UNIVERSITY OF CALIFORNIA SAN DIEGO

Artificial Neural Network Application for Detecting Seizure Focus using Neuroimaging

A Thesis submitted in partial satisfaction of the requirements

for the degree Master of Science

in

Biology

by

Jun Xian Rao

Committee in charge:

Professor Carrie McDonald, Chair
Professor Mikio Aoi, Co-Chair
Professor Kyle Hasenstab
Professor Liam Muller

2022

The Thesis of Jun Xian Rao is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

TLE	Temporal Lobe Epilepsy
MTL	Mesial Temporal Lobe
VEEG	Video-Electroencephalogram
MRI	Magnetic Resonance Imaging
ML	Machine Learning
DL	Deep Learning
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
IRB	Institutional Review Board
Bonn	University of Bonn
Cleve	Cleveland Clinic
Emory	Emory University
MUSC	Medical University of South Carolina
Rush	Rush University
UCSD	University of California San Diego
UCSF	University of California San Francisco
CAT12	Computational Anatomy Toolbox
SPM	Statistical Parametric Mapping
TPM	Tissue Probability Maps
MNI	Montreal Neurological Institute
ROI	Regions of Interest

QC	Quality Control
EDA	Exploratory Data Analysis
FDR	False Discovery Rate
EDA	Exploratory Data Analysis
Conv	2D Convolutional Layer
ReLU	Rectified Linear Unit Activation
BatchNorm	Batch Normalization Layer
MaxPool	Max-Pooling Layer
PLS	Partial Least Square
LR	Logistic Regression
FDA	U.S. Food and Drug Administration
Grad-CAM	Gradient-weighted Class Activation Mapping
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic Curve
AO	Age of Onset
DURILL	Duration of Illness
MTS	Mesial Temporal Sclerosis
ASM	Anti-seizure Medication
Adam	Adaptive Moment Estimation
SGD	Stochastic Gradient Descent
STD	Standard Deviation
TPR	True Positive Rate (L-TLE Classified as L-TLE)
FPR	False Positive Rate (R-TLE Classified as L-TLE)

TNR	True Negative Rate (R-TLE Classified as R-TLE)
FNR	False Negative Rate (L-TLE Classified as R-TLE)
GAN	Generative Adversarial Network

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Carrie McDonald for the continuous support of my master's study and research. I would like to thank her for allowing me to volunteer here at her lab and eventually complete my master's project here. Her guidance helped me in all the time of research and through the writing of this thesis. I could not have imagined having a better advisor and mentor for my research career.

Besides my advisor, I would like to acknowledge and give my warmest thanks to Dr. Erik Kaestner for his mentorship throughout my time in the lab. He guided me from the beginning and taught me how to be the researcher I am today. He had taught me everything from conducting research to writing reports and everything along the way. He had been an instrumental mentor and will always be.

I would also like to thank Dr. Kyle Hasenstab for his help on this master's project and in my data science career. I want to thank him for all the mentorship he provided for me to become a Machine Learning researcher. I would not be able to complete this project without him taking the time to meet with me weekly to discuss project findings and next steps.

I would also like to thank my committee members, Professor Mikio Aoi and Professor Liam Muller, who have been the best professors I have had the opportunity to take or work for and who kindly accepted to be my committee members.

In the end, I would like to thank all the members of McDonald Lab, Adam Schadler, Dr. Donatello Arienzo, Dr. Alena Stasenko, Dr. Anny Reyes, Daniel Asay, Anna Christina Macari, and Akshara Balachandra, for all the help over the years and making my time at the lab the best that I could ever ask for.

ABSTRACT OF THE THESIS

Artificial Neural Network Application for Detecting Seizure Focus using Neuroimaging

by

Jun Xian Rao

Master of Science in Biology

University of California San Diego, 2022

Professor Carrie McDonald, Chair

Professor Mikio Aoi, Co-Chair

Temporal lobe epilepsy (TLE) is the most common form of focal epilepsy in adults. Many of those affected do not respond well to anti-seizure medications and require neurosurgical intervention. For these patients, locating the seizure focus to the left or right

hemisphere is a critical first step in surgical planning. Through the use of Convolutional Neural Networks (CNN) that can survey medical diagnostic images in finer detail and extract more convoluted information than humans, we proposed that it will aid in the localization of epileptogenic regions (i.e., seizure generating regions) in the brain. We explored different pre-trained and novel CNN architectures trained on 364 patients with TLE from seven different epilepsy centers to ensure generalizability. Our findings show that a CNN model outperforms a more standard linear model like a Logistic Regression (LR) in performing the task of side-of-onset classification, with the best CNN model outperforming the average LR model by 17.39% in classification accuracy. The model also shows that information important for determining side-of-onset is not limited to the mesial temporal lobe (MTL) but is also located in extra-temporal regions like the parietal lobe, precentral gyrus, and cingulate gyrus. This study shows that the classification problem of Left versus Right TLE patients benefits from a more complex, nonlinear model and whole-brain information and therefore medical examination of TLE would benefit from incorporating machine learning to aid in clinician-led localization of the epileptogenic zone.

INTRODUCTION

Epilepsy is the abnormal firing of neurons that causes chronic seizures, behavior changes, and even cognitive impairments such as memory loss in patients. It is a devastating neurological disorder affecting 1.5 to 5% of the worldwide population, of which the majority of adults with focal seizures have temporal lobe epilepsy (TLE) (Bell et al., 2014). TLE is focal epilepsy where the seizure starts in the temporal lobe of the brain. This is important because our temporal lobe is critically involved in memory, language, and emotions among other processes. The effects of having a seizure in this region of the brain can range from having mood swings, and memory loss to oral and motor automatisms (Engel, 1996). The most common and most available treatment for epilepsy is anti-seizure medications. But one in three individuals with epilepsy does not respond to anti-seizure drugs and therefore requires further treatment, often neurosurgical intervention (French et al., 2007). More invasive treatments like surgery to remove the seizure focus are available but require further diagnoses to localize this region. Currently, Video-EEG (VEEG) is the primary method to localize seizures where the neural network in the patient's brain that is generating seizures is identified. But VEEG is supported by a range of additional measures during surgical decision-making, and neuroimaging is one of the most important supporting tools (French et al., 2007). Improvements in locating which neurons are capable of inducing seizures, and therefore are potential neurosurgical targets, would be useful in matching patients to these individualized treatments.

Structural imaging plays a critical role in supporting TLE diagnosis and localization of seizure onset. Specifically, MRI can complement VEEG by detecting structural changes in the brains of TLE patients that often co-localize with the seizure focus. Structural MRI is a non-

invasive, and rapid imaging tool to visualize the gray and white matter abnormalities in the patient's brain. It is used during diagnosis to look for observable damage, like lesions, in the brain. However, about 40% of the patients have MRI scans that appear normal upon visual inspection, even for experienced epilepsy neuroradiologists (So et al., 2015). Studies have shown that, although undetectable to a human observer, micro-changes such as thinning in specific cortical regions or changes in how neurons connect will affect the severity of the disease (French et al., 2007). Therefore, tools that can help identify subtle structural changes in the brain and locate epileptic neurons that cause the seizure could improve treatment and surgery outcomes and allow us to have a better understanding of TLE.

Machine Learning Approaches to Medical Imaging

Recent developments in machine learning (ML) are enabling us to analyze complicated data such as neuroimaging with complex relationships. With the improvements in automation and advances in computational power, we have obtained large amounts of data and the ability to analyze them. The field of ML came into existence to leverage the amount of data available (Ramesh et.al., 2004). Here, the goal is to develop computer algorithms that can learn from the data, identify patterns, make predictions or decisions, or even create statistical models to understand the world around us. Recent improvements in ML have caused it to generate interest in the biomedical research field as a clinical decision support tool (Deo, 2015). Many ML methodologies developed to classify everyday objects in images, such as artificial neural networks, have been explored for their use in medical image analysis.

In artificial neural networks, biological neurons are imitated to create networks that work similarly to biological neural networks in the brain and can learn to make decisions (LeCun, et al.,

2015). They are similar because artificial neural networks are a specific form of ML where human input in the learning process is limited, and the algorithm learns the pattern itself. There are many forms of artificial neural networks, and their applications are endless, from speech recognition to natural language processing, where a machine learns to understand and replicate human language, to image classification. It can provide a bridge between the traditional biology of doing bench work and the growing computational resources and processing power (Harmet, Tremblay, 2017). They allow us to perform experiments and analyze them in quantities unimaginable to biologists even a decade ago. A specific form of the artificial neural network model that has a growing interest in the medical imaging field is Convolutional Neural Network (CNN) (Litjens et al., 2021). CNN works by breaking down images into significant features and using these features to make decisions on the input. These deep learning models are shown to be useful in image processing and be a valuable tool in medical imaging analysis because they can extract features within the MRI scans that are not easily detectable or comprehensible to the human observer (Abbasi et al., 2019).

The main components of CNNs are the convolutional layers where important features of the image are extracted. They work by passing in signal information from a previous layer through filters that adjust what kind of signal passes through, which then will cause specific activations in the current layer and that information will be passed on to the next layer of neurons through more filters. These activations are the so-called features, and the combinations of these features create a feature map of a specific layer. Layers can also have different numbers of filters and different filter sizes and learn to extract different features out of the input (O'Shea et al., 2015). The process of feature extraction allows the program to represent the original input image in compact and separate information bits. One feature map could be about the brightness of the image while

another could be about edges present in the image and another about the shapes of objects in the image. Taking into account all these features, the algorithm then decides based on all the features present. We know that TLE causes volumetric differences in the hippocampus of the patient, and possibly extra-temporal structural changes (Farid et al. 2012). CNN could be a pivotal tool in providing us with more information about the structural features that are affected by the disease.

Machine Learning and Epilepsy

Previous research has shown that structural MRI images can be used to train a CNN model to detect the presence of epilepsy (i.e., classify TLE patients against controls) (Gleichgerrcht et al., 2021). This trained CNN model was able to perform better than trained medical personnel in identifying epilepsy patients and showed that extratemporal pathology may help to differentiate the brains of patients with TLE from those of healthy controls. In this project, we will expand upon prior research by addressing the more complex question of whether a CNN model applied to structural MRI data can be used to lateralize the seizure focus (i.e., determine left versus right temporal lobe seizure onset). The goal for my CNN model is to provide the ability to examine differences at a much finer spatial resolution, potentially revealing novel patterns of disease location and providing enhanced classification and detection. My CNN model approach is intended to be used to complement VEEG in the diagnosis and treatment of TLE. Another goal of this project is to use this CNN model to learn more about the location and extent of pathology in TLE and how it differs between patients with Left versus Right TLE.

MATERIALS

Patient Dataset

Institutional Review Board (IRB) approval for data sharing was obtained from each of the participating sites. 364 patients across seven sites were included in this study, of which, 200 have Left Temporal Lobe Epilepsy (L-TLE) and 164 have Right Temporal Lobe Epilepsy (R-TLE). Sites included the University of Bonn (L-TLE n=24, R-TLE n=12), Cleveland Clinic (L-TLE n=26, R-TLE n=6), Emory University (L-TLE n=42, R-TLE n=47), Medical University of South Carolina (L-TLE n=27, R-TLE n=16), Rush University (L-TLE n=25, R-TLE n=28), University of California San Diego (L-TLE n=34, R-TLE n=40), and University of California San Francisco (L-TLE n=22, R-TLE n=15). Side of onset diagnosis was determined at the site by a medical team of specialists including neurosurgeons, neurologists specializing in epilepsy, neuroradiologists, and neuropsychologists using converging approaches including imaging, neurophysiology, semiology, and neuropsychological methodologies (e.g., video-EEG).

MRI preprocessing

Patient scans included in this study contain both 1.5T and 3T structural MRI brain scans. Scanner manufacturers and imaging acquisition parameters from each site are listed as follows:

- Bonn 3T: Siemens Magnetom Trio 3T scanner, 8-channel head coil, isotropic voxel size of 1mm, TR = 650ms, TE = 3.97ms, TI = 650ms, flip angle 10°
- Cleveland 1.5T: 1.5T, isotropic voxel size of 1.25mm, TR = 11ms, TE = 4.6ms, flip angle 20°

- Cleveland 3T: 3T, isotropic voxel size of 1mm, TR = 1860ms, TE = 3.4ms, TI = 1100ms, flip angle 10°
- Emory 3T: Siemens Prisma 3T scanner, 12-channel head coil, isotropic voxel size of 0.8mm, TR = 2300ms, TE = 2.75ms, TI = 1100ms, flip angle 8°
- MUSC 3T: Siemens Skyra 3T scanner, 12-channel head coil, isotropic voxel size of 1mm, TR = 2050-2250ms, TE = 2.5-18ms, flip angle 10°
- Rush 3T: Siemens Verio 3T scanner, 8-channel head coil, isotropic voxel size of 0.6mm, TR = 2300ms, TE = 3.29ms, TI = 900ms, flip angle 8°
- UCSD 1.5T: 1.5T, 8-channel head coil, isotropic voxel size of 1mm, TR = 10.73mm, TE = 4.87ms, TI = 1000ms, flip angle 8°
- UCSD 3T: GE Discovery MR750, 8-channel head coil, isotropic voxel size of 1mm, TR = 8.132ms, TE = 3.172mm, TI = 600ms, flip angle 8°
- UCSD 3T: GE Discovery MR751, 8-channel head coil, isotropic voxel size of 1mm, TR = 8.1ms, TE = 3.156mm, TI = 600ms, flip angle 8°
- UCSF 3T: GE Discovery MR751, 8-channel head coil, isotropic voxel size of 1mm, TR = 8.1ms, TE = 3.156mm, TI = 600ms, flip angle 8°

METHODS

Preprocessing

Images were preprocessed using the Computational Anatomy Toolbox (CAT12) (Gaser et al., 2016) extension of the standard voxel-based morphometry software Statistical Parametric Mapping (SPM12). We first segmented images into gray, white, and CSF tissue types using the Tissue Probability Maps (TPMs) provided by SPM. Images were then spatially registered to the Montreal Neurological Institute brain atlas (MNI) space using the Geodesic Shooting method. Hippocampal volumes were also computed from the ROIs provided by CAT12 as a comparative reference.

Quality control of scans and output images was performed using the automated retrospective QC tools provided by CAT12, which accounts for noise, inhomogeneity, and image resolution to rate scan quality. All scans rated below satisfactory (<75) were removed from the study.

Study Design

First, we performed an Exploratory Data Analysis (EDA) to understand the MRI data and the difference between the model classes. The study is divided into three parts. The first part explores the different CNN architectures to identify the best architecture for further finetuning and analysis. The second part finetunes the best architecture and comparison to reference models. The last part obtains the feature importance map of the model to study where the CNN is focusing to make its classification.

Exploratory Data Analysis

To understand the data and differences between the left and right TLE groups, a two-sample Welch's t-test is done on the left and right TLE groups by comparing each voxel intensity across the whole gray matter mask. The p-value is False Discovery Rate (FDR) adjusted and threshold at 0.01 to create a p-value map to visualize the difference in brain structure between the two groups at the voxel level.

Baseline CNN model

The CNN contains three Conv-ReLU-BatchNorm-MaxPool layers. The 2D Convolutional (Conv) layers contain 8, 16, and 32 filters respectively. Max pooling was performed with a stride of 2. Convolutional layers are followed by a Softmax dense layer with 2 outputs, one for Left TLE and one for Right TLE.

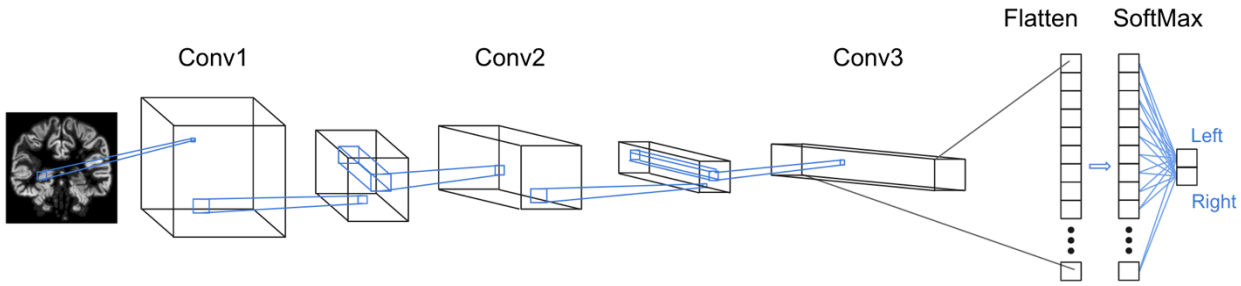

Figure 1: **Baseline Model Architecture.** Single coronal slice input image. The first convolutional layer has 3 by 3 2D convolution with ReLU activation and batch normalization. It has a shape of 160 by 160 by 8. This is followed by a 2 by 2 max-pooling layer with a shape of 80 by 80 by 8. The second convolutional layer has 3 by 3 2D convolution with ReLU activation and batch normalization. It has a shape of 80 by 80 by 16. This is followed by a 2 by 2 max-pooling layer with a shape of 40 by 40 by 16. The last convolutional layer has 3 by 3 2D convolution with ReLU activation and batch normalization. It has a shape of 40 by 40 by 32. It is then flattened and a fully connected dense layer with SoftMax activation outputs the model prediction for the left or the right side of seizure onset.

Input Selection

After building a working baseline CNN model, inputs to the model are selected using grid search. It was shown that coronal slices of the brain work best for a similar task as much of the atrophy is located in the MTL region which can be best shown in the coronal view (Gleichgerrcht et. al., 2021). A grid search of all coronal slices is performed to identify the optimal slice.

Novel Architectures

Multiple Input Slice Architecture

With multiple slices producing models with good performances, we tested a 2.5D CNN approach where we input 4 additional neighboring slices (2 anterior slices and 2 posterior slices) to give the model more information about the general region of the temporal lobe.

Transfer Learning Using ResNet50

State-of-the-art models trained on the ImageNet dataset were used as transfer learning models in our study to test their potential for classifying between left and right TLE patients. The ResNet50 model was selected for its residual convolutional blocks that allow for a deeper network without the tradeoffs (He, et. al., 2015). The model was downloaded from the Keras database (https://keras.io/api/applications/) with weights being pre-trained on the ImageNet dataset. The last layer of the model was removed and a new softmax dense layer was added to the output of the two classes. These models were also trained using the Adam optimizer and the learning rate was set at the default setting of 0.01.

Models Incorporating Hippocampal Volume

Previous research has shown that hippocampal volume is a good predictor of the side of onset in TLE patients (Farid, et. al, 2012). We have decided to include hippocampal volume information into the model in two different methods.

The first method of inputting hippocampal volume into the deep learning model is as two additional channels. One channel where every pixel has the value of the left hippocampal volume and the second channel with the same setup but now with the value of the right hippocampal volume. This is a straightforward method to combine the hippocampal volumes into the input images.

The second method is to extract the feature map from the last convolutional layer of the multiple input slice CNN, input it into a Partial Least Square (PLS) model to reduce the dimension and extract the 20 most important features, and lastly combine these 20 features with the two hippocampal volumes into a Logistic Regression (LR) model to classify between left and right TLE patients.

Model Training and Hyperparameter Search

CNN Training

Training of the CNN models in each of the experiments in this study is done with the same parameters. Cross-validation is not implemented in this study due to the low sample size, instead, 100 models were trained with randomized patients at each run and the average performance of the models are compared. The data are first separated into a training set with 70% of the patients, a validation set with 15% of the patients, and a testing set with the remaining 15% of the patients. Each split of the data maintains the class imbalance to guarantee equal comparisons. In the final

stage of the study after the models are tested and tuned, the same patient splits are used to train the three models (the 2.5D CNN with hippocampal volumes, the LR model using hippocampal volumes only, and the random model) so that they can be compared side-by-side.

Model Tuning and Hyperparameter Search

With the input of the model selected, the model hyperparameters are optimized using grid search. Stochastic gradient descent is compared with the newer Adam optimizer, where the Adam optimizer produces similar performance but with reduced training times. The binary cross-entropy loss function was chosen since there were only two classes in this study. We then added a regularizer to the last convolutional layer to stop certain weights from becoming too heavy in the model. The learning rate and lambda of the regularizer are tuned together as they are dependent on each other. During the model training process, an early stopping mechanism was implemented that stops the training of the model when validation loss has not decreased in 15 epochs.

Reference Models

To test the validity of using a deep learning approach to tackle the problem of classification of left versus right TLE patients, we also trained an LR model using only the two hippocampal volumes as inputs. This tests if a linear model can produce similar or even better results than the more complex nonlinear deep learning model. Hippocampal volumes were chosen as input because they are used in FDA-cleared medical imaging programs such as NeuroQuant (Stelmokas, et. al. 2018). We also trained a random model where the training set labels are randomly shuffled to not allow the model to learn any information about the classes (the validation and testing sets are not shuffled). This will test if the model is picking up class imbalance (when the CNN model performs

similarly to the random model) or learning features that are important to differentiate the two classes (when the CNN model outperforms the random model).

Grad-CAM Maps of Feature Importance

To visualize the important features discovered by the model, we created Grad-CAM maps of the last convolutional layer of the 2.5D CNN model with hippocampal volumes. We first average all the Grad-CAM maps of a single patient from all 100 models, then we average all those averages along with all subjects from the same group to produce a Grad-CAM map of what the CNN model will focus on when presented with either a left TLE patient or right TLE patient.

Model Testing and Statistical Analysis

The performance metrics of each of the three algorithms (the 2.5D CNN model with hippocampal volumes, the LR model using only hippocampal volumes, and the random model) are compared to evaluate the performance of each algorithm compared to the others. Paired t-tests are used to compare the single value performance metrics (e.g., accuracy, sensitivity, specificity, area under the curve (AUC)). The diagram of the Receiver Operating Characteristic Curve (ROC) curves is produced with the mean curve between all the runs and a 95% confidence interval curve for each algorithm. The confusion matrix from each run of the 2.5D CNN model and the LR model is also combined into their respective average confusion matrix for the evaluation of their performance in each class and to compare between algorithms.

RESULTS

Patient Demographics

Two hundred Left TLE patients and 164 Right TLE patients were included in this study. Shown below in **Table 1** are the patient demographics for each group. There appears to be no significant difference in age, sex, handedness, education, age of onset (AO), duration of illness (DURILL), mesial temporal sclerosis (MTS) status, antiseizure medications (ASM), surgery status, and Engel (surgery outcome score) between the two sides of epilepsy onset.

Table 1: **Demographics table.** The table shows the age, sex, handedness, education, AO, DURILL, MTS status, ASMs, Surgery status, and Engel score breakdown between the two classes. The last column shows the t-test outcome for the difference between the two classes for each demographic criteria.

	L-TLE	R-TLE	Test
N	200	159	-
Age	40.2 (15)	37.6 (13)	t(307)=1.66; p=0.099
Sex (F/M)	119/81	94/65	FET=1.02; p=1
Handedness (L/R)	48/128	31/115	FET=1.39; p=0.24
Education	13.8 (2.4)	13.8 (2.4)	t(278)=0.263; p=0.79
AO	21 (15)	19.2 (14)	t(357)=1.15; p=0.25
DURILL	18.4 (16)	17.5 (14)	t(179)=0.366; p=0.71
MTS (no/yes)	130/66	104/55	FET=1.04; p=0.91
ASMs	2.24 (0.91)	2.36 (0.98)	t(312)=-1.11; p=0.27
Surgery (no/yes)	54/139	43/112	FET=1.01; p=1
Engel (I/II/III/IV)	73/14/8/4/6	59/6/7/2/2	FET; p=0.61

Voxel-Wise Analysis of Epilepsy Onset

After preprocessing and transforming all the scans to the same imaging space, we initially performed a voxel-wise analysis between the two groups of patients to understand the differences in their brain anatomy at the level of a single voxel. The masked p-value map of the voxels that are statistically significant after FDR correction and $p < 0.01$ are shown in **Figure 1**. We can see from the axial (left), coronal (middle), and sagittal (right) views of the brain that most of those highlighted regions are within the hippocampus bilaterally. It should also be noted that the highlighted regions are qualitatively more overlapping the left hippocampus than the right hippocampus. There are a few voxels with significant differences outside of the hippocampal area shown in the sagittal and axial views. However, those regions are still within the temporal lobe of the patient's brain.

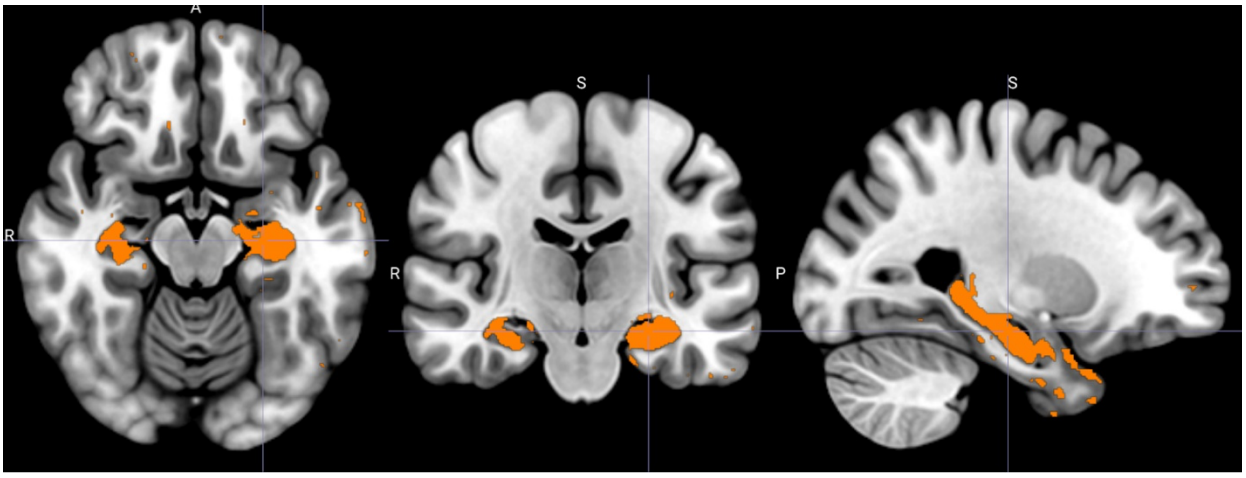

Figure 2: **Voxel-wise t-test of gray matter intensity between left and right TLE patients.** The brain is displayed in the following order: axial (left), coronal (center), and sagittal (right) view. The figure shows significant voxels after running a two-sample t-test comparing voxel intensity between left and right TLE patients' gray matter maps. The voxels are FDR corrected with a p-value cutoff of 0.01. Most of the significant voxels are in the hippocampus with a small number of voxels outside of the hippocampus being significant and they are still within the temporal lobe.

Optimal Slice

The model used in this study implements 2-dimensional convolutional layers while the scans are in 3-dimensional space, thus we tested the performance of the model using different slices as input. The result of the per slice performance is shown in **Figure 2**. We can see that there is a sharp increase in model accuracy after slice 78 and peaks at slice 87. From slice 87, the model accuracy slowly declines and by slice 141, the model has similar accuracy to when slice 78 is used as input. Looking at the single slice accuracy of the models, we determine that the top 5 sequential slices for the model are from slice 84 to slice 96. **Figure 3** shows the location of each coronal slice selected in the final model. We can see that it starts at the posterior hippocampus (at slice 84) and toward the middle of the hippocampus (at slice 96).

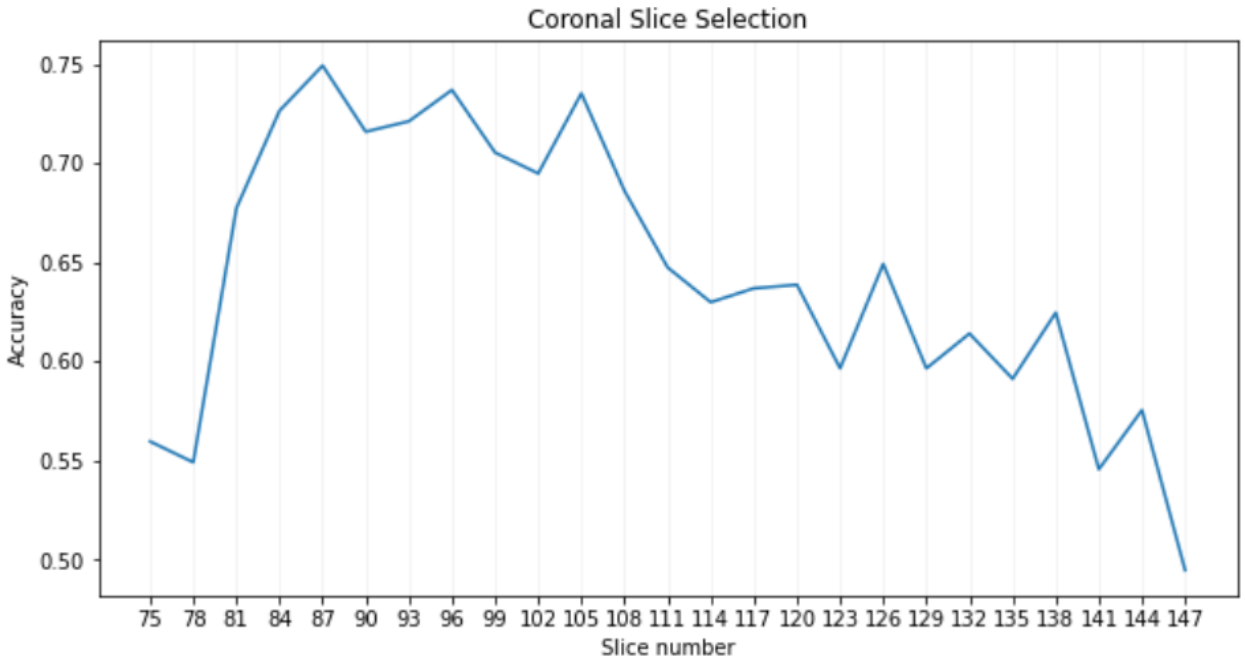

Figure 3: **Slice accuracy plot.** The plot of the accuracy obtained by a model using the slice as input. There is a sharp increase after slice 78, peaks at slice 87, and then slowly decline in accuracy in later slices.

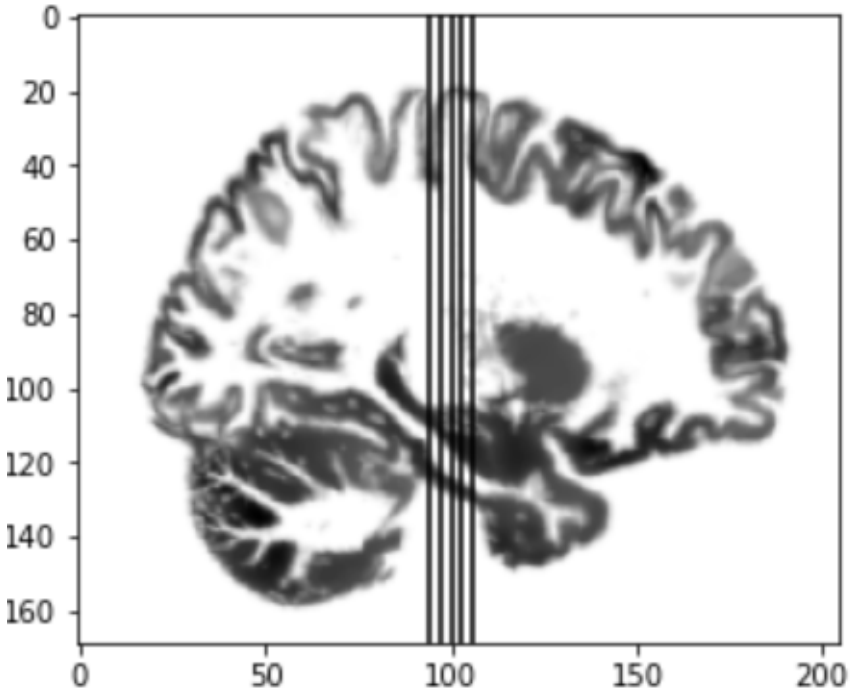

Figure 4: **Locations of the 5 slices chosen as input for the model.** Slice 84, represented by the left-most vertical red line in the figure, is at the posterior hippocampus. The last of the five slices used, slice 96, is in the middle of the hippocampus.

Hyperparameters used

A grid search was performed to determine the best hyperparameters for the model's training (see **Table 2**). Two optimizers were tested, the more established Stochastic Gradient Descent (SGD) and the new optimizer Adam. Both optimizers yielded similar model performance and the final model was trained using the Adam optimizer for its quicker training time.

Table 2: **Hyperparameter results.** This table contains the hyperparameters that will result in the best-performing models for the dataset and model architecture.

Optimizer	Learning rate	Epsilon (Adam)	Regularization Delta (L2)
SGD (Default)	1e-2	N/A	N/A
SGD	1e-3	N/A	1e-3
Adam	1e-4	1e-8	1e-2

Comparison Between CNN Architectures

The first experiment in this study is exploring the best CNN architecture for the task of classifying left and right TLE patients. **Table 3** shows the model performances of the different model architectures. A Welch's two-sample t-test is performed between the accuracy of each model architecture and the final 2.5D CNN model architecture we proposed, and the p-value is then Bonferroni corrected with alpha = 0.01, n = 100, p-adjusted = 1e-5.

We begin with the simple 2-dimensional (2D) CNN model with layer specifications listed in the Methods section. The average accuracy of the model is 67.34% and we can see that there is a significant difference ($p<1e-5$) in the performance of the 2D CNN and the 2.5D CNN. The second boxplot from **Figure 4** is the accuracy of the 2D CNN and we can see that it ranges from 50.90% to 85.50% with a median of 67.30% and no outliers or skewness.

We then implemented transfer learning using ResNet50 trained on the ImageNet dataset and fine-tuned it to perform the classification task of left and right TLE patients. This architecture has an average accuracy of 54.69% and has performances that are significantly lower than the 2.5D CNN architecture ($p<1e-5$). The third boxplot from **Figure 4** is of the transfer learning model where we see a median accuracy of 54.55%, no spread (where most of the models have a 54.04% accuracy), and outliers at 61.82%.

The ensemble method was one of the ways we incorporated hippocampal volume data into the deep learning model. It has an average accuracy of 71.13% and performs significantly differently from the 2.5D CNN model ($p<1e-5$). The fourth boxplot from **Figure 4** shows the distribution of accuracies of the ensemble model where it has a median accuracy of 70.91%, with accuracies ranging from 56.36% to 83.64% and no clear skewness or outliers.

The final architecture implemented is a multi-slice 2D CNN with hippocampal volumes as extra input channels that we call the 2.5D CNN model. This model obtained an average accuracy of 77.96%. The first boxplot from **Figure 4** is of the 2.5D CNN model where we see a median accuracy of 78.20%, with accuracies ranging from 63.60% to 89.1% and no outliers

Table 3: **Model accuracy distribution statistics for different model architectures.** The table shows the model performance statistics over 100 runs with their mean, standard deviation (std), minimum (min), lower quartile (25%), median (50%), upper quartile (75%), and maximum (max). It contains data for the 2D CNN, transfer learning using ResNet50, ensemble, and the proposed 2.5D hybrid CNN architectures.

	2.5D CNN	2D CNN	ResNet50	Ensemble Model
count	100.000000	100.000000	100.000000	100.000000
mean	0.779610	0.673430	0.546909	0.711273
std	0.051428	0.073976	0.010233	0.062252
min	0.636000	0.509000	0.545455	0.563636
25%	0.745000	0.618000	0.545455	0.672727
50%	0.782000	0.673000	0.545455	0.709091
75%	0.818000	0.727000	0.545455	0.763636
max	0.891000	0.855000	0.618182	0.836364

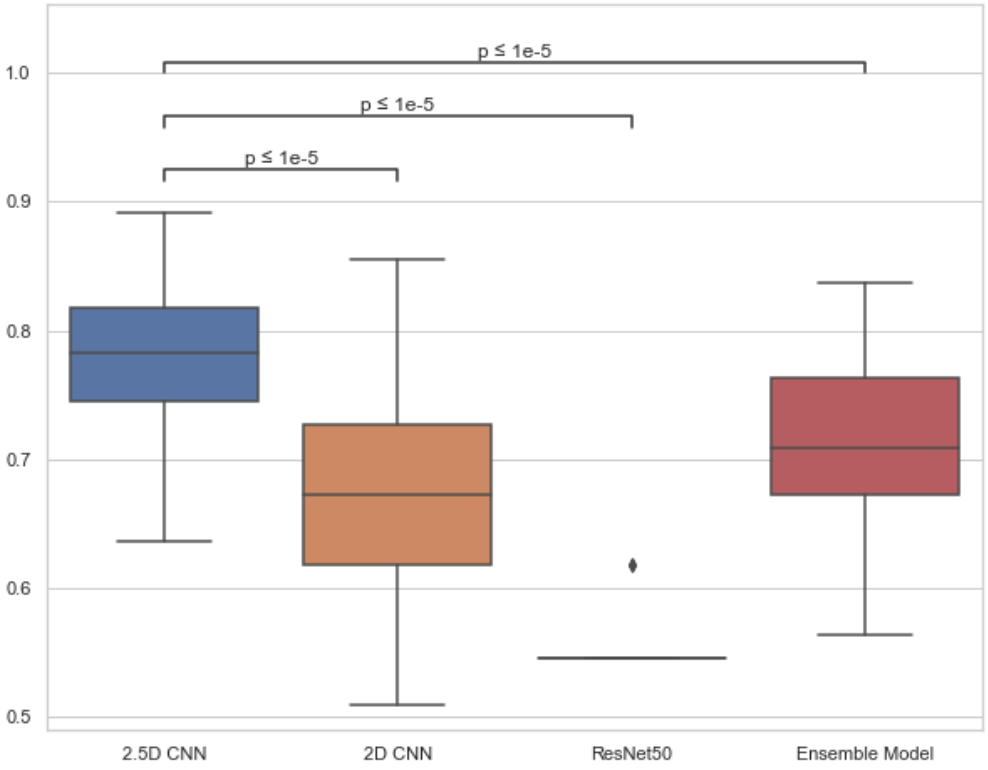

Figure 5: **Distribution of model accuracies of different CNN architectures.** Boxplot of the performances of the different models tested in this study. It includes 2.5D CNN, 2D CNN, transfer learning using ResNet50, and ensemble learning. There is a significant difference between the 2.5D CNN model and the other models where the p-values are significant after Bonferroni correction (alpha=0.01 and n=100, p-adjusted=1e-5).

Performance of 2.5D Model Relative to the Reference Models

The second experiment in this study is comparing the final CNN architecture to the reference models, a randomized model, and a logistic regression model trained on only hippocampal volume. This enabled us to establish its significance both against chance (randomized model) and against simple approaches (logistic regression model trained on only hippocampal volume). **Table 4** shows the distribution of different model accuracies. A Welch's paired t-test is performed between the accuracy of each model architecture and the final 2.5D CNN model architecture we proposed, and the p-value is then Bonferroni corrected with alpha = 0.01, n = 100, p-adjusted = 1e-5. Refer to the previous section for the distribution of 2.5D CNN model accuracies. The first comparison is to the Random model where during training, the train labels are shuffled to ensure that the CNN will not learn anything from the data. It has an average accuracy of 51.65%. The second boxplot from **Figure 5** is of the Random model where we see a median accuracy of 56.36%, with accuracies ranging from 21.82% to 67.27%, a few outliers at the low end, and a left-skewed distribution.

The next reference model is a logistic regression model (LR Model) trained using only the value of the left and right hippocampal volume; the resulting model has an average accuracy of 71.71%. The third boxplot from **Figure 5** is of the logistic regression model where we see a median accuracy of 71.82%, with accuracies ranging from 58.18% to 83.64%, and with a few outliers at both ends.

Table 4: **Model accuracy distribution statistics for the 2.5D CNN and reference models.** The table shows the model performance statistics over 100 runs with their mean, standard deviation (std), minimum (min), lower quartile (25%), median (50%), upper quartile (75%), and maximum (max). It contains data for the 2.5D hybrid CNN (2.5D CNN), the Random Model, and the LR model.

	2.5D CNN	Random Model	LR Model
count	100.000000	100.000000	100.000000
mean	0.779610	0.516545	0.717091
std	0.051428	0.080573	0.053296
min	0.636000	0.218182	0.581818
25%	0.745000	0.472727	0.690909
50%	0.782000	0.536364	0.718182
75%	0.818000	0.563636	0.745455
max	0.891000	0.672727	0.836364

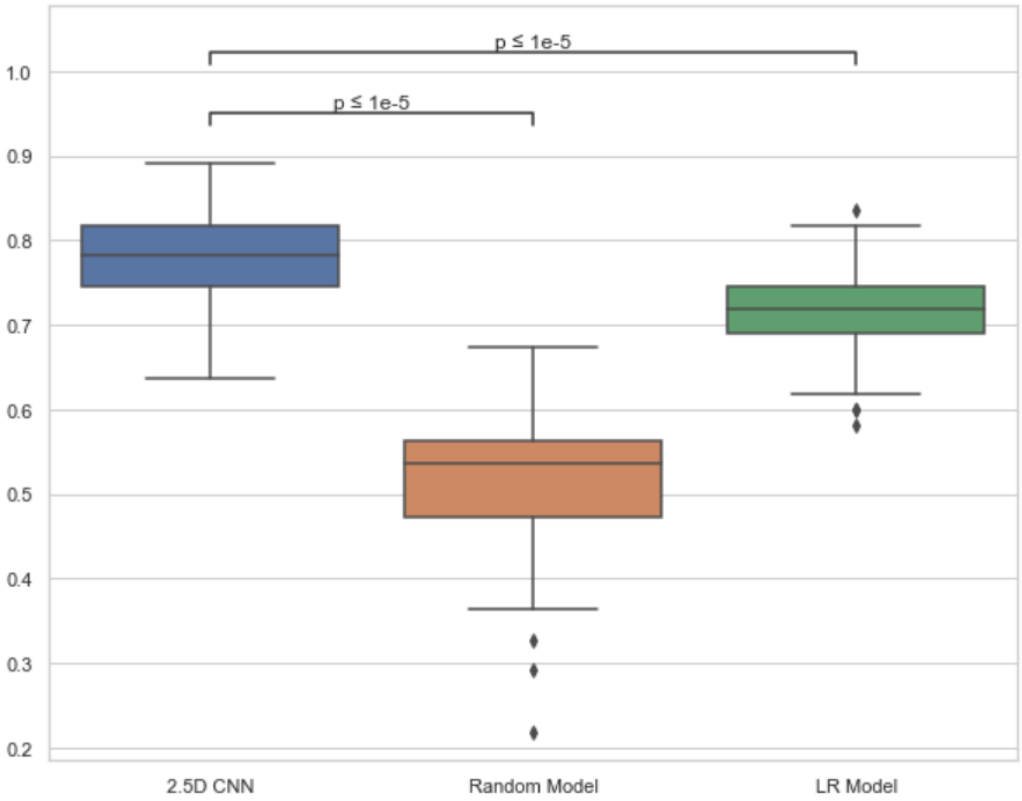

Figure 6: **Distribution of model accuracies of 2.5D CNN and reference models.** Boxplot of the performances of the different models tested in this study. It includes the 2.5D CNN, the Random model, and the LR model. There is a significant difference between the 2.5D hybrid CNN model and the other models where the p-value is significant after Bonferroni correction (alpha=0.01 and n=100, p-adjusted=1e-5).

A pairwise comparison was performed between the 2.5D CNN and random model and then another one between the 2.5D CNN and logistic regression model. **Figures 6a** and **6b** show the distribution of values after a pair-wise subtraction of the 2.5D CNN accuracy by the respective reference model (random model for **6a** and logistic regression model for **6b**). The distribution of the pairwise difference in accuracy between the 2.5D CNN and the random model shown in **Figure 6a** shows that the 2.5D CNN always outperforms the random model and outperforms it on average by 26.31%. The distribution of the pairwise difference in accuracy between the 2.5D CNN and the logistic regression model shown in **Figure 6b** shows that the 2.5D CNN outperforms the logistic regression model most of the time where a few subject splits have logistic regression models that outperform the 2.5D CNN model. The 2.5D CNN model outperforms the logistic regression model on average by 6.25%.

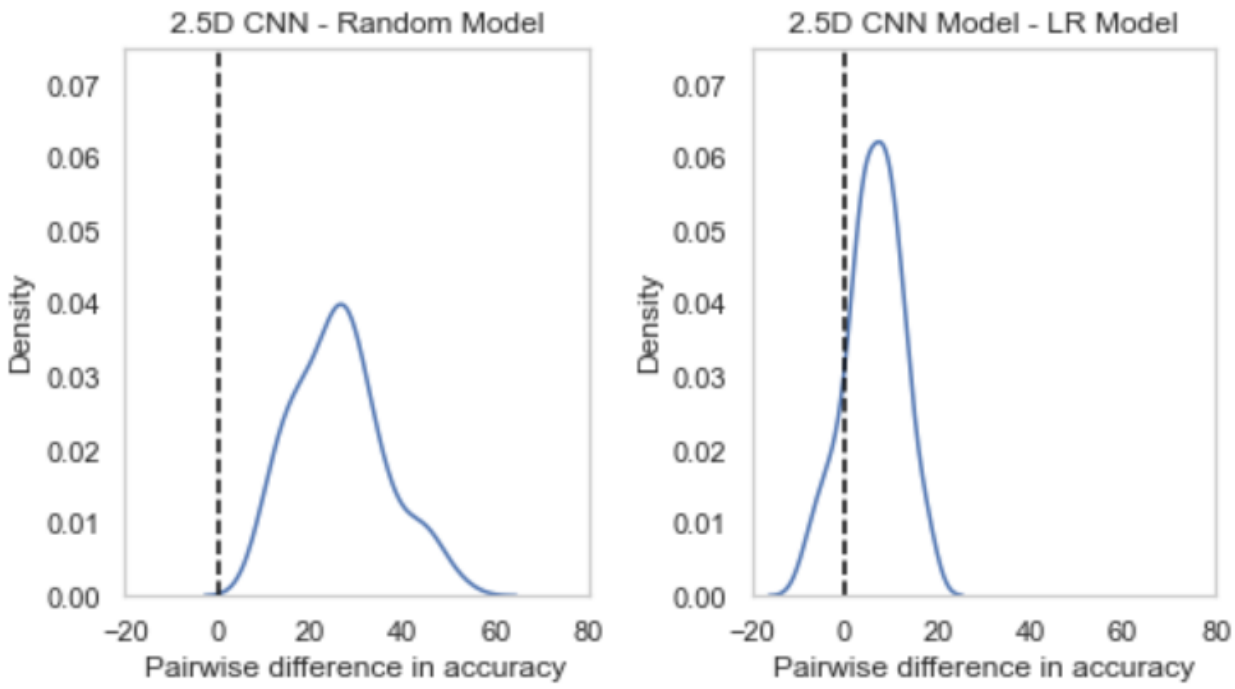

Figure 7a/7b: **The distribution of the pairwise difference in accuracy between the 2.5D CNN and the random model (left) and the logistic regression model (right) respectively.** A pair-wise subtraction of the 2.5D CNN model accuracy with the respective reference model accuracy at each run.

Then we analyze the specific model performances of each model. **Table 5** shows the different measures of model performance of the 2.5D CNN model and the two reference models. The 2.5D CNN model has a true positive rate (TPR) (defined as left classified as left) of 0.45, a false positive rate (FPR) (defined as right classified as left) of 0.09, a false negative rate (FNR) (defined as left classified as right) of 0.13, a true negative rate (TNR) (defined as right classified as right) of 0.33, a sensitivity of 0.78, a specificity of 0.78, a precision of 0.83, a negative rate of 0.72, and an AUC of 0.86. The random model has a TPR of 0.41, an FPR of 0.14, an FNR of 0.35, a TNR of 0.11, a sensitivity of 0.54, a specificity of 0.44, a precision of 0.75, a negative rate of 0.23, and an average AUC of 0.49. The LR model has a TPR of 0.43, an FPR of 0.11, an FNR of 0.17, a TNR of 0.28, a sensitivity of 0.72, a specificity of 0.72, a precision of 0.79, a negative rate of 0.63, and an average AUC of 0.84.

Table 5: **The measures of model performance of the different models.** The table contains data on the true positive rate (TPR) (L-TLE classified as Left), false positive rate (FPR) (R-TLE classified as Left), false negative rate (FNR) (L-TLE classified as Right), true negative rate (TNR) (R-TLE classified as Right), average accuracy, sensitivity, specificity, precision, negative rate and average area under the curve (AUC) of the 2.5D hybrid CNN, the Random model, and the Logistic Regression model.

	2.5D CNN	Random Model	LR Model
TPR	0.829333	0.751333	0.793000
FPR	0.281200	0.766400	0.374000
FNR	0.170667	0.248667	0.207000
TNR	0.718800	0.233600	0.626000
Accuracy	0.774067	0.492467	0.709500
Sensitivity	0.829333	0.751333	0.793000
Specificity	0.718800	0.233600	0.626000
Precision	0.746788	0.495036	0.679520
Negative Rate	0.808125	0.484379	0.751501
Average AUC	0.856940	0.491480	0.836320

Finally, we have the ROC of the three different algorithms in **Figure 7**. The red curve (2.5D hybrid CNN) has the highest average AUC while the logistic regression model's ROC (green) starts with a higher true positive rate. However, the red 2.5D hybrid CNN ROC curve soon went above the green Logistic Regression ROC curve until the very end. We can also see that there is a lot of overlap between the ROC ranges between the 2.5D hybrid CNN model and the Logistic Regression model with the 2.5D hybrid CNN model having slightly higher AUC over all the runs. The blue ROC curve is of the random model, and it is around the diagonal meaning it has not learned anything and truly is a random model with an AUC of around 0.5.

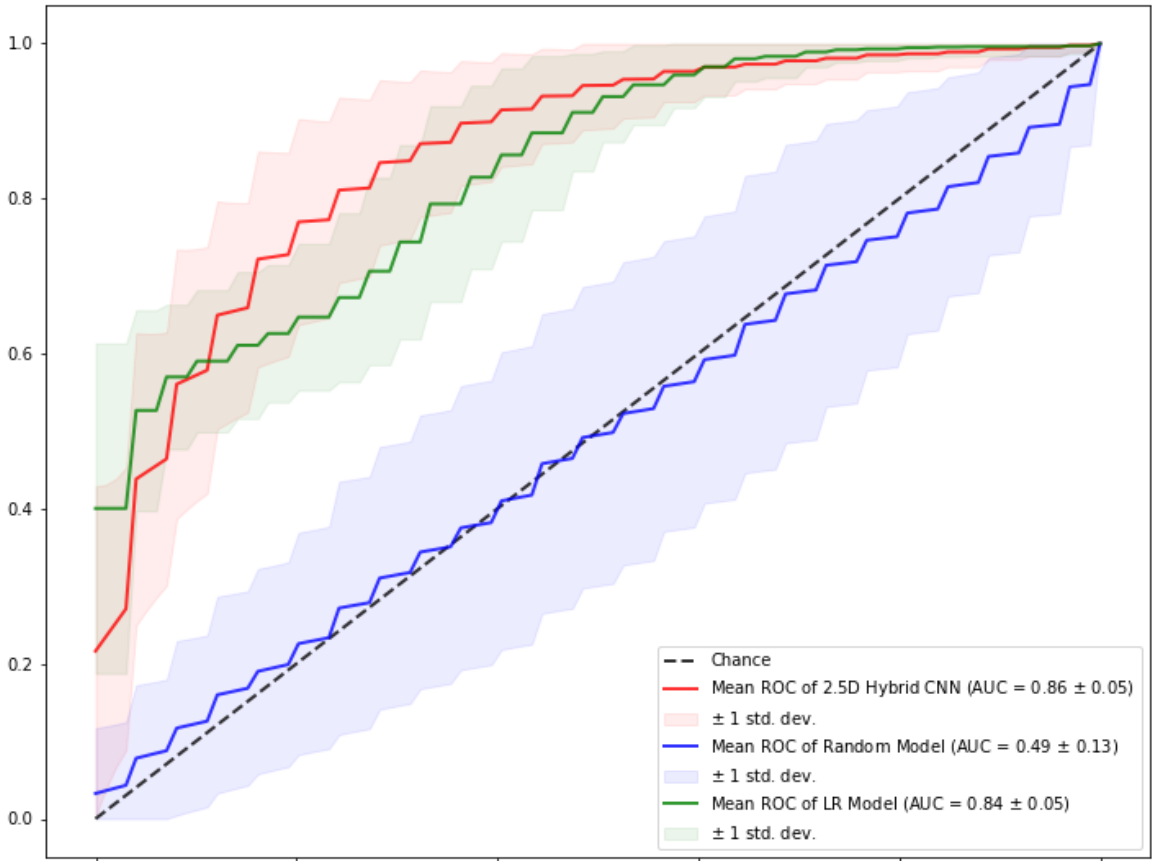

Figure 8: **Receiver Operating Characteristic Curve (ROC) of the three models.** The red curve is for the 2.5D hybrid CNN model with the shaded region being the one standard deviation range for the ROC. The green curve is for the LR model with the shaded region being the one standard deviation range for the ROC. The blue curve is of the Random model with the shaded region being the one standard deviation range for the ROC.

Grad-CAM Map of the 2.5D CNN Model

The last experiment of this study is to analyze the neuroanatomical regions that the CNN model focuses on to complete the classification task. **Figure 8** shows the Grad-CAM feature map of the last convolutional layer of the 2.5D CNN model revealing the pixels from the input images that have the highest weight in the decision of the model overlaid on top of the input image to give reference to brain structures. The red regions show high attention from the model while the blue regions are regions of low attention by the model. The model showed strong information content not only in the MTL, including the hippocampus, but the entire temporal lobe, as well as extra-temporal regions (i.e., parietal lobe, precentral gyrus, and cingulate gyrus), which contained important information for the model to decide on side of epilepsy onset.

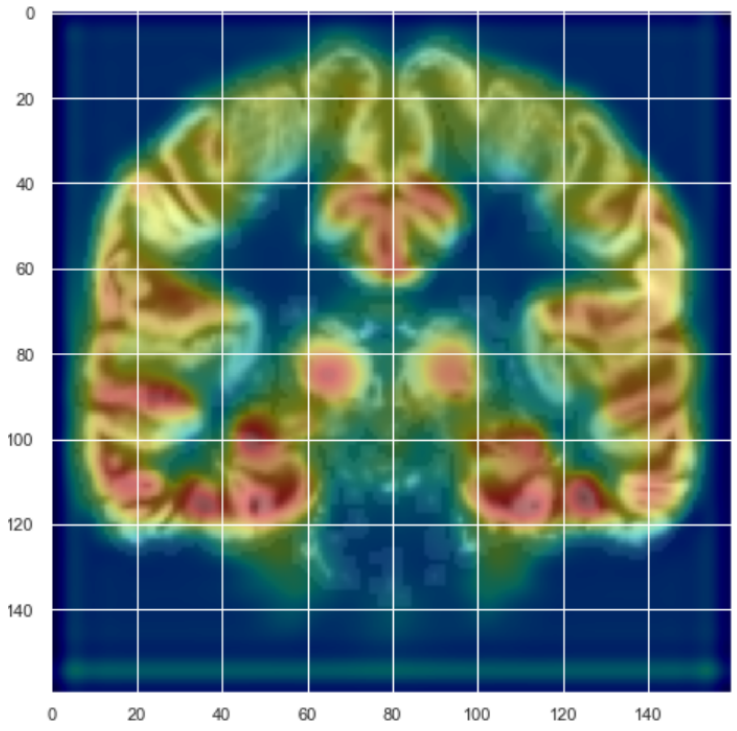

Figure 9: **Grad-CAM feature map of the 2.5D hybrid CNN model.** The red regions indicate high neuronal activation in the CNN model, meaning features in those regions are important for the model in completing the classification task. The Grad-CAM map is overlaid on top of an input image to give reference to regions of activation. The mesial temporal region of the brain is shown to have high information content. Extra-temporal regions, like the parietal lobe, precentral gyrus, and cingulate gyrus are shown to also contain information for the classification task.

DISCUSSION

This study aimed to explore how well artificial neural networks could be used to lateralize the epileptic focus in patients with TLE using whole-brain, structural MRI data. We first developed and tested four novel model architectures and found that the 2.5D hybrid CNN model significantly outperformed the other models. We then tested and confirmed this approach is valid and provided improvements to the pre-existing approach. We then analyzed the characteristics of the model and examined its feature importance maps to conclude that although much of the deterministic information is in the MTL, there are extra-hippocampal and extra-temporal regions that the model uses to make its decision and thus provide additional information regarding neuroanatomical changes in left vs right TLE.

Model Architecture Comparison

We see that the final 2.5D CNN model with hippocampal volume as additional channel inputs performs the best compared to the other tested artificial neural network architectures. This final architecture produces models that have an average accuracy of 77.96%. This compares favorably compared to the simple 2D CNN architecture, which has an average accuracy of 67.34% and is significantly lower than the 2.5D CNN architecture. This means that there is more information in additional inputs of the 2.5D hybrid model compared to the basic 2D CNN model. The model is likely to take advantage of the additional spatial contextual information provided by the additional slices to make a more informed decision on the side of epilepsy onset (Vu, et. al., 2020). Additionally, our 2.5D model included hippocampal information due to previous work demonstrating that hippocampal volume is a good indicator of the side of epilepsy onset in TLE patients (Fraid, et. al. 2012).

We explored whether transfer learning would benefit the model. ResNet50 was chosen for its ability to control for the accuracy being saturated and degrade as model depth increases, we want to implement a deeper network with a small sample size thus we chose ResNet50 and expected the residual blocks will help solve the problem of degradation (He, et. al., 2015). However, the transfer learned ResNet50 architecture has an average accuracy of 54.69% and is also significantly lower than the 2.5D CNN architecture. As we see from the results, it was not able to be trained and has a steady accuracy of 54.69% which is close to the group proportions of 56.95% Lefts to 45.05% Rights. The model was only able to learn to guess Left due to that group taking up a slightly higher proportion of the whole dataset.

The ensemble model that combines CNN feature maps and hippocampal volumes has an average accuracy of 71.13% and is significantly lower than the 2.5D CNN architecture. We decided on two methods to input that information into the model, the first is the simpler hybrid CNN architecture described above. And the second is the more complex ensemble model which first trains a multi-slice CNN model and takes the feature importance map of that model and reduces it with a partial least square model to the 20 most important features and then combines those features with the hippocampal volume to be inputted into a logistic regression classifier. The resulting ensemble model performs worse than the simpler 2.5D hybrid CNN model. Its performance is comparable to the simple logistic regression model using only hippocampal volumes. This means that the ensemble model is using little to no information from important features derived from the CNN and mainly classifying using the hippocampal volumes. This also means the 2.5D hybrid CNN takes advantage of more extensive features derived during training from the images to make the classification. The main takeaway from these experiments is that the more information about the patient and less complexity in the model architecture, the better the

overall model performance. This could likely be due to the limited dataset size limiting our ability to leverage a more complex model.

With an artificial neural network architecture selected, we fine-tuned the model to obtain the best hyperparameters for training the model to allow the model to reach its highest performance. We found that our model performs better with a lower learning rate which controls how fast the model is moving along the gradient. A faster learning rate gives it the ability to get out of a local minimum while a slower learning rate allows the model to converge faster. The higher performance with a slower learning rate can point to the data being more complex and therefore needs a slower learning rate to converge. We also implemented L2 regularization to prevent specific neurons from getting too heavy and dominant. This in theory controls for overfitting of the data by dispersing the weight of the neurons so it cannot rely on a few neurons but on all the neurons to complete the task. And the minimization of the gap between the training set accuracy and the validation set accuracy proves that the L2 regularization is working to stop overfitting (if overfitting, there will be a huge gap between the training set accuracy and validation set accuracy because it is overfitted to the training set). We also changed from the Stochastic Gradient Descent (SGD) optimizer to the Adaptive Moment Estimation (Adam) optimizer. As noted previously, the Adam optimizer is an improved gradient-based optimizer that has a changing learning rate so that each parameter can learn at its own rate. This will speed up the training time by allowing adaptive learning rates which in turn help the model converge faster (Kingma and Ba 2015). With the best architecture and its hyperparameters optimized, we can finally compare it to other reference models to analyze the deep neural network's potential in this classification task.

Model Performance Comparison

The first thing we tested using the current 2.5D hybrid is whether its performance reflected a true improvement from random chance. We tested this by shuffling the training data (training and validation set) labels before using them to train a model using the current architecture. This produces the Random Model, and we can see that it has an average accuracy of 51.65% on the correctly labeled leave out test set giving it a significant difference in performance compared to the 2.5D hybrid model with correctly labeled training data. Since the 2.5D hybrid CNN model is trained and tested using the same data split as the Random Model on each run, we can do a pair-wise comparison of the two models, and we see that the pairwise difference in accuracies is always higher for the 2.5D hybrid CNN and the difference ranges for 0% to about 60%. The near 0.50 sensitivity (0.54) and specificity (0.43) show that any classification this model makes has about a 50% chance of being correct (slightly better chances for a left prediction but this could likely be due to the class imbalance). This shows that our 2.5D hybrid model is learning from the data and not just guessing using class imbalance. This gives us more confidence in analyzing the feature importance maps because we know that the highlighted features do contain useful information for the model to make a better decision on the side of epilepsy onset than random guessing.

Next, we tested if our 2.5D CNN model has added additional information to what is currently used clinically. This was accomplished with a Logistic Regression (LR) model trained on only hippocampal volume, an informative biomarker often used currently in epilepsy lateralization (Fraid, et. al., 2012). Now we want to test if using only hippocampal volume will give us similar results to a deep learning approach and thereby eliminate the need for a complex deep learning model. The LR model will try to draw a hyperplane that separates the left and right TLE patients based solely on their left and right hippocampal volume data. The LR model has an

average accuracy of 71.71% and is significantly different from the 2.5D hybrid CNN model. Since the 2.5D hybrid CNN model is trained and tested using the same data split as the LR model on each run, we can do a pair-wise comparison of the two models and we see that the pairwise difference in accuracies is mostly higher for the 2.5D hybrid CNN but for a few runs, the LR model outperformed the 2.5D CNN model. Looking at their pairwise differences, we can see that there is less difference in performance compared to the Random model with the LR model being on average 6.25% less accurate than the 2.5D hybrid CNN. Looking at the model performance measurements (i.e., TPR, FPR, etc.) we can conclude that the LR model performs more poorly than the CNN model where it has more incorrect predictions for both left and right cases compared to the CNN model. However, it is interesting to note that the average AUC is higher for the LR model compared to the CNN model. This means that the LR on average is better at obtaining a high true positive rate without sacrificing the FPR. Overall, the more complex 2.5D hybrid CNN model identifies a nonlinear solution to the problem of classifying the side of epilepsy onset in TLE patients. And a whole-brain approach that not only focuses on the hippocampus and the Mesial Temporal Lobe will provide additional benefits to a model classifying between left and right TLE.

CNN Properties

We have now shown that the 2.5D hybrid CNN model is an appropriate tool to classify left versus right TLE patients, we will now explore the properties of the model and what biological knowledge we can gain from it. Looking at the accuracy of the model, we can see a clear difference in performance when we compare its accuracy in correctly identifying left TLE subjects (82.93%) and correctly identifying right TLE subjects (71.88%), we see that it is better at predicting left TLE

patients than right TLE patients. This could be due to the slight class imbalance, but the slight imbalance should not cause such a big discrepancy in accuracy between the two classes. It is interesting especially when a previous study shows that atrophy is distributed evenly between the two sides of onset groups (Liu, et. al. 2015) so we should not expect one class to be easier than the other class as they should both have equal amounts of atrophy that indicate that they belong to their class. A confounding variable here could be the misclassifications of some of the scans where they could also have bilateral atrophy. It is well known that epilepsy can cause widespread cortical changes, even when the seizure focus is restricted to one temporal lobe.

We also looked at the Grad-Cam maps derived from the CNN model to better understand which regions show the features that the CNN model focuses on to make its predictions. From the Grad-CAM maps, we can observe that a lot of attention was focused on the MTL. However, regions throughout the temporal lobe appear important, as well as regions within the parietal lobe, the precentral gyrus, and the cingulate gyrus. This shows that atrophy is not just localized in the mesial temporal lobe but throughout the brain. Another interesting finding from the Grad-CAM map is that when shown Left TLE patients, the model is not just focusing on the atrophy in the left hemisphere but on atrophy within both hemispheres. In conclusion, deep learning models for the task of classifying the side of epilepsy onset in TLE show promise and support prior studies revealing the presence of extra-temporal pathology that is subtle, yet consistent in left and right TLE.

Limitations and Future Directions

This study is limited by the small sample size, potential labeling issues, and image quality issues between imaging sites. First, the sample size of brain scans is relatively small, especially

for developing a deep learning model. We have 364 patients, including 200 left TLE and 164 right TLE patients. State-of-the-art computer vision algorithms are typically trained on tens of thousands of scans that are carefully curated. It is hard to develop a deep neural network with more training parameters before the model overfits with so little training data. Small datasets also are prone to outliers having big influences on the overall model performance. The second limitation is that there could have been some misclassifications in the image labels. Our dataset comes from seven different epilepsy centers, and we use the class label given to us by each site based on expert opinions. This could lead to different standards of labeling and possible misclassification and contamination of the model's performance. The third constraint is that image intensities and quality varied across imaging sites. Differences in imaging protocols can lead to differences in scan quality and/or comparability that can present challenges for CNN models.

One experiment to implement in the future is to use the "sliding window" method where the input images are offset by a few slices and inputted as a new sample to improve the limited dataset and therefore increase the accuracy of the model. This method is successfully implemented by the group that developed the U-Net model architecture that became a successful segmentation tool using deep learning (Ronneberger et al., 2015). To deal with possible misclassification of patients, we could have relied on post-surgical outcome data (i.e., seizure-free or not seizure-free) to verify that the labels (i.e., side of seizure onset) were correct. If the patient becomes seizure-free after their labeled hemisphere has been operated on, we can then be more confident that the labeling is correct. Another experiment to test is a generative adversarial network (GAN) that augments the input to another image space (Mao et. al., 2020). This could help with normalizing the images between sites that have completely different scanning protocols and prevent the model from detecting imaging site differences and focusing on site of onset differences.

Conclusion

This study not only shows that artificial intelligence is capable of performing the specific task of detecting epileptic focus but that it can extract and take advantage of information from medical images that humans and traditional (i.e., linear) approaches cannot access. And additionally, it shows that in TLE, not only are the well-studied and expected structural features of the MTL important, but extra-temporal regions are informative as well and thus warrant further research into the significance of these regions. With a bigger sample size and more computational power and time, artificial intelligence can be a powerful tool in biomedical research and diagnosis.

REFERENCES

Abbasi, B., & Goldenholz, D. M. (2019). Machine learning applications in epilepsy. *Epilepsia*, *60*(10), 2037–2047. https://doi.org/10.1111/epi.16333

Bell, G. S., Neligan, A., & Sander, J. W. (2014). An unknown quantity-the worldwide prevalence of epilepsy. *Epilepsia*, *55*(7), 958–962. https://doi.org/10.1111/epi.12605

Cendes, F., & McDonald, C. R. (2022). Artificial intelligence applications in the imaging of epilepsy and its comorbidities: Present and future. *Epilepsy Currents*, 153575972110686. https://doi.org/10.1177/15357597211068600

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*(20), 1920–1930. https://doi.org/10.1161/circulationaha.115.001593

Engel, J. (1996). Introduction to temporal lobe epilepsy. *Epilepsy Research*, *26*(1), 141–150. https://doi.org/10.1016/s0920-1211(96)00043-5

Farid, N., Girard, H. M., Kemmotsu, N., Smith, M. E., Magda, S. W., Lim, W. Y., Lee, R. R., & McDonald, C. R. (2012). Temporal lobe epilepsy: Quantitative mr volumetry in detection of hippocampal atrophy. *Radiology*, *264*(2), 542–550. https://doi.org/10.1148/radiol.12112638

French, J. A. (2007). Refractory epilepsy: Clinical overview. *Epilepsia*, *48*(s1), 3–7. https://doi.org/10.1111/j.1528-1167.2007.00992.x

Gaser, C., & Dahnke, R. (2016). CAT-A Computational Anatomy Toolbox for the Analysis of Structural MRI Data.

Gleichgerrcht, E., Munsell, B., Keller, S. S., Drane, D. L., Jensen, J. H., Spampinato, M. V., Pedersen, N. P., Weber, B., Kuzniecky, R., McDonald, C., & Bonilha, L. (2021). Radiological identification of temporal lobe epilepsy using artificial intelligence: A feasibility study. *Brain Communications*, *4*(2). https://doi.org/10.1093/braincomms/fcab284

Hamet, P., & Tremblay, J. (2017). Artificial Intelligence in medicine. *Metabolism*, *69*. https://doi.org/10.1016/j.metabol.2017.01.011

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.90

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015). https://doi.org/10.1038/nature14539

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on Deep Learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu, M., Bernhardt, B. C., Bernasconi, A., & Bernasconi, N. (2015). Gray matter structural compromise is equally distributed in left and Right Temporal Lobe epilepsy. *Human Brain Mapping*, *37*(2), 515–524. https://doi.org/10.1002/hbm.23046

Mao, X., & Li, Q. (2020). Generative Adversarial Networks (gans). *Generative Adversarial Networks for Image Generation*, 1–7. https://doi.org/10.1007/978-981-33-6048-8_1

Ramesh, A. N., Kambhampati, C., Monson, J. R., & Drew, P. J. (2004). Artificial intelligence in medicine. *Annals of the Royal College of Surgeons of England*, *86*(5), 334–338. https://doi.org/10.1308/147870804290

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

So, E. L., & Ryvlin, P. (2015). MRI-negative epilepsy: Evaluation and surgical management. *Cambridge University Press.* https://doi.org/10.1017/CBO9781139525312

Sone, D., & Beheshti, I. (2021). Clinical application of machine learning models for Brain Imaging in epilepsy: A Review. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.684825

Stelmokas, J., Yassay, L., Giordani, B., Dodge, H. H., Dinov, I. D., Bhaumik, A., Sathian, K., & Hampstead, B. M. (2017). Translational MRI volumetry with NeuroQuant: Effects of version and normative data on relationships with memory performance in healthy older adults and patients with mild cognitive impairment. *Journal of Alzheimer's Disease*, *60*(4), 1499–1510. https://doi.org/10.3233/jad-170306

O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. *ArXiv*. https://doi.org/arXiv:1511.08458

Vu, M. H., Grimbergen, G., Nyholm, T., & Löfstedt, T. (2020). Evaluation of multislice inputs to convolutional neural networks for medical image segmentation. *Medical Physics*, *47*(12), 6216–6231. https://doi.org/10.1002/mp.14391

Wythoff, B. J. (1993). Backpropagation Neural Networks. *Chemometrics and Intelligent Laboratory Systems*, *18*(2), 115–155. https://doi.org/10.1016/0169-7439(93)80052-j