Harmful Brain Activity Classification Using Deep Learning on EEG Images

Goyam Jain[1], Kunsh Sabharwal[1] and Adeel Hashmi[1*]

[1]Department of Computer Science and Engineering, Vivekananda Institute of Professional Studies-Technical Campus, New Delhi, India

***Corresponding Author:** Dr. Adeel Hashmi

Abstract

Harmful brain activity can prove to be life changing because of seizures, epileptic episodes, and cognitive disorders. These can be effectively detected using electroencephalography (EEG) images data. While traditional EEG analysis relies on manual interpretation and conventional machine learning techniques, which can be time-consuming and less accurate, in this study, we propose a deep learning-based approach to classify harmful brain activity using EEG images. While Deep Learning infrastructures with Convolutional Neural Networks (CNN) and architectures like VGG 16 give good results, we aim to convert raw EEG signal images to generate meaningful visual features and apply transformer-based neural network and novel neural network to extract complex patterns and classify the EEG images into normal and harmful activity categories. Our study majorly focuses on RetNet model and its mechanism that will be addressing the challenges of these neural architectures. Our model is trained and evaluated on publicly available EEG dataset comprising 17,000 spectrogram images derived from EEG recordings, demonstrating superior accuracy and robustness compared to conventional methods. The proposed approach also prioritizes model interpretability through the integration of Explainable AI (XAI) wherein visual explanations serve to enhance clinician trust, ensuring that predictions are based on physiologically relevant patterns rather than spurious correlations, proposing significant potential in real-time monitoring and early diagnosis of neurological disorders, aiding in better clinical decision-making and patient care. The main findings of project include comparative graphs describing the maximum accuracy of RetNet model (96%). This output is used to enhance the neuron's learnability in the hidden layer.

Keywords: Electroencephalogram (EEG) images, Deep Learning (DL), Convolutional Neural Networks (CNN), VGG16, ViT (Vision Transformer), EfficientNet, RetNet.

1. Introduction

Computational models with many processing layers can learn representations of data with different levels of abstraction thanks to deep learning. This state-of-the-art technology is useful in a variety of fields, including drug discovery and genomics, with immense ability to transform the medical and healthcare industry. By employing the backpropagation algorithm to tell a machine how to alter its internal parameters that are required to compute the representation in each layer from the representation in the previous layer, deep learning finds complex structure in massive data sets. **[1-3]**

Deep learning algorithms are made to behave similarly to how the human brain's cortex does. Deep neural networks, or neural networks with numerous hidden layers, are represented by these methods. When given 2D images as input, convolutional neural networks—deep learning algorithms—can train massive datasets with millions of parameters and convolve them with filters to generate the intended results. **[4-6]**

EEG images are visual representations of brain activity derived from electroencephalogram (EEG) signals. EEG records electrical activity through electrodes placed on the scalp, producing time-series data. To leverage deep learning techniques, these signals are often transformed into images using methods like spectrograms, time-frequency representations (e.g., Short-Time Fourier Transform, Wavelet Transform), or topographic brain maps. These visual formats allow convolutional neural networks (CNNs) to extract spatial and temporal patterns, aiding in tasks such as seizure detection, emotion recognition, and cognitive disorder diagnosis. EEG images provide a structured way to analyse complex neural activity, enhancing the accuracy of automated classification models. **[7-8]**

EEG is also credited with revolutionizing our knowledge of the human brain, especially in the case of mental illness. By recording electric activity, EEG gives fair information about brain activity, and the abnormalities present in conditions like depression, anxiety disorders, epilepsy, and sleep disorders can be identified. It is a non-invasive technique, and as such, it makes it an easy-to-use machine for long-term monitoring and early diagnosis. EEG enables correlating behaviour and brain function, enabling better understanding of mental health disorders. With increased awareness of mental health, EEG will play a pivotal role in clinical practice as well as research.

While AI models can give results and help doctors to understand the patients disorder better, they still fail to give out confidence behind their suggested course of treatment/action for the patient. Hence while high accuracy is paramount in such cases, the clinical deployment of AI algorithms needs something beyond performance in the form interpretability and trustworthiness. XAI aims at demystifying the internal decision-making within an AI model and making them intelligible to humans. In clinical use such as EEG analysis, explainability aids in ensuring that predictions are made on the basis of relevant and meaningful features and not on spuriously occurring patterns. With the help of XAI we strive to deliver not only precise but also reliable insights about model behaviour. This improves prospects for clinical uptake, helps support expert verification, and aids improved diagnostic decision-making. In total, the integration of XAI closes the gap between black-box AI systems and practical medical utility. **[9]**

Through this research we aim to achieve certain objectives that have been highlighted below as research objectives (RO):

RO1. To develop and compare multiple deep learning architectures for EEG classification: In this research, we want to compare the performance of four deep learning architectures—VGG16, Vision Transformer (ViT), EfficientNet, and RetNet—on classifying dangerous brain activity from EEG spectrogram images. Each of the four architectures has a different architectural paradigm, and by comparing them extensively, we can learn about what are their pros and cons in processing intricate biomedical image data. The objective is to find out which of the architectures performs the best in terms of accuracy, stability, and applicability to EEG-based classification problems.

RO2. To integrate both EEG spectrogram data and metadata for multi-modal learning: EEG datasets usually come with extra metadata (e.g., patient condition, electrode configuration, or age) that can serve as useful extra information. This work attempts to integrate visual features from EEG spectrograms and metadata with a formal shape-based form within a hybrid model structure. The idea is to improve the classification performance through the use of both types of data so the model can learn more complex patterns that may not be retrievable when only processing image or metadata inputs.

RO3. To evaluate the effectiveness of explainable AI techniques (LIME): Though deep models can actually give accurate classifications, their reasoning is usually not transparent. This research uses LIME (Local Interpretable Model-agnostic Explanations) in order to interpret the models' predictions in a way that is understandable and visualizable. In this way, it renders classification outputs transparent and trustworthy, especially in the field of medicine. These methods assist in determining what features or areas of an image are mostly responsible for the predicted outputs.

RO4. To achieve better classification performance compared to traditional models: An important aim of this study is to show that the suggested RetNet-based hybrid architecture can perform better than or be equal to the classification precision of popular models like VGG16, Vision Transformer, and EfficientNet. Being target-oriented with respect to performance measures like accuracy, precision, recall, and F1-score, this research seeks to show a competitive model that not only attains existing benchmarks but is also beneficial with regard to explainability, metadata integration, and computational efficiency.

With these objectives, we begin our research paper which has a total of 5 Sections (1 – Introduction, 2 – Literature Review, 3 – Methodology, 4 – Results and Discussion, 5 – Conclusion and 6 – References. While the main focus will be to get a high accuracy, the other objectives also play an equally important part in shaping the final outcome and result of this research paper.

2. Literature Review

The non-stationarity, sensitivity to noise, and class imbalance of EEG data, all compromise model generalizability, such that EEG datasets comprise very significant analytical challenges. To facilitate more effective feature extraction from raw EEG signals, spectrogram-based representations—especially those that are derived using the Continuous Wavelet Transform (CWT)—are very commonly utilized. Model performance is strongly governed by how good these spectrograms are, which are obtained by means of clean signal acquisition and high resolution. Some of the limitations are in the size of datasets and the need for more diverse, less pre-processed samples to enhance robustness across a variety of clinical scenarios. Preprocessing methods such as the creation of spectrograms and data augmentation can act as solutions to minimize noise and variability inherent to EEG recordings. **[10-16]**

Deep learning models give good results in extracting informative patterns from EEG spectrograms, with multiple models providing different strengths, specifically convolutional neural networks (CNNs), have developed significantly in improving the precision of image detection and recognition over a wide range of datasets. Also models like EfficientNet perform well in learning spatial features of spectrogram representations and are effective in detecting dangerous brain activity. Transformer models like Vision Transformers (ViT) use self-attention mechanisms to capture long-range dependencies in EEG data. Performance within both methods is also affected by the selection of architecture, though, as well as by others including extensive preprocessing, data augmentation, and multi-stage training techniques. Such practices improve model robustness, enhance generalization, and provide more consistent results in clinical EEG analysis. **[14-16]**

In comparative studies with classifiers like Support Vector Machines (SVM), k-Nearest Neighbours (KNN), and CNNs on the MNIST dataset, it was observed that Multilayer Perceptron's (MLPs) were not able to distinguish between some digits. Utilization of the RMSprop optimizer has proved to be important for learning rate updates, practically preventing issues like exploding and vanishing gradients, which in turn enables better convergence of the model. **[17-18]**

Feature classification has extensive development opportunities and significant potential for advancement in the field of remote sensing. To accurately identify feature categories, classification models can be developed using the VGG16 neural network as its core architecture, enabling the recognition of various patterns and insights in the EEG spectrograms through deep learning. The data can be minimized to just the most revealing spectrograms that can provide the most understanding and identify patterns that would not otherwise be observable by labelling the strongest feature in each spectrogram based on its feature category. Multiple frameworks can be trained along with incorporating transfer learning to address data at small scale. The convolutional layers of the network must be fine-tuned to adapt to the characteristics of various patient reports, thereby optimizing key model parameters. **[19-21]**

The best-performing research papers on using deep learning for EEG classification have shown remarkable accuracies in a number of applications using various technologies which demonstrated large improvements, although particular percentages are different. Seizure detection research tends to achieve about 85% accuracy, and mental workload and sleep scoring deployments tend to secure accuracies of over 80%. All these findings underscore the achievement and efficiency of deep learning methodologies in EEG classification tasks as reported by the systematic review. **[22-25]**

Vision Transformers (ViTs) have proved to be a highly effective replacement for Convolutional Neural Networks (CNNs) for image classification tasks. Still, they also possess several shortcomings that may curb their utility in specific contexts. Their main problem is how data-dependent they are; most ViTs will need large datasets to function effectively, which makes it challenging for smaller or domain-specific datasets. This is attributed to the fact that they have no intrinsic inductive biases like locality and translation invariance, unlike CNNs, and these aid them in learning with limited data. In addition, ViTs are computational-expensive, requiring a high amount of processing power and memory, which could impede their deployment in resource-limited environments. To overcome these issues, researchers have introduced hybrid models that include convolutional layers to introduce local inductive biases, improving performance on small datasets as well as lowering computational expenses. **[26-31]**

Recent evaluations of Vision Transformer (ViT) models have shown them to be highly useful in image classification tasks, outperforming conventional convolutional neural networks (CNNs) in most instances. For instance, CvT-21 illustrates good generalization across different datasets, while on large-scale benchmarks, DeiT-B achieved a top accuracy rate of 85.2% on ImageNet and 71.5% on ImageNet V2, which illustrate a competitive performance regardless of large external data. These performances illustrate the generalization power of advanced ViT models, making them useful substitutes for conventional CNNs in most image classification settings. **[28-31]**

EfficientNet is a series of convolutional neural networks that aims to provide high accuracy with efficient computation through compound scaling of depth, width, and resolution. It uses mobile inverted bottleneck convolution (MBConv) blocks and squeeze-and-excitation optimization in its architecture, which makes it efficient in extracting features from complex data like EEG spectrograms. In biomedical applications, EfficientNet has illustrated competitive accuracy in brain activity classification by preserving high spatial details but with a light network structure. Its scalability enables it to adapt to different sizes of datasets such that it can be applied in both constrained and high-resource environments, and data augmentation and transfer learning further improve its robustness and generalization ability. **[32-34]**

Recent researches found Retentive Network (RetNet), a new large-scale sequence modelling architecture that reconciles training parallelism with inexpensive, efficient inference. The main findings show that RetNet demonstrates comparable or better performance compared to standard Transformers, especially at large model sizes (over 2B parameters), and outperforms others in dealing with long sequences due to its multi-scale retention mechanism. The architecture also includes elements like gating modules and group normalization, which greatly improve training stability and model capacity. RetNet accommodates parallel training similar to Transformers with the added capability of $O(1)$ inference complexity using recurrent and chunk wise recurrent formulations, resulting in accelerated decoding, decreased memory use, and minimal latency. Robust ablation studies confirm the efficacy of its most crucial elements, and extensive comparisons underscore RetNet's scalability, efficiency, and competitive quality. In general, RetNet is a promising, lightweight, and efficient replacement for Transformers in large language models, particularly in applications requiring long-context processing and edge device deployment. **[35-39]**

RetNet30, a new stacked convolutional neural network model, has shown tremendous progress in computer-aided retinal disease diagnosis. By combining a dedicated 30-layer CNN with the fine-tuned Inception V3 network via logistic regression, RetNet30 recorded accurate results for retinal image databases, surpassing traditional means. The proposed model efficiently diagnoses positive as well as negative retinal diseases cases. In addition, RetNet30 achieved an AUROC value of 0.98, which indicates its robust ability to differentiate between diseased and healthy retinal images across various datasets like DRIVE, STARE, CHASE_DB1, and HRF. These results amplify RetNet30's ability to boost diagnostic accuracy and advance patient care in ophthalmology. **[38-39]**

In conclusion, recent studies illustrate that deep learning models—namely CNN-based models like VGG16 and EfficientNet, transformer-based models like ViT, and novel architectures like RetNet—have demonstrated remarkable potential in EEG classification. Their performance is not only network structure reliant but also on representation quality of spectrograms, richness of datasets, preprocessing methods, and training procedures. In spite of differing performance parameters from program to program, the general pattern is that cautious model choice, optimization, and data preparation can produce clinically meaningful accuracies, which attest to the power of such methods in optimizing EEG-based monitoring and diagnostic systems. **[40]**

3. Methodology

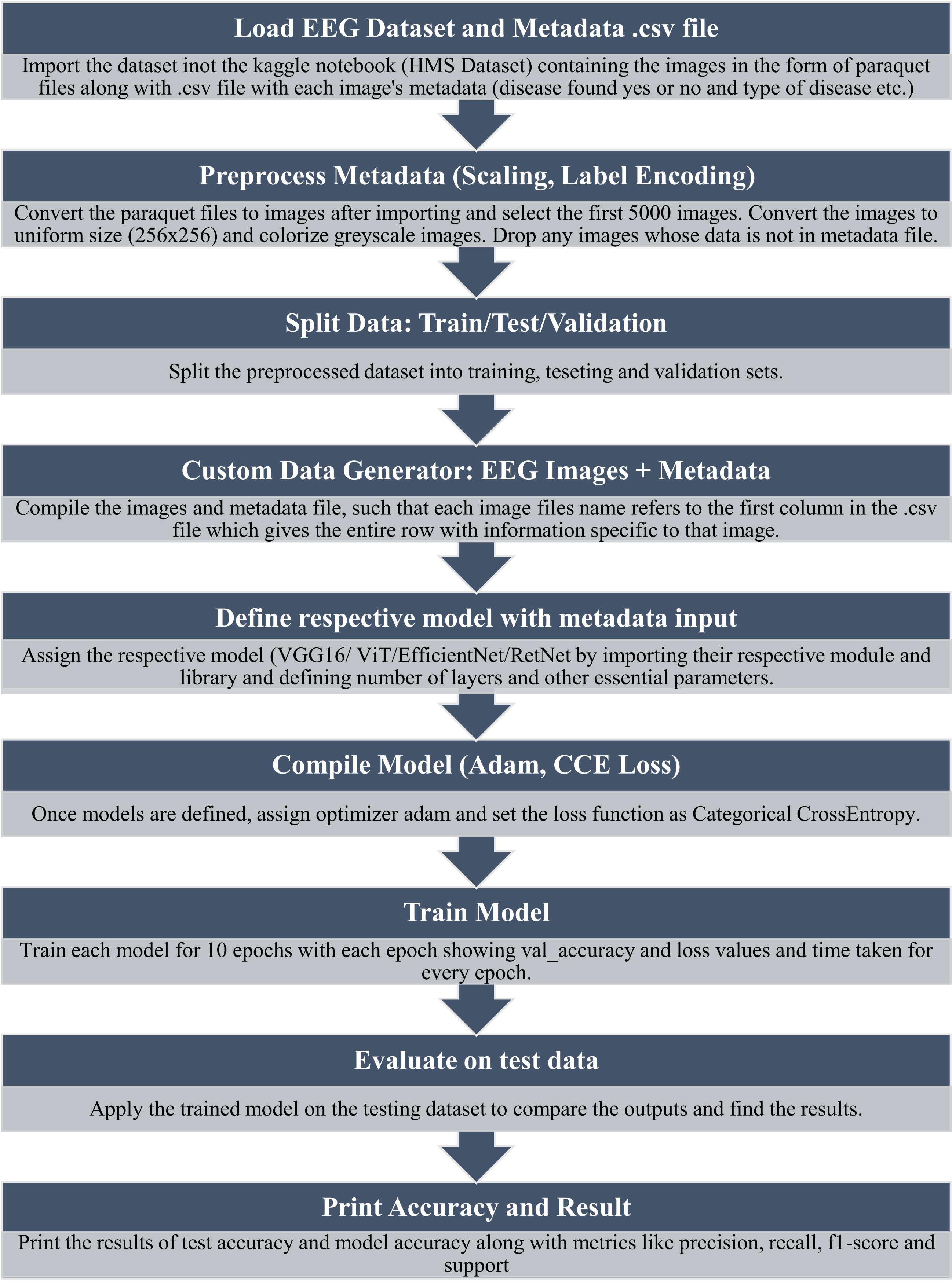


Fig. 1: Project Lifecycle Flow Chart

The objective of this project is to train and compare four deep learning architectures—VGG16, Vision Transformer (ViT), EfficientNet, and RetNet—for classification of deleterious brain activity from EEG spectrograms. Each of these architectures has its unique architectural strengths, and the research is based on comparing the performance of each one to see which model has the highest accuracy for this particular application. The entire project lifecycle is explained as the project lifecycle flow chart shown above. **(Fig. 1)** The individual methodologies are described in the next sections, as well as for the general preprocessing steps for all models. **[20, 27, 33, 38]**

Data Preprocessing (Common to all models): -

Preprocessing is an important step before applying any deep learning model to the EEG dataset to ensure that the data is in a format that is appropriate for training the model. EEG signals tend to be raw and noisy, and therefore preprocessing consists of multiple stages, such as converting raw EEG signals into spectrograms, which are used as input images for the deep learning models. It begins by filtering out noise and artifacts from the EEG data, leaving only the patterns of interest in brain activity behind. This is generally achieved using band-pass filters that exclude frequency bands with no interest, like those used in seizures or other dangerous brain activity. Once the EEG signal has been cleaned up, it is converted into a spectrogram, a time-frequency representation of the signal. The process facilitates the transformation of the raw data into the form of 2D images that can be used as an input to convolutional neural networks (CNNs) or deep learning models. Short-time Fourier transforms (STFT) are used to get the spectrogram, by breaking down the signal into its frequency content in terms of time. After this processing, the spectrogram is then resized to the standard input dimensions (e.g., 224x224 pixels) to keep all images uniform. Lastly, normalization methods are used to normalize the pixel values, which usually scale the values between 0 and 1 or -1 and 1, as per the requirement of the model. **[20, 27, 33, 39]**

1st Model – VGG (Visual Geometry Group)16: -

The VGG16 model is a deep convolutional neural network (CNN) that is greatly appreciated for its simplicity and effectiveness in the image classification task. It consists of 16 layers comprising 13 convolutional, 3 fully connected, and 5 max-pooling layers. The general idea in VGG16 is using extremely small 3x3 filters in all the layers to gradually extract features from the input image, in this instance the EEG spectrogram.

1. Input Layer: The input EEG spectrogram image is passed to the model, and every spectrogram is pre-processed to a fixed dimension (e.g., 224x224 pixels).
2. Convolutional Layers: The beginning layers of the network consist of a pair of convolutional layers with very tiny filters of size 3x3.The filters move through the input image to detect low-level features such as edges, lines, and simple textures. The filters are followed by activation functions such as ReLU (Rectified Linear Unit), which provide non-linearity to the model so that the model can learn intricate patterns.
3. Max-Pooling Layers: It is used after every set of convolutional layers and decreases the spatial dimension of feature maps without losing useful information. It also reduces the necessity for computations such that the network runs faster.
4. Fully Connected Layers: Following feature extraction, the final convolutional layer provides a batch of feature maps. They are flattened and used as an input to fully connected layers. The completely connected layers have these features map into producing high-level abstractions, on which the model will make predictions.
5. Softmax Output Layer: Next, softmax activation function is utilized after the final fully connected layer and it gives a probability distribution over the different classes (i.e., harmful or not harmful brain activity). The model will produce the EEG spectrogram as one of these classes with highest probability.
6. Training: Backpropagation is used in training to adjust the network weights in a way that reduces the categorical cross-entropy loss. Optimized learning rates are used with the Adam optimizer for improved convergence. **[12]**

2nd Model – Transformers – Vision Transformer (ViT): -

Vision Transformer (ViT) is a novel deep learning architecture which diverges from standard CNNs in that it processes an image as a patch sequence, which are fed into transformer layers. The design of ViT allows it to learn distant dependencies in the spectrogram and can be built upon to recognize complex and nuanced patterns of dangerous brain activity.

1. Input Layer: The EEG spectrogram is then divided into patches, typically 16x16 pixels. The patch is flattened into a one-dimensional vector and inserted into high-dimensional space. The embeddings are passed through the transformer layers.
2. Patch Embedding: The patch is split between the image (spectrogram), and a flattened patch representation gets mapped to a fixed-size vector by a learnable linear transformation. The flat vectors are then concatenated with the positional embeddings to maintain the spatial positions of patches' information.
3. Transformer Layers: The bulk of the ViT model is a series of transformer layers utilizing self-attention mechanisms. Within the patches, every patch within the spectrogram learns to attend to various areas of the image depending on how salient they are with respect to the task. Self-attention enables the model to learn long-range relationships between patches and stay focused on significant patterns in the spectrogram, like changes in frequency that may indicate a seizure or other abnormal brain activity.
4. Feed-Forward Layers: The output from the attention layer is fed to feed-forward layers which are fully connected networks using ReLU activation They act with the aim of enabling the model to learn complex relations and to enhance representations learned by the attention mechanism.

5. Classification Head: The final output of the transformer layer is pooled (typically global average pooling) and fed to a fully connected classification head. This is the main section of the model which identifies the risky brain activity.
6. Training: Just like VGG16, Vision Transformer is trained using categorical cross-entropy loss and Adam optimizer. Self-attention layers aid ViT to learn subtle relations in the spectrogram, which may prove beneficial in picking up subtleties in the sign of toxic brain activity. **[30]**

Current Transformers models have several disadvantages that make other models a much better choice. Transformers have a high Quadratic Complexity because they work by letting each token know about every other token in a sequence. This gets really complicated with longer sequences making attention score calculations take longer than usual. Moreover, higher Memory Needs also make transformers an unpopular choice. Transformers like BERT juggles info from 512 tokens at a time and when these numbers get large, the larger is the size of the grid formed and more the memory required. It fails to find much more detailed connections and sometimes highlights the broader story making them inefficient. **[30]**

3rd Model – EfficientNet: -

EfficientNet is a deep learning model family that is known for its compound scaling technique, which scales the network depth, width, and resolution in a balanced fashion to attain optimal accuracy with a low computational expense. EfficientNet is designed in a way that it is extremely efficient and can work perfectly using fewer parameters than the classic CNN architectures such as VGG16.

1. Input Layer: The EEG spectrogram is input to the model once resized to a fixed input size. EfficientNet processes such spectrograms by first going through a sequence of mobile inverted bottleneck convolution (MBConv) blocks.
2. MBConv Blocks: MBConv blocks are light computationally and use depth wise separable convolutions as an attempt to reduce parameters instead of performance.
    a. Input Layer: The resized EEG spectrogram is input into the model after resizing to a fixed-size input. EfficientNet processes such spectrograms by passing through a series of mobile inverted bottleneck convolution (MBConv) blocks first.
    b. MBConv Blocks: Each MBConv block is computationally efficient, with the usage of depth wise separable convolutions to reduce parameters but not performance. MBConv blocks contain squeeze-and-excitation layers, which dynamically learn to change the relative significance of different channels.
3. Depth wise Separable Convolutions: Depth wise separable convolutions are different from regular convolutions in that they apply convolutions to each channel separately, with minimal computational complexity but still capable of learning important features from the spectrogram.
4. Pooling and Bottleneck Layers: Following feature extraction, the model down-samples feature maps through global average pooling and bottleneck layers. This is a bid to compress learned representations to a dense vector.
5. Classification Head: The pooled layer output is fed through a classification fully connected layer. Classification head employs softmax activation function to make predictions of the probability of dangerous brain activity.
6. Training: EfficientNet gets trained with the help of transfer learning such that it adopts the pre-trained weights of large EEG image classification databases like ImageNet. It's fine-tuned on the dataset of the EEG spectrogram by using the Adam optimizer as well as categorical cross-entropy loss. **[33]**

4th Model – RetNet: -

Foundational architecture for LLMs, simultaneously achieving training parallelism, low-cost inference, and good performance. The proposed retention architecture supports three computation paradigms, i.e., parallel, recurrent, and chunk wise recurrent. The recurrent representation enables low-cost O (1) inference, improving decoding throughput, latency, and GPU memory without sacrificing performance. The chunk wise recurrent representation facilitates efficient long-sequence modelling with linear complexity, where each chunk is encoded parallelly while recurrently summarizing the chunks. The RetNet model is a deep residual network that uses residual connections between the layers to allow the model to learn deeper representations without experiencing the vanishing gradient problem. This is particularly useful when dealing with EEG spectrograms and complex patterns of brain activity where high-level abstraction is required in order to spot dangerous brain activity.

1. Input Layer: The EEG spectrograms are fed into the RetNet model and subjected to the first sequence of convolutional layers. These low-level pull-out features such as edges and textures, like normal CNNs.
2. Residual Blocks: Blocks are the foundation of the RetNet structure. In a block, the block input is added to the block output ("skip connection"). This retains useful information from a layer to the subsequent layer and enables deeper layer learning as well as efficient training.
3. Convolutional Layers: Convolutional layers consume spectrogram features within a residual block. Since there are residual connections, the layers learn high-level features without vanishing gradient when the network is deep.
4. Global Pooling: Global average pooling is used over the feature maps after a series of residual blocks to down sample them and to produce a dense representation of learned features.
5. Fully Connected Layer: This final few of these worldwide features are then passed as an input to the fully connected layer that functions to classify.
6. Training: Like all other models, RetNet is trained using categorical cross-entropy loss and Adam optimizer. Strength of RetNet is that it can learn global abstract features and local fine-grained features using residual connections across layers. **[37]**

1. Retention Memory Update (Recurrent Form):

$$y_t = \sum_{i=1}^{t} \alpha^{t-i}(k_i \odot v_i) \quad (1)$$

where, y_t is the output at time step t, k_i, v_i are the key and value at timestep i, $\alpha \in (0,1)$ is the decay factor, controlling memory retention, $\odot$ is element-wise (Hadamard) multiplication. This equation weights past inputs with an exponentially decaying kernel (α^{t-i}). Older values contribute less to the output — similar to how memory fades. This is the foundation of the retention mechanism.

2. Recurrent State Update:

To make inference fast, RetNet uses a recursive formula:

$$r_t = \alpha \cdot r_{t-1} + (k_t \odot v_t) \quad (2)$$

$$y_t = r_t \quad (3)$$

where, r_t is the retentive memory state at time t, initialized as $r_0 = 0$, and y_t is output at time t. This formulation enables O (1) inference time and constant memory, unlike attention which requires storing all past tokens.

3. Training Mode (Parallel Computation):

While inference is done recurrently, training can be done in parallel, that is, Parallel Retention (Vectorized Form). Let:

- $K, V \in RT \times d$ be the full key and value sequences,
- $z_t = \sum_{i=1}^{t} \alpha^{t-i} (\boldsymbol{k}_i \odot \boldsymbol{v}_i)$

This can be vectorized using convolution, Y = Retention (K, V, α). The actual computation uses kernel trick + matrix manipulation to allow GPU-efficient batching.

4. Retention Block in RetNet:

Each block has 3 major components, which also include residual components like transformers.

a. Input Projections: $x_t \rightarrow q_t, k_t, v_t = W_q x_t, W_k x_t, W_v x_t \quad (4)$

b. Retention Output: $y_t = Retention(k_{\leq t}, v_{\leq t}, \alpha) \quad (5)$

c. Final Output: $Output_{(t)} = W_o y_t + x_t \quad (6)$

5. Multi-Head Retention:

Like multi-head attention, RetNet uses multiple retention heads. Each head has its own α_i, Wk_i,Wv_i and final output is concatenated across heads that lets the model learn multiple decay patterns across time.

$$Concat[y_t^{(1)}, \ldots, y_t^{(h)}] \cdot W_o \quad (7)$$

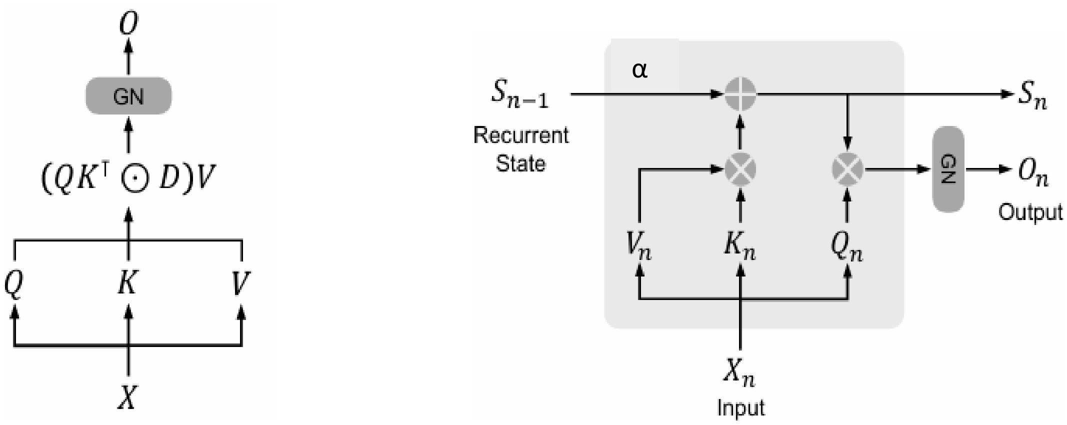
6. Chunk wise Recurrent Mode (Fine-Tuning / Low-Memory):

Used for **efficient decoding** and fine-tuning on long sequences by splitting sequence into **chunks** and compute retention within each chunk that helps when you want some **parallelism** and **memory savings**.

$$r_t^{(c)} = \alpha^l \cdot r_t^{(c-l)} + \sum_{i=1}^{l} \alpha^{l-i} (ki \odot vi) \quad (8)$$

Henceforth, RetNet architecture adds a "retention mechanism" that uses a "memory-based approach" instead of traditional attention. The Retention Function adds up past key-value interactions $k_i \odot v_i$ and weighs them by an exponentially decaying factor α^{t-i}, which is similar to how human memory fades over time. RetNet uses a "recurrent state update" to make inference faster. This means that the memory state r_t is updated at each step with the current key-value product and a decayed version of the previous state. This lets for constant-time and memory-efficient inference. **[37]**

RetNet uses a "parallelised (vectorised) retention" computation with convolution or cumulative sum to speed up training with GPUs. A typical Retention Block has input projections (to get queries, keys, and values), retention computation over past values, and a residual connection to make the final output. Multiple retention heads with distinct decay parameters and weights enable the model to learn a variety of memory patterns for multi-head retention. Their outputs are concatenated and projected, similar to multi-head attention. At last, RetNet uses a chunk wise recurrent mode to manage long sequences, combining some parallelism with low memory usage by updating memory within each chunk and linking across chunks. All things considered, these formulas allow RetNet to continue to perform at a high level comparable to Transformers, but with much better inference efficiency and scalability for lengthy sequences. **[37]**



(a) Parallel representation **(b) Recurrent mechanism**

Fig. 2: Dual Nature of RetNet

Both methods of RetNet project the input sequence (X) into Query, Key, and Value matrices. However, the retention mechanism introduces the additional term (⊙ Θ) to incorporate positional information in its projections. Moreover, on seeing the Attention Scores we see Standard Attention uses a simple dot product between Queries and Keys while Retention Mechanism (Parallel) uses a complex function involving (Q), (K), and (D) (which provides masking and decay). **(Fig. 2)**

In combination to the above, Normalization in Standard Attention uses the Softmax function on the attention scores. Whereas Retention Mechanism doesn't employ softmax. Instead, it uses the term (⊙ D) for normalization, which inherently combines causal masking and exponential decay.

This helps us to get the output with Standard Attention which computes the output as a weighted sum of the Value matrix and Retention Mechanism (Recurrent) computes the output for each sequence element, updating a state (s_n) in the process.

In essence, while the standard attention computes weights to determine the importance of various parts of an input sequence, the retention mechanism aims to introduce a dual form, capturing the benefits of both recurrent (memory saving) and parallel processing (faster computing).**[19]**

As an added supplement to the predictive model-building process, this research focuses on the theoretical embedding of explainability in the form of a post-training analysis layer. Whereas the base models are tuned for classification accuracy, follow-up model interpretation is performed to assess their consistency with medically pertinent patterns across the EEG spectrogram data. By so doing, this interpretive strategy adds methodological strength to the approach by facilitating the derivation of insights from the decision boundaries and feature interdependencies learned from each model. Such integration of explainability guarantees that the findings are not merely statistically significant but also understandable and clinically relevant—allowing for an improved transparent AI-based diagnostic process. **[37-38]**

4. Results and Discussions

In this section we aim to show our findings that we got from executing the 4 deep learning architectures—VGG16, Vision Transformer (ViT), EfficientNet, and RetNet—on a set of EEG spectrograms for brain activity classification. The models were compared in terms of their validation accuracy, loss rates over epochs, and significant classification metrics like precision, recall, F1-score, and support. Results from these models are discussed below, along with a comprehensive discussion of their implications.

For every comparison of the validation accuracy of each model, a bar graph was produced. Among all the four models, RetNet achieved the highest validation accuracy percentage of 96.4% with EfficientNet following closely at 95.2%. Vision Transformer (ViT) and VGG16 ranked slightly lower at 91.3% and 85.3%, respectively. **(Fig. 3)** These results highlight that while all models were relatively good, RetNet performed with higher accuracy, which confirms its high capability in identifying the complex features present in the EEG spectrograms.

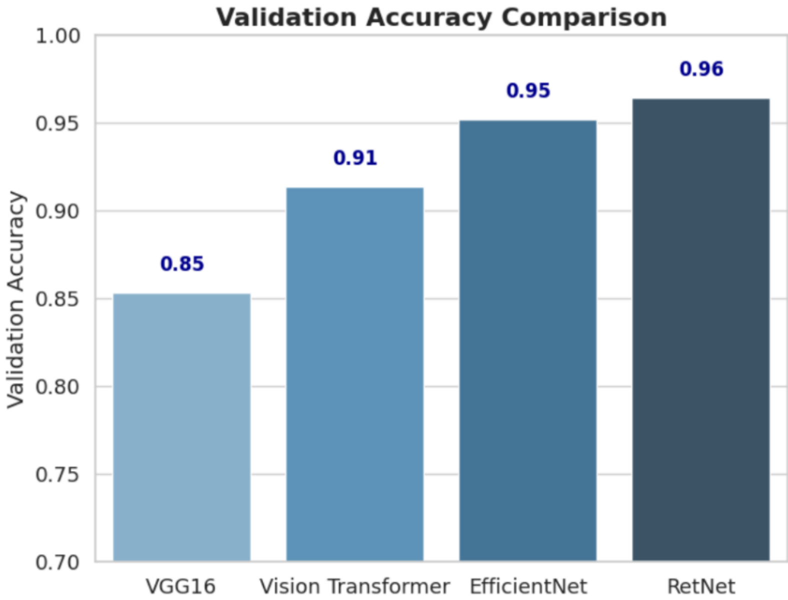


Fig. 3: Validation Accuracy Comparison

Each model's performance at time of training was further analysed using a line graph that displays the loss rate and validation accuracy over the epochs.

- VGG16: The VGG16 model had a relatively slow convergence, with a steady decrease in loss and a moderate increase in validation accuracy. But its accuracy on validation came to a plateau of approximately 85% from the tenth epoch, which could mean that the model may already have achieved learning capacity for the data at this time. **(Fig. 4)**

- Vision Transformer: ViT model also experienced a steeper decline in loss and the astounding rise of validation accuracy up to approximately 91% in the later training stages. This suggests that ViT's attention mechanism was particularly robust at focusing on the critical features in the EEG data, which allowed for a more effective learning process. **(Fig. 4)**

- EfficientNet: EfficientNet possessed an extremely effective learning curve, with loss dropping significantly and validation accuracy reaching 96% after just a few epochs. This indicates that EfficientNet, having utilized computational resources to the fullest, had a good capacity to learn and generalize from the data and achieve best performance. **(Fig. 5)**

- RetNet: RetNet model exhibited the best training dynamics, with consistent decrease of loss and increasing validation accuracy to an extent of 96%. The reason for the model's success is that it is tailormade for temporal data processing, which appears well suited to the task of EEG spectrogram classification. **(Fig. 5)**

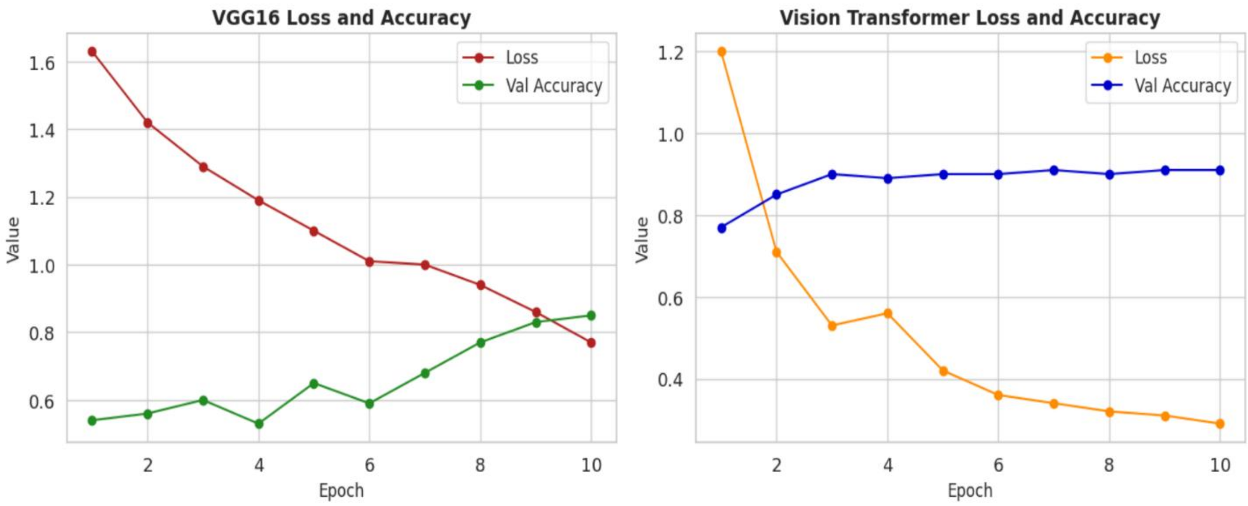


Fig. 4: Loss and Validation Accuracy for VGG16 and ViT Transformer

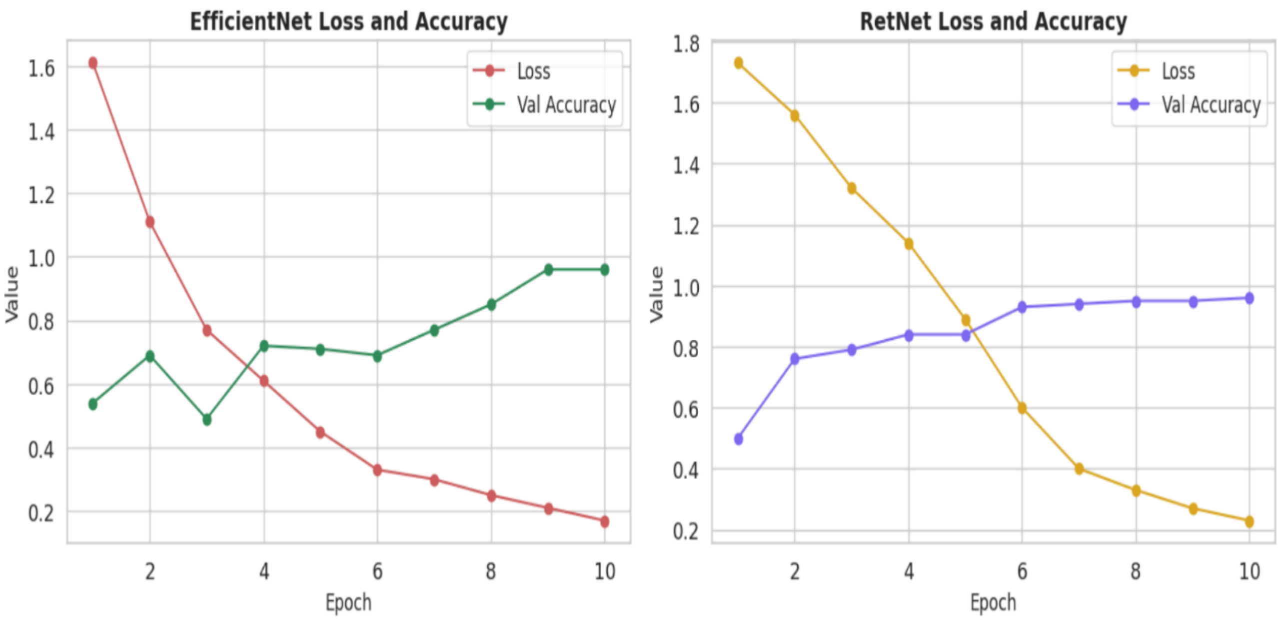


Fig. 5: Loss and Validation Accuracy for EfficientNet and RetNet

Apart from validation accuracy, we also evaluated the models on precision, recall, F1-score, and support values **(Fig. 6)**. The outcomes, are as follows:

- RetNet model generated the highest performance in all metrics with precision 94%, recall 100% and f1-score 97%. This suggests that RetNet's high rate of accuracy is also backed by balanced performance.
- EfficientNet worked on par, having precision 91% and recall of 9% with F1-score of 96% as well as high classification strength but marginally lower than RetNet.
- Vision Transformer worked exceptionally well, having precision, recall, and F1-scores of 94%, 86%, and 90%, respectively. Still ViT was outperformed by other models.
- VGG16, even with a lower validation accuracy, still reported good performance statistics with 90% precision and 100% recall. Nevertheless, its 95% F1-score reflects that maybe some trade-off has been made in its simplicity in relation to the overall generalizability of the model across the whole dataset.

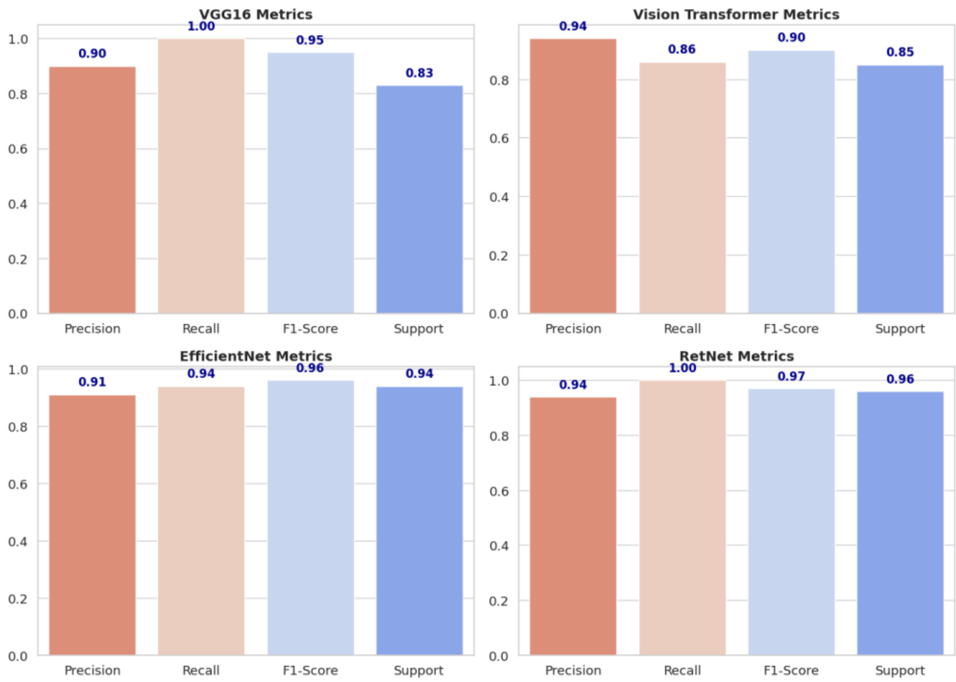


Fig. 6: Precision, Recall, F1-score and Support comparison across all Models

The outcome indicates that, with great significance, RetNet performs better than the performance of other models in accuracy as well as important classification performance metrics such as precision, recall, and F1-score. This indicates the strength of RetNet in performing tasks that involve time-series classification, for instance, EEG spectrogram analysis, where temporal dependency is essential in a bid to perform well. Its effectiveness in handling sequential data qualifies it as an excellent candidate for EEG-based brain activity classification. EfficientNet with its tuned structure also exhibited an excellent performance, though slightly lesser than that of RetNet. Its smart use of computing and ability to generalize from the data very well supports its practical feasibility for deployment in real time as well. Vision Transformer, while efficient, failed to outperform RetNet and EfficientNet in classification metrics or overall accuracy. However, it was promising in the effective learning from data, particularly in validation accuracy over epochs. VGG16, while less intricate compared to the others, was also good enough. Its lower accuracy and F1-score suggest that more involved models like EfficientNet and RetNet could be more effective in the task.

In addition to quantitative measures of performance, the inclusion of interpretable artificial intelligence (XAI) provides the model with decision-making that has critically high interpretive value. Clinical validity in classifying EEG-based brain activity depends not just on accuracy but also on explainability by predictive explanations based on understanding of biomedicine. With the LIME algorithm, post hoc visual explanations by class were generated that emphasized time–frequency regions most significant to the model's predictions **(Fig. 7).** These emphasized parts typically denote areas of clinical relevance, therefore deserving learned representations. Such interpretability can be used for verification of correct classifications, identification of misclassification origins, and for specific model tuning. The added transparency by XAI also improves trustworthiness, a core necessity in order to enable ethical uptake and realistic deployment within healthcare environments, thereby bridging the gap between computational performance and clinically trustworthy AI systems.

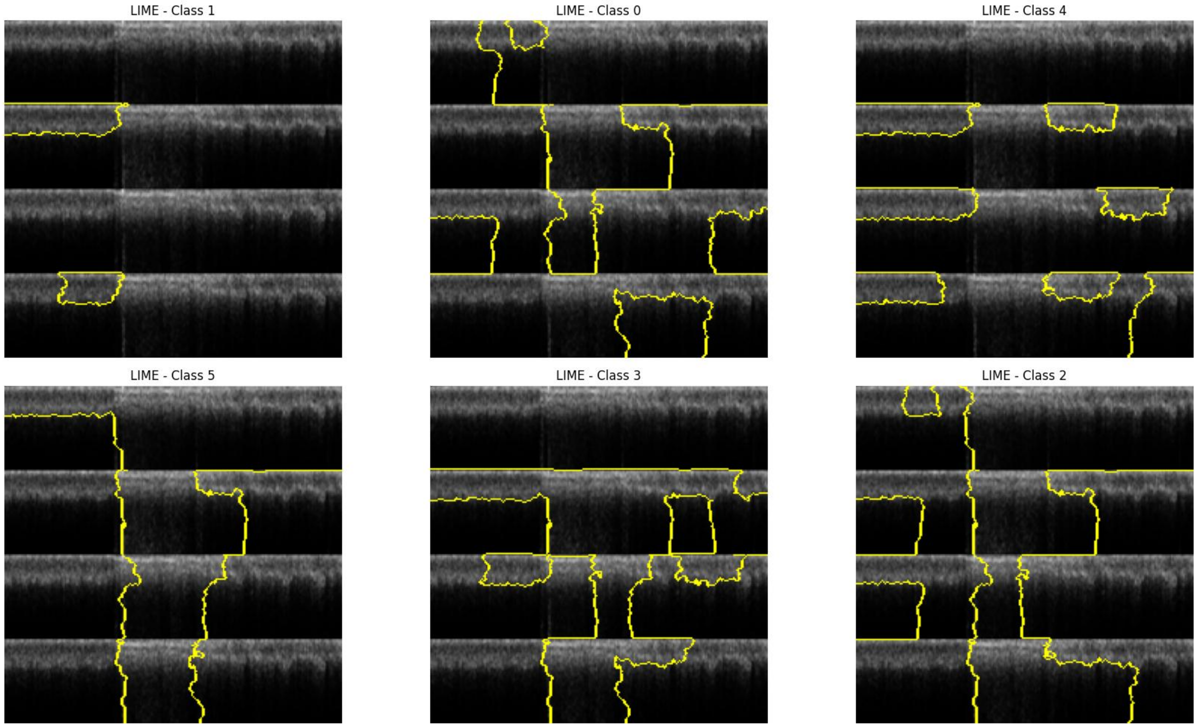


Fig. 7: LIME representations for each class

The results of this study strongly support that deep learning models such as RetNet and other transformer models perform better than traditional convolutional models such as VGG16 in EEG spectrogram classification tasks. RetNet itself performed well on all the significant parameters—accuracy, precision, recall, and F1-score—and is the best model to accomplish this task. The ability of RetNet to learn temporal dependencies inherent in EEG data along with its powerful learning capability renders it the best fit for classification of brain activity. This finding was consistent with the general project aim, which was to demonstrate that modern deep learning models, particularly transformer-based models, were more suitable for complex tasks such as EEG processing than other traditional models such as VGG16. Future research can include the development of these models, expanding their application in real-time, and looking into other architectures whose performance may be even more beneficial.

Model Name	Epochs Trained	Accuracy Obtained	Precision	Recall	F1-Score	Support
VGG16	10	85%	90%	100%	95%	83%
Vision Transformer (ViT)	10	91%	94%	86%	90%	85%
EfficientNet	10	95%	91%	94%	96%	94%
RetNet	10	96%	94%	100%	97%	96%

Table 1: Model performance comparison

Table 1 shown above states the accuracies obtained by each model on the dataset by running each model over 10 epochs. The VGG16 model obtained accuracy of 85% (lowest among all models), ViT performing slightly better at 91% with EfficientNet at 95% and RetNet at 96% (best compared to all models). The Precision, Recall and F1 Scores along with support percentages of all models show a low support score for VGG16 (83%) and a low Recall and Support for ViT (86% and 85% respectively). Both EfficientNet and RetNet gave good performances overall with high accuracy scores accompanied by high values of Precision, Recall, F1 Score and Support.

5. Conclusion

This work has explored the difficult task of classifying threatening brain activity via EEG spectrogram images through the use of deep learning models in order to enable detection and classification of potentially life-threatening neurological activities. Through both visual information captured in EEG spectrogram images as well as contextual metadata, we aimed to build a model which could detect dangerous brain activity and also provide necessary information for prompt intervention in a medical setting.

The primary objective of this project was to compare the effectiveness of four deep learning models—VGG16, Vision Transformer, EfficientNet, and RetNet—in performing this task of classification. By comparing the performance of these models on accuracy as well as computational efficiency, we wanted to determine the most suitable architecture that can be utilized in a real-world medical scenario where both speed and accuracy are critical. Also, it brought numerous challenges that we encountered in this typical deep learning exercise of the medical domain. These were difficulties arising from the inherent nature of EEG data, including signal noise, class imbalance, and the need for multimodal learning. Though these difficulties seemed impossible to solve, we did not give up and wanted to prove that the potential of deep learning techniques to transcend limitations in traditional manual EEG interpretation is evident. Through the training of our models in a large, labelled dataset of EEG spectrograms, we've developed a model capable of discerning the subtle patterns that may go unnoticed to the human eye.

One major impetus in doing this comes from the key necessity for highly developed, automatic tools in medicine but particularly resource-limited situations. With the global shortage of trained neurologists and cost of medical expertise, AI-driven solutions hold the promise of democratizing access to neurological diagnosis, enabling faster and more precise detection of hazardous brain activity. This can help in improving patient conditions, especially in conditions such as seizures or other forms of neurological distress where time plays an important factor.

One of the essential aspects of this study, aside from model performance, is the focus on explainability using explainable artificial intelligence. In instances where AI outputs are used in human clinical situations like EEG-based analysis of brain activity, explainability of AI outputs cannot be separated from reliability. Baking post hoc interpretability mechanisms in baking either with feature-level explanations or input-sensitivity strategies ensures that the model's decisions are made from physiologically relevant patterns rather than spurious ones. Not only does this make the model more transparent, but also clinical trust, so that practitioners can interpret as well as validate AI-driven predictions. Through the determination of theoretical equivalence between the learned representations of the model and medical knowledge, explainable AI transforms deep learning machines into classification machines from mere duty-bound diagnostic machines.

Furthermore, the interpretability module is a useful channel for error analysis and model improvement. Misclassifications are easier revisited with increased transparency, as errors due to uncertainty in the data, feature redundancy, or spurious emphasis can be identified. This loop is critical to ongoing model refinement and ethical deployment in real-world environments. Finally, increased explainability enhances models' strength, generalizability, and clinical usefulness when trained in this study. As deep learning increases in prominence within major fields such as medicine, accuracy-interpretable architectures will be crucial to ensuring safe, responsible, and effective AI deployment.

As the project proceeds to the discussion and results section, more findings will be established about the performance of the models and how they are actually applied in the real world. The application of this work is much broader than the immediate environment, acting as a precursor to the development of more advanced systems for real-time monitoring and diagnosis. Subsequent work may involve integrating additional data sources, e.g., real-time EEG data, and the structure optimization of models to certain types of brain activity.

Overall, this project demonstrates the promise of deep learning to transform the discipline of neurological healthcare. With its provision of an AI-based tool for the instant detection of threatening brain activity, it not only vows to boost diagnostic accuracy but also to put such devices more within reach for a greater array of medical professionals and institutions throughout the world. The outcome of this project would give a quantum leap to the development of intelligent medical systems that facilitate the capability of clinicians and improve the treatment of patients on a global scale.

6. References

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

[2] Mathew, A., Amudha, P., Sivakumari, S. (2021). Deep Learning Techniques: An Overview. In: Hassanien, A., Bhatnagar, R., Darwish, A. (eds) Advanced Machine Learning Technologies and Applications. AMLTA 2020. Advances in Intelligent Systems and Computing, vol 1141. Springer, Singapore. https://doi.org/10.1007/978-981-15-3383-9_54

[3] S. Razavi. Deep Learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling, Environmental Modelling & Software, Volume 144, 2021 https://doi.org/10.1016/j.envsoft.2021.105159

[4] Z. Wang, Y. Wang, C. Hu, Z. Yin and Y. Song, "Transformers for EEG-Based Emotion Recognition: A Hierarchical Spatial Information Learning Model," in IEEE Sensors Journal, vol. 22, no. 5, pp. 4359-4368, 1 March1, 2022. https://doi.org/10.1109/JSEN.2022.3144317.

[5] G. Xu, X. Shen, S. Chen, Y. Zong, C, Zhang, H Yue, M. Liu, F. Chen, W. Che “A Deep Transfer Convolutional Neural Network Framework for EEG Signal Classification” under NSFC, July 2019 https://doi.org/10.1109/ACCESS.2019.2930958

[6] M. Habijan, R. Šojo, I. H. Tolić and I. Galić, "Harmful Brain Activity Classification Using Ensemble Deep Learning," 2024 International Symposium ELMAR, Zadar, Croatia, 2024, pp. 109-112 https://doi.org/10.1109/ELMAR62909.2024.10694603.

[7] Rajwal, S., Aggarwal, S. Convolutional Neural Network-Based EEG Signal Analysis: A Systematic Review. *Arch Computat Methods Eng* **30**, 3585–3615 (2023) https://doi.org/10.1007/s11831-023-09920-1

[8] B. Chakravarthi, Sin-Chun Ng, M. R. Ezilarasan and Man-Fai Leung, EEG-based emotion recognition using hybrid CNN and LSTM classification https://doi.org/10.3389/fncom.2022.1019776

[9] M. Gagliardi, D. Maurmo, T. Ruga, E. Vocaturo and E. Zumpano, BrAInVision: A hybrid explainable Artificial Intelligence framework for brain MRI analysis, Image and Vision Computing, Volume 161, September 2025, 105629 https://doi.org/10.1016/j.imavis.2025.105629

[10] Z. J. Wang et al., "CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 1396-1406, Feb. 2021 https://doi.org/10.1109/TVCG.2020.3030418

[11] R. Chauhan, K. K. Ghanshala and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 278-282 https://doi.org/10.1109/ICSCCC.2018.8703316.

[12] J. Tao, Y. Gu, J. Sun, Y. Bie and H. Wang, "Research on vgg16 convolutional neural network feature classification algorithm based on Transfer Learning," 2021 2nd China International SAR Symposium (CISS), Shanghai, China, 2021, pp. 1-3 https://doi.org/10.23919/CISS51089.2021.9652277.

[13] T. Kaur and T. K. Gandhi, "Automated Brain Image Classification Based on VGG-16 and Transfer Learning," 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, 2019, pp. 94-98 https://doi.org/10.1109/ICIT48102.2019.00023.

[14] Rashid, M., Mustafa, M., Sulaiman, N., Islam, M.N. (2024). EEG and EMG-Based Multimodal Driver Drowsiness Detection: A CWT and Improved VGG-16 Pipeline. In: Hassan, M.H.A., Omar, M.N., Johari, N.H., Zhong, Y. (eds) Proceedings of the 2nd Human Engineering Symposium. HUMENS 2023. Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-99-6890-9_27

[15] L. Li, "Deep Learning-based EEG Signal Identity Recognition Using VGGNet," 2024 4th

International Conference on Neural Networks, Information and Communication Engineering (NNICE), Guangzhou, China, 2024, pp. 1092-1095
DOI 10.1109/NNICE61279.2024.10498553.

[16] C. M. Michel, M. M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, R. G. de Peralta, EEG source imaging, Clinical Neurophysiology, Volume 115, Issue 10, 2004, Pages 2195-2222, ISSN 1388-2457.
https://doi.org/10.1016/j.clinph.2004.06.001.

[17] J. Wang, W. Hang, S. Liang, Q. Wang, B. Chen and J. Qin, "Convolutional Retentive Network for EEG Decoding," *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1-5
https://doi.org/10.1109/ICASSP49660.2025.10890343.

[18] Cohen, Michael X. Where does EEG come from and what does it mean?, Trends in Neurosciences, Volume 40, Issue 4, 208 – 218
https://www.cell.com/trends/neurosciences/fulltext/S0166-2236(17)30024-3

[19] H. Altaheri, G. Muhammad, M. Alsulaiman “Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: a review.” In conf. Neural Comput & Applic 35, 14681–14722 (2023).
https://doi.org/10.1007/00521-021-06352-5

[20] G. Dai, J. Zhou, J. Huang and N. Wang, HS-CNN: a CNN with hybrid convolution scale for EEG motor imagery classification, Published 6 January 2020 • © 2020 IOP Publishing Ltd
https://doi.org/10.1088/1741-2552/ab405f

[21] A. Craik, Y. He, J. L Contreas “Deep learning for electroencephalogram (EEG) classification tasks: a review” in Journal of Neural Engineering, V.16(iii) April 2019
https://doi.org/10.1088/1741-2552/ab0ab5

[22] Gao, Z., Dang, W., Wang, X. *et al.* Complex networks and deep learning for EEG signal analysis. *Cogn Neurodyn* **15**, 369–388 (2021)
https://doi.org/10.1007/s11571-020-09626-1

[23] S. S. Bhatti, A. Yadav, M. Monga and N. Kumar, "Comparative Analysis of Deep Learning Approaches for Harmful Brain Activity Detection Using EEG," 2024 IEEE 8th International Conference on Information and Communication Technology (CICT), Prayagraj UP, India, 2024, pp. 1-6, doi: 10.1109/CICT64037.2024.10899480.

[24] S. Ganesan, Y. N. Kiran and S. Ram, Harmful Brain Activity Classification of Spectrograms with Transfer Deep Learning
https://doi.org/10.21203/rs.3.rs-4294555/v2

[25] M. -A. Li and D. -Q. Xu, "A Transfer Learning Method based on VGG-16 Convolutional Neural Network for MI Classification," *2021 33rd Chinese Control and Decision Conference (CCDC)*, Kunming, China, 2021, pp. 5430-5435
https://doi.org/10.1109/CCDC52312.2021.9602818.

[26] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open, Volume 3, 2022, Pages 111-132, ISSN 2666-6510.
https://doi.org/10.1016/j.aiopen.2022.10.001.

[27] C. Chen, H. Wang, Y. Chen, Z. Yin, X. Yang, H. Ning, Q. Zhang, W. Li, R. Xiao and J. Zhao, Understanding the brain with attention: A survey of transformers in brain sciences
https://doi.org/10.1002/brx2.29

[28] H. Adeli, S. Minni and N. Kriegeskorte, Predicting brain activity using Transformers
https://doi.org/10.1101/2023.08.02.551743

[29] J. Xie et al., "A Transformer-Based Approach Combining Deep Learning Network and Spatial-Temporal Information for Raw EEG Classification," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 2126-2136, 2022,
https://doi.org/10.1109/TNSRE.2022.3194600.

[30] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, L. Wang, 2025. Vision Transformers for Image Classification: A Comparative Survey. Technologies, 13(1), p.32.
https://doi.org/10.3390/technologies13010032

[31] S. Dongre, S. Mehta, 2024, June. RetViT: Retentive Vision Transformers. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-8). IEEE.
https://doi.org/10.1109/ICCCNT61001.2024.10726145

[32] Md Manowarul Islam, Md Alamin Talukder, Md Ashraf Uddin, A. Akhter and M. Khalid, BrainNet: Precision Brain Tumor Classification with Optimized EfficientNet Architecture
https://doi.org/10.1155/2024/3583612

[33] Y. A. Sadoon, M. Khalil and D. Battikh, Predicting Epileptic Seizures Using EfficientNet-B0 and SVMs: A Deep Learning Methodology for EEG Analysis
https://doi.org/10.3390/bioengineering12020109

[34] U. K. Naik M and S. R. Ahamed, Wavelet-based Autoencoder and EfficientNet for Schizophrenia Detection from EEG Signals
https://doi.org/10.48550/arXiv.2407.17540

[35] G.K. Erabati, H.Araujo, 2024. Retformer: Embracing point cloud transformer with retentive network. IEEE Transactions on Intelligent Vehicles.
https://doi.org/10.1109/TIV.2024.3417260

[36] H. Yang, Z. Li, Y. Chang and Y. Wu, A survey of Retentive Network
https://doi.org/10.48550/arXiv.2506.06708

[37] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Eang, F. Wei "Retentive Network: A Successor to Transformer for Large Language Models" in Cornell University, submitted on 17 Jul 2023.
https://doi.org/10.48550/arXiv.2307.08621

[38] D, U., N., N. & Sebasthiyar, A. Early diagnosis of diabetic retinopathy using retinal network. Multimed Tools Appl (2025).
https://doi.org/10.1007/s11042-025-20682-9

[39] Z. Lin, R. Cui, L. Ning, J. Peng, Temporal Features-Fused Vision Retentive Network for Echocardiography Image Segmentation. Sensors. 2025; 25(6):1909.
https://doi.org/10.3390/s25061909

[40] K. Subramaniam, A. Naganathan, 2024. RetNet30: A Novel Stacked Convolution Neural Network Model for Automated Retinal Disease Diagnosis. International Journal of Imaging Systems and Technology, 34(5), p.e23187.
https://doi.org/10.1002/ima.23187