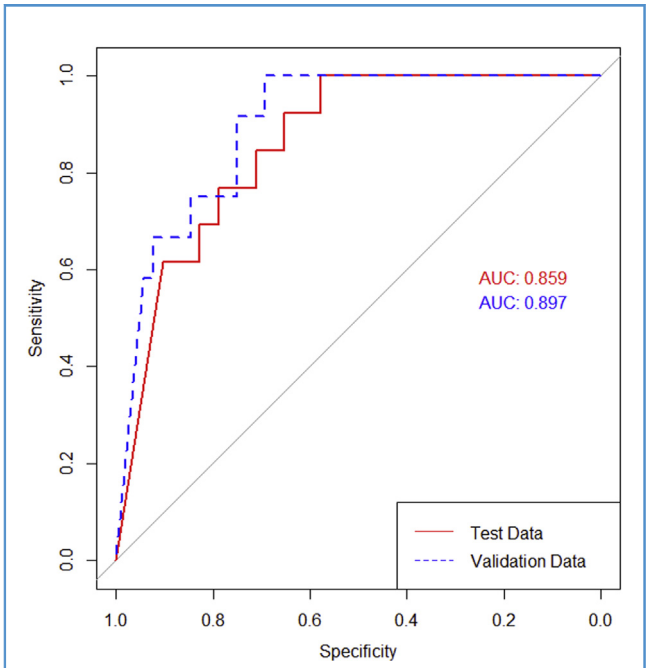

Figure 3. Validation and test data receiver operating characteristic (ROC) curve for convolutional neural network (CNN) model. *Blue dashed line* represents ROC of the CNN model on the validation data set with area under the curve (AUC) of 0.897. *Red solid line* represents ROC of the CNN model on the testing data set with area under the curve of 0.859.

Table 4. Subgroup Analysis by Institution

Institution	Number of Patients	Accuracy (%)
Epilepsy group St. Louis Children's Hospital	23	86.95
Epilepsy group Children's Memorial Hermann Hospital	2	100.0*
Kennedy Krieger Institute	14	28.57†
NeuroImage sample	3	100.0*
New York University Child Study Center	17	94.12
Oregon Health Science University	19	94.73
Peking University	27	92.59
University of Pittsburg	11	9.091†
Washington University in St. Louis (control group St. Louis Children's Hospital)	13	73.08

Number of patients included only patients present in validation and test data sets. Patients in the training data sets were not included in this analysis.
*Sample size at these institutions was too small, and model performance was not significantly different from any other institutions, including institutions, because of low power.
†Model performance on data obtained at these institutions was significantly different from other institutions determined by logistic regression.

our knowledge, CNN has not been used to classify epilepsy using rfMRI temporal latency data. In this study, we evaluate the ability of rfMRI latency-trained CNN to classify pediatric patients with epilepsy from healthy controls.

The overlap of control and epilepsy latency z score normal distributions in 36 seed regions ranged from 0.44 to 0.58, indicating that there was a difference in latency z scores between control and epilepsy groups. We hypothesized that CNN could detect these differences in rfMRI latency data to classify patients with epilepsy from healthy controls. Overlap for the PCC seed region was of particular interest because the PCC has been shown to be a driver of the default mode network, and PCC-network dysfunction has been implicated in many neurologic diseases.[17,43-45]

Before testing the final data set, we evaluated 4000 CNN models using different combinations of parameters using validation data

Table 5. Subgroup Analysis by Age

Group	Patient Age (years)	Number of Patients	Accuracy (%)
Epilepsy	≤11	7	85.71
	>11	18	88.89
Control	≤11	69	72.46
	>11	35	74.29

Number of patients included only patients present in validation and test data sets. Patients in the training data sets were not included in this analysis. Model performance difference between subgroups was not significant.

Table 6. Subgroup Analysis by Cause

Cause	Number of Patients	Accuracy (%)
Cortical malformation	5	80.00
Inflammation	5	100.0
Stroke	1	100.0
Tumor	5	100.0
No pathology available	9	77.78

Number of patients included only patients present in validation and test data sets. Patients in the training data sets were not included in this analysis. Model performance difference between subgroups was not significant.

(refer to sections on "Learning Rate," "Dropout Rate," and "Epochs for Hyperparameter Ranges"). The CNN model with the best validation sensitivity used a dropout rate of 50% and a learning rate of 1e-4 and was trained for 181 epochs. The CNN model classified 78% of patients in the validation data set correctly with a 92% sensitivity, 75% specificity, area under the receiver operating characteristic curve of 0.90, and F1 score of 0.61. After the unseen test results were classified, the same CNN model classified 74% of patients in the test data set correctly with an 85% sensitivity, 71% specificity, area under the receiver operating characteristic curve of 0.86, and F1 score of 0.59. The difference between classification results using test and validation data indicates the presence of some potential overfitting, but performance differences classifying validation and test data were <10%. There is no standard validation-test difference to our knowledge that indicates adequate model generalizability because it may differ between applications. We determined for this study that <10% difference would be acceptable. The sensitivity of the CNN models was of highest interest, indicating how well the models classified the epilepsy group (positive label, $n = 63$). Machine learning models tend to perform better on the more common classification case, which in our study was the healthy controls (negative label, $n = 259$). As a result, we chose to use sensitivity to evaluate our models after validation before classification of unseen test data. Therefore, the relatively high sensitivity of 92% and 85% for validation and test data sets, respectively, was encouraging.

Table 7. Subgroup Analysis by Surgical Outcome

Surgical Outcome (Engel Class)	Number of Patients	Accuracy (%)
I	14	92.86
II	4	100.0
III	3	33.33*
IV	2	100.0
Not available	2	100.0

Number of patients included only patients present in validation and test data sets. Patients in the training data sets were not included in this analysis.
*Model performance on data with surgical outcome was significantly different from other institutions determined by logistic regression.

Subgroup analysis showed that model performance was not influenced by patient age determined by comparing performance of the model on patients older than 11 years, the mean age of the study group, compared with patients younger than or aged 11 years (see Table 5). The model incorrectly predicted only 3 of 22 patients with epilepsy in the test ($n = 2$) and validation data sets ($n = 1$). All 3 patients were from the SLCH epilepsy group. One patient's epilepsy was caused by a cortical malformation with an Engel class III surgical outcome. The other 2 patients had no pathology information available and had an Engel class of I and III. Overall, the model performed uniformly well on all epilepsy subgroups (cause and surgical outcome), with the exception of Engel class III. The model predicted only 1 of 3 patients with surgical outcome of Engel class III correctly, but there was only a small sample of these patients for subgroup analysis (see Table 7). Subgroup analysis by patient institution showed that patient data obtained from the Kennedy Krieger Institute and the University of Pittsburg were classified at a significantly lower rate than were data obtained from other institutions (see Table 4). Initially, we believed that these differences in model performance on data from these institutions could be caused by differences in the scanner used at these institutions. The scanner used at the Kennedy Krieger Institute was a 3T Phillips scanner compared with 7 of 9 institutions that used a 3-T Siemens scanner (6 of 9 used a 3-T Siemens Trio scanner specifically). See Supplementary Table 1 for full scanner details by institution. However, this explanation is less likely, because patients obtained from the University of Pittsburg were imaged on a 3-T Siemens Trio scanner like most patients in the study population. The sample size for CMHH and the NeuroImage sample were only 2 and 3 patients, respectively. Although the model correctly classified all patients from these institutions, the sample size was too low to show any significant differences with the Kennedy Krieger Institute and the University of Pittsburg.

Our study was limited primarily by the relatively small amount of rfMRI data for patients with epilepsy compared with rfMRI data for healthy controls. As previously discussed, machine learning models trained with unbalanced data (more training data on one class compared with another) typically perform better on the more represented data class. In this study, healthy controls made up approximately 80% of the data. Therefore, we expected all of our models to have higher specificity compared with sensitivity, which was observed in some trained models. Our final model correctly predicted 85% of patients with epilepsy in the test data set, and 92% of patients with epilepsy in the validation data set. However, this finding is likely to the result of using sensitivity as the criteria for selecting a model for evaluation of final unseen test data. The relatively low difference between validation and test results and the relatively high sensitivity of the model on the test data are encouraging signs that high sensitivity is generalizable to new data.

In addition, our smaller epilepsy data set limits our ability to train CNN models to classify subgroups (i.e., surgical outcomes and cause) to provide additional prognostic information to clinicians. This situation is to the result of a combination of the previously mentioned limitations of data size and bias in unbalanced data sets. To perform subgroup classifications with CNN, our study population would be limited to only patients with epilepsy ($n = 63$). In addition, some patients with epilepsy had incomplete data, further reducing our study population for subgroup analysis. In the case of Engel class outcomes, outcomes were greatly skewed to Engel class I in the data set (I, $n = 29$; II, $n = 7$; III, $n = 5$; and IV, $n = 4$). As a result, the models we trained tended to incorrectly classify most patients as Engel class I. Similar results were observed when attempting to classify other subgroups. As a result, the scope of this study was limited to binary classification of healthy controls and patients with epilepsy.

In our current data set, we include only pediatric patients with refractory epilepsy and healthy control patients. Therefore, classification of refractory epilepsy from nonrefractory epilepsy was not studied. We are actively collecting data of pediatric patients with nonrefractory epilepsy to perform this analysis in the future. Although, we believe that the CNN model in this study could be a helpful adjunct, a model that can also classify patients with nonrefractory epilepsy is desirable.

In this study, we used a simple CNN architecture. Many different CNN architectures have shown higher performance compared with simple architectures such as the one used in this study.[46,47] We chose to implement a simple CNN architecture, to show the power of the algorithm with the current data set. As we gather more data, we will also explore more complex architectures. To address the fact that deep learning requires large quantities of data, we can also explore transfer learning. In transfer learning, models previously trained on large quantities of unrelated data are trained with a new smaller data set of interest.

We used Texas Advanced Computing Center's Stampede2 supercomputer to train and evaluate our CNN models using serial computation and a basic configuration. Each model was trained and evaluated over 200 epochs, which took 48 hours and was the maximum run time on a computing node on Stampede2. In future studies, we would try to increase the efficiency of the program through parallelizing bottlenecks in the code. This strategy will be especially important as we accumulate more patients in our multiinstitutional efforts, which will increase processing time as well. Increasing the number of training cycles may increase the performance of our algorithm as well as that of many of the models with lower learning rates that may require more epochs to converge to a global minimum.

We are obtaining more epilepsy data from Children's Memorial Hermann Hospital and other institutions to include in our data set. With more data and exploration of transfer learning, we will be able to address many of our limitations and perform additional clinically relevant subgroup classifications. In future studies, we will explore more complex CNN architectures, which may improve model performances as well as allow classification of individual voxels as epileptic foci.[48-50]

CONCLUSIONS

CNNs trained on rfMRI temporal latency images were able to classify pediatric patients with epilepsy from healthy controls. This method may become a useful adjunct to help reduce the delay of referral to specialized epilepsy centers. As an image-based test, rfMRI can be performed at most imaging centers. The data could be processed at specialized centers, triggering referrals for

patients at high risk for epilepsy for earlier specialized care. Continued data acquisition will improve the performance of this scalable algorithm. In addition, more data will allow additional clinically valuable classifications including subgroup analysis and refractory versus nonrefractory epilepsy.

DATA AVAILABILITY STATEMENT

The control data set analyzed in this study can be found in the ADHD-200 repository hosted by the 1000 Functional Connectome Project (http://fcon_1000.projects.nitrc.org/indi/adhd200/). Patient and program information may be obtained on request with institutional review board approval.

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

Ryan D. Nguyen: Writing - original draft, Methodology, Investigation, Formal analysis, Software, Data curation, Validation, Project administration. **Emmett H. Kennady:** Writing - original draft, Formal analysis, Data curation. **Matthew D. Smyth:** Funding acquisition, Writing - review & editing, Methodology, Validation. **Liang Zhu:** Formal analysis, Software, Data curation, Validation. **Ludovic P. Pao:** Writing - review & editing, Methodology, Formal analysis. **Shannon K. Swisher:** Writing - original draft, Project administration. **Alberto Rosas:** Writing - original draft, Software, Data curation. **Anish Mitra:** Writing - review & editing, Methodology, Formal analysis. **Rajan P. Patel:** Funding acquisition, Writing - review & editing. **Jeremy Lankford:** Funding acquisition, Writing - review & editing. **Gretchen Von Allmen:** Funding acquisition, Writing - review & editing. **Michael W. Watkins:** Funding acquisition, Writing - review & editing. **Michael E. Funke:** Funding acquisition, Writing - review & editing. **Manish N. Shah:** Funding acquisition, Writing - review & editing, Methodology, Investigation, Resources, Supervision, Project administration.

ACKNOWLEDGMENTS

The authors would like to thank Marc E. Raichle, M.D. and Abraham Z. Snyder, Ph.D., M.D. for their advice and support in rfMRI analysis. The authors would also like to thank and acknowledge the 1000 Functional Connectome Project, participating institutions, and associated funding sources for the use of the ADHD 200 data set. In addition, the authors would like to thank the Texas Advanced Computing Center for their advice, support, and use of Stampede2 supercomputers in the training and evaluation of the CNN hyperparameter grid searches.

REFERENCES

1. Aaberg KM, Gunnes N, Bakken IJ, et al. Incidence and prevalence of childhood epilepsy: a nationwide cohort study. *Pediatrics*. 2017;139:e20163908.
2. Kwan P, Brodie MJ. Early identification of refractory epilepsy. *N Engl J Med*. 2000;342:314-319.
3. Smith SJ. EEG in the diagnosis, classification, and management of patients with epilepsy. *J Neurol Neurosurg Psychiatry*. 2005;76(suppl 2):ii2-ii7.
4. Mihara T, Inoue Y, Matsuda K, et al. Recommendation of early surgery from the viewpoint of daily quality of life. *Epilepsia*. 1996;37(suppl 3):33-36.
5. Westerveld M, Sass KJ, Chelune GJ, et al. Temporal lobectomy in children: cognitive outcome. *J Neurosurg*. 2000;92:24-30.
6. Berg AT, Baca CB, Loddenkemper T, Vickrey BG, Dlugos D. Priorities in pediatric epilepsy research: improving children's futures today. *Neurology*. 2013;81:1166-1175.
7. Baulac M, de Boer H, Elger C, et al. Epilepsy priorities in Europe: a report of the ILAE-IBE Epilepsy Advocacy Europe Task Force. *Epilepsia*. 2015;56:1687-1695.
8. Dlugos DJ. The early identification of candidates for epilepsy surgery. *Arch Neurol*. 2001;58:1543-1546.
9. Gilliam F, Kuzniecky R, Meador K, et al. Patient-oriented outcome assessment after temporal lobectomy for refractory epilepsy. *Neurology*. 1999;53:687-694.
10. Guerrini R. Epilepsy in children. *Lancet*. 2006;367:499-524.
11. Del Gaizo J, Mofrad N, Jensen JH, et al. Using machine learning to classify temporal lobe epilepsy based on diffusion MRI. *Brain Behav*. 2017;7:e00801.
12. Rajpoot K, Riaz A, Majeed W, Rajpoot N. Functional connectivity alterations in epilepsy from resting-state functional MRI. *PLoS One*. 2015;10:e0134944.
13. Torlay L, Perrone-Bertolotti M, Thomas E, Baciu M. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. *Brain Inform*. 2017;4:159-169.
14. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118.
15. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402-2410.
16. Jirsa VK, Proix T, Perdikis D, et al. The virtual epileptic patient: individualized whole-brain models of epilepsy spread. *Neuroimage*. 2017;145(pt B):377-388.
17. Pizoli CE, Shah MN, Snyder AZ, et al. Resting-state activity in development and maintenance of normal brain function. *Proc Natl Acad Sci U S A*. 2011;108:11638-11643.
18. Proix T, Bartolomei F, Guye M, Jirsa VK. Individual brain structure and modelling predict seizure propagation. *Brain*. 2017;140:641-654.
19. Bharath RD, Panda R, Raj J, et al. Machine learning identifies "rsfMRI epilepsy networks" in temporal lobe epilepsy. *Eur Radiol*. 2019;29:3496-3505.
20. Boerwinkle VL, Mirea L, Gaillard WD, et al. Resting-state functional MRI connectivity impact on epilepsy surgery plan and surgical candidacy: prospective clinical work. *J Neurosurg Pediatr*. 2020:1-8.
21. Boerwinkle VL, Mohanty D, Foldes ST, et al. Correlating resting-state functional magnetic resonance imaging connectivity by independent component analysis-based epileptogenic zones with intracranial electroencephalogram localized seizure onset zones and surgical outcomes in prospective pediatric intractable epilepsy study. *Brain Connect*. 2017;7:424-442.
22. Mitra A, Snyder AZ, Blazey T, Raichle ME. Lag threads organize the brain's intrinsic activity. *Proc Natl Acad Sci U S A*. 2015;112:E2235-E2244.
23. Mitra A, Snyder AZ, Hacker CD, Raichle ME. Lag structure in resting-state fMRI. *J Neurophysiol*. 2014;111:2374-2391.
24. Mitra A, Snyder AZ, Constantino JN, Raichle ME. The lag structure of intrinsic activity is focally altered in high functioning adults with autism. *Cereb Cortex*. 2017;27:1083-1093.
25. Shah MN, Mitra A, Goyal MS, et al. Resting state signal latency predicts laterality in pediatric medically refractory temporal lobe epilepsy. *Childs Nerv Syst*. 2018;34:901-910.
26. Shah MN, Nguyen RD, Pao LP, et al. Role of resting state MRI temporal latency in refractory pediatric extratemporal epilepsy lateralization. *J Magn Reson Imaging*. 2019;49:1347-1355.
27. Mennes M, Biswal BB, Castellanos FX, Milham MP. Making data sharing work: the FCP/INDI experience. *Neuroimage*. 2013;82:683-691.
28. Kowerko D. *calc_overlap_twonormal(s1,s2,mu1,mu2,xstart,xend,xinterval)*. *MathWorks*. MathWorks; 2015.

https://la.mathworks.com/matlabcentral/fileexchange/49823-calc_overlap_twonormal-s1-s2-mu1-mu2-xstart-xend-xinterval?s_tid=FX_rc2_behav. Accessed May 4, 2020.

29. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *Neuroimage*. 2012; 62:782-790.

30. Abadi M, Barham P, Chen J, et al. *TensorFlow: a system for large-scale machine learning. Paper presented at: 12th USENIX Conference on Operating Systems Design and Implementation*. USENIX Association; November 2-4, 2016. Savannah, GA, USA.

31. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *J Mach Learn Res*. 2010;9:249-256.

32. Nair V, Hinton GE. *Rectified linear units improve restricted Boltzmann machines. Paper presented at: 27th International Conference on International Conference on Machine Learning*. ICML; June 21-24, 2010. Haifa, Israel.

33. Kingma DP, Ba J. *Adam. A Method for Stochastic Optimization. Paper presented at: International Conference on Learning Representations*. San Diego: dblp; May 7-9, 2015.

34. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929-1958.

35. Bottou L. Stochastic gradient descent tricks. *Lect Notes Comput Sci*. 2012;7700:430-445.

36. Baldi P, Sadowski P. Understanding dropout. *Advances in Neural Information Processing Systems 26*. December 11-15, 2013.

37. Chicco D. Ten quick tips for machine learning in computational biology. *BioData Mining*. 2017;10:35.

38. Engel J, Ness PV, Rasmussen T, L. O. *Surgical Treatment of the Epilepsies*. New York: Raven Press; 1993.

39. Elger CE, Helmstaedter C, Kurthen M. Chronic epilepsy and cognition. *Lancet Neurol*. 2004;3: 663-672.

40. Sperling MR. Sudden unexplained death in epilepsy. *Epilepsy Curr*. 2001;1:21-23.

41. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adeli H. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Comput Biol Med*. 2018;100:270-278.

42. Truong ND, Nguyen AD, Kuhlmann L, et al. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Netw*. 2018;105:104-111.

43. Fransson P, Marrelec G. The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: evidence from a partial correlation network analysis. *Neuroimage*. 2008;42:1178-1184.

44. Leech R, Sharp DJ. The role of the posterior cingulate cortex in cognition and disease. *Brain*. 2014;137:12-32.

45. Raichle ME. The brain's default mode network. *Annu Rev Neurosci*. 2015;38:433-447.

46. He K, Zhang X, Ren S, Sun J. *Deep residual learning for image recognition*. 2016 IEEE *Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, Nevada: IEEE; June 27-30, 2016:770-778.

47. Szegedy C, Wei L, Yangqing J, et al. *Going deeper with convolutions*. 2015 IEEE *Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, Massachusetts: IEEE; June 7-12, 2015:1-9.

48. Dong H, Yang G, Liu F, Mo Y, Guo Y. *Automatic brain tumor detection and segmentation using U-net based fully convolutional networks*. *21st Annual Conference, Medical Image Understanding and Analysis 2017*. Edinburgh, UK: Springer; July 11-13, 2017:506-517.

49. Kamnitsas K, Bai W, Ferrante E, et al. *Ensembles of multiple models and architectures for robust brain tumour segmentation*. *BrainLes 2017: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. New York City, New York: Springer, Cham; 2018: 450-462.

50. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015*. 2015;9351:234-241.

Conflict of interest statement: The authors declare that the article content was composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received 31 August 2020; accepted 24 December 2020

Citation: World Neurosurg. (2021) 149:e1112-e1122. https://doi.org/10.1016/j.wneu.2020.12.131

Journal homepage: ***www.journals.elsevier.com/world-neurosurgery***

Available online: ***www.sciencedirect.com***

1878-8750/$ - see front matter

SUPPLEMENTARY DATA

Supplementary Table 1. Healthy Control and Epilepsy Cohort Resting-State Functional Magnetic Resonance Imaging Parameters by Institution

Institution	Scanner	Repetition Time (milliseconds)	Echo Time (milliseconds)	Flip Angle (°)	Field of View (mm)	Number of Slices	Slice Thickness (mm)
Healthy control cohort							
Brown University	Siemens Trio 3 T	2000	25	90	192	35	3
Kennedy Krieger Institute	Philips 3 T	2500	30	75	256	47	3
NeuroImage	Siemens Avanto 1.5 T	1940	40	80	224	37	3
New York University Langone Medical Center	Siemens Allegra 3 T	2000	15	90	240	33	4
Oregon Health Science University	Siemens Trio 3 T	2500	30	90	240	36	3.8
Peking	Siemens Trio 3 T	2000	30	90	200	33	3.5
	Siemens Trio 3 T	2000	30	90	200	33	3
	Siemens Trio 3 T	2000	30	90	220	30	4.5
Pittsburg	Siemens Trio 3 T	1500	29	70	200	29	4
Washington University	Siemens Trio 3 T	2500	27	90	256	32	4
Epilepsy cohort							
Washington University	Siemens Trio 3 T	2070	25		256	36	
Memorial Hermann Hospital	Siemens Trio 3 T	3000	30	90	144	45	3