BRAIN
ORIGINAL ARTICLE

Redefining diagnostic lesional status in temporal lobe epilepsy with artificial intelligence

Ezequiel Gleichgerrcht,[1] Erik Kaestner,[2] Reihaneh Hassanzadeh,[3,4] Rebecca W. Roth,[1] Alexandra Parashos,[5] Kathryn A. Davis,[6] Anto Bagić,[7] Simon S. Keller,[8,9] Theodor Rüber,[10,11] Travis Stoub,[12] Heath R. Pardoe,[13,14] Patricia Dugan,[14] Daniel L. Drane,[1] Anees Abrol,[3] Vince Calhoun,[3,4] Ruben I. Kuzniecky,[15] Carrie R. McDonald[2,16] and Leonardo Bonilha[17]

Despite decades of advancements in diagnostic MRI, 30%–50% of temporal lobe epilepsy (TLE) patients remain categorized as 'non-lesional' (i.e. MRI negative) based on visual assessment by human experts. MRI-negative patients face diagnostic uncertainty and significant delays in treatment planning. Quantitative MRI studies have demonstrated that MRI-negative patients often exhibit a TLE-specific pattern of temporal and limbic atrophy that might be too subtle for the human eye to detect. This signature pattern could be translated successfully into clinical use via advances in artificial intelligence in computer-aided MRI interpretation, thereby improving the detection of brain 'lesional' patterns associated with TLE.

Here, we tested this hypothesis by using a three-dimensional convolutional neural network applied to a dataset of 1178 scans from 12 different centres, which was able to differentiate TLE from healthy controls with high accuracy (85.9% ± 2.8%), significantly outperforming support vector machines based on hippocampal (74.4% ± 2.6%) and whole-brain (78.3% ± 3.3%) volumes. Our analysis focused subsequently on a subset of patients who achieved sustained seizure freedom post-surgery as a gold standard for confirming TLE.

Importantly, MRI-negative patients from this cohort were accurately identified as TLE 82.7% ± 0.9% of the time, an encouraging finding given that clinically these were all patients considered to be MRI negative (i.e. not radiographically different from controls). The saliency maps from the convolutional neural network revealed that limbic structures, particularly medial temporal, cingulate and orbitofrontal areas, were most influential in classification, confirming the importance of the well-established TLE signature atrophy pattern for diagnosis. Indeed, the saliency maps were similar in MRI-positive and MRI-negative TLE groups, suggesting that even when humans cannot distinguish more subtle levels of atrophy, these MRI-negative patients are on the same continuum common across all TLE patients. As such, artificial intelligence can identify TLE lesional patterns, and artificial intelligence-aided diagnosis has the potential to enhance the neuroimaging diagnosis of TLE greatly and to redefine the concept of 'lesional' TLE.

1 Department of Neurology, Emory University, Atlanta, GA 30329, USA
2 Department of Radiation Medicine & Applied Sciences, University of California San Diego, San Diego, CA 92093, USA
3 Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA 30303, USA
4 School of Electrical and Computer Engineering (ECE), Georgia Institute of Technology, Atlanta, GA 30332, USA
5 Department of Neurology, Medical University of South Carolina, Charleston, SC 29425, USA
6 Department of Neurology, University of Pennsylvania, Philadelphia, PA 19104, USA

Received February 24, 2024. Revised October 24, 2024. Accepted November 26, 2024. Advance access publication January 22, 2025

7 Department of Neurology, University of Pittsburgh, Pittsburgh, PA 15213, USA
8 Department of Pharmacology and Therapeutics, University of Liverpool, Liverpool L69 7ZX, UK
9 The Walton Centre NHS Foundation Trust, Liverpool L9 7LJ, UK
10 Department of Neuroradiology, University Hospital Bonn, Bonn, Germany 53127
11 Department of Epileptology, University Hospital Bonn, Bonn, Germany 53127
12 Department of Neurological Sciences, Rush University, Chicago, IL 60612, USA
13 The Florey Institute of Neuroscience and Mental Health, Victoria VIC 3010, Australia
14 Department of Neurology, New York University Grossman School of Medicine, New York, NY 10017, USA
15 Department of Neurology, School of Medicine at Hofstra/Northwell, Hempstead, NY 10075, USA
16 Department of Psychiatry, University of California San Diego, San Diego, CA 92093, USA
17 Department of Neurology, University of South Carolina, Columbia, SC 29203, USA

Correspondence to: Ezequiel Gleichgerrcht, MD, PhD Emory University Brain Health Center,
12 Executive Park Drive, Suite 250, Atlanta, GA 30325, USA
E-mail: ezegleich@gmail.com

Keywords: machine learning; artificial intelligence; temporal lobe epilepsy; structural MRI

Introduction

MRI is integral in the diagnostic work-up of epilepsy patients, offering high sensitivity to focal neuropathologies such as hippocampal sclerosis, the predominant cause of temporal lobe epilepsy (TLE). Despite advancements in MRI technology and post-processing techniques,[1-7] ~30%–50% of TLE cases remain 'non-lesional',[8] showing no detectable abnormalities on MRI (i.e. MRI negative). This diagnostic ambiguity not only delays treatment[9] but also hinders healthcare planning and surgical evaluations, potentially diminishing the success rates of surgical interventions.[10,11]

The designation of lesional versus non-lesional TLE still remains almost exclusively dependent on the subjective qualitative interpretation of MRI by human experts. Except for hippocampal volumetry[12,13] and T2 relaxometry,[14,15] which are used routinely in only some centres, the neuroimaging diagnosis of TLE does not leverage other quantifiable aspects of imaging. This is in sharp contrast to neuroimaging research advances, which have consistently revealed a signature pattern of atrophy in TLE involving structures in the limbic regions,[16] notably the rhinal cortices,[17] the thalamus,[18,19] the anterior cingulate[20] and temporal neocortex.[21] Indeed, this pattern of quantifiable atrophy is present reliably across most patients with TLE,[16] whether MRI-positive or MRI negative, i.e. with or without visually detectable abnormalities.[22]

Although TLE is characterized by this consistent spatial distribution of brain changes, the individual variability impedes the direct application of neuroimaging research in clinical settings. This crucial gap might now be resolved with artificial intelligence (AI) applied to image interpretation. Given the growing availability of large multicentre studies, such as ENIGMA-Epilepsy[22-26] or Connectome Abnormalities to Predict Epilepsy Surgery (CAPES),[27] machine learning and out-of-sample prediction can be improved significantly to train models that are sensitive to crucial neuroimaging features. Recently, we demonstrated that convolutional neural networks (CNN) can distinguish TLE from controls using T1-weighted scans with specificity and sensitivity of 91% and 82%, respectively, including a high proportion of MRI-negative cases.[28] Furthermore, to confirm that CNNs are not simply detecting only non-specific limbic atrophy or age-related changes, but are sensitive to the signature pattern in TLE, we demonstrated that CNNs have an accuracy of 90% and precision of 86% in discriminating patients with TLE from Alzheimer's disease after controlling for age.[29]

Given the promising findings in TLE thus far, it is crucial to test whether AI can enhance the detection of MRI-negative cases by leveraging specific neuroanatomical patterns associated with TLE. If MRI-negative patients can be identified accurately by AI despite being labelled non-lesional by humans, this could be transformative for the diagnosis of TLE and could redefine our understanding of 'lesional' epilepsy. An AI-aided decision support tool could be impactful by changing the imaging status for many patients from MRI negative to MRI positive. Additionally, the analysis of saliency maps, highlighting key brain regions influencing model classification, offers an interpretable and consistent neuroanatomical framework for evaluation.

This study explores the utility of a machine learning classifier, specifically evaluating its performance and the generated saliency maps, using MRI data from 589 healthy controls (HC) and 589 TLE patients from across 12 cohorts, encompassing both MRI-negative and MRI-positive patients. We aimed to test the ability of the classifies to categorize MRI-negative patients accurately. Additionally, we compared the saliency maps of lesional and non-lesional TLE patients to ascertain whether the classifier used similar neuroanatomical features across these human-labelled groups. Importantly, our analysis included patients who attained surgical freedom post-surgery, providing a robust, human-independent gold standard to validate the TLE diagnosis.

Materials and methods

Figure 1 provides a general overview of the participants and methods used in this study. To discriminate between healthy controls (HC) and patients with TLE (see 'Participants' section) based on structural MRI scans (see 'Imaging acquisition and pre-processing' section), we trained an artificial neural network (see 'Artificial neural network' section). To ensure the gold standard of TLE, we focused the analyses on patients who underwent resection or ablation of the medial temporal region and who remained seizure free ≥1 year after surgery. This group provided the strongest available gold standard for confirmation of TLE and minimized the risk of TLE mimics, TLE plus syndromes, or multifocal disease. Using

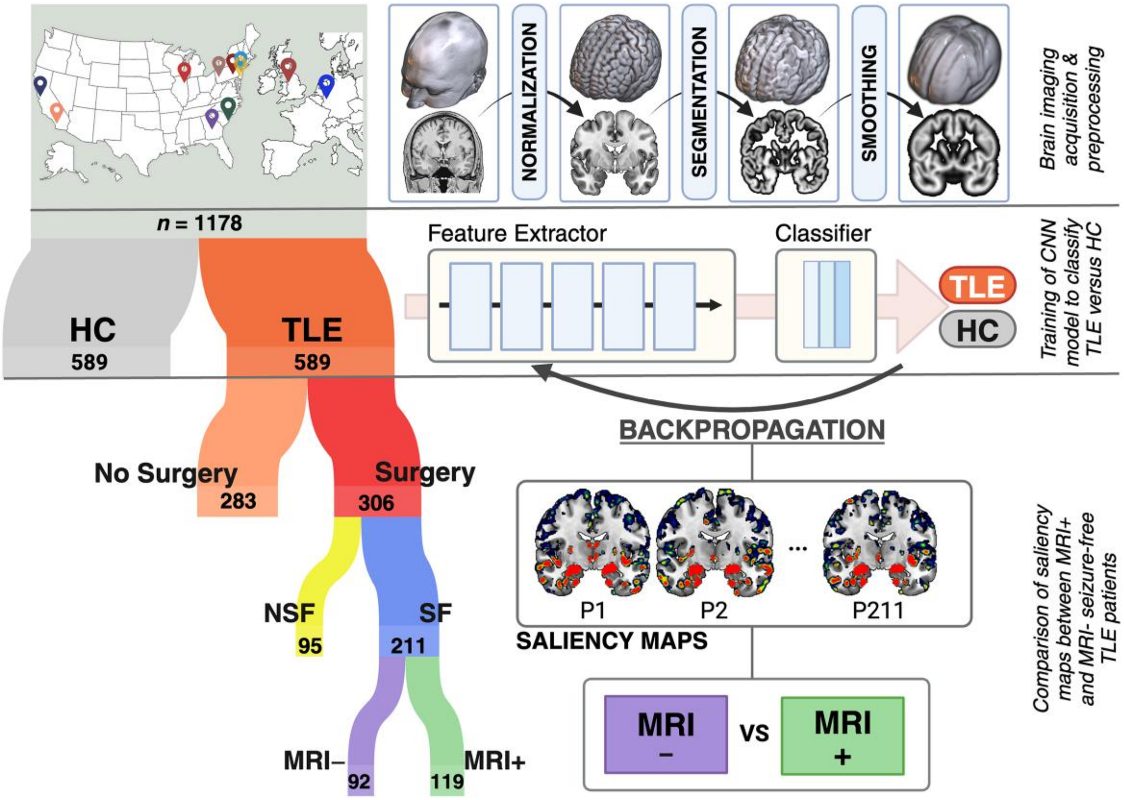

Figure 1 Schematic overview of the study population and methods. A large cohort of healthy controls (HC) and patients with temporal lobe epilepsy (TLE) was recruited from across 11 sites in the USA and Europe, in addition to the Human Connectome Project. Their T1-weighted MRI scans were normalized, segmented and smoothed, then entered into a three-dimensional convolutional neural network (CNN) model, which consisted of a feature extractor with five fully convolutional layers (ConvNet) followed by max pooling (excluding the fifth layer), in addition to a classifier with two outputs. Using vanilla backpropagation, we generated individual saliency maps for the TLE patients for whom seizure outcome was available ≥1 year after surgery. For this group, saliency maps were compared between those who had been deemed lesional (MRI+) or non-lesional (MRI–) by human experts.

backpropagation, we generated saliency maps at the individual level (see ‘Saliency maps’ section) to identify the brain regions that mostly contributed to the classification. These saliency maps were then compared between cases classified by human experts as lesional (MRI-positive) or non-lesional (MRI-negative) (see ‘Statistical analyses’ section).

Participants

The study encompassed 1178 participants from 12 cohorts, including Emory University, Medical University of South Carolina (MUSC), New York University (NYU), Northwell University, Rush University, University of Bonn, University of California San Diego (UCSD), University of California San Francisco (UCSF), University of Liverpool, University of Pennsylvania (Penn) and University of Pittsburgh (Pitt), in addition to the Human Connectome Project (HCP). The Institutional Review Board of each participating institution provided local approval for research studies in epilepsy in accordance with the Declaration of Helsinki. Table 1 summarizes the basic demographic information for each cohort.

Patients with temporal lobe epilepsy

Patients were included if they were ≥18 years old and had a diagnosis of unilateral drug-resistant TLE as defined by the International League Against Epilepsy.[30] Diagnostic confirmation was made at each site and involved at least clinical assessment, 3 T MRI and long-term video EEG monitoring. On an individual basis, supplemental neuroimaging techniques, such as PET and subtraction ictal single-photon emission computed tomography coregistered to MRI, in addition to intracranial EEG monitoring were obtained to clarify the epilepsy syndrome. Patients with the presence of other major neurological diseases, mass-occupying lesions or additional identified epileptic foci were excluded from the study. Patients with TLE ($n = 589$) were 54.8% female and 38.0 [standard deviation (SD) = 12.2] years old on average.

Surgical outcome data ≥1 year post-operatively were available for ~52% of the sample. Of the 306 patients who had undergone surgery, ~69% had remained seizure free. Of these 211 seizure-free patients, 56.4% ($n = 119$) had been deemed ‘MRI-positive’ and 43.6% ($n = 92$) had been labelled ‘MRI-negative’ by the treating site (Fig. 1; end branches of the participant tree). This ‘lesional’/‘non-lesional’ determination was the result of multidisciplinary patient conferences at each centre, where scans were reviewed by neurologists, neurosurgeons and neuroradiologists for the presence of radiographic signs of underlying hippocampal sclerosis or focal cortical dysplasia. Any cases with space-occupying lesions had been excluded prior to inclusion in the study.

Of the seizure-free patients, 32.3% had undergone laser ablation of hippocampal structures. Of the remainder of patients who underwent resection, 38.8% were selective amygdalohippocampectomies, and the rest were standard anterior temporal lobectomies. There were no significant differences between the distribution of

Table 1 Summary of demographic and clinical variables for temporal lobe epilepsy patients based on MRI lesional status

Variable	MRI negative ($n = 92$)	MRI positive ($n = 119$)	Statistical comparison
Age, years (mean ± SD)	37.1 (12.2)	38.1 (14.1)	$t(209) = 0.53$, $P = 0.6$
Sex, % male	44.6	51.2	$\chi^2 = 0.36$, $P = 0.55$
Age at onset, years (mean ± SD)	21.6 (13.9)	15.0 (10.8)	$t(209) = 167.2$, $P < 0.001$
Race, % C/W:% AA/B:% A/PI	69.8:30.2:0	74.1:24.1:1.8	$\chi^2 = 1.2$, $P = 0.55$
History of TBI, %	21.7	11.1	$\chi^2 = 1.87$, $P = 0.17$
History of febrile seizures, %	19.6	20.5	$\chi^2 = 0.16$, $P = 0.90$
Side of TLE, % left	39.5	40.1	$\chi^2 = 1.8$, $P = 0.18$
Surgery type, % resection	68.5	65	$\chi^2 = 0.21$, $P = 0.89$

AA/B = African American/Black; A/PI: Asian/Pacific Islander; SD = standard deviation; C/W = Caucasian/White; TBI = traumatic brain injury; TLE = temporal lobe epilepsy.

ablative versus resective surgeries in each seizure-free TLE group ($\chi^2 = 0.21$, $P = 0.89$). For those with available pathology results, 64.1% showed hippocampal sclerosis (HS), 12.1% showed focal cortical dysplasia (FCD), 15.2% showed a combination of both HS and FCD, and 8.4% had inconclusive or generic findings (e.g. Chaslin gliosis with no overt signs of HS or FCD). There were no significant differences in the distribution of neuropathology findings between MRI-positive and MRI-negative patients ($\chi^2 = 6.10$, $P = 0.11$).

Healthy controls

Healthy controls ($n = 589$) were recruited at each site based on the absence of a major neurological or psychiatric disorder and the absence of a brain structural abnormality (either known *a priori* or discovered through research scanning). The controls derived from the HCP cohort were chosen randomly from the latest release of 1200 subjects (https://www.humanconnectome.org/study/hcp-young-adult). HC were 53.1% female and 37.4 (SD = 14.4) years old on average. Age [$t(1176) = 0.67$, $P = 0.5$] and sex ($\chi^2 = 0.86$, $P = 0.36$) were not statistically different between the groups.

Imaging acquisition and pre-processing

Participants were scanned locally at each centre using 3 T MRI. Supplementary Table 1 summarizes the scanner type and acquisition protocol at each site. Before input into the CNN, T1-weighted MRI images were lightly preprocessed using the open-source *nii_preprocess* package (https://github.com/neurolabusc/nii_preprocess/) for MATLAB. Briefly, we initially normalized all T1-weighted images into standard stereotaxic Montreal Neurological Institute 152 space (113 × 137 × 113) using the normalize function from the software package Statistical Parametric Mapping (SPM12, version 7771; Functional Imaging Laboratory, Wellcome Trust Centre for Neuroimaging Institute of Neurology, University College London; http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) with the following parameters: bias regularization = 0.0001, bias full width at half maximum (FWHM) = 60, tissue probability map = TPM.nii, voxel size = 1 mm × 1 mm × 1 mm, and fourth degree b-spline interpolation. We subsequently segmented the normalized outputs into grey and white matter maps using the CAT12 extension toolbox (version 2000; http://www.neuro.uni-jena.de/cat12/)[31] using default parameters. To minimize individual variability in positioning of gyri and sulci, we subsequently smoothed grey and white segmented images using the smooth function of SPM with 3D FWHM at 10 mm. Voxels with ≥20% probability of being grey matter were included in the analyses.

For use in the analysis of saliency maps, the normalized images were also used to compute region of interest (ROI) volumes using *nii_preprocess* parcellated based on the Automated Anatomical Labeling atlas.[32]

Artificial neural network

Model architecture

We designed a 3D fully convolutional network model[33] to classify TLE versus HC participants. The model contained a feature extractor and a classifier. The feature extractor contained four sequential convolutional layers (3 × 3 × 3, 3 × 3 × 3, 2 × 2 × 2 and 2 × 2 × 2), each followed by batch normalization, maximum pooling and rectified linear unit (ReLU) activation, and a fifth convolutional layer (1 × 1 × 1) with batch normalization and ReLU activation. The classifier consisted of an average pooling layer (3 × 3 × 3), a dropout layer (dropout probability = 0.5) and a 1 × 1 × 1 convolutional layer with output dimension = 2 (TLE versus HC) (Supplementary Fig. 1). To remove the potential influence of scanning differences across centres, we used ComBat harmonization to centre voxel intensity across sites, as previously done in large multisite epilepsy studies.[23,24] We did not find a meaningful improvement in model performance on the harmonized data (see Supplementary material); therefore, the remaining analyses were performed on non-harmonized data to minimize the pre-processing of training/validation and testing inputs into the model.

Model execution

We used 10 runs of stratified 5-fold cross-validation to divide the data into training, validation and test sets (i.e. 50 total iterations). To ensure a consistent distribution of classes across all folds, we used stratified sampling, then repeated the stratified 5-fold cross-validation to probe the stability of the model. Given the equal number of TLE and HC, our dataset avoided potential biases caused by group imbalance of a majority class during training. To retain the multicentric nature of this dataset, we also ensured that all the three datasets (i.e. training, validation and test sets) had input images from all the sites. Each 5-fold cross-validation run comprised 1-fold as test samples and the remaining 4-folds as training/validation samples. This training/validation portion of the data was divided into training:validation sets at an 80:20 ratio. For every stratified 5-fold cross-validation iteration, we trained the model exclusively on the training set, identified the best-performing model on the validation set, and subsequently applied it to the unseen test cohort. Therefore, the test cohort was exclusively an unseen set, which was not included in any part of the learning process.

Grid search optimization

To tune hyper-parameters for optimal performance, we adopted a grid search approach using Stochastic Gradient Descent and Adam optimizers varying an array of hyperparameters as summarized in Supplementary Table 2. All models underwent training for 400 epochs. To counteract overfitting, we incorporated early stopping with a train patience of 40 epochs.

Saliency maps

We used vanilla backpropagation[34] on our optimally selected model from the grid search to generate saliency maps for each seizure-free TLE patient. Saliency maps identify crucial brain regions at the voxel level that are most influential in distinguishing between TLE and HC. This process involves a forward pass in which the model computes probabilities for each class (TLE versus HC), followed by a backward pass calculating the gradient of the target class probability with respect to the input image. These gradients highlight the input features (voxels) most affecting the class probability, thus forming the saliency map where higher values indicate greater importance. In other words, for each patient, for each of the 10 runs, we yielded: (i) the class prediction (i.e. whether the machine classified the patient as TLE or HC); (ii) the probability with which the machine classified the participant into that category; and (iii) the saliency map. To ensure robustness, we averaged saliency maps across all cross-validation folds.

Statistical analyses

Model performance evaluation and benchmarking

We evaluated the performance of the 3D CNN model in classifying TLE versus HC by computing accuracy and F1 scores for the test fold across the 50 model iterations (i.e. 10 runs with 5-folds each). For comparison of CNN performance with widely available features, we then used ROI volumes as input features to train a linear and non-linear support vector machine (SVM) to discriminate between TLE and HC. This volume-based classification served as a benchmark to gauge how much more discriminatory power the 3D CNN model could achieve relative to a model trained exclusively on hippocampal volumes ('SVM hipp') or a model trained on whole-brain region-based volumetrics ('SVM all'). The SVM models used linear and radial basis function kernels with the C parameter in the range of [0.0001, 0.001, 0.01, 0.1, 1, 10, 100]. We contrasted mean accuracy and F1 values for the three models (CNN, SVM based on whole-brain volumes and SVM based on hippocampal volumes) using a one-way ANOVA with Tukey's HSD *post hoc* test for pairwise comparisons.

Comparison of MRI-positive versus MRI-negative temporal lobe epilepsy patients

We compared MRI-positive and MRI-negative patients on both accurate identification (i.e. model performance) and pattern of information (i.e. model saliency). To compare model performance, we calculated the accuracy of correctly identified patients for each group, averaged across the test sets from 50 rounds, for the CNN and the two SVM models. We then tested the effect of group (MRI positive versus MRI negative) and model architecture (CNN versus SVM all versus SVM hipp), in addition to their interactions, on accuracy using a 2 × 3 two-factor ANOVA followed by Tukey's *post hoc* test.

In addition to comparing the binary classification, we examined whether the 3D CNN model-defined probability with which patients were classified as TLE differed between the groups. Each patient had 10 predictions, each with an associated probability derived (i.e. one for each run in which the patient was part of the test dataset). We averaged the probabilities associated with each class prediction. For instance, if a TLE patient was correctly classified as being TLE on 7 of 10 runs, we averaged the probability across those 7 runs and the probability across the 3 runs in which the classifier (incorrectly) predicted HC group. We then ran a 2 × 2 two-way ANOVA with class predicted (TLE versus HC) and MRI status (MRI positive versus MRI negative) on the classification probability.

For a comparison of voxel-based saliency, we performed a two-tailed *t*-test at each voxel to compare the mean saliency of MRI-positive versus MRI-negative groups based on individual saliency weights from the CNN classifier. Then, to quantify the spatial similarity between the group-averaged saliency maps of MRI-positive and MRI-negative patients while controlling for spatial autocorrelation, we used BrainSMASH[35] (Brain Surrogate Maps with Autocorrelated Spatial Heterogeneity), a Python-based computational platform for statistical testing of spatially autocorrelated brain maps. BrainSMASH yields a null map based on the spatial (Euclidian) distance of the voxels, which was then used to contrast the similarity between averaged maps for MRI-positive and MRI-negative patients (see Supplementary material).

Saliency values were *z*-transformed for visualization purposes. To perform ROI-based analyses, we then clustered voxels based on the Automated Anatomical Labeling atlas[32] and averaged the saliency values of all voxels within each of the 116 ROI of the atlas. We applied *t*-tests iteratively between MRI-positive versus MRI-negative groups across all 116 ROIs. Given the large number of statistical contrasts, voxel and ROI statistical comparison *P*-values were adjusted using the Benjamini–Hochberg method,[36] which is a widely used approach for controlling the false discovery rate. We chose this approach over more conservative models to reduce the risk of Type II errors (false negatives) compared with more stringent methods (e.g. Bonferroni), because our main goal was to identify potential voxels or ROIs that were differentially weighted in MRI-positive versus MRI-negative cases. In addition, we carried out Pearson correlations of saliency weights for all ROI pairs.

Results

When benchmarking the overall performance of our model, we found that the 3D CNN model had a mean accuracy of 85.9% ± 2.8% (median = 85.2%, range = 79.2%–89.8%) and significantly outperformed [$F(2,147) = 170.9$, $P < 0.001$] the SVM models trained on hippocampal (74.4% ± 2.6%, median = 74.9%, range = 67%–89.8%) and whole-brain (78.3% ± 3.3%, median = 78.5%, range = 70.8%–85.5%) volumes. Likewise, mean F1 scores were significantly higher [$F(2,147) = 217.5$, $P < 0.001$] for the 3D CNN model (84.1% ± 3.5%, median = 84.3%, range = 76.3%–90.1%) than SVM models trained on hippocampal (67.9% ± 4.5%, median = 68.8%, range = 52.2%–76.9%) and whole-brain (77.3% ± 3.6%, median = 77.3%, range = 68.5%–85.5%) volumes. Figure 2 summarizes the distribution of metrics across all 50 iterations for each model.

Next, we compared MRI-positive and MRI-negative detection performance in our surgically validated cohort, as summarized by Fig. 3A. We found a significant effect of group [$F(1294) = 230.0$,

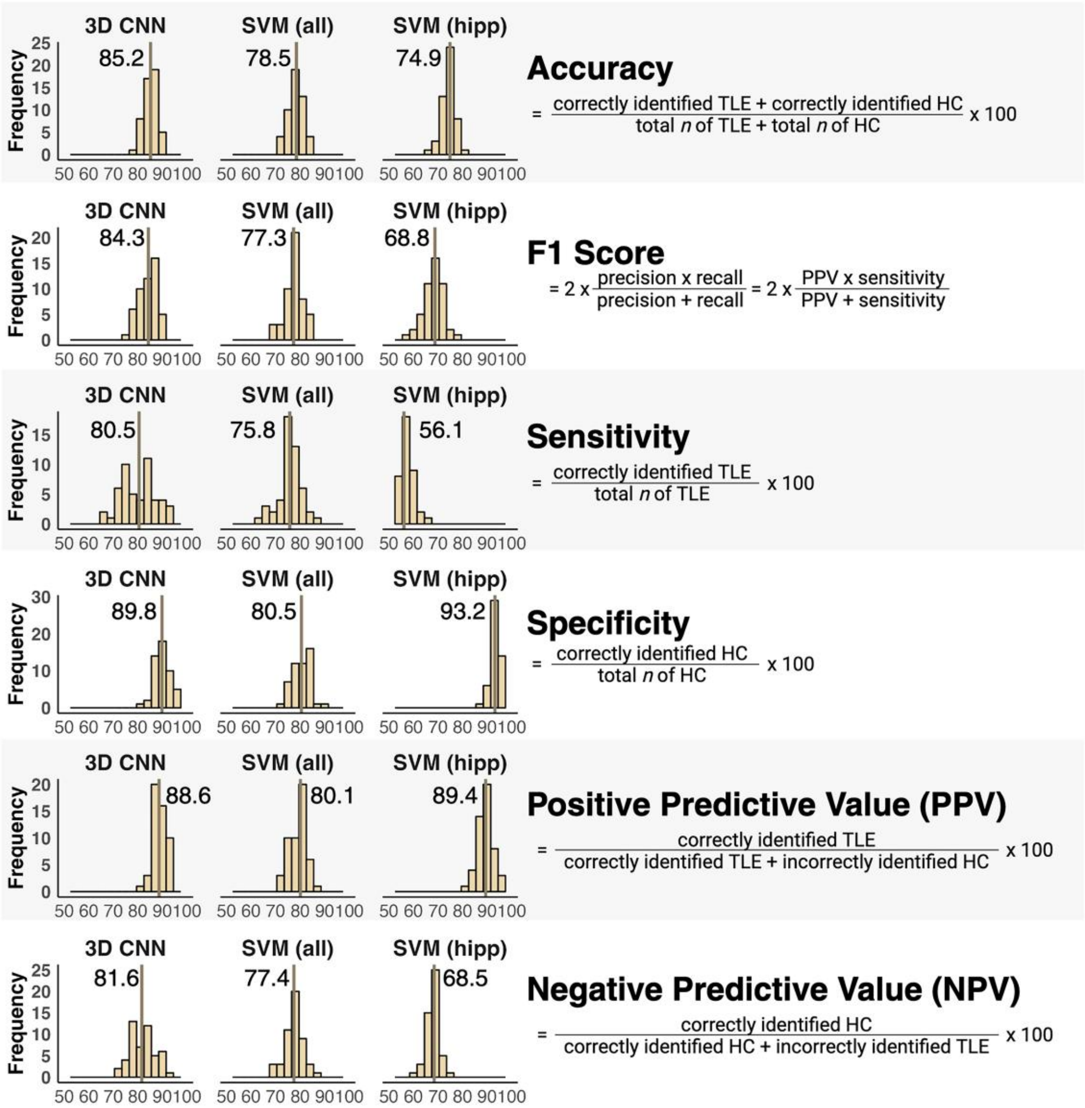

Figure 2 Distribution of performance metrics across the three models evaluated in this study. A 3D convolutional neural network (CNN) based on smoothed grey matter maps and support vector machines (SVM) based on whole-brain volumes (all) or hippocampal volumes (hipp). Vertical lines represent the median of the 50 iterations for each model (10 runs of 5-folds each), with the median value attached. A significantly higher accuracy and F1 score were noted for the 3D CNN model than for both SVM approaches. HC = healthy control; SD = standard deviation; TLE = temporal lobe epilepsy.

$P < 0.001$], with higher accuracy for MRI-positive than MRI-negative patients. We also observed a significant effect of model [$F(2294) = 171.2$, $P < 0.001$], with whole-brain SVM showing higher performance than hippocampus-based SVM ($P < 0.001$) and with CNN outperforming both SVM models ($P < 0.001$). There was also a significant interaction between TLE group and model [$F(2294) = 138.3$, $P < 0.001$]. Specifically, the 3D CNN accuracy to classify patients as TLE was 82.7% (SD = 9.3%) for MRI-negative and 89.5% (SD = 6.6%) for MRI-positive patients. In contrast, the SVM model performance within the surgically validated group showed an accuracy of 73.4% (SD = 9.0%) in MRI-negative and 73.9% (SD = 8.7%) in MRI-positive patients when trained on whole-brain volumes and 42.5% (SD = 12.6%) in MRI-negative and 82.5% (SD = 6.7%) in MRI-positive patients when trained exclusively on hippocampal volumes.

Investigating 3D CNN model certainty, when patients were classified as TLE, the mean probability of belonging to that class was 92% (SD = 10.0%) for MRI-negative and 94.2% (SD = 8.5%) for MRI-positive. In contrast, when classified as HC, the mean probability of belonging to that class was 78.1% (SD = 10.1%) for MRI-negative and 76.7% (SD = 10.6%) for MRI-positive patients. Therefore, there was an effect of predicted class on mean individual classification probability [$F(1280) = 146.3$, $P < 0.001$] but no main effect of MRI lesional group [$F(1280) = 3.0$, $P = 0.16$] and no significant interaction between predicted class and MRI lesional group [$F(1280) = 1.9$, $P = 0.19$] (Fig. 3B).

The assessment of potential saliency differences between surgically validated MRI-positive and MRI-negative patients at the whole-brain voxel-wise level revealed that both MRI-positive and MRI-negative groups demonstrated a similar distribution of saliency weights throughout the brain. Positive saliency values (i.e. supporting the classification of TLE) were seen predominantly in limbic structures, including most prominently over medial temporal regions followed by orbitofrontal and cingulate areas and (to a lesser extent) over the temporal neocortex (posterior > anterior) and the premotor regions. Negative saliency values (i.e.

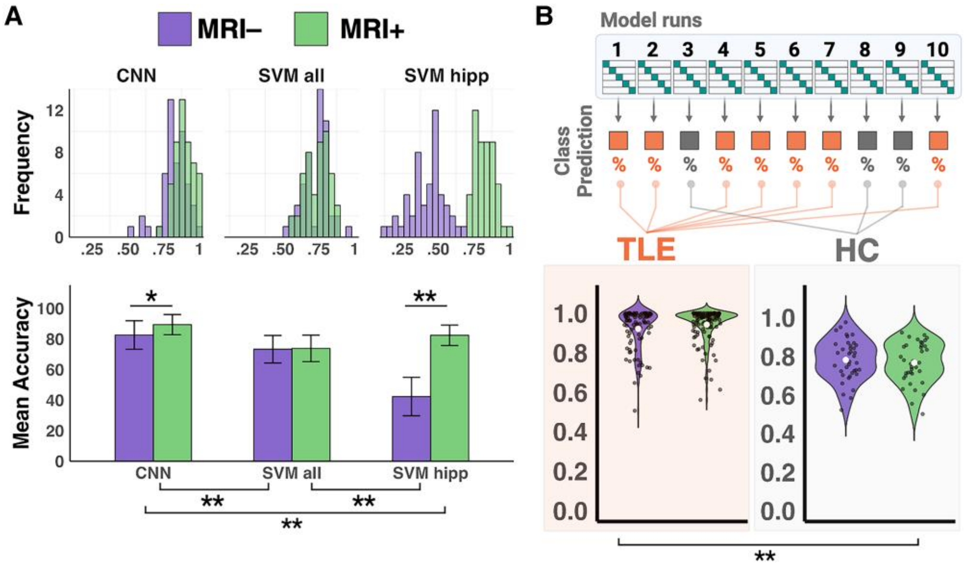

Figure 3 Performance across models. The convolutional neural network (CNN) model was run for a total of 50 iterations (10 runs with 5-folds), each one yielding a predicted class [temporal lobe epilepsy (TLE) versus healthy controls (HC)] and the individual probability of belonging to such class. (**A**) *Top*: Histogram plots show the run-level distribution of accuracy scores for classifying surgically validated patients correctly as TLE for MRI− (purple) and MRI+ (green) groups. *Bottom*: Bar plots reveal that the mean (± standard deviation) accuracy was significantly higher for MRI+ than MRI− patients in the CNN (*$P < 0.01$) and the support vector machine (SVM) model trained on hippocampal values (**$P < 0.001$). Overall, the CNN model was significantly more accurate than both SVM models (**$P < 0.001$). (**B**) The top diagram provides a visual summary of performance of the CNN model on a hypothetical patient. The model is run 10 times, with each patient appearing in the testing fold once per run. Each run will yield a binary classification (orange = TLE, grey = HC). Using backpropagation, a percentage probability can be computed for the likelihood of belonging to the predicted class. We averaged the probabilities of each class across runs of similar prediction to compare the performance of the model on each TLE group based on the predicted class. The violin plots reveal that the mean individual probability of belonging to TLE for patients was significantly higher than the probability of belonging to the HC group (**$P < 0.001$), as expected, but there were no significant differences of such probabilities between MRI− (violet) and MRI+ (green) patients.

supporting the classification of HC) clustered predominantly around the temporal poles and the basal ganglia. As shown by Fig. 4, medial temporal, prefrontal, orbitofrontal and subcortical clusters of voxels had higher saliency values in MRI-positive compared with MRI-negative groups. Overall, differences in whole-brain saliency values when comparing MRI-positive and MRI-negative patients were scattered and weak. When correcting for familywise errors, there were no significant differences between the groups on independent voxel saliency weights. The observed Pearson correlation coefficient between the group-averaged saliency maps of MRI-positive and MRI-negative groups was $R = 0.997$, indicating a high degree of spatial similarity. To assess the statistical significance of this observed correlation, we generated 1000 spatially autocorrelated null maps using BrainSMASH. The null distribution of correlation coefficients was then compared with the observed correlation, which yielded a P-value of 0.00099, indicating that the observed spatial similarity is highly unlikely to be attributable to chance and is statistically significant. Furthermore, there were no significant differences between MRI-positive and MRI-negative patients in the median voxelwise correlation coefficient between individual and group-averaged saliency maps for all seizure-free TLE patients (see Supplementary Fig. 2).

Likewise, at the ROI level, there was a comparable distribution of regional saliency values favouring the classification of TLE between the two groups (Fig. 5). The cortical regions with the highest saliency values were all medial temporal structures: for MRI-positive patients, left amygdala $z = 5.6$, right amygdala $z = 5.3$, left hippocampus $z = 3.1$, right hippocampus $z = 2.8$, left parahippocampus $z = 1.7$ and right parahippocampus $z = 1.1$; and for MRI-negative patients, right amygdala $z = 6.9$, right hippocampus $z = 5.7$, left amygdala $z = 3.0$, left hippocampus $z = 1.7$, right parahippocampus $z = 1.6$ and left parahippocampus $z = 1.5$. Supplementary Table 3 shows that there were no significant differences between MRI-positive and MRI-negative patients for any of the ROIs. Indeed, the z-scores across ROIs were strongly correlated between the groups ($r = 0.86$, $P < 0.001$), as shown by Supplementary Fig. 3.

Discussion

The quest to improve diagnostic precision and efficiency for patients with drug-resistant focal epilepsy is a significant and ongoing challenge that can markedly influence treatment trajectories and outcomes in epilepsy care. Our study sought to address a crucial gap in determining the pathological status of TLE as determined by structural MRI, particularly in ‘non-lesional’ cases (i.e. MRI-negative cases) where traditional imaging analysis by human experts did not reveal a clear pathological marker. By using robust machine learning techniques to analyse T1-weighted MRI data from surgically confirmed TLE, we aimed: (i) to demonstrate the ability of a 3D CNN to identify MRI-negative TLE patients in a large, heterogeneous multicentre dataset; and (ii) to compare the salient neuroanatomical features that are used by the machine to determine whether patterns associated with MRI-positive and MRI-negative patients are similar.

Our analysis involved two main steps. First, our approach used a 3D CNN trained to differentiate between TLE and HC, which the model achieved with high accuracy. Across all patients, we benchmarked the performance of the model against machines trained on hippocampal and whole-brain volumes, finding that 3D CNN outperformed them by an average of 7.6 (whole-brain approach) to 11.5 (hippocampal focused approach) percentage points of

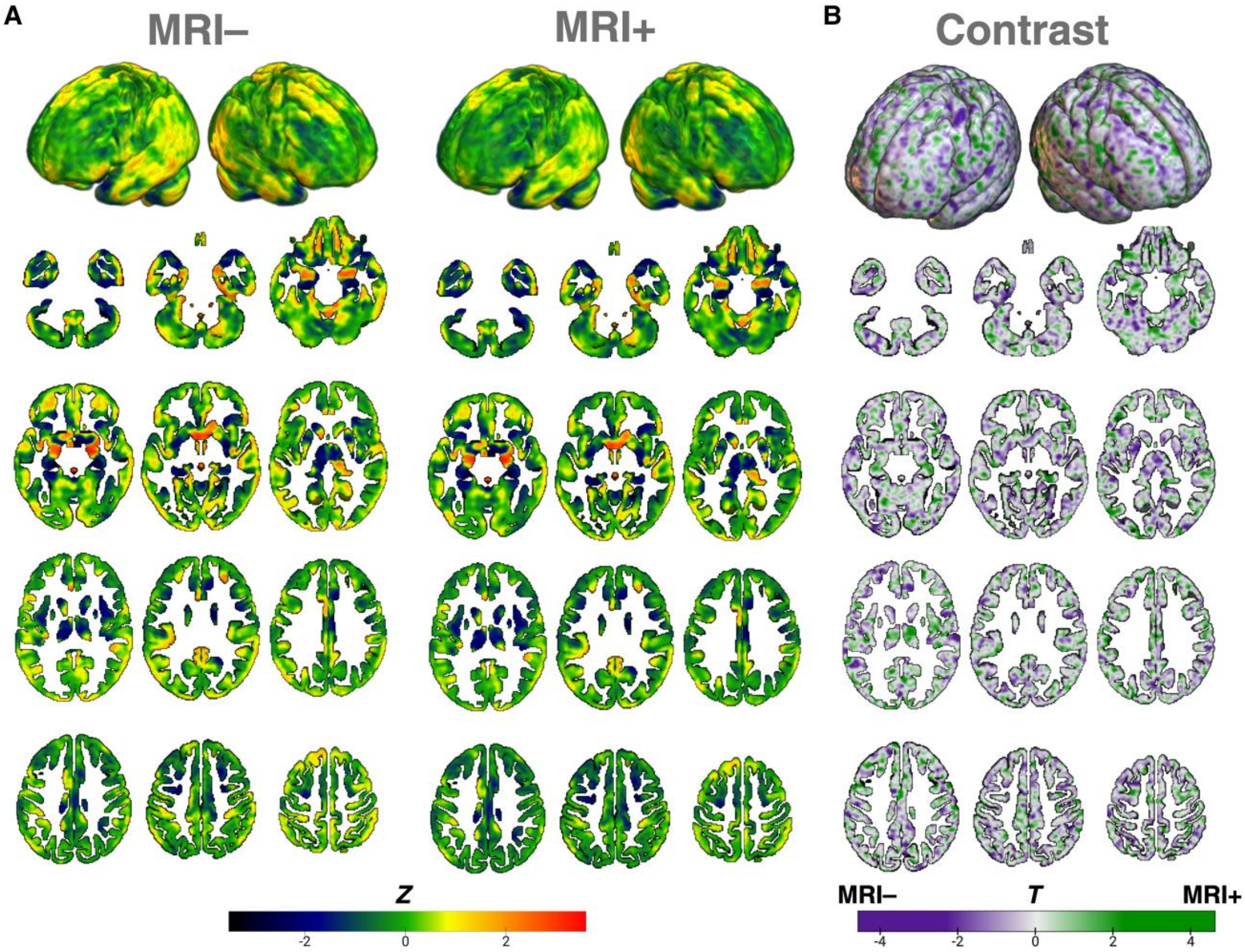

Figure 4 Voxel-based saliency weights. (**A**) Voxel-based saliency weights (z-transformed) for patients deemed to have unremarkable MRI (MRI−) and patients deemed to have lesional MRI (MRI+) by human expert consensus. *Top* row: 3D renders; *subsequent rows*: sequential axial slices from ventral (temporal tip and cerebellar cortex) to dorsal (frontal) brain in radiological convention. (**B**) The voxel-wise statistical contrast between the saliency weights of each group, with *T*-values colour-coded based on the group favoured (purple for MRI− and green for MRI+). No voxel or cluster of voxels survived familywise error correction.

accuracy. In other words, as demonstrated in previous efforts,[28,29,37] the artificial neural network is capable of extracting features in the grey matter maps beyond mere brain or regional volumes to support the accurate identification of TLE.

Our second step involved an analysis of performance and saliency maps in the subset of patients who went through surgery and demonstrated sustained seizure freedom ≥1 year post-operatively. We believe this group is the best available gold standard to confirm a diagnosis with TLE, i.e. it constitutes a surgically validated TLE group. We acknowledge that a portion of the non-seizure-free patients might also have TLE, because the potential reasons for surgical failure are multiple; yet, focusing on the seizure-free TLE patients provides the closest group available to a 'ground truth' in epilepsy. The model outperformed human experts at determining when a scan belonged to a patient with TLE in this group by almost 30% points of accuracy, with only 56.4% of analysed surgically validated patients deemed to have an MRI with features of underlying TLE pathology by human expert consensus. This is especially important because the classification by human experts had been made with availability of clinical context and with the entire set of MRI sequences (e.g. additional T2-weighted scans) and (frequently) supporting modalities, such as PET and EEG findings. This contrasts with the present machine learning algorithm, which was trained exclusively on T1-weighted grey matter maps. Overall, the machine's higher accuracy scores are consistent with prior studies using similar approaches in smaller, more homogeneous cohorts.[28] They are also superior to machines trained on ROI level data.[23]

As expected, we observed that machine learning models performed better for MRI-positive than MRI-negative patients, with the 3D CNN seeing 6.8% higher accuracy for lesional cases. However, the machines trained on regional volumetric data showed divergent performances. Hippocampus-focused models revealed a marked discrepancy based on MRI lesional status, classifying MRI-negative patients at chance level. Similar to humans, relying heavily on hippocampal volumes leads to poor sensitivity to detect TLE when hippocampal volume loss is not apparent. When trained on whole-brain regional volumetric data instead, the performance of the model was comparable for both lesional TLE groups, confirming that a whole-brain level approach provides incremental information beyond the medial temporal structures. We propose that the improved performance of CNN over the whole-brain volume-based SVM reflects the additional advantage of neural network architectures in leveraging voxel-level data for

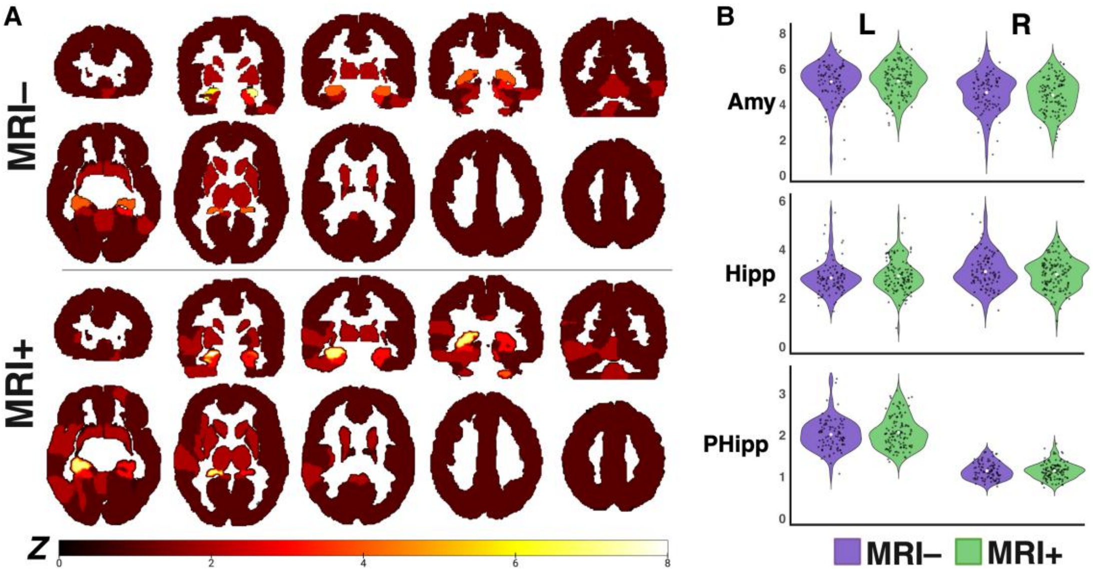

Figure 5 Region of interest-based saliency weights. (**A**) Region of interest-based saliency weights (z-transformed) for MRI− (*top*) and MRI+ (*bottom*) patients with temporal lobe epilepsy in radiological convention. For each group, the *top* row displays coronal slices from anterior to posterior, and the *bottom* row displays axial slices from ventral to dorsal. There was a comparable distribution of salient regions between the two groups. (**B**) The mean z saliency weights for the three top regions of interest (Amy = amygdala; Hipp = hippocampus; PHipp = parahippocampus) on the left (L) and right (R) sides of the brain for both MRI− (violet) and MRI+ (green) patients with temporal lobe epilepsy. No significant differences were found for any of these contrasts.

disease classification, as predicted by prior ROI-based analyses with limited disease classification accuracy.[23]

The saliency maps generated by the CNN model revealed that limbic structures, particularly the medial temporal regions but also cingulate and orbitofrontal areas, were most influential in the classification task, aligning with the known pathology of TLE.[33] However, the high saliency of premotor and lateral temporal structures favouring a classification of TLE also corroborates existing literature that emphasizes the role of extra medial and extra temporal pathology in TLE.[22,23,38]

Our hypothesis that AI-generated saliency maps would emphasize similar neuroanatomical features in both lesional and non-lesional TLE groups was rooted in the premise that many non-lesional patients demonstrate a visually elusive yet quantifiably consistent pathological signature of epilepsy. Therefore, we believed that the AI would detect this subtle, yet diagnostically crucial, pattern not readily apparent to the human eye. These findings were confirmed by showing that the saliency maps of both TLE groups were highly similar even when controlling for spatial autocorrelation. We confirmed that there was consistency of salient neuroanatomical features identified in both MRI-positive and MRI-negative groups. Additionally, the model-derived individual probability of being correctly classified as TLE (or even incorrectly classified as HC) by the machine was not significantly different based on MRI lesional status. The similarity in distribution of saliency weights throughout the brain across both groups also suggests a more nuanced understanding of TLE pathology that extends beyond the binary presence or absence of visible lesions. This observation is in line with prior literature showing that morphometric analysis reveals similar patterns of atrophy in MRI-positive and MRI-negative patients not otherwise identified visually.[10,38]

Our results add to the growing literature using machine learning and AI to 'unmask' lesions in the MRI scans of patients who had been deemed non-lesional by humans. We note, however, that the model introduced here does not intend to segment the boundaries of a specific lesion in the brain, but rather to determine the presence of epilepsy even when human eyes do not readily pick up on radiographic findings to support such a clinical conclusion. Emerging new work by other teams is diligently offering new avenues to detect precise lesion boundaries. For example, a seminal deep learning model trained on a multicentre cohort of 148 patients with histologically confirmed FCD across nine sites showed sensitivity of 83% in detecting a lesion in an independent cohort of FCD cases.[39] Subsequently, the Multi-centre Epilepsy Lesion Detection (MELD) Project introduced a pipeline to output individual patient reports showcasing 33 neuroimaging features and their saliency to a neural network classifier to detect FCD automatically. The cohort was largely heterogeneous, with overall sensitivity of 59% but which increased to 85% among patients who were seizure free after resection of FCD type IIB.[40] Although these methods have mostly relied on heavily supervised models relying on manual annotations by humans, which can be time-consuming and prone to biases and errors, future work might enhance such pipelines by offering a more vaguely supervised approach. In our case, the model is not trained to detect specific pathological patterns (e.g. to distinguish FCD from hippocampal sclerosis) but to classify the presence or absence of disease (i.e. TLE) at large.

Although our findings in TLE are promising, it is important to recognize the limitations of our approach. First, the generalizability of our results is dependent on the diversity of the dataset and the applicability of the model to broader clinical settings. We addressed this challenge by pooling participants from across 12 cohorts with diverse demographic representations. However, replicating this study in an even larger sample derived from a more global multicentre dataset will be important. As we increase the availability of raw MRI data across centres, at least two major challenges emerge. On the one hand, acquisition protocols across sites are inconsistent, because the data used in this study are derived from the clinical pipeline. Here, we tackled this issue by ensuring the presence of studies from all cohorts across all datasets (training, validation and testing). We implemented this approach for its simplicity because the main focus of our analysis was the comparison of saliency maps based on MRI lesional status; however, alternative methods are also worth considering, such as harmonizing the imaging data to eliminate potential confounders while preserving disease-related brain signals. In our analysis, and as previously reported by other studies with multisite training cohorts in the hundreds,[23,24] we did not find an advantage on model performance from harmonizing the data. However, the use of alternative novel neuroimaging harmonization techniques should be investigated further to test their role in improving model accuracies. On the other hand, as the multicentre samples increase, the overlap of clinical variables available and consistently coded across sites diminishes given the retrospective collection of these data. This, for example, limited our ability to stratify analyses based on pathology type: although the overall histopathological label might be available for resective cases, the granular description of FCD subtype or the precise anatomical location was not consistently accessible at the time of analysis. Encompassing all FCD types and locations as a homogeneous group would provide a narrow view of the diversity of radiographic presentations (size, anatomical distribution and signal changes) of this disease. Initiatives such as ENIGMA-Epilepsy will soon allow for the aforesaid proposed global analysis as they continue to collect voxel-level data with well-described clinical samples. Likewise, projects such as MELD[40] are moving towards lesion-agnostic models that provide comprehensive detection and segmentation of specific MRI lesions.

Another current limitation in the generalizability of our results is the focus on TLE. Future directions extending our model to non-temporal foci will also be crucial. We introduce the present analysis as an initial milestone in rethinking how lesional status might be defined in focal epilepsy. However, the next natural step towards achieving a generalizable model would involve training a hierarchical machine using subnetworks to fine-tune downstream classification branches. For example, one could first train the machine to detect controls versus epilepsy, then TLE versus non-TLE, then explore how those maps differ between MRI-positive and MRI-negative patients. This would refine subtasks and enhance the distinguishability of saliency maps at both the subject and group levels. Likewise, future models could try to distinguish focal from generalized onset epilepsies, particularly in cases where non-invasive data fail to provide a definitive answer. Here, concentrating on TLE allowed for a relatively homogeneous cohort of patients for whom we could readily identify a gold-standard diagnostic test subset, represented by those patients with TLE who underwent laser ablation or temporal resection and remained seizure free ≥1 year after surgery. Although this is not a definite confirmation of TLE (e.g. seizures can still re-emerge after the first year), we believe this to be the closest approximation to a ground truth available through human research in epilepsy.

Furthermore, although saliency maps provide a visual representation of the focus of the model, they tend to capture the local sensitivity of the model to changes in its input without necessarily accounting for non-linear structures. As such, saliency maps can be noisy without providing an exhaustive explanation of the features influencing model outputs.[41] Accordingly, there are limitations inherent to the exact clinical significance of saliency patterns, which require further validation. For example, future studies should use techniques such as occlusion sensitivity-based saliency plots to study the differential contribution of different brain areas to classification accuracy. Likewise, future analyses could consider the value of targeting with intracranial EEG those areas of high saliency in cases classified as TLE by the machine but deemed MRI negative by humans. This would help to supplement neuroimaging patterns with neurophysiological data to probe the biological signature of these areas. Additionally, it will be crucial for future studies to incorporate the role of white matter data to further enhance disease classification and to compare how these might inform classification differently in MRI-positive versus MRI-negative patients. Overall, we believe our research paves the way for several future studies integrating saliency analysis and clinical translation. Prospective validation of our model in a clinical setting would be crucial to ascertain its utility and reliability in real-world diagnostic scenarios. Further research should also explore integrating our model with other diagnostic tools for a more comprehensive assessment of TLE, including multimodal inputs for training and testing.

Although these results represent an early step in determining the potential redefinition of lesional status in TLE cases, if similar saliency maps are consistently found in independent cohorts and further refined through hierarchical models like the one presented here, our findings could have potential implications for radiological practice. Specifically, by identifying the regions highlighted in the saliency maps, radiologists could be trained to focus on areas that the model deems crucial for distinguishing between MRI-positive and MRI-negative TLE patients. This could enhance the detection of subtle changes that might not be apparent immediately through traditional inspection. In essence, the saliency maps serve as a guide to areas warranting closer scrutiny, potentially improving diagnostic accuracy and aiding in early detection.

Conclusion

We have demonstrated that a 3D AI based on voxel-level structural MRI data can classify TLE patients with high accuracy and that the neuroanatomical markers used for such classification are similar regardless of whether humans were able to identify pathological imaging markers visually. Our study represents a significant advancement in the application of machine learning to neuroimaging in epilepsy by highlighting the potential of AI in identifying neuroanatomical features associated with TLE, even in non-lesional cases. Taken together, these results offer a promising direction for enhancing diagnostic accuracy and potentially improving surgical outcomes, potentially redefining the 'lesional' concept in TLE.

Data availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available owing to their collection as part of inter-institutional data user agreements that do not readily stipulate outward sharing.

Acknowledgements

Portions of some figures in this manuscript were created using BioRender.com

Funding

This study was supported by an National Institute of Neurological Disorders and Stroke (NINDS) grant R01NS110347. E.G. was supported by Georgia CTSA KL2 and UL1 grants (KL2TR002381 and UL1TR002378). V.D.C., A.A. and R.H. were supported by RF1AG063153. C.R.M. was supported by NINDS R01s NS124585, NS122827 and NS120976. K.A.D. was supported by NINDS R02 NS116504. S.S.K. was supported by UK Medical Research Council (grant numbers MR/S00355X/1 and MR/K023152/1). Healthy control data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The funding sources did not have any role in study design, data collection or analysis, interpretation of results, writing of this manuscript or the decision to submit for publication.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Bernasconi A, Antel SB, Collins DL, et al. Texture analysis and morphological processing of magnetic resonance imaging assist detection of focal cortical dysplasia in extra-temporal partial epilepsy. *Ann Neurol*. 2001;49:770-775.
2. Bernasconi A, Bernasconi N, Bernhardt BC, Schrader D. Advances in MRI for ‘cryptogenic’ epilepsies. *Nat Rev Neurol*. 2011;7:99-108.
3. De Ciantis A, Barba C, Tassi L, et al. 7T MRI in focal epilepsy with unrevealing conventional field strength imaging. *Epilepsia*. 2016; 57:445-454.
4. Huppertz HJ, Wellmer J, Staack AM, Altenmuller DM, Urbach H, Kroll J. Voxel-based 3D MRI analysis helps to detect subtle forms of subcortical band heterotopia. *Epilepsia*. 2008;49:772-785.
5. Strandberg M, Larsson EM, Backman S, Kallen K. Pre-surgical epilepsy evaluation using 3T MRI. Do surface coils provide additional information? *Epileptic Disord*. 2008;10:83-92.
6. Wang I, Oh S, Blümcke I, et al. Value of 7T MRI and post-processing in patients with nonlesional 3T MRI undergoing epilepsy presurgical evaluation. *Epilepsia*. 2020;61:2509-2520.
7. Zhu H, Scott J, Hurley A, Gaxiola-Valdez I, Peedicail JS, Federico P. 1.5 versus 3 tesla structural MRI in patients with focal epilepsy. *Epileptic Disord*. 2022;24:274-286.
8. Tellez-Zenteno JF, Hernandez-Ronquillo L. A review of the epidemiology of temporal lobe epilepsy. *Epilepsy Res Treat*. 2012; 2012:630853.
9. Mann L, Rosenow F, Strzelczyk A, et al. The impact of referring patients with drug-resistant focal epilepsy to an epilepsy center for presurgical diagnosis. *Neurol Res Pract*. 2023;5:65.
10. Muhlhofer W, Tan YL, Mueller SG, Knowlton R. MRI-negative temporal lobe epilepsy—What do we know? *Epilepsia*. 2017;58: 727-742.
11. Tellez-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.
12. Pardoe HR, Pell GS, Abbott DF, Jackson GD. Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? *Epilepsia*. 2009;50:2586-2592.
13. Jack CR Jr, Sharbrough FW, Twomey CK, et al. Temporal lobe seizures: Lateralization with MR volume measurements of the hippocampal formation. *Radiology*. 1990;175:423-429.
14. Vos SB, Winston GP, Goodkin O, et al. Hippocampal profiling: Localized magnetic resonance imaging volumetry and T2 relaxometry for hippocampal sclerosis. *Epilepsia*. 2020;61:297-309.
15. Jackson GD, Connelly A, Duncan JS, Grunewald RA, Gadian DG. Detection of hippocampal pathology in intractable partial epilepsy: Increased sensitivity with quantitative magnetic resonance T2 relaxometry. *Neurology*. 1993;43:1793-1799.
16. Bonilha L, Elm JJ, Edwards JC, et al. How common is brain atrophy in patients with medial temporal lobe epilepsy? *Epilepsia*. 2010;51:1774-1779.
17. Bonilha L, Kobayashi E, Rorden C, Cendes F, Li LM. Medial temporal lobe atrophy in patients with refractory temporal lobe epilepsy. *J Neurol Neurosurg Psychiatry*. 2003;74:1627-1630.
18. Bernasconi N, Duchesne S, Janke A, Lerch J, Collins DL, Bernasconi A. Whole-brain voxel-based statistical analysis of gray matter and white matter in temporal lobe epilepsy. *Neuroimage*. 2004;23:717-723.
19. Natsume J, Bernasconi N, Andermann F, Bernasconi A. MRI volumetry of the thalamus in temporal, extratemporal, and idiopathic generalized epilepsy. *Neurology*. 2003;60:1296-1300.
20. Keller SS, Roberts N. Voxel-based morphometry of temporal lobe epilepsy: An introduction and review of the literature. *Epilepsia*. 2008;49:741-757.
21. Bonilha L, Rorden C, Castellano G, et al. Voxel-based morphometry reveals gray matter network atrophy in refractory medial temporal lobe epilepsy. *Arch Neurol*. 2004;61:1379-1384.
22. Whelan CD, Altmann A, Botia JA, et al. Structural brain abnormalities in the common epilepsies assessed in a worldwide ENIGMA study. *Brain*. 2018;141:391-408.
23. Gleichgerrcht E, Munsell BC, Alhusaini S, et al. Artificial intelligence for classification of temporal lobe epilepsy with ROI-level MRI data: A worldwide ENIGMA-Epilepsy study. *Neuroimage Clin*. 2021;31:102765.
24. Hatton SN, Huynh KH, Bonilha L, et al. White matter abnormalities across different epilepsy syndromes in adults: An ENIGMA-Epilepsy study. *Brain*. 2020;143:2454-2473.
25. Lariviere S, Rodriguez-Cruces R, Royer J, et al. Network-based atrophy modeling in the common epilepsies: A worldwide ENIGMA study. *Sci Adv*. 2020;6:eabc6457.
26. Sisodiya SM, Whelan CD, Hatton SN, et al. The ENIGMA-Epilepsy working group: Mapping disease from large data sets. *Hum Brain Mapp*. 2020;43(1):113-128.
27. Gleichgerrcht E, Keller SS, Drane DL, et al. Temporal lobe epilepsy surgical outcomes can be inferred based on structural connectome hubs: A machine learning study. *Ann Neurol*. 2020;88:970-983.
28. Gleichgerrcht E, Munsell B, Keller SS, et al. Radiological identification of temporal lobe epilepsy using artificial intelligence: A feasibility study. *Brain Commun*. 2022;4:fcab284.
29. Chang AJ, Roth R, Bougioukli E, et al. MRI-based deep learning can discriminate between temporal lobe epilepsy, Alzheimer’s disease, and healthy controls. *Commun Med (Lond)*. 2023;3:33.
30. Berg AT, Berkovic SF, Brodie MJ, et al. Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE Commission on Classification and Terminology, 2005–2009. *Epilepsia*. 2010;51:676-685.
31. Hong J, Feng Z, Wang SH, et al. Brain age prediction of children using routine brain MR images via deep learning. *Front Neurol*. 2020;11:584682.

32. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*. 2002;15:273-289.
33. Peng H, Gong W, Beckmann CF, Vedaldi A, Smith SM. Accurate brain age prediction with lightweight deep neural networks. *Med Image Anal*. 2021;68:101871.
34. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*. [Preprint] http://arxiv.org/abs/1312.6034v2
35. Burt JB, Helmer M, Shinn M, Anticevic A, Murray JD. Generative modeling of brain maps with spatial autocorrelation. *Neuroimage*. 2020;220:117038.
36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 1995;57:289-300.
37. Kaestner E, Rao J, Chang AJ, et al. Convolutional neural network algorithm to determine lateralization of seizure onset in patients with epilepsy: A proof-of-principle study. *Neurology*. 2023;101:e324-e335.
38. Mueller SG, Laxer KD, Barakos J, Cheong I, Garcia P, Weiner MW. Widespread neocortical abnormalities in temporal lobe epilepsy with and without mesial sclerosis. *Neuroimage*. 2009;46:353-359.
39. Gill RS, Lee HM, Caldairou B, et al. Multicenter validation of a deep learning detection algorithm for focal cortical dysplasia. *Neurology*. 2021;97:e1571-e1582.
40. Spitzer H, Ripart M, Whitaker K, et al. Interpretable surface-based detection of focal cortical dysplasias: A multi-centre epilepsy lesion detection study. *Brain*. 2022;145:3859-3871.
41. Kim B, Seo J, Jeon S, Koo J, Choe J, Jeon T. Why are saliency maps noisy? Cause of and solution to noisy saliency maps, eds. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*: IEEE; 2019:4149–4157.