Enhancing MRI Analysis in Temporal Lobe Epilepsy: Sequential Classification and Segmentation of Hippocampal Structures

Vihan Bhattacharjee
Bellarmine College Prep
San Jose, USA
e-mail: v.bhattacharjee27@bcp.org

Kamalasree Sudhakar
AI Club
Mountain View, USA
e-mail: kamalasree.s@aiclub.ai

Sindhu Ghanta
AI Club
Mountain View, USA
e-mail: sindhu.ghanta@aiclub.ai

***Abstract*—Temporal Lobe Epilepsy is a chronic neurological disorder affecting 50 million people worldwide, with diagnosis often complicated by overlapping symptoms with other conditions. Traditional diagnostic methods, like electroencephalography and manual Magnetic Resonance Imaging analysis, face challenges of subjectivity and inaccuracies. This study explores machine learning approaches to address these limitations, utilizing MobileNet and U-Net models to improve diagnostic accuracy and efficiency in Temporal Lobe Epilepsy. MobileNet, a lightweight classification model, successfully distinguishes between normal and abnormal hippocampi, achieving a classification accuracy of 94%. This is further enhanced by the U-Net segmentation network, which achieves an Intersection over Union score of 0.902 in identifying hippocampal abnormalities associated with Temporal Lobe Epilepsy. The integration of these models significantly improves diagnostic precision and efficiency, offering substantial support to radiologists in clinical settings. The promising results underscore the potential of advanced machine learning techniques in enhancing diagnostic processes. Future research will aim to increase model robustness by diversifying training datasets and exploring alternative algorithmic approaches, thereby improving the models' applicability and reliability across various clinical environments.**

Keywords-Temporal Lobe Epilepsy; Machine Learning; MobileNet; U-Net; Hippocampal Abnormalities.

I. Introduction

Temporal Lobe Epilepsy (TLE) is a chronic neurological disease caused by the abnormal discharge of neurons in the brain [1]. TLE affects about 50 million people worldwide [2]. It is the fourth most common neurological disorder, and it impacts people of all ages, including adults, children, and newborns [3].

Seizures caused by TLE can cause brain injury, physical harm, and even a shorter lifespan. Specifically, TLE can lead to significant memory loss due to damage in the hippocampus, as well as impacting attention, learning, and language processing, all of which are vital for daily tasks [4]. Additionally, recent research has indicated that individuals with TLE are at higher risk of depression, anxiety, and mood instability, stemming from dysfunction in the limbic system [5]. The most alarming consequence of TLE is the increased risk of mortality from sudden unexpected death in epilepsy (SUDEP), vascular diseases, suicide, pneumonia, and underlying causes like brain tumors: all of these factors cause individuals with epilepsy to have a mortality rate as much as 11.6% higher than expected [6]. That is why it is essential to ensure early and accurate diagnosis of TLE to increase the chance for patients to receive proactive treatment and restore their quality of life. Currently, epilepsy diagnosis involves multiple steps that combine clinical observation with testing to identify both the seizure type and its underlying cause. Electroencephalography (EEG) serves as the primary diagnostic tool for epilepsy, capturing the brain's electrical signals to detect the irregular patterns characteristic of the condition. However, interpretations of EEG scans are subjective, and radiologists may misdiagnose TLE as other conditions, such as a primary psychiatric disorder (i.e., schizophrenia) [7], which has many similar symptoms as TLE. Additionally, diagnosing epilepsy presents significant challenges due to its non-specific symptoms that sometimes look similar to those of other medical conditions. The diagnostic process primarily depends on reviewing the patient's medical history and documenting seizure episodes from the witnesses, which can be incomplete or inaccurate. That is why previous studies have demonstrated that prior seizures experienced by patients are often not diagnosed as epilepsy, impacting as many as 38% of epilepsy patients [8]. To overcome these limitations in diagnosing TLE, researchers have explored alternative diagnostic methods, particularly by examining the state of the hippocampus [9], a structure located deep within the temporal lobe. The hippocampus plays a crucial role in memory processing, emotion regulation, spatial navigation, and learning [10]. In TLE, the hippocampus often shows damage or structural changes, such as hippocampal sclerosis, in which nerve cells are lost and scarring develops. Hippocampal sclerosis is present in between 30.5% and 45% of all epilepsy syndromes and in 56% of cases of TLE [11]. Therefore, it is valuable for radiologists to detect hippocampal abnormalities in potential TLE patients.

Magnetic Resonance Imaging (MRI) with efficient spatial navigation capabilities is key for evaluating the state of the hippocampus [12]. The process of identifying and marking the hippocampus in MRI scans relies heavily on manual techniques, which require a significant time investment and are subject to inter-observer differences. For example, in temporal lobe epilepsy, hippocampal sclerosis can sometimes appear as very mild shrinkage or subtle signal changes, which may be overlooked even by experienced radiologists [13]. Additionally, processing MRI images presents challenges due to the substantial computational resources needed and the extensive time required for completion [1].

Automation can assist radiologists in enhancing their detection process. Machine learning and deep learning algorithms can extract complex features from MRI images that are often

 ISBN: 978-1-68558-362-0

difficult for the human eye to distinguish, enabling precise detection of hippocampal sclerosis and other hippocampal abnormalities. Models trained on MRI data have immense potential to improve the accuracy of detecting hippocampal abnormalities; for instance, deep learning models can perform automated segmentation of the hippocampus, measure its volume precisely, and detect microstructural changes [14]. However, there are many considerations researchers must make when constructing these models, including data quality, data generalizability, and preprocessing methods. This study explores the implementation of the U-Net segmentation algorithm to localize hippocampal abnormalities, as well as the MobileNet classification model to serve as a prior step in distinguishing healthy from non-healthy hippocampal tissues, providing a comprehensive approach to automated TLE diagnosis.

The paper is organized and begins with an introduction in Section I and exploration of related work in Section II. An explanation of the various stages of methodology was in Section III, followed by results in Section IV. Section V contains the discussion. Finally, a conclusion about the implications of the study is present in Section VI.

II. Related work

Several researchers have explored the potential of utilizing computer vision deep learning algorithms to detect minute or microstructural abnormalities in hippocampal MRIs to assist in TLE diagnosis. Segmentation models have shown especially promising results, as they are able to precisely delineate the hippocampal region with minimal human intervention. Experiments with segmentation are often conducted by incorporating transfer learning to pre-trained Convolutional Neural Networks (CNN) to minimize computational resources while maximizing accuracy. However, some common limitations in these studies are insufficient hyperparameter optimization, high computational resource requirements for training complex models, and class imbalance due to limited availability of hippocampal-specific slices.

For example, Chaouch et al. [15] studied hippocampal atrophy in mesial temporal lobe epilepsy via CNN deep learning segmentation. While these segmentation models show success in isolating the particular volume and location of hippocampus atrophy, as seen through a Dice score of 0.86 and sensitivity of 93%, the datasets used to train these models have a vast imbalance of relevant shots of the hippocampus: out of 3050 slices used in the model, only 459 actually contain hippocampal regions, making approximately 80% of the data obsolete and rendering the model liable to class imbalance issues and overfitting.

Additionally, Chang et al. [16] utilized a custom CNN architecture with 3 main layers to diagnose TLE and Alzheimer's disease from MRI brain scans by analyzing both the hippocampal region, as well as the broader structure of the brain. Their model achieves an accuracy of 90.45% in correctly identifying the presence of TLE, and an F1 score of 0.85. However, a lack of exhaustive hyperparameter tuning, as acknowledged by the researchers themselves, likely constrained the model's accuracy and stability, as the most optimal epochs and learning rates have the potential to significantly improve model performance. Similarly, Jiang et al. [17] created a novel CNN architecture, HS-Net, to accurately detect the presence of hippocampal sclerosis characteristic of TLE, achieving results like an Area Under the Curve (AUC) of 0.894 and an accuracy of 82.88 . Nonetheless, the model fails to precisely delineate the location or structure of the impacted hippocampus, and thus only serves to classify the presence of hippocampal sclerosis without actually automating the diagnostic process.

Most notably, Widodo et al. [12] utilizes the U-Net model to segment the hippocampus in MRI images. They utilized transfer learning to train a 3D U-Net model with data from a range of patients with various cognitive states (normal, mild cognitive impairment, and Alzheimer's). They achieved an Intersection Over Union (IoU) of 0.75 and a sensitivity of 0.80. While this work demonstrates the effectiveness of U-Net architectures for hippocampal segmentation, it does not address class imbalance within the dataset or test how hyperparameter selection influences model performance. Similar to the approach in Widodo's study, this study utilizes the U-Net architecture for segmenting the hippocampus in MRI images. Furthermore, to overcome the flaws present in previous studies, a wider range of hyperparameters is used to test the architecture's efficacy and the dataset is balanced in preprocessing to ensure the model is trained upon an equal number of healthy and unhealthy scans. Most significantly, this study also conducts experiments with the CNN MobileNet to conduct classification on the MRI images and classify them as "healthy" vs "unhealthy" before segmentation; this creates a more real-world workflow that saves computational resources and only utilizes segmentation in necessary cases. Despite conducting experiments with a limited number of models, the proposed framework has the potential to localize hippocampal abnormalities more effectively and efficiently than traditional methods and other experimental approaches.

III. Methods

Dataset Analysis: This study utilized publicly available data from Jafari-Khouzani et al., a 2.58 GB dataset containing 50 brain T1-weighted MRI volumes with hippocampus labels [18]. The decision to use a dataset of T1-weighted MRI images rather than alternatives like T2 was driven by several key factors. T1-weighted MRI images provide high anatomical detail and clearly distinguish between gray and white matter, with white matter appearing lighter and gray matter darker [19]. This makes them ideal for detecting structural abnormalities within the hippocampus. Additionally, T1 scans produce sharper, higher-resolution images and are employed in procedures that require highly precise data [20].

Specifically, 40 of the images belong to patients who have temporal lobe epilepsy. Several of the patients had atrophic hippocampi, allowing for the model to train with real-world data with more challenging segmentation, while the remaining 10 images were from subjects without epilepsy and displayed normal hippocampi. The dataset also contained

labels that detailed the precise anatomical boundaries of the hippocampus within each volumetric scan. Due to the dataset's class imbalance, there may have been a risk of overfitting the model to become biased towards the unhealthy class. This data imbalance was mitigated by random undersampling, where the number of scans within the healthy class were made the same as the scans in the unhealthy class, equating both data subsets.

Data Preprocessing: The preprocessing phase of this study involved several systematic steps to prepare the dataset for model training. The first step began with doing a detailed dataset analysis to understand the distribution between the healthy and non-healthy image cases, and separating the images per class distribution. Figure 1 shows an unprocessed MRI scan. Image adjustments were done to load the MRI image and its corresponding label file to the model.

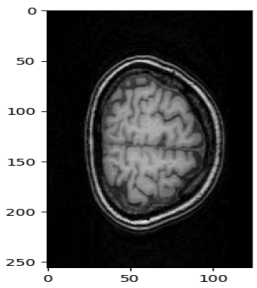


Figure 1. Sample 3D MRI scan from the dataset

Segmentation: To preprocess the images for segmentation, the first step was to convert the 3D MRI volumes into 2D image slices to simplify processing and model training. To convert the 3D volumetric data to 2D, individual image slices were extracted, which were cross-sectional images of the 3D images. The 2D slices were extracted in three orientations: axial (horizontal) slices, coronal (frontal) slices, and sagittal (lateral) slices. The MRI scans were divided into slices to increase the size of the dataset, facilitating more robust training; lower computational resources necessary for the models; and to leverage pre-trained segmentation models, which are traditionally trained on 2D slices.

After the slices were isolated, all of them were normalized to convert the wide-ranging MRI voxels into pixels with a standard range of intensity. Min-Max normalization was used to ensure that all the pixels contained a standard intensity from 0 to 255 [21]. The segmentation labels were also converted to binary masks, such as 0 for the background and 255 for the hippocampus region. The normalized images and the corresponding masks were saved in two separate folders as .png files. Figure 2 illustrates the methodology used for conducting the experiments of this research.

Classification: After resizing the previously gathered MRI slices to 224 pixels, each image was converted to a 3-channel RGB format to adhere to the requirements of MobileNet. Then, pixel values were rescaled from [0,255] to [-1,1], to be assigned to the labels of "unhealthy" and "healthy." Finally, class labels were converted to vectors (e.g., [0,1]) to be assigned to either health category. This ensured that every slice could be easily classified as either healthy or unhealthy during subsequent training.

Data Splitting: Then the preprocessed dataset was split into sets of training, validation, and test by using a random sampling approach. Approximately 10% of the dataset was allocated to validation, and 10% to testing. The remaining 80% was used for training. This ensured the proper evaluation of model performance. Figure 3 illustrates the completed MRI scan and the corresponding mask after the preprocessing stage.

Model Training and Validation: The study was structured around three sequential experiments designed to overcome the limitations inherent in using solely segmentation or classification models. Using only classification would lack the necessary anatomical localization for clinical treatment, while applying segmentation to every MRI slice is computationally inefficient, particularly for large datasets. This integrated approach minimizes computational resources and energy consumption while retaining critical diagnostic detail. Transfer learning was utilized across all models, leveraging pre-trained weights to accelerate convergence and improve performance given the dataset size.

The first experiment was Hippocampus segmentation using U-Net. This experiment focused on detecting the precise anatomical delineation of the hippocampus, a key step for volumetric analysis in TLE diagnosis. The U-Net CNN was selected due to its proven efficacy in biomedical image segmentation. U-Net models provide optimal results even with small tumor lesions or relatively unclear organ boundaries, making it especially valuable for medical image segmentation, which often faces these challenges [22]. U-Net is the optimal choice for researchers due to its ability to obtain accurate results, even when trained on a limited dataset [23], as the model automatically incorporate alterations, such as flips and rotations to expand the size of datasets, and incorporates end-to-end training, which enables researchers to train U-Net models directly on broad medical images, rather than handcrafted features, reducing the need for excessive data. Additionally, Lu et al. [24] finds that U-Net's divide-and-conquer encoder strategy, which divides an input image into five different feature maps, facilitates efficient segmentation and localization; this feature is crucial for clinical environments, where rapid decisions need to be made with accurate data. While other models may exhibit strong performances, Turk et al. [25] shows how U-Net achieves a comparable accuracy to other experimental models; for instance, in Lu et.al's study, the traditional U-Net model achieved a 97.49% accuracy, compared to 98.00% of a modified model. Furthermore, the U-Net model required the least training time and computational cost, compared to other deep learning models like V-Net and SegNet, making it more practical in healthcare settings.

During the training process, hyperparameters were meticulously tuned; learning rates ranged from 0.000001 to 0.05, while epochs were assessed from 10 to 50. The results of model training were analyzed with loss and the IoU metric, defined as Area of Overlap/Area of Union.

The second experiment was classification using MobileNet,

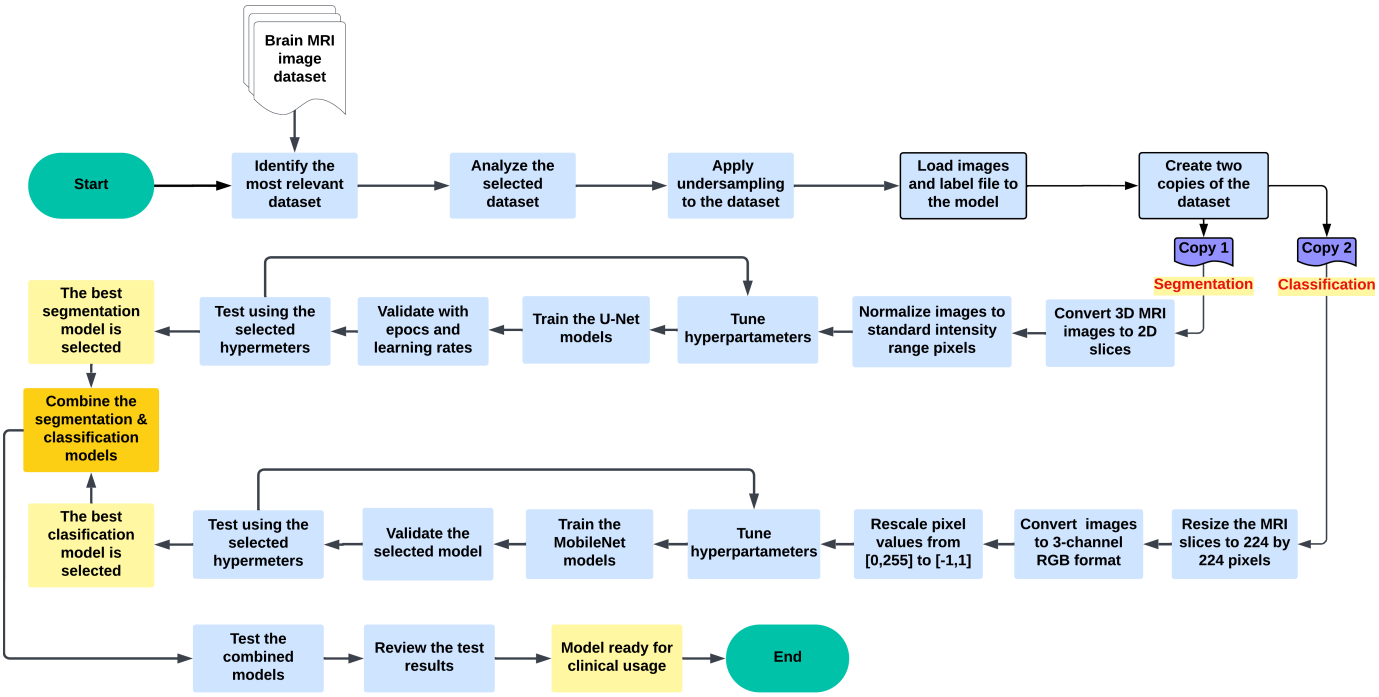
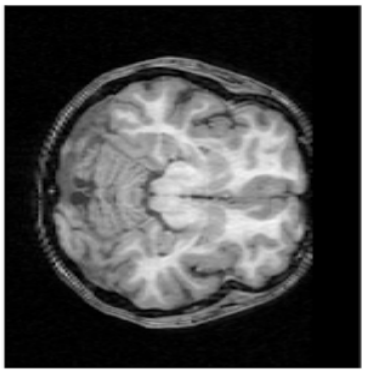


Figure 2. Flowchart of image Data Processing and Machine Learning stages



(a) Preprocessed 2D MRI scan

(b) Preprocessed mask

Figure 3. Preprocessed data for segmentation

a Convolutional Neural Network commonly used in image classification. The objective of this experiment was to create an efficient binary screening tool capable of differentiating between healthy and unhealthy hippocampus slices. The MobileNet architecture was chosen for its lightweight design and computational efficiency, making it ideal for the high-volume classification required in the initial screening phase.

Despite requiring fewer parameters and having significantly fewer computational needs than other similar CNNs, Ochoa-Ornelas et al.'s study [26] demonstrated how the model could achieve an accuracy of 98.77% on a dataset containing medical imagery. Additionally, MobileNet produces similar accuracies to other heavier models like Xception on medical imagery classification, as evidenced by Prakash et al. [27].

The model was trained to classify MRI slices into "healthy" or "unhealthy" categories. Similar to the segmentation experiment, hyperparameters were consistently monitored with epochs and learning rates in the same ranges. Final results were quantified by analyzing accuracy metrics and detailed confusion matrices to evaluate the rate of misclassification.

The final experiment simulated the real-world application, integrating the optimal classification and segmentation models sequentially. The MobileNet model (Experiment 2) first processed the entire testing set to identify slices likely exhibiting TLE pathology. Subsequently, the high-precision U-Net model (Experiment 1) was applied only to the slices flagged as "unhealthy" by the MobileNet model. This combined method validated the system's ability to minimize computational resources while providing accurate localization of the hippocampus, establishing a methodology with enhanced efficiency and accuracy. The final performance of this integrated

 ISBN: 978-1-68558-362-0

system was tested using the testing dataset. Final analysis of accuracy via sample images revealed high confidence scores as well as accurate segmentation.

Results Analysis: After validation and testing were completed for both the U-net and MobileNet models, the results were analyzed using IoU metrics, loss, and confusion matrices. In segmentation, the IoU value is a valuable metric that evaluates how much the predicted mask overlaps with the actual location of the hippocampus. For classification, the IoU compares the overlap between predicted and true positive classes. A higher IoU value indicates a more accurate model performance. Additionally, evaluation of loss for segmentation or confusion matrices for classification was conducted for each model to gain a greater understanding of specific types of errors or inaccuracies.

IV. Results

This section outlines findings from the segmentation and classification experiments. The results include analyses of IoU metrics, loss metrics, confusion matrices, and graphical and tabular comparisons of model performance.

In the segmentation experiment, the model was trained with the U-Net algorithm. There were 35 experiments using a combination of seven learning rates (0.000001-0.05) and epochs (10-50), as shown in Table I and II. IoU quantifies the degree of overlap between the pixels within the model's predicted mask and the ground truth images; a higher IoU indicates superior segmentation accuracy, reflecting the model's robustness across a variety of scans. IoU was used to measure the performance of the model because it provides a rigorous and interpretable metric that captures both false positives and false negatives, ensuring that the evaluation reflects the true spatial accuracy of the segmentation. Additionally, IoU also captures both over-segmentation and under-segmentation errors, providing a more comprehensive assessment than simple pixel accuracy.

TABLE I. Hippocampus Segmentation U-Net IoU Results

Epochs / Learning Rate	0.000001	0.00001	0.0001	0.001	0.005	0.01	0.05
10	0.0064	0.0193	0.5159	0.8028	0.7991	0.8242	0.0175
20	0.0074	0.024	0.7015	0.865	0.2977	0.8486	0.8401
30	0.6543	0.6802	0.73	0.8233	0.8709	0.8794	0.8588
40	0.8755	0.879	0.8886	**0.8961**	0.8944	0.8944	0.8776
50	0.0121	0.373	0.876	0.874	0.8707	0.8763	0.8642

TABLE II. Hippocampus Segmentation U-Net Loss Results

Epochs / Learning Rate	0.000001	0.00001	0.0001	0.001	0.005	0.01	0.05
10	0.3911	0.1297	0.0055	0.0022	0.0025	0.0023	0.0972
20	0.289	0.123	0.0031	0.002	0.0312	0.0024	0.0024
30	0.0039	0.0034	0.0028	0.0025	0.0031	0.003	0.0031
40	0.0023	0.0023	0.0027	**0.0064**	0.0072	0.0057	0.0036
50	0.195	0.009	0.0023	0.0029	0.0027	0.0031	0.0024

In the U-Net experiment, the highest IoU during validation of the model was achieved at epoch 40 with a learning rate of 0.01, yielding an IoU of 0.8961. When solely this combination of learning rate and epoch was used upon an unseen test dataset, the model yielded an IoU of 0.902, demonstrating the model's consistent yet accurate performance across a wide range of data. However, these same hyperparameters also yielded a relatively high loss of 0.0064, as shown in Table II, indicating that while the model correctly captured the shape of the target regions, it did so with uncertain probability outputs that were penalized by the loss function. This loss analysis is crucial to gauge the accuracy of the U-Net model, because it reveals how confidently the model delineates the location of the hippocampus. The loss function directly measures the model's probabilistic errors, demonstrating instances when confidence is low despite the model making a correct prediction.

(a) Sample image from test dataset (b) Ground truth hippocampus mask (c) U-Net predicted Mask

Figure 4. Successful Segmentation of Hippocampal Abnormalities

The image shown in Figure 4 highlights an example of how the segmentation done for the sample hippocampus image is accurate. The model successfully identifies both hippocampal regions in the correct locations. The accurate results suggest that the model can determine the anatomical structure due to its training.

In the classification experiment, the model was trained with the MobileNet algorithm. Similarly, 35 experiments were also conducted with the same range of learning rates and epochs, as shown in Table III. For these sets of experiments, accuracy was evaluated, along with results from confusion matrices, as shown in Figure 4. While accuracy measures what percent of the model's prediction accurately matches each image's corresponding class label, confusion matrices give further insight into the false positives and negatives the model may exhibit. These trends demonstrate the model's weaknesses and potential class imbalances, ultimately helping to identify areas where architectural adjustments or additional data are needed. Validation revealed that at 20 epochs and a learning rate of 0.01, the model reached its highest accuracy of 0.9483, reflecting that in over 94% of instances, the model was able to accurately identify the class of the image as "healthy" or "unhealthy". When these specific hyperparameters were tested on an unseen test dataset, the model yielded an accuracy of 0.94. Another metric that was analyzed was confusion matrices, which demonstrated that the model had significantly more true positives than true negatives or false positives. The model detected true positives at a rate of 87.36% and true negatives at 96.55%, as shown in Figure 5, indicating the model's ability to distinguish accurately between different images of the hippocampi.

TABLE III. CLASSIFICATION MOBILENET ACCURACY RESULTS

Epochs / Learning Rate	0.000001	0.00001	0.0001	0.001	0.005	0.01	0.05
10	0.5345	0.7874	0.8908	0.9195	0.9138	0.9195	0.8851
20	0.7701	0.8506	0.9138	0.9368	0.9425	**0.9483**	0.9195
30	0.7701	0.8678	0.9195	0.9368	0.9368	0.9425	0.5
40	0.7089	0.8678	0.931	0.9368	0.9425	0.931	0.9195
50	0.7759	0.8793	0.9368	0.9253	0.9425	0.9368	0.9253

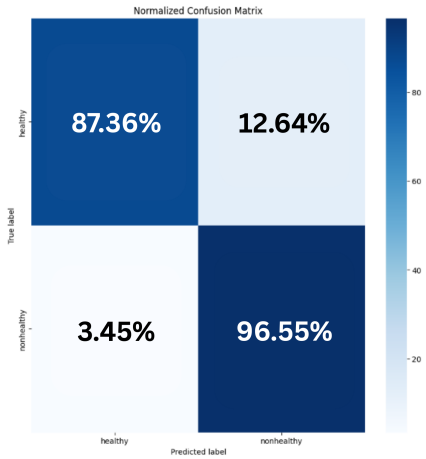


Figure 5. Confusion Matrix for MobileNet Classification on Test Dataset

V. DISCUSSION | EVALUTION

Segmentation Performance: Results from hyperparameter tuning demonstrate that the U-Net segmentation model accurately localizes the precise location of the hippocampus. Table I shows that learning rates between 0.0001 and 0.01 consistently yielded high IoU scores. At 10 epochs, only a few learning rates (0.001-0.01) achieved acceptable segmentation quality, with lower rates failing to converge, and higher rates showing instability. The results at 20 epochs showed some improvement, though the 0.005 learning rate experienced a dramatic collapse with IoU of 0.2977, demonstrating the model's instability in this range. The transition started at 30 epochs where learning rates 0.000001 and 0.00001 started showing improvement in IoU, with values above 0.6543. These results indicate the model requires comprehensive training, even for low learning rates, to gather sufficient information for accurate predictions.

A more accurate performance occurred at 40 epochs, with all IoU being greater than 0.8755. Similar to the other epochs, the learning rates at the lower end of the range contain the lowest IoU scores, demonstrating the model's lack of convergence at those rates. The best configuration, with the highest IoU, occurs when the optimal learning rate is 0.001 and the epoch is 40, leading to an IoU of 0.8961. However, at 50 epochs, the model's failure to converge at lower learning rates is apparent, with the lowest IoU being 0.0121. Results from higher learning rates do not show any improvement from 40 epochs, suggesting that higher levels of epochs do not increase performance and may introduce mild overfitting or training instability. Figure 6 represents the hyperparameter tuning process in a graphical format, plotting the learning rate and epoch versus IoU. This graph also corroborates the previous IoU analysis.

Additionally, loss results from this experiment also support the IoU findings, reflecting strong underlying prediction quality despite the differences in probability calibration. The loss metric captures the model's confidence while making predictions. At 10 epochs, there is generally very high loss, especially for very low learning rates (0.000001) and very high learning rates (0.05), demonstrating that the model is underfitting and is unadjusted to the dataset. The results at 20 epochs show a decrease in loss, demonstrating the model's growing accuracy. At 30 and 40 epochs, the model exhibits a stability in loss, with most loss values from 0.02 to 0.03. This suggests 30 and 40 epochs is sufficient for the model to achieve optimal convergence without overfitting. At 50 epochs, improvements in loss are marginal and are not offset by the additional computational resources necessary. At the hyperparameters with the high IoU output, the model yielded a loss of 0.0064, a relatively high loss, especially when compared to the model's performance with the same epoch and different learning rates. A relatively high loss occurring despite a high IoU may indicate the model's uncertainty while making pixel-level predictions. This lack of confidence may occur if the dataset contains ambiguous or noisy scans, making certain predictions difficult. Although augmenting the dataset would increase the model's confidence in prediction and decrease loss across the board, ambiguous data reflects real-world conditions present in clinical facilities, which may contain various scans with different qualities. However, clinicial validation by radiologists is necessary to fully validate the efficacy of the model in real-world settings. Overall, this selection of learning rates and epochs helped to balance convergence, stability, and overfitting, helping to create a model that could robustly and accurately localize abnormal hippocampal tissue.

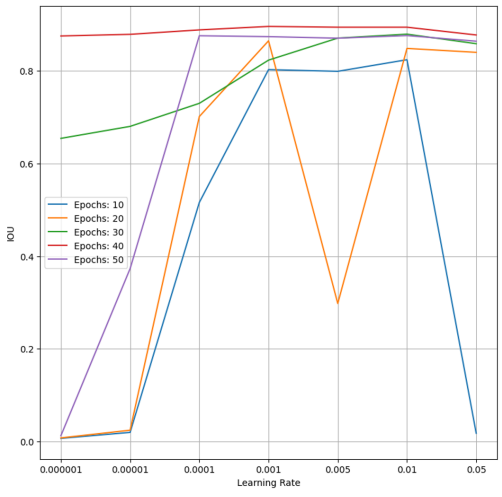


Figure 6. Graph of Segmentation IoU Results

The most optimal combination of hyperparameters yielded

an IoU of 0.902 on an unseen test dataset, highlighting the model's ability to generalize patterns accurately beyond the test dataset; the relative consistency between the validation and test IoU results demonstrates the model's stability and lack of overfitting.

Classification Performance: For the second experiment of classification, the MobileNet model demonstrated excellent balanced performance in distinguishing healthy from non-healthy hippocampal tissues. The performance of the model was recorded according to a simple accuracy metric, where the higher the score, the more accurate the model was in identifying healthy or unhealthy hippocampal scans. Similar to the segmentation experiment, hyperparameters were closely analyzed with the same ranges in learning rate and epochs. Table II demonstrates how all combinations of hyperparameters achieved relatively high accuracy scores, with the majority being greater than 0.77. Across all epochs, the lowest learning rate of 0.000001 had the lowest accuracy score, conveying how this small learning rate hindered gradient updates, preventing the model from capturing meaningful updates in each epoch. Lower-end learning rates (0.00001 and 0.0001) had varied results, but larger epochs achieved higher scores due to the model's ability to create more gradient updates with more epochs. The model achieved the best results with learning rates ranging from 0.001 to 0.01, indicating that moderately high learning rates allowed MobileNet to converge efficiently and extract discriminative features reliably for classification. The highest accuracy of 0.9483 was achieved at 20 epochs and a learning rate of 0.01, indicating how despite containing a lower number of epochs, the larger learning rate allowed the model to effectively adjust its gradients and learn patterns effectively. The largest learning rate of 0.05 fared slightly poorly compared to the previous epochs most likely due to training instability, with a sudden decrease in performance at 30 epochs. This model's accuracy results indicate how the MobileNet model fairs better with larger learning rates as they enable faster convergence and more effective gradient updates. The graph in Figure 7 highlights the accuracy results of the classification experiment, by showing the trends across the hyperparameters.

The normalized confusion matrix, as shown in Figure 5, corroborates the accuracy results and shows the classification model's performance in distinguishing between "healthy" and "non-healthy" cases. The model exhibits a high accuracy in identifying true positive cases, with 87.36% accuracy, as well as true negative cases, with a 96.55% accuracy. This indicates that the model is able to accurately discern if the MRI scan is healthy or unhealthy, with minimal error. The model fares slightly better when detecting healthy scans, as evidenced by its higher accuracy for true negatives. Conversely, the slightly lower true positive accuracy indicates that a small proportion of unhealthy scans may be misclassified as healthy, highlighting areas where further refinement during training could enhance sensitivity without compromising overall reliability.

These results can be analyzed in terms of precision and recall metrics. Precision measures the proportion of true positive predictions to total positive predictions, focusing on decreasing the occurrence of false positive predictions. Recall, however, measures the proportion of true positive predictions to total predictions and focuses on minimizing false negatives. The precision score for "healthy" cases was 0.96, but for "unhealthy" cases, it was 0.88. However, the recall score for "healthy" cases was 0.87, but for "unhealthy" cases, it was 0.97. These results illustrate how the model prioritized minimizing false negative predictions over isolating false positive cases. A false positive diagnosis may lead to unnecessary medical intervention, leading to additional procedures and tests that contain certain risks. In clinical settings, however, a false negative diagnosis may pose a more significant danger to patients, leading to delayed treatment and intervention, reducing the chances of optimal patient recovery. These findings demonstrate how the MobileNet algorithm has the potential to be deployed in a clinical setting to detect structural abnormalities in hippocampal MRIs, as it is able to accurately classify unhealthy from healthy scans. Further experiments will be conducted under the purview of radiologists to further validate the results.

Combination of Both Experiments: When a dual approach of the models was employed, with classification occurring before segmentation, the overall architecture yielded success. When tested on an unseen test image, the architecture exhibited a confidence score of 1 and was able to accurately segment the location of the impacted hippocampus on a prediction mask. The overall results from all three experiments demonstrate how a dual deep learning approach with classification and segmentation is able to successfully detect hippocampal abnormalities, and ultimately detect the presence of TLE. This approach has the potential to assist physicians with diagnosing TLE, as these complementary models produce clinically meaningful outputs, which guide further evaluation and treatment planning.

Despite the model's potential for success, there were some limitations within the methodology of the study which may have constrained the model's reliability. The analysis was conducted using a single-source dataset, which only contained 50 scans; this limited amount of data may have limited the model's exposure to the diversity of data found in typical

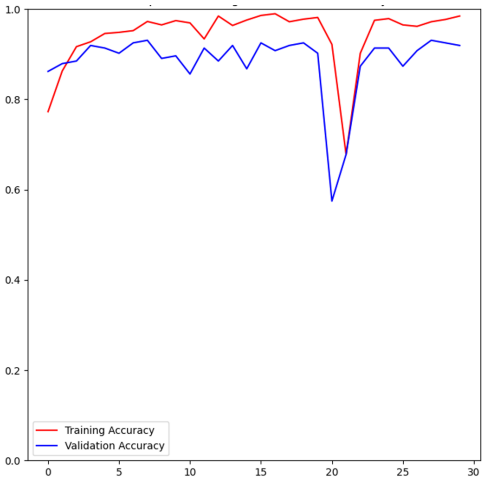


Figure 7. Graph of Classification Accuracy Results

clinical settings. Additionally, a more detailed analysis of other varieties of classification and segmentation models is necessary to determine the best possible architectures to achieve the highest accuracies and greatest generalizability across varied datasets.

For the second experiment of classification, the MobileNet model demonstrated excellent balanced performance in distinguishing healthy from non-healthy hippocampal tissues. The performance of the model was recorded according to a simple accuracy metric, where the higher the score, the more accurate the model was in identifying healthy or unhealthy hippocampal scans. Similar to the segmentation experiment, hyperparameters were closely analyzed with the same ranges in learning rate and epochs. Table II demonstrates how all combinations of hyperparameters achieved relatively high accuracy scores, with the majority being greater than 0.77. Across all epochs, the lowest learning rate of 0.000001 had the lowest accuracy score, conveying how this small learning rate hindered gradient updates, preventing the model from capturing meaningful updates in each epoch. Lower-end learning rates (0.00001 and 0.0001) had varied results, but larger epochs achieved higher scores due to the model's ability to create more gradient updates with more epochs. The model achieved the best results with learning rates ranging from 0.001 to 0.01, indicating that moderately high learning rates allowed MobileNet to converge efficiently and extract discriminative features reliably for classification. The highest accuracy of 0.9483 was achieved at 20 epochs and a learning rate of 0.01, indicating how despite containing a lower number of epochs, the larger learning rate allowed the model to effectively adjust its gradients and learn patterns effectively. The largest learning rate of 0.05 fared slightly poorly compared to the previous epochs most likely due to training instability, with a sudden decrease in performance at 30 epochs. This model's accuracy results indicate how the MobileNet model fairs better with larger learning rates as they enable faster convergence and more effective gradient updates.

The normalized confusion matrix as shown in Figure 5 exhibits a high accuracy in identifying true positive cases, with 87.36% accuracy, as well as true negative cases, with a 3.45% accuracy. These results can be analyzed in terms of precision and recall metrics. Precision measures the proportion of true positive predictions to total positive predictions, focusing on decreasing the occurrence of false positive predictions. Recall, however, measures the proportion of true positive predictions to total predictions and focuses on minimizing false negatives. The precision score for "healthy" cases was 0.96, but for "unhealthy" cases, it was 0.88. However, the recall score for "healthy" cases was 0.87, but for "unhealthy" cases, it was 0.97. These results illustrate how the model prioritized minimizing false negative predictions over isolating false positive cases. In clinical settings, a false negative diagnosis may pose a more significant danger to patients, leading to delayed treatment and intervention. These findings demonstrate how the MobileNet algorithm has the potential to be deployed in a clinical setting to detect structural abnormalities in hippocampal MRIs.

VI. Conclusion and Future Work

This study underscores the potential of integrated deep learning models as powerful tools for identifying hippocampal abnormalities associated with TLE. By achieving an IoU of 0.902 with the U-Net segmentation model and a test accuracy of 0.94 with the MobileNet classifier, the study highlights the strong performance of both architectures in accurately localizing unhealthy hippocampal tissue. These results, supported by IoU, accuracy and confusion matrix metrics, reveal the capability of the combined architecture to minimize misdiagnoses and provide consistent and unbiased analysis suitable for clinical decision support. However, several opportunities remain for future research. The study utilized a relatively limited data set with slice-level labels, which can constrain the generalizability of the model across diverse patient populations and imaging conditions. A dataset containing volumetric scans from a variety of centers could reduce the risk of bias. Furthermore, a broader exploration of hyperparameters, such as optimizer selection, learning rate schedules, and augmentation strategies, as well as the incorporation of 3D volumetric data, could significantly enhance model robustness and diagnostic precision. Additionally, expanding the number of algorithms that were trained and tested upon would further enhance understanding of which models are most effective for segmentation and classification. Expanding the framework to encompass additional epilepsy subtypes or other neurological disorders may further improve versatility.

Overall, this study presents a promising approach to advancing TLE diagnostics through the integration of deep learning into neuroimaging workflows. By enabling efficient, accurate, and scalable detection of hippocampal abnormalities, the proposed models can assist clinicians in early identification and assessment of epilepsy-related pathology. Future advancements, like training the models from a larger variety of data sources, validation by industry professionals, and rigorous evaluation in real-world clinical environments, will be pivotal in shaping the future of deep learning-driven diagnostic imaging.

References

[1] A. Chaouch, A. Ben Abdallah, R. Ayari, M. H. Bedoui, and M. Aissi, "Dl segmentation and 3d assessment of hippocampal atrophy in mesial temporal lobe epilepsy", in *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2022, pp. 1–6. DOI: 10.1109/ATSIP55956.2022.9805895.

[2] E. Kadhim, S. Al-Jumaili, and O. N. Uçan, "Diagnosis of epileptic seizures and hypoxic-ischemic encephalopathy using artificial intelligence based on eeg signal: A review", in *2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, 2024, pp. 1–6. DOI: 10.1109/ISAS64331.2024.10845267.

[3] C. R. Arias, R. M. Durón, and A. V. Delgado-Escueta, "Identification of new epilepsy syndromes using machine learning", in *2019 IEEE 39th Central America and Panama Convention (CONCAPAN XXXIX)*, 2019, pp. 1–4. DOI: 10.1109/CONCAPANXXXIX47272.2019.8977043.

[4] E. Tramoni-Negre, I. Lambert, F. Bartolomei, and O. Felician, "Long-term memory deficits in temporal lobe epilepsy", *Revue Neurologique*, vol. 173, no. 7-8, pp. 490–497, Jul. 2017, Epub 2017 Aug 31. DOI: 10.1016/j.neurol.2017.06.011.

[5] D. Sone et al., "Contribution of the μ-opioid receptor system to affective disorders in temporal lobe epilepsy: A bidirectional relationship?", *Epilepsia*, vol. 64, no. 2, pp. 420–429, 2022. DOI: 10.1111/epi.17463.

[6] K. D. Laxer et al., "The consequences of refractory epilepsy and its treatment", *Epilepsy & Behavior*, vol. 37, pp. 59–70, Aug. 2014, Open Access article under CC BY-NC-ND license. Review article., ISSN: 1525-5050. DOI: 10.1016/j.yebeh.2014.05.031.

[7] R. K. Sadangi, B. Patra, R. Maity, N. P. Maity, and B. Charan Sutar, "A review on the diagnosis of epilepsy using ml", in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2023, pp. 1–6. DOI: 10.1109/ICCCNT56998.2023.10306863.

[8] N. Alessi, P. Perucca, and A. M. McIntosh, "Missed, mistaken, stalled: Identifying components of delay to diagnosis in epilepsy", *Epilepsia*, vol. 62, no. 7, pp. 1494–1504, 2021. DOI: 10.1111/epi.16929.

[9] S. J. Czuczwar, Ed., *Epilepsy*. Brisbane, Australia: Exon Publications, Apr. 2022, Open Access. Available from NCBI Bookshelf, ISBN: 978-0-6453320-4-9. DOI: 10.36255/exon-publications-epilepsy.

[10] D. Villamizar-Torres, A. C. Cepeda Trillos, and A. Vargas-Moreno, "Mesial temporal sclerosis and epilepsy: A narrative review", *Acta Epileptologica*, vol. 6, no. 1, p. 28, Sep. 2024, Open Access, ISSN: 2524-5333. DOI: 10.1186/s42494-024-00172-5.

[11] M. Thom, "Review: Hippocampal sclerosis in epilepsy: A neuropathology review", *Neuropathology and Applied Neurobiology*, vol. 40, no. 5, pp. 520–543, Aug. 2014, Open Access. Highly cited: 428+ citations, ISSN: 0305-1846. DOI: 10.1111/nan.12150.

[12] R. S. S. Widodo, I. K. E. Purnama, and R. F. Rachmadi, "Volumetric hippocampus segmentation using 3d u-net based on transfer learning", in *2024 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2024, pp. 1–6. DOI: 10.1109/CIVEMSA58715.2024.10586572.

[13] E. H. Middlebrooks et al., "Radiologic classification of hippocampal sclerosis in epilepsy", *American Journal of Neuroradiology*, Feb. 2024, Published online ahead of print. DOI: 10.3174/ajnr.A8214.

[14] J. Jiang et al., "Automated detection of hippocampal sclerosis using real-world clinical mri images", *Computers in Biology and Medicine*, vol. 162, p. 107 457, 2023. DOI: 10.1016/j.compbiomed.2023.107457.

[15] A. Chaouch et al., "Deep learning brain mri segmentation and 3d reconstruction: Evaluation of hippocampal atrophy in mesial temporal lobe epilepsy", in *Good Practices and New Perspectives in Information Systems and Technologies (WorldCIST 2024)*, ser. Lecture Notes in Networks and Systems, vol. 986, Springer, Cham, 2024, pp. 243–253. DOI: 10.1007/978-3-031-60218-4_22.

[16] A. J. Chang, R. Roth, E. Bougioukli, et al., "Mri-based deep learning can discriminate between temporal lobe epilepsy, alzheimer's disease and healthy controls", *Communications Medicine*, vol. 3, no. 1, pp. 1–13, 2023. DOI: 10.1038/s43856-023-00262-4.

[17] J. Jiang et al., "Automated detection of hippocampal sclerosis using real-world clinical mri images", *Frontiers in Neuroscience*, vol. 17, p. 1 180 679, 2023. DOI: 10.3389/fnins.2023.1180679.

[18] K. Jafari-Khouzani, K. Elisevich, S. Patel, and H. Soltanian-Zadeh, *Mri dataset for hippocampus segmentation (hippseg_2011)*, https://www.nitrc.org/projects/hippseg_2011/, Accessed: 2025-11-29, 2017.

[19] O. F. Kaneko, N. J. Fischbein, J. Rosenberg, M. Wintermark, and M. M. Zeineh, "The "white gray sign" identifies the central sulcus on 3t high-resolution t1-weighted images", *American Journal of Neuroradiology*, vol. 38, no. 2, pp. 276–280, 2017. DOI: 10.3174/ajnr.A5014.

[20] A. Sudhyadhom, I. U. Haq, K. D. Foote, M. S. Okun, and F. J. Bova, "A high resolution and high contrast mri for differentiation of subcortical structures for dbs targeting: The fast gray matter acquisition t1 inversion recovery (fgatir)", *NeuroImage*, vol. 47, T44–T52, 2009, International Brain Mapping Intraoperative Surgical Planning Society (IBMISPS), ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2009.04.018.

[21] X. Pei et al., "Robustness of machine learning to color, size change, normalization, and image enhancement on micrograph datasets with large sample differences", *Materials & Design*, vol. 232, p. 112 086, 2023. DOI: 10.1016/j.matdes.2023.112086.

[22] S. Guo and L. Li, "Medical image segmentation based on 3d u-net", in *2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP)*, IEEE, 2020, pp. 279–283. DOI: 10.1109/ICSIP49896.2020.9277322.

[23] N. Siddique, P. Sidike, C. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications", *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021, arXiv preprint arXiv:2011.01118. DOI: 10.1109/ACCESS.2021.3086020.

[24] H. Lu, Y. She, J. Tie, and S. Xu, "Half-unet: A simplified u-net architecture for medical image segmentation", *Frontiers in Neuroinformatics*, vol. 16, p. 911 679, Jun. 2022. DOI: 10.3389/fninf.2022.911679.

[25] F. Turk and M. Kılıçaslan, "Lung image segmentation with improved u-net, v-net and seg-net techniques", *PeerJ Computer Science*, vol. 11, e2700, Feb. 2025. DOI: 10.7717/peerj-cs.2700.

[26] R. Ochoa-Ornelas, A. Gudiño-Ochoa, J. A. García-Rodríguez, and S. Uribe-Toscano, "Enhancing early lung cancer detection with mobilenet: A comprehensive transfer learning approach", *Franklin Open*, vol. 10, p. 100 222, 2025, Available online 22 January 2025. DOI: 10.1016/j.fraope.2025.100222.

[27] M. Prakash M and K. GN, "A comparative analysis of mobilenet and xception architecture for classification of endoscopy images", *International Journal of Computing and Artificial Intelligence*, vol. 3, no. 2, pp. 60–65, 2022, ISSN: 2707-658X. DOI: 10.33545/27076571.2022.v3.i2a.56.